UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SANDRELLY LUIZ COUTINHO

ANÁLISE DA INFLUÊNCIA DE ATRIBUTOS NO DESEMPENHO DE VENDAS DE APARTAMENTOS EM EMPREENDIMENTOS NO RECIFE: UMA ABORDAGEM COM APLICAÇÃO DE MINERAÇÃO DE DADOS, COM A METODOLOGIA CRISP-DM, FOCADA EM PROBLEMAS DE DECISÃO BINÁRIA E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (X-AI).

Recife

2025

SANDRELLY LUIZ COUTINHO

ANÁLISE DA INFLUÊNCIA DE ATRIBUTOS NO DESEMPENHO DE VENDAS DE APARTAMENTOS EM EMPREENDIMENTOS NO RECIFE: UMA ABORDAGEM COM APLICAÇÃO DE MINERAÇÃO DE DADOS, COM A METODOLOGIA CRISP-DM, FOCADA EM PROBLEMAS DE DECISÃO BINÁRIA E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (X-AI).

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Orientador (a): Prof. Dr. Paulo Jorge Leitão Adeodato

Coorientador (a): Prof. Dr. Bruno Campello de Souza

Recife

2025

Coutinho, Sandrelly Luiz.
Análise da influência de atributos no desempenho de vendas de apartamentos em empreendimentos no Recife: uma abordagem com aplicação de mineração de dados, com a metodologia CRISP- DM, focada em problemas de decisão binária e inteligência artificial explicável (X-AI) / Sandrelly Luiz Coutinho. - Recife, 2025.
282f.: il.

Dissertação (Mestrado)- Universidade Federal de Pernambuco, Centro de Informática, Pós-Graduação em Ciências da Computação, 2025.
Orientação: Paulo Jorge Leitão Adeodato.
Coorientação: Bruno Campello de Souza.

1. Inteligência Artificial Explicável (X-AI); 2. Mineração de Dados - CRISP-DM; 3. Mercado imobiliário. I. Adeodato, Paulo Jorge Leitão. II. Souza, Bruno Campello de. III. Título.

UFPE-Biblioteca Central

Sandrelly Luiz Coutinho

ANÁLISE DA INFLUÊNCIA DE ATRIBUTOS NO DESEMPENHO DE VENDAS DE APARTAMENTOS EM EMPREENDIMENTOS NO RECIFE: UMA ABORDAGEM COMAPLICAÇÃO DE MINERAÇÃO DE DADOS, COM A METODOLOGIA CRISP-DM,FOCADA EM PROBLEMAS DE DECISÃO BINÁRIA E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI).

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 31/07/2025.

BANCA EXAMINADORA

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE

_____

Prof. Dr. Kellyton dos Santos Brito
Departamento de Computação / UFRPE

_____

Prof. Dr. Paulo Jorge Leitão Adeodato
Centro de Informática/UFPE
(orientador)

*Dedico este trabalho a **Camila Mendes,** por estar sempre ao meu lado, e às minhas filhas **Beatriz** e **Luiza** que ficaram um pouquinho sem o pai neste período. Aos meus pais, que me deram a educação e a força necessárias para eu estar aqui neste momento, e aos meus irmãos, que trazem alegria sempre. E, finalmente ao menu amigo e orientador **Paulo Adeodato** que estava perto mesmo nos momentos de maior aperreio.*

AGRADECIMENTOS

Agradeço a todos que apoiaram a minha decisão em voltar à academia, mesmo depois de 20 anos de formado nesse lugar tão especial que é o Centro de Informática da UFPE, que na minha época de graduação ainda era o velho e bom D.I.

Agradeço de forma especial ao amigo, e hoje professor do CIN, Sérgio Soares, que me incentivou a voltar ao CIN em plena pandemia de COVID-19 para fazer algumas aulas de disciplinas isoladas.

Agradeço especialmente ao professor, meu orientador e amigo, Paulo Adeodato por me aceitar em duas disciplinas isoladas e, então, me aceitar como orientando.

E, por fim, à minha família, que entende que nada vem sem muito esforço e dedicação.

RESUMO

A presente dissertação propõe o desenvolvimento de um modelo preditivo e explicável para classificar o sucesso de empreendimentos imobiliários do tipo apartamento, lançados no Recife. Parti-se da análise de dados históricos de vendas, atributos físico-comerciais das unidades e variáveis econômicas de contexto. A abordagem combina técnicas de Mineração de Dados com métodos de Inteligência Artificial Explicável (X-AI), sendo estruturada segundo a metodologia CRISP-DM, que orienta de forma sistemática as etapas desde a compreensão do negócio até a modelagem e a avaliação dos resultados. Dada a reconhecida escassez de bases e a conhecida fragmentação e falta de organização de bases no mercado imobiliário brasileiro, adotou-se como fonte principal a plataforma RE.AI.s, consolidada há mais de 25 anos no setor. Para o município do Recife, foco deste estudo, foi extraído um conjunto robusto contendo 300 empreendimentos e 28.729 unidades habitacionais, com registros mensais de vendas entre janeiro de 2021 e maio de 2025, período temporal ideal para a pesquisa. A garantia da qualidade dos dados seguiu uma abordagem híbrida, combinando métodos estatísticos e o conhecimento empírico do mercado local. Essa validação foi complementada pela colaboração com a equipe de dados RE.AI.s e das construtoras participantes, conferindo maior confiabilidade. Para capturar a complexidade da performance comercial, adotou-se uma abordagem inovadora com a criação de duas classes-alvo binárias distintas, inicialmente definidas no nível do empreendimento e posteriormente propagadas para as unidades: (i) Velocidade de Vendas, que classifica empreendimentos que atingem ou superam 30% de suas vendas nos primeiros três meses pós-lançamento, indicando atratividade imediata; e (ii) Resiliência de Vendas, que identifica empreendimentos que mantêm menos de 20% de unidades não vendidas após 18 meses, refletindo a solidez e aderência do produto a longo prazo. Para avaliar a capacidade preditiva dos dados, foram aplicados modelos supervisionados de classificação, com ênfase em Árvores de Decisão e Regressão Logística. O grande diferencial do estudo reside no foco em Inteligência Artificial Explicável (X-AI), que visa não apenas prever, mas também tornar transparentes e compreensíveis as razões por trás das classificações do modelo.Os resultados demonstram que o uso de atributos no nível da unidade aumentou significativamente a precisão dos modelos e permitiu a

construção de uma ferramenta de apoio à decisão baseada em ciência de dados e alta tecnologia. Além disso, usamos uma abordagem diferenciada, com foco na explicabilidade e na influência dos atributos na performance das vendas (em vez da precificação), visando classificar com sucesso o potencial de lançamento de novos empreendimentos. Em um setor ainda marcado por decisões empíricas e heurísticas individuais, esta abordagem representa uma inovação metodológica com potencial de impacto prático. Como continuidade, sugere-se o aprofundamento da análise espacial, de modelos geográficos, e o uso de outros modelos de inteligência artificial para aprimorar a acurácia preditiva.

Palavras-chave: Mercado Imobiliário; Recife; Mineração de Dados; CRISP-DM; Inteligência Artificial Explicável (X-AI).

ABSTRACT

This dissertation proposes the development of a predictive and explainable model to classify the success of apartment developments launched in Recife. It starts by analyzing historical sales data, physical-commercial attributes of the units, and contextual economic variables. The approach combines Data Mining techniques with Explainable Artificial Intelligence (X-AI) methods, structured according to the CRISP-DM methodology, which systematically guides the stages from business understanding to modeling and results evaluation. Given the recognized scarcity, fragmentation, and lack of organized databases in the Brazilian real estate market, the RE.AI.s platform was adopted as the main data source, a system consolidated over 25 years in the sector. For the municipality of Recife, the focus of this study, a robust dataset was extracted containing 300 developments and 28,729 residential units, with monthly sales records between January 2021 and May 2025, an ideal timeframe for this research.

Data quality assurance followed a hybrid approach, combining statistical methods with empirical knowledge of the local market. This validation was complemented by collaboration with the RE.AI.s data team and the participating construction companies, ensuring greater reliability. To capture the complexity of commercial performance, an innovative approach was adopted with the creation of two distinct binary target classes, initially defined at the development level and subsequently propagated to the individual units: (i) Sales Velocity, which classifies developments achieving or exceeding 30% of their sales within the first three months post-launch, indicating immediate attractiveness; and (ii) Sales Resilience, which identifies developments that maintain less than 20% unsold units after 18 months, reflecting the product's long-term solidity and market fit.

To evaluate the data's predictive capability, supervised classification models were applied, with an emphasis on Decision Trees, Logistic Regression, and Rule Induction. The key differentiator of this study lies in its focus on Explainable Artificial Intelligence (X-AI), which aims not only to predict but also to make the reasons behind the model's classifications transparent and comprehensible.

The results demonstrate that using attributes at the unit level significantly increased model accuracy and allowed for the construction of a decision support tool based on

data science and high technology. Furthermore, we employed a differentiated approach focusing on explainability and the influence of attributes on sales performance (rather than pricing), aiming to successfully classify the potential of new development launches. In a sector still largely driven by empirical decisions and individual heuristics, this approach represents a methodological innovation with the potential for practical impact. For future work, we suggest further spatial analysis, geographic models, and the use of other artificial intelligence models to enhance predictive accuracy.

Keywords: Real Estate Market; Recife; Data Mining; CRISP-DM; Explainable Artificial Intelligence (X-AI).

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O mercado imobiliário brasileiro, particularmente no segmento primário, representa um setor estratégico para o desenvolvimento urbano e econômico das cidades. A atividade promovida por construtoras e incorporadoras é responsável por uma parcela significativa do Produto Interno Bruto (PIB) nacional e atua como um termômetro sensível às variações macroeconômicas, políticas habitacionais e dinâmicas de consumo das famílias. Nesse contexto, compreender os fatores que influenciam o sucesso de vendas de novos empreendimentos é um desafio crucial, tanto para a sustentabilidade financeira dos agentes privados quanto para a formulação de políticas públicas de moradia.

Com o avanço das tecnologias de ciência de dados e inteligência artificial (IA), surge a oportunidade de transformar o grande volume de informações disponíveis sobre o mercado imobiliário em sistemas preditivos capazes de apoiar a tomada de decisão. Ao empregar abordagens de mineração de dados e modelos explicáveis de aprendizagem de máquina, é possível não apenas prever o desempenho comercial de um empreendimento, mas também compreender os fatores que mais influenciam o comportamento de compra dos consumidores.

Esta dissertação propõe uma abordagem analítica voltada à predição por classificação binária do sucesso de vendas de empreendimentos imobiliários, com ênfase em modelos explicáveis de IA. A metodologia adotada baseia-se no modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), utilizando dados reais do mercado imobiliário da cidade do Recife/PE, em dois níveis de granularidade: empreendimentos e unidades habitacionais. O trabalho busca responder à seguinte questão central: **quais características mais influenciam o sucesso comercial de um empreendimento imobiliário e como modelos explicáveis de IA podem construir previsões confiáveis que possam ser facilmente compreendidas por profissionais do setor.**

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

O lançamento de um novo empreendimento imobiliário envolve uma série de decisões estratégicas de alto risco, que vão desde a aquisição do terreno até a

definição do produto, do preço e da campanha de comercialização. Uma das etapas mais críticas nesse processo é a previsão de desempenho comercial após o lançamento, pois um erro nessa estimativa pode comprometer severamente o fluxo de caixa da incorporadora, a rentabilidade do projeto e a credibilidade da marca no mercado.

Historicamente, a avaliação do potencial de vendas de um novo empreendimento tem sido baseada em métodos empíricos, experiência de mercado e análises comparativas com projetos anteriores. Embora essas abordagens tenham valor, elas são frequentemente limitadas pela subjetividade e pela incapacidade de captar padrões complexos em grandes volumes de dados. Além disso, o mercado imobiliário é caracterizado por forte heterogeneidade espacial, flutuações econômicas e mudanças no perfil do consumidor, o que dificulta a generalização de experiências passadas para novos contextos.

Outro desafio importante é a natureza fragmentada, dispersa e, muitas vezes, pouco confiável dos dados disponíveis. Informações sobre lançamentos, vendas, preços e características das unidades costumam estar distribuídas entre diversas fontes, nem sempre padronizadas ou atualizadas. Isso impõe dificuldades adicionais para análises quantitativas mais rigorosas e consistentes.

Além disso, o ciclo de vida de um empreendimento imobiliário é notadamente longo. Desde o momento do lançamento até a completa comercialização das unidades, podem transcorrer vários anos, durante os quais o projeto está exposto a riscos diversos: variações nas taxas de juros, mudanças em políticas públicas de habitação, alterações no perfil da demanda, entrada de concorrentes diretos na mesma região, entre outros. Essa exposição prolongada torna ainda mais relevante a capacidade de prever, desde as etapas iniciais, o potencial de sucesso do projeto.

Nos últimos anos, a crescente digitalização do setor e a maior disponibilidade de bases de dados estruturadas abriram novas possibilidades para análises fundamentadas. Ao mesmo tempo, a popularização de ferramentas de inteligência artificial, como algoritmos de classificação supervisionada e técnicas de explicação de modelos (X-AI), permite desenvolver soluções preditivas que não apenas estimam o desempenho comercial de um empreendimento, mas também tornam essas estimativas compreensíveis e úteis para a tomada de decisão.

Nesse cenário, o presente trabalho se propõe a preencher uma lacuna na literatura e na prática do setor, ao construir um modelo explicável de predição do sucesso comercial de empreendimentos, considerando tanto a velocidade de vendas nos primeiros meses quanto a resiliência da comercialização ao longo do tempo. O estudo adota uma perspectiva orientada por dados, explorando atributos nos níveis de empreendimento e de unidade habitacional, e utilizando métricas binárias de sucesso como critério para treinamento e avaliação dos modelos. O objetivo é oferecer previsões com boa precisão e de maneira transparente, contribuindo para decisões mais embasadas por parte das construtoras, incorporadoras e analistas de mercado.

1.2 MOTIVAÇÃO E JUSTIFICATIVA

A motivação para o desenvolvimento deste trabalho nasce da combinação entre a vivência prática no mercado imobiliário e o interesse científico por métodos de análise baseados em inteligência artificial e mineração de dados. Com mais de duas décadas de experiência na atuação direta com diretores e gestores de construtoras e incorporadoras, observa-se de forma recorrente a dificuldade em estimar com segurança o desempenho comercial de novos projetos, sobretudo em cenários de incerteza econômica ou intensa competição regional.

Mesmo diante da crescente digitalização do setor e do volume cada vez maior de informações disponíveis, as decisões estratégicas relacionadas ao lançamento e comercialização de empreendimentos ainda são majoritariamente orientadas por heurísticas, experiências anteriores e avaliações subjetivas. Isso pode resultar em decisões mal calibradas quanto ao produto, preço, prazo ou posicionamento de mercado, com impactos significativos no risco financeiro e na velocidade de absorção das unidades.

No entanto, o sucesso de vendas de um empreendimento está diretamente relacionado à sua capacidade de atender à demanda real do mercado — ou seja, construir o que o consumidor deseja adquirir. A falta de entendimento aprofundado sobre o perfil do cliente, suas preferências e restrições de escolha é um dos principais fatores que comprometem essa adequação entre oferta e demanda. Neste trabalho, parte-se da premissa de que o comportamento de compra dos

consumidores pode ser inferido a partir do desempenho de vendas observado. Por isso, adota-se como verdade de referência (*ground truth*) o próprio histórico de vendas dos empreendimentos como indicador objetivo do grau de aderência da oferta ao desejo do mercado.

Do ponto de vista acadêmico, o problema da previsão de sucesso em vendas no mercado imobiliário primário ainda é pouco explorado na literatura brasileira, sobretudo quando se trata da aplicação de modelos preditivos com múltiplos níveis de granularidade (empreendimentos e unidades) e com preocupação explícita com a interpretabilidade das decisões dos algoritmos. Grande parte dos estudos existentes concentra-se em estimativas de preços ou agrupamentos de perfil de consumidor, mas poucos abordam diretamente a tomada de decisão estratégica por parte das incorporadoras no momento do lançamento de um projeto.

Além disso, a maior parte dos modelos preditivos aplicados a problemas reais sofre de baixa transparência. Quando utilizados em contextos sensíveis, como o investimento imobiliário, isso pode comprometer a confiança e a adoção das soluções por parte dos tomadores de decisão. Dessa forma, técnicas de inteligência artificial explicável (X-AI) assumem um papel central, pois permitem não apenas realizar predições com boa precisão, mas também comunicar de forma clara os fatores que fundamentam essas predições.

Justifica-se, portanto, o desenvolvimento de um modelo de predição binária de sucesso em vendas que seja ao mesmo tempo robusto do ponto de vista analítico, aplicável ao contexto real de mercado e transparente para os agentes envolvidos. Tal abordagem pode contribuir para reduzir a incerteza nos lançamentos imobiliários, aumentar a eficiência na alocação de recursos e aprimorar a inteligência comercial das empresas do setor. Além disso, os benefícios potenciais da aplicação desse modelo extrapolam os limites das construtoras e incorporadoras, impactando positivamente todo o ecossistema envolvido — incluindo imobiliárias, corretores de imóveis, arquitetos e demais profissionais que orbitam o ciclo de vida de um empreendimento.

1.3 OBJETIVOS E HIPÓTESES

1.3.1 Objetivo geral

Desenvolver e avaliar um modelo de predição binária do sucesso de vendas de empreendimentos residenciais no mercado imobiliário primário, utilizando técnicas de inteligência artificial explicável (X-AI) e dados históricos de vendas, a fim de apoiar a tomada de decisão estratégica de construtoras e demais agentes do setor.

1.3.2 Objetivos específicos

- Estruturar e integrar uma base de dados consolidada com informações de empreendimentos residenciais lançados na cidade do Recife/PE entre 2022 e 2025, contemplando atributos nos níveis de empreendimento e de unidade habitacional.
- Definir critérios objetivos para classificação binária do sucesso comercial de empreendimentos, considerando diferentes perspectivas temporais de desempenho (ex.: velocidade de vendas e resiliência ao longo do tempo).
- Aplicar técnicas de preparação de dados e seleção de atributos relevantes, com foco na interpretabilidade e aderência ao contexto do mercado imobiliário.
- Treinar, ajustar e comparar diferentes algoritmos de classificação supervisionada — incluindo modelos explicáveis como ***Decision Tree*** (Árvore de Decisão), ***Random Forest*** (Floresta Aleatória) e ***Logistic Regression*** (Regressão Logística) — avaliando seu desempenho preditivo com métricas apropriadas.
- Empregar métodos de inteligência artificial explicável (X-AI) para identificar os fatores mais influentes na predição do sucesso de vendas e comunicar os resultados de forma compreensível para usuários não especialistas.
- Analisar os resultados à luz das características do mercado local e discutir as contribuições práticas e teóricas do modelo desenvolvido.

1.3.3 Hipóteses da pesquisa

A pesquisa parte das seguintes hipóteses:

- **H1**: É possível construir um modelo preditivo com qualidade, bom desempenho e transparência para classificar o sucesso de vendas de empreendimentos residenciais com base em dados históricos e atributos estruturais.
- **H2**: O desempenho comercial de um empreendimento (velocidade e resiliência de vendas) está fortemente associado à sua capacidade de atender ao perfil de demanda do mercado, o que pode ser inferido a partir do comportamento de compra dos consumidores.
- **H3**: Modelos explicáveis de classificação podem fornecer interpretações acessíveis a fim de fortalecer a confiança dos tomadores de decisão e, então facilitar a adoção prática de soluções baseadas em IA no setor imobiliário.

1.4 METODOLOGIA DA PESQUISA

A metodologia desta pesquisa seguiu o modelo **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), amplamente adotado em projetos de ciência de dados e mineração de conhecimento, por sua flexibilidade, estrutura iterativa e compatibilidade com diferentes domínios de aplicação (CHAPMAN et al., 2000, p. 15-22). O processo foi dividido em seis fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação. Esta dissertação abordou integralmente as cinco primeiras etapas, com foco analítico e experimental, não contemplando, neste momento, uma fase de implantação em ambiente produtivo. No entanto, o objetivo é disponibilizar a solução desenvolvida ao mercado, ampliando seu impacto prático e operacional.

O estudo teve como unidade de análise o mercado imobiliário primário da cidade do Recife/PE, com recorte temporal entre janeiro de 2022 e maio de 2025. Foram utilizados dois conjuntos principais de dados: (i) um *dataset* de empreendimentos residenciais lançados no período, com informações agregadas sobre vendas, características do projeto e perfil da construtora; e (ii) um *dataset* de unidades habitacionais vinculadas a esses empreendimentos, contendo atributos físicos, valores, disponibilidade comercial e indicadores derivados.

Ambos os conjuntos de dados passaram por um processo robusto de preparação, envolvendo a seleção de dados válidos, o tratamento de valores

ausentes, a consolidação com fontes externas (como dados econômicos e indicadores de construtoras), a criação de novos atributos explicativos e a transformação das variáveis em formatos compatíveis com algoritmos de aprendizagem de máquinas. A preparação dos dados foi orientada por boas práticas de engenharia de atributos, com ênfase em garantir a interpretabilidade dos resultados (KUHN; JOHNSON, 2019, p.42).

A classe-alvo da modelagem foi definida com base em critérios binários de sucesso comercial, construídos a partir da análise do comportamento de vendas: velocidade (ex.: percentual vendido nos primeiros meses) e resiliência (ex.: percentual vendido após período mais longo). Esses critérios funcionaram como *proxy* do grau de aderência entre a oferta do empreendimento e o desejo de compra do consumidor, em linha com a premissa de que o desempenho de vendas reflete o ajuste ao perfil de demanda.

Na etapa de modelagem, foram aplicados algoritmos de classificação supervisionada com foco na explicabilidade dos resultados. Dentre os modelos considerados, destacam-se *Decision Tree*, *Random Forest* e *Logistic Regression*, todos reconhecidos por sua capacidade de gerar decisões interpretáveis por humanos. Os modelos foram avaliados quanto à sua performance preditiva por meio de métricas como *AUC-ROC*, *F1-score*, *Acuracy* (acurácia), *Precision* (precisão) e *Recall* (revocação).

Por fim, técnicas de inteligência artificial explicável (X-AI) foram empregadas para identificar os fatores mais relevantes nas decisões dos modelos, com o objetivo de fornecer aos profissionais do setor imobiliário não apenas uma classificação automatizada, mas também subsídios claros para interpretação e confiança nos resultados. A análise dos achados considerou o contexto local do mercado e buscou contribuir tanto para a prática profissional quanto para o avanço científico no uso de IA no setor habitacional.

1.5 CONTRIBUIÇÕES

Esta pesquisa buscou oferecer contribuições relevantes tanto para o avanço do conhecimento científico na área de ciência de dados aplicada ao setor imobiliário

quanto para a prática de mercado de construtoras, incorporadoras e demais agentes envolvidos no ciclo de vida de empreendimentos residenciais.

Do ponto de vista acadêmico, o trabalho propôs um modelo de predição binária do sucesso de vendas de empreendimentos imobiliários com base em dados reais e atualizados de mercado, estruturado conforme a metodologia CRISP-DM. A abordagem adotou múltiplos níveis de granularidade — empreendimentos e unidades habitacionais — o que permitiu uma análise mais refinada e contextualizada dos fatores que influenciam o desempenho comercial. Além disso, o estudo dá ênfase à utilização de algoritmos explicáveis de aprendizado de máquina, contribuindo para o debate sobre transparência, confiança e interpretabilidade em aplicações de inteligência artificial (X-AI).

Uma contribuição metodológica importante esteve na definição dos critérios objetivos de sucesso, a partir de métricas de velocidade e resiliência de vendas, os quais foram utilizados como verdade de referência (*ground truth*) no processo de modelagem. Essa escolha esteve ancorada na premissa de que o comportamento de compra dos consumidores é o reflexo mais fiel da adequação do produto à demanda real do mercado.

Do ponto de vista prático, os resultados obtidos podem subsidiar decisões mais embasadas na fase de planejamento e lançamento de novos empreendimentos, reduzindo o risco comercial e aumentando a eficiência na alocação de recursos. A explicação dos fatores mais influentes nas previsões dos modelos permite que construtoras, incorporadoras e analistas compreendam não apenas o resultado da classificação, mas também as razões por trás desse resultado — o que amplia a aplicabilidade das soluções no ambiente de negócios.

Além disso, a contribuição do trabalho estende-se para todo o ecossistema que gravita em torno da produção habitacional, incluindo imobiliárias, corretores, arquitetos, profissionais de marketing e empresas de tecnologia do setor. Ao fornecer ferramentas analíticas mais transparentes e orientadas por dados, o modelo desenvolvido pode apoiar a tomada de decisão em diferentes etapas do ciclo de vida do empreendimento, desde a concepção do produto até a sua comercialização plena.

Espera-se, ainda, que a partir do desenvolvimento deste trabalho seja possível manter uma base de dados estruturada e continuamente atualizada, que

permita o aperfeiçoamento contínuo dos modelos preditivos, a expansão da análise para outras localidades e tipologias de produto, e o fortalecimento de uma cultura orientada por dados no mercado imobiliário.

1.6 ORGANIZAÇÃO DO DOCUMENTO

Este documento está estruturado em dez capítulos, além das referências bibliográficas, apêndices e anexos, com o objetivo de apresentar de forma clara e sequencial o desenvolvimento desta dissertação.

- **Capítulo 1: Introdução** – Apresenta o contexto do problema de pesquisa, a motivação e justificativa para o estudo, os objetivos e hipóteses propostas, a metodologia geral da pesquisa, as contribuições esperadas e a organização do documento.
- **Capítulo 2: Revisão da Literatura** – Aborda os principais trabalhos e conceitos existentes na literatura que fundamentam o estudo, contextualizando a pesquisa no cenário acadêmico e identificando lacunas.
- **Capítulo 3: Fundamentação Teórica** – Detalha os conceitos teóricos essenciais para a compreensão do trabalho, incluindo mineração de dados, a metodologia CRISP-DM, modelos de classificação, técnicas de avaliação de modelos e o campo da Inteligência Artificial Explicável (X-AI), além das ferramentas utilizadas.
- **Capítulo 4: Metodologia do Projeto** – Descreve a aplicação prática da metodologia CRISP-DM ao problema de negócio, detalhando o contexto do domínio, o modelo conceitual, o ambiente experimental e, de forma aprofundada, as etapas de entendimento e preparação dos dados.
- **Capítulo 5: Desenvolvimento e Avaliação dos Modelos Preditivos** – Apresenta a arquitetura da solução, a preparação específica dos dados para modelagem, a seleção dos algoritmos de aprendizagem de máquina, o processo de treinamento e validação, o refinamento dos modelos para as hipóteses de Velocidade e Resiliência de Vendas.
- **Capítulo 6: Resultados e Discussões** – Consolida a apresentação dos resultados obtidos e as abordagens de extração de conhecimento e análise de interpretabilidade (X-AI), realiza uma análise crítica do desempenho dos

modelos, discute as descobertas e *insights* para o mercado imobiliário, valida as hipóteses de pesquisa e compara os achados com estudos anteriores na literatura.

- **Capítulo 7: Conclusão e Trabalhos Futuros** – Sintetiza os principais achados da pesquisa, discute as limitações do estudo e propõe direções para trabalhos futuros.
- **Referências Bibliográficas** – Lista todas as fontes citadas ao longo do documento.
- **Apêndices e Anexos** – Inclui materiais complementares relevantes para o estudo.

2. REVISÃO DA LITERATURA

O presente capítulo tem como objetivo fundamentar teórica e metodologicamente os conceitos, técnicas e abordagens que sustentam a proposta desta pesquisa, integrando conhecimentos das áreas de mercado imobiliário, mineração de dados, aprendizado de máquina e inteligência artificial explicável (XAI). A construção de um modelo preditivo aplicado à identificação de empreendimentos imobiliários com maior ou menor probabilidade de sucesso em vendas exige não apenas domínio técnico sobre os algoritmos utilizados, mas também uma compreensão crítica sobre as variáveis envolvidas, os desafios inerentes ao tratamento de dados reais e a importância da interpretabilidade no suporte à tomada de decisão.

Inicialmente, são discutidos os principais aspectos que caracterizam o mercado imobiliário primário brasileiro, com foco na dinâmica de lançamentos residenciais e nos fatores que influenciam o desempenho de vendas. Em seguida, são apresentados os fundamentos do aprendizado supervisionado aplicado à classificação binária. A metodologia CRISP-DM é abordada como estrutura orientadora para o processo de modelagem preditiva, desde a compreensão do problema até a preparação dos dados e avaliação dos modelos. A seguir, discute-se o papel da inteligência artificial explicável na construção de modelos que não apenas performem bem, mas também sejam compreensíveis e auditáveis por especialistas de negócio. Por fim, são analisados trabalhos correlatos que exploram o uso de técnicas de mineração de dados e aprendizado de máquina para predição de desempenho no setor imobiliário.

Essa revisão proporciona a base teórica necessária para justificar as escolhas metodológicas adotadas neste estudo, bem como para posicionar a contribuição desta dissertação no contexto da literatura científica existente.

2.1 MERCADO IMOBILIÁRIO PRIMÁRIO E FATORES DE SUCESSO

O mercado imobiliário primário compreende os empreendimentos disponibilizados pela primeira vez ao mercado, geralmente no contexto de lançamentos residenciais realizados por construtoras e incorporadoras. Trata-se de

um segmento estratégico da cadeia produtiva da construção civil, com forte impacto na economia urbana, geração de empregos e dinamismo do setor financeiro. No Brasil, esse mercado é caracterizado por ciclos de expansão e retração, fortemente influenciados por políticas públicas de financiamento, indicadores macroeconômicos como taxa de juros e inflação, e mudanças no perfil de demanda habitacional (LOCATELLI et al., 2017).

O sucesso comercial de um empreendimento no mercado primário é condicionado por um conjunto de fatores multidimensionais que envolvem características do produto imobiliário, posicionamento de mercado, atributos da localização, reputação da construtora e condições de oferta e demanda no momento do lançamento. Elementos como tipologia da unidade, metragem, quantidade de dormitórios, vagas de garagem, padrão de acabamento e valor por metro quadrado afetam diretamente a atratividade do imóvel perante o público-alvo. Além disso, variáveis contextuais como bairro, proximidade a polos de comércio e serviços, infraestrutura urbana e perfil socioeconômico da região exercem influência significativa sobre o ritmo de vendas (ALMEIDA; AMANO; TUPY, 2022).

Estudos recentes sugerem que a percepção de valor por parte dos consumidores e a adequação do empreendimento à demanda latente são os principais vetores de velocidade de comercialização. Empreendimentos mal posicionados, com preço inadequado ou lançados em momentos de saturação da oferta tendem a apresentar desempenho abaixo do esperado, independentemente de suas características construtivas. A literatura também mostra que fatores como caminhabilidade, segurança e acesso a transporte público podem exercer influência significativa na formação de valor e, portanto, no sucesso de vendas (DE NADAI; LEPRI, 2018).

Apesar da importância crítica desses fatores, o processo decisório no setor imobiliário brasileiro ainda é fortemente orientado por práticas intuitivas, julgamentos empíricos e *benchmarking* informal entre gestores, corretores e incorporadores. Historicamente, decisões como a escolha do bairro, tipologia ou momento de lançamento têm sido guiadas por experiências passadas ou percepções subjetivas de mercado, com pouca sistematização analítica ou uso estruturado de dados históricos. Essa abordagem, embora muitas vezes eficiente em contextos de baixa complexidade, tende a ser limitada em cenários mais competitivos e voláteis.

Nesse contexto, a incorporação de técnicas de análise de dados e modelos preditivos surge como uma resposta metodológica promissora, capaz de reduzir vieses cognitivos, antecipar riscos comerciais e aumentar a precisão das decisões de lançamento. A transição de uma cultura decisória baseada exclusivamente em experiência para uma cultura orientada por dados (*data-driven*) representa uma evolução estratégica do setor, promovendo maior eficiência, escalabilidade e adaptação às novas exigências de mercado.

2.2 MINERAÇÃO DE DADOS E APRENDIZADO DE MÁQUINA APLICADOS AO DOMÍNIO IMOBILIÁRIO

A crescente disponibilidade de dados sobre empreendimentos, vendas, características urbanas e perfil de clientes tem potencializado o uso de abordagens quantitativas para a tomada de decisão no setor imobiliário. Nesse cenário, a Mineração de Dados (*Data Mining*) emerge como um processo estruturado para transformar dados brutos em conhecimento acionável (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p.37-54). Essa etapa, parte central do *pipeline* conhecido como KDD (*Knowledge Discovery in Databases*), engloba seleção, pré-processamento, mineração e interpretação — conforme descrito por Fayyad et al.

A mineração de dados vai além da análise estatística descritiva tradicional, pois automatiza a identificação de padrões, correlações e regularidades relevantes a partir de grandes volumes de dados.

2.2.1 Problemas de classificação supervisionada

Os problemas de classificação — em particular, os de **classificação binária** — são amplamente utilizados em contextos preditivos onde se deseja estimar a possibilidade de ocorrência de um evento, por exemplo, essa abordagem pode ser especialmente relevante para o setor imobiliário, quando se busca identificar empreendimentos com maior possibilidade de ter sucesso ou falha em atingir suas metas comerciais.

2.2.2 Modelos de classificação e aplicações no setor imobiliário

Diversos estudos recentes demonstram o uso efetivo de técnicas de aprendizado de máquina no contexto imobiliário:

- Em um estudo sistemático, Moreno-Foronda, Sánchez-Martínez & Pareja-Eastaway (2023) compararam modelos de **r**egressão linear hedônica com algoritmos de *machine learning* (como árvores de decisão e redes neurais). Concluíram que os modelos baseados em ML apresentaram maior capacidade preditiva, embora os modelos lineares fossem mais interpretáveis em termos de impacto individual das variáveis.
- Jha et al. (2020) desenvolveram um classificador para prever se o preço final de venda seria maior ou menor que o valor pedido, utilizando *Logistic Regression, Random Forest* e *XGBoost*. O *XGBoost* apresentou melhor desempenho no conjunto de dados estudado
- Pastukh & Khomyshyn (2025) aplicaram métodos de *ensemble learning* — como *Random Forest*, *Gradient Boosting* e *Extra Trees* — para previsão de preços imobiliários, obtendo alta precisão e baixa taxa de erro, mostrando a robustez desses algoritmos em modelos de avaliação massiva de imóveis.

Modelos de **árvores de decisão**, como CART ou C4.5, são especialmente valorizados por sua **clareza interpretativa**: cada caminho da árvore representa uma regra de decisão facilmente comunicável a especialistas do setor. Já os algoritmos baseados em **regras de indução** (como RIPPER ou PART), embora menos comuns, também fornecem regras “if-then” isoladas que podem ser úteis para explicabilidade operacional.

Essas técnicas, quando aliadas a boas práticas de preparação de dados (como *encoding*, normalização e seleção de variáveis), viabilizam modelos preditivos robustos, generalizáveis e explicáveis, capazes de apoiar decisões estratégicas de lançamento, precificação e segmentação no mercado imobiliário.

2.3 METODOLOGIA CRISP-DM

O CRISP-DM (*Cross-Industry Standard Process for Data Mining*) é reconhecido como o modelo padrão de fato para projetos de mineração de dados, sendo neutro em relação à indústria, ferramenta ou técnica utilizada e amplamente adotado na prática acadêmica e profissional (SHEARER, 2000, p 13-22). Sua primeira versão foi publicada formalmente em 2000 por Chapman et al., definindo um percurso iterativo composto por seis fases interconectadas e complementares:

2.3.1 Fases do CRISP-DM

As seis fases que compõem o ciclo de vida do CRISP-DM são:

1. *Business Understanding* (Entendimento do Negócio)

Estabelece os objetivos gerais da avaliação preditiva, define os resultados esperados e constrói um plano alinhado ao contexto do negócio imobiliário.

2. *Data Understanding* (Compreensão dos Dados)

Abrange a coleta, exploração inicial e avaliação da qualidade dos dados disponíveis, permitindo identificar lacunas, *outliers* e características relevantes antes da modelagem.

3. *Data Preparation* (Preparação dos Dados)

Corresponde à seleção, limpeza, transformação, criação de variáveis e codificação dos atributos. É nesta fase que ocorre a extração de *features* e tratamentos como o agrupamento de categorias, *one-hot encoding* e discretizações utilizadas no estudo.

4. *Modeling* (Modelagem)

Aplica-se técnicas de aprendizagem de máquinas adequadas ao problema, calibrando parâmetros e realizando validação cruzada para otimizar desempenho.

5. *Evaluation* (Avaliação)

Avalia a qualidade dos modelos sob perspectiva de negócio, considerando métricas como precisão, sensibilidade, AUC-ROC, e recursos para aumentar a interpretabilidade, verificando se os resultados atendem aos objetivos definidos na fase 1.

6. *Deployment* (Implantação)

Implanta o modelo em ambiente produtivo ou o operacionaliza para subsídio à tomada de decisão — por exemplo, através de relatórios, dashboards ou APIs de predição.

Essa estrutura é concebida para ser iterativa; decisões tomadas em fases posteriores muitas vezes exigem revisões ou retornos a etapas anteriores, o que garante maior robustez e refinamento contínuo ao longo do projeto.

2.3.2 Adequação do CRISP-DM ao estudo

No contexto deste trabalho, o CRISP-DM se mostrou particularmente adequado pela sua flexibilidade e alinhamento com requisitos do setor imobiliário, tais como:

- Enfoque explícito em Entendimento do Negócio, permitindo relacionar diretamente os objetivos preditivos (velocidade de vendas) às necessidades das construtoras e incorporadoras.
- Ênfase na qualidade dos dados e engenharia de atributos, refletida na fase de preparação, crucial para lidar com características diversas (temporais, geográficas, booleanas e contínuas).
- Possibilidade de interpretação dos modelos, facilitando a adoção de técnicas explicáveis e seu uso estratégico na decisão imobiliária.
- Natureza iterativa do processo, que se adapta ao ciclo de análise, avaliação e recalibração contínua quando novos dados se tornam disponíveis.

A metodologia CRISP-DM, portanto, forneceu o arcabouço conceitual e operacional que sustentou a estrutura utilizada neste estudo, desde o levantamento dos objetivos comerciais até a aplicação dos modelos preditivos e sua interpretação pelos tomadores de decisão.

2.4 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (X-AI)

2.4.1 Emergência da X-AI e a "caixa-preta"

Com o avanço das técnicas de inteligência artificial, especialmente modelos complexos como redes neurais profundas e *ensembles*, tornou-se evidente um desafio central: a falta de transparência e confiança – o problema da chamada **caixa-preta**. Esses modelos, embora altamente precisos, podem gerar decisões sem justificativas compreensíveis para os *stakeholders*. Essa característica limita

sua adoção em domínios sensíveis e críticos, como saúde, finanças e, no caso deste estudo, o setor imobiliário que demanda **confiabilidade, auditabilidade e explicação dos resultados**.

A *Explainable Artificial Intelligence (X-AI)* surgiu como uma disciplina voltada a mitigar esses problemas, propondo métodos que tornem os modelos de IA mais interpretáveis, confiáveis e conformes com requisitos de transparência e regulação (ANGIELOV; RUGGIERI et al., 2018).

2.4.2 Interpretabilidade vs. Explicabilidade

Embora frequentemente usados de forma intercambiável, os termos **interpretabilidade** e **explicabilidade** têm nuances distintas:

- **Interpretabilidade** refere-se à facilidade com que um ser humano pode compreender o funcionamento do modelo em sua totalidade (por exemplo, uma árvore de decisão simples).
- **Explicabilidade** diz respeito à capacidade de entender por que um modelo tomou uma decisão específica em uma instância individual, independentemente da compreensão global do modelo completo.

A X-AI combina ambos os aspectos: desde técnicas intrínsecas (modelos interpretáveis por *design*) até métodos *post-hoc* que geram explicações locais ou globais após o treinamento dos modelos.

2.4.3 Abordagens e técnicas de X-AI

Alcance Global vs. Local

De acordo com Guidotti et al. e revisões do estado da arte, os métodos de X-AI são classificados com base no **alcance da explicação**:

- **Explicação global**: visa revelar padrões globais e a lógica interna do modelo como um todo.
- **Explicação local**: focada em explicar a predição de uma única instância.

Essa distinção é especialmente útil para orientar a escolha das técnicas em função da necessidade de transparência ou aplicação corporativa.

SHAP (*SHapley Additive exPlanations*)

O SHAP baseia-se na **teoria de jogos cooperativos**, atribuindo a cada atributo (feature) um valor de contribuição para uma predição específica, com base nos *valores de Shapley*. Essa abordagem fornece explicações **locais e globais coerentes**, com propriedades axiomaticamente fundamentadas como consistência, eficiência e simetria — amplamente aplicada em análises com dados tabulares e modelos complexos.

LIME (*Local Interpretable Model-agnostic Explanations*)

O LIME é uma técnica **agnóstica ao modelo**, que cria um modelo auxiliar simples local (como uma regressão linear) ao redor de uma instância específica, permitindo entender quais variáveis influenciaram aquela predição ("como se o modelo fosse linear naquele ponto"). O LIME é especialmente útil para explicações locais e pode servir como suporte operacional em decisões individuais.

2.4.4 Relevância da X-AI no mercado imobiliário

No contexto imobiliário, a aplicabilidade da X-AI é especialmente relevante:

- **Confiança dos usuários**: os gestores das construtoras e incorporadoras precisam entender por que um modelo considera determinada unidade ou empreendimento com alta probabilidade de sucesso.
- **Auditabilidade**: permite justificar decisões diante de *stakeholders* internos ou externos (reguladores, investidores).
- ***Insights* estratégicos**: explicações locais podem revelar atributos-chave que impactam diretamente a velocidade de vendas (ex: bairro, metragem, preço/m²).
- **Uso de XAI como ferramenta reflexiva**: ao integrar SHAP ou LIME em *dashboards* de decisão, é possível atribuir *scores* de risco e oportunidade de forma explorável e interativa.

Essa combinação — modelos preditivos robustos com explicabilidade operativa — permite que o setor imobiliário avance além de decisões baseadas em

intuição, promovendo maior **transparência, precisão e alinhamento com as necessidades de negócio**.

2.5 ESTUDOS RELACIONADOS

Diversos estudos têm demonstrado a eficácia de modelos de aprendizagem de máquinas e mineração de dados para predizer o desempenho de vendas ou valor de imóveis, tanto em contextos internacionais quanto brasileiros.

2.5.1 Estudos internacionais

- **Predição de oportunidades de mercado em Madrid**: Baldominos et al. (2018) aplicaram técnicas de regressão, *k-nearest neighbors*, SVM e redes neurais para identificar anúncios imobiliários subvalorizados em tempo real no distrito de Salamanca, com ênfase em engenharia de atributos e comparação de técnicas.
- **Comparação entre regressão hedônica e ML**: Jha et al. (2020) criaram classificadores para prever se o preço de venda seria superior ao valor inicial anunciado, testando *Logistic Regression, Random Forest, Voting Classifier* e *XGBoost*. O algoritmo *XGBoost* atingiu melhor desempenho no conjunto de dados da Flórida, destacando sua robustez.
- **XGBoost e SHAP em mercados instáveis**: Xu & Nguyen (2022) aplicaram regressões, *Random Forest* e *XGBoost* para prever preços em subúrbios de Chicago após a pandemia. Utilizaram valores de Shapley (SHAP) para interpretar a importância dos atributos no modelo.

2.5.2 Trabalhos aplicados no Brasil

- **Fatores de decisão na compra de imóveis residenciais**: Souza (2010) investigou, por meio de uma pesquisa de levantamento com 470 respondentes na cidade de São Paulo, quais atributos influenciam a decisão de compra de apartamentos residenciais. Utilizando análise fatorial exploratória, o autor identificou sete fatores principais — exigências de

suporte gerais, necessidades complementares, conforto espacial, suporte ao prédio, aproveitamento da natureza, localização e privacidade — que explicam 57% da variância dos dados. O estudo evidenciou que qualidade construtiva, conforto interno e localização são os atributos mais valorizados, oferecendo subsídios relevantes para incorporadoras e profissionais de marketing imobiliário na definição de projetos e estratégias de segmentação.

- **Redes neurais aplicadas à precificação imobiliária**: Brito, Felzemburgh e Jardim (2025) desenvolveram um modelo de aprendizado de máquina baseado em redes neurais artificiais para estimar preços de venda de apartamentos em Salvador (BA). Utilizando dados coletados via web scraping e tratados metodologicamente, o modelo alcançou desempenho robusto (R^2 = 0,87; MAE = R$ 915,44; MAPE = 18,06%), superando métodos tradicionais de regressão hedônica. A análise espacial dos erros mostrou distribuição homogênea, evidenciando a capacidade da rede em capturar padrões complexos de valorização imobiliária sem vieses geográficos.
- **Modelos clássicos e redes neurais na avaliação imobiliária**: Silva, Santana e Rocha (2023) aplicaram regressão linear, regressão polinomial, Perceptron e redes neurais multicamadas para prever preços de imóveis no Norte de Minas Gerais. As bases de dados foram criadas por web scraping e tratadas com normalização e k-means. O modelo MLP apresentou o melhor desempenho (EMQ = $2{,}7\times10^{-2}$; $R^2 \approx 0{,}87$), superando os métodos tradicionais e evidenciando o potencial do aprendizado de máquina na precificação imobiliária regional.

Esses estudos evidenciam que:

- Modelos de aprendizado de máquina tendem a superar a precisão de técnicas tradicionais ao prever preço e performance de vendas.
- A explicabilidade (via SHAP, LIME, árvores ou regras) é fundamental para tornar os resultados úteis ao contexto imobiliário.
- Projetos com dados reais, tanto no exterior quanto no Brasil, validam a aplicação prática da abordagem adotada nesta dissertação.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 INTRODUÇÃO

Esta seção apresenta os principais conceitos teóricos que sustentam a metodologia adotada neste estudo. São discutidas as bases da mineração de dados, a metodologia CRISP-DM, os modelos de classificação utilizados, os métodos de avaliação dos modelos, os fundamentos da Inteligência Artificial Explicável (X-AI) e as ferramentas computacionais empregadas no desenvolvimento do projeto.

3.2 MINERAÇÃO DE DADOS

A mineração de dados (*data mining*) se consolidou no cruzamento entre estatística, aprendizagem de máquinas e bancos de dados. Seu objetivo é descobrir automaticamente padrões, tendências e informações úteis a partir de grandes volumes de dados. Ela é um passo chave no processo mais amplo conhecido como Descoberta de Conhecimento em Bases de Dados (*KDD — Knowledge Discovery in Databases*), que inclui a seleção, o pré-processamento, a transformação, a mineração propriamente dita e a interpretação dos dados. Para ilustrar isso de forma clara e científica, o diagrama abaixo representa o processo de KDD, adaptado de Fayyad, Piatetsky-Shapiro & Smyth (1996), mostrando as cinco etapas principais:

1.Seleção dos Dados (*Selection*)

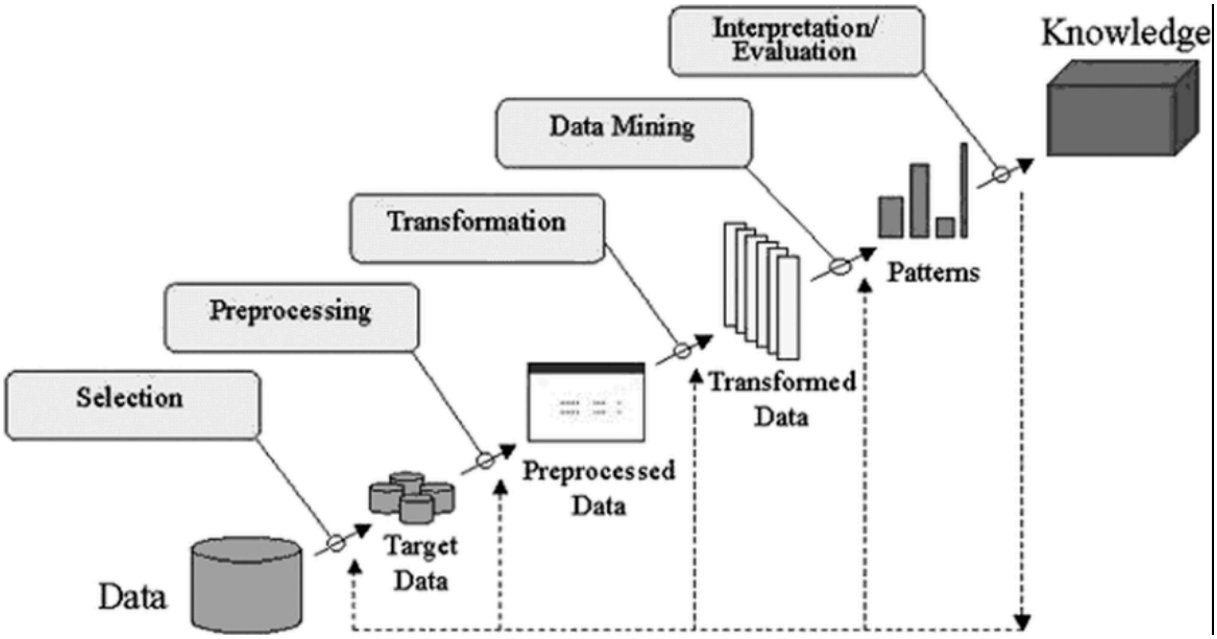
2.Pré-processamento (*Preprocessing*)

3.Transformação (*Transformation*)

4.Mineração de Dados (*Data Mining*)

5.Interpretação e Avaliação (*Interpretation/Evaluation*)

Figura 3.1 – Etapas do processo de KDD



Fonte: Fayyad, Piatetsky-Shapiro & Smyth, 1996

No contexto da pesquisa científica e da análise de negócios, a mineração de dados é especialmente valiosa por sua capacidade de transformar dados brutos em conhecimento aplicável e acionável. Ao invés de simplesmente descrever os dados, seu propósito é gerar modelos descritivos e preditivos que permitam compreender comportamentos passados e antecipar eventos futuros. Isso é particularmente relevante no mercado imobiliário, onde padrões de compra, dinâmica de oferta e demanda, sazonalidade e outros fatores econômicos influenciam significativamente o desempenho de empreendimentos.

A mineração de dados usa diferentes tipos de tarefas analíticas. Entre as mais comuns estão:

- **Classificação**: atribuição de categorias a instâncias com base em variáveis explicativas (ex: prever se um imóvel terá “alta” ou “baixa” velocidade de vendas);
- **Regressão**: estimação de valores numéricos contínuos (ex: valor de venda de uma unidade);

- **Agrupamento (clustering)**: segmentação de dados não rotulados com base em semelhanças;
- **Associação**: descoberta de regras que descrevem co-ocorrência entre atributos (ex: clientes que compram imóveis de alto padrão tendem a comprar em certas regiões);
- **Detecção de anomalias**: identificação de registros que se desviam significativamente do padrão esperado.

Em aplicações reais, essas tarefas são frequentemente combinadas, e os modelos resultantes precisam ser interpretáveis, validados estatisticamente e conectados aos objetivos do domínio.

Na presente dissertação, usamos a mineração de dados para estruturar um problema de classificação binária do sucesso de vendas de empreendimentos imobiliários. Essa tarefa envolveu a transformação de atributos de diferentes granularidades (nível de empreendimento e de unidade), o enriquecimento dos dados com variáveis externas (econômicas e regionais) e a criação de variáveis-alvo com base em critérios comerciais de velocidade e resiliência de vendas. Ao longo do processo, nos preocupamos em garantir a qualidade dos dados por meio de procedimentos de seleção, tratamento de valores ausentes, valores atípicos e integração de fontes heterogêneas — sempre seguindo as boas práticas recomendadas na literatura (Pyle, 1999; Witten et al., 2016).

É importante destacar que, embora a mineração de dados seja tradicionalmente orientada à performance preditiva, o avanço recente das aplicações exige cada vez mais atenção à explicabilidade dos modelos. Nesse sentido, esta dissertação também incorporou os princípios da Inteligência Artificial Explicável (X-AI), garantindo que os padrões descobertos sejam não apenas eficazes, mas também compreensíveis e úteis para os tomadores de decisão no mercado imobiliário.

A mineração de dados não é apenas um processo de escavação cega de padrões. Seu poder está em ser dirigida por hipóteses de negócio e validada por conhecimento do domínio (Zhu & Ferreira, 2014, p. 93).

Essa perspectiva orientada ao problema de negócio é o que fundamentou o uso da metodologia CRISP-DM na sequência deste trabalho, como será detalhado na seção seguinte.

3.3 METODOLOGIA CRISP-DM

A metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) é reconhecida como o modelo de processo mais amplamente utilizado em projetos de mineração de dados. Proposta por Chapman et al. (2000), ela oferece uma abordagem estruturada, iterativa e independente de domínio específico, permitindo que projetos de descoberta de conhecimento em dados avancem de forma controlada, transparente e reaplicável. Seu caráter cíclico favorece o refinamento contínuo e é particularmente adequado para problemas de negócios complexos, como a análise de vendas no setor imobiliário.

O CRISP-DM divide o processo de mineração de dados em seis fases principais:

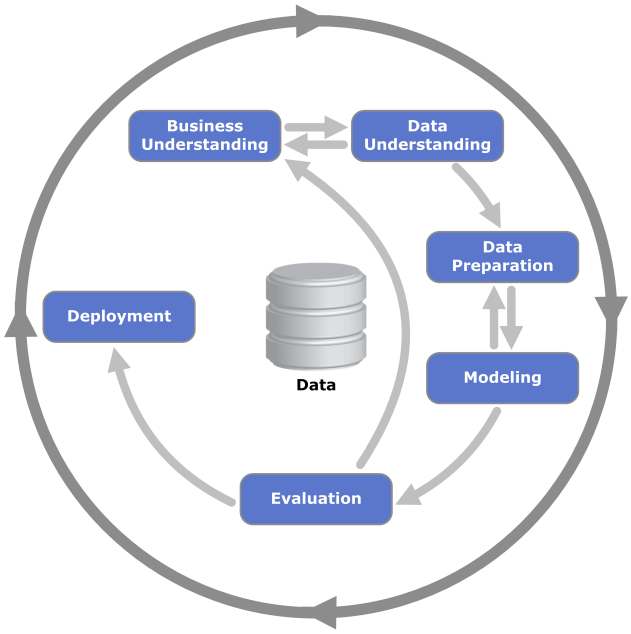
- Compreensão do Negócio (*Business Understanding*): Esta fase envolve a definição clara dos objetivos do negócio e a tradução desses objetivos em problemas de mineração de dados. No contexto desta dissertação, isso significou modelar o sucesso de empreendimentos imobiliários com base em critérios como velocidade de vendas e resiliência.
- Compreensão dos Dados (*Data Understanding*): Aqui, a etapa inclui a coleta inicial, descrição, exploração e validação dos dados disponíveis. Nesta fase, foram analisados os atributos dos empreendimentos e das unidades, a origem das bases e a sua qualidade estrutural e semântica.
- Preparação dos Dados (*Data Preparation*): Essa fase envolve a limpeza, integração, transformação, enriquecimento e formatação dos dados para a modelagem. A dissertação dedicou atenção especial a esta etapa, incluindo a criação de atributos derivados, a normalização e a codificação categórica.
- Modelagem (*Modeling*): Esta etapa trata da seleção e aplicação de algoritmos preditivos, como árvores de decisão e regressão logística, adaptando parâmetros e estratégias conforme a natureza da variável-alvo e dos dados. Também foram exploradas práticas de *cross-validation* e balanceamento de classes.
- Avaliação (*Evaluation*): Nessa fase, verifica-se a qualidade dos modelos produzidos à luz dos objetivos do negócio. Métricas como *AUC-ROC, F1-*

score, Acuracy, Precision e *Recall* foram utilizadas para validar a capacidade dos modelos de identificar empreendimentos com alta atratividade ou resiliência de vendas.

- Implantação (*Deployment*): Embora a implantação em um sistema produtivo não tenha sido o escopo central deste trabalho, os resultados e os modelos desenvolvidos serão disponibilizados em um repositório para reprodutibilidade, e por certo, será integrada futuramente à plataforma RE.AI.s.

Chapman et al. (2000) destacam que o CRISP-DM é um processo iterativo, no qual descobertas em fases posteriores podem levar ao retorno e ao ajuste de decisões anteriores. No caso deste estudo, por exemplo, a análise exploratória levou à redefinição de atributos e da própria construção da classe-alvo, reforçando o ciclo adaptativo do método.

Figura 3.2 – Fases do processo CRISP-DM



Fonte: Chapman et al. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

"O CRISP-DM é suficientemente genérico para ser aplicado em qualquer domínio, mas suficientemente detalhado para ser seguido passo a passo por cientistas de dados e analistas" (Chapman et al., 2000, p. 15).

3.4 MODELOS DE CLASSIFICAÇÃO

A classificação é uma das tarefas mais relevantes da mineração de dados, especialmente em contextos onde o objetivo é prever uma variável categórica a partir de atributos preditivos. Em termos formais, trata-se de um problema de aprendizado supervisionado, no qual um modelo é treinado a partir de exemplos rotulados para posteriormente atribuir classes a novas observações com base em seus atributos (Han et al., 2012).

Como técnica fundamental em mineração de dados, a **classificação binária** é amplamente empregada quando o objetivo é prever uma variável categórica binária — ou seja, com apenas duas classes possíveis — a partir de atributos observáveis.

o problema de classificação está centrado na previsão do sucesso comercial de empreendimentos imobiliários, com base em critérios de sucesso previamente definidos: **Velocidade de Vendas** e **Resiliência de Vendas**. Cada empreendimento (ou unidade habitacional) é rotulado como pertencente à classe “1” (sucesso) ou “0” (insucesso), conforme métricas temporais de comercialização.

A escolha dos algoritmos de classificação foi orientada não apenas pelo desempenho preditivo, mas também pela **capacidade de explicabilidade dos modelos**, um aspecto fundamental no domínio do mercado imobiliário, onde a transparência das decisões é essencial para a adoção prática por analistas, gestores e incorporadores. Por isso, optou-se por três modelos clássicos e complementares:

- ***Decision Tree*** (Árvore de Decisão), pela sua interpretabilidade gráfica e lógica hierárquica;
- ***Random Forest*** (Floresta Aleatória), por agregar várias árvores de decisão em um modelo conjunto mais estável e robusto;
- ***Logistic Regression*** (Regressão Logística), por sua robustez estatística e coeficientes diretamente interpretáveis;

As subseções a seguir apresentam uma breve fundamentação teórica de cada modelo, discutindo seu funcionamento, vantagens, limitações e adequação ao problema abordado neste estudo.

3.4.1 Decision Tree

Decision Tree é um modelo de aprendizado utilizado para tarefas de classificação e regressão. Sua estrutura em forma de árvore binária permite representar decisões sequenciais baseadas nos atributos de entrada, particionando o espaço de dados em subconjuntos homogêneos com relação à variável-alvo (Han et al., 2012). Cada nó interno da árvore representa um teste condicional sobre um atributo, e cada folha contém uma decisão de classe. O processo de classificação consiste em percorrer a árvore da raiz até uma folha, seguindo os ramos de acordo com os valores dos atributos da instância.

No contexto da classificação binária, como adotado nesta dissertação, as árvores de decisão são particularmente valiosas por sua capacidade de produzir modelos altamente interpretáveis, que facilitam o entendimento das relações entre os atributos (como área, valor, região, padrão do imóvel etc.) e o resultado predito (sucesso ou insucesso de vendas). Essa característica torna as árvores úteis não apenas para predição, mas também como ferramenta de descoberta de conhecimento e explicação de padrões.

A construção da árvore ocorre de forma recursiva, aplicando critérios de partição baseados em métricas de impureza como o Ganho de Informação, a Razão de Ganho ou o Índice de Gini. O objetivo é maximizar a pureza dos subconjuntos gerados a cada divisão, reduzindo a heterogeneidade da variável-alvo.

Ao final da construção, a árvore pode ser podada (*pruning*) para reduzir o risco de *overfitting*, removendo subdivisões que não apresentam ganho estatístico significativo. Isso garante maior generalização em dados não vistos.

3.4.2 Random Forest

Random Forest é um algoritmo de aprendizado de máquina do tipo *ensemble*, desenvolvido por Breiman (2001), que combina o resultado de múltiplas árvores de decisão independentes para melhorar a precisão e a capacidade de generalização do modelo. A ideia central consiste em reduzir a variância característica das árvores individuais por meio da agregação (*bagging — bootstrap aggregating*), isto é, o treinamento de cada árvore em uma amostra aleatória do conjunto de dados, obtida com reposição.

Cada árvore é construída com um subconjunto distinto de observações e de atributos, garantindo diversidade entre os modelos individuais. Durante a etapa de predição, as saídas das árvores são combinadas por votação majoritária (no caso de classificação) ou média (em regressão), resultando em um modelo mais estável, robusto a ruídos e menos suscetível a *overfitting*.

3.4.3 Logistic Regression

A *Logistic Regression* é um dos modelos estatísticos mais utilizados em tarefas de classificação binária supervisionada, especialmente quando há interesse tanto na predição quanto na interpretação dos efeitos de cada variável sobre a probabilidade de um determinado desfecho. Ao contrário da regressão linear, que modela uma variável contínua, a regressão logística estima a probabilidade de ocorrência de um evento binário, como sucesso (classe 1) ou insucesso (classe 0), a partir de uma combinação linear dos atributos preditores.

O modelo assume a seguinte forma funcional, utilizando a função sigmoide para mapear a combinação linear de atributos para um intervalo de probabilidade [0, 1]:

Fórmula 6.1 - Regressão Logística

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

onde:

- Y é a variável-alvo binária (por exemplo, sucesso de vendas);
- X=(x1 ,x2 ,…,xk) são os atributos do empreendimento ou da unidade;
- βi são os coeficientes estimados pelo modelo.

A transformação logística (função sigmoide) assegura que os valores preditos estejam sempre no intervalo [0, 1], permitindo interpretar os resultados como **probabilidades de ocorrência** do evento de interesse. A decisão final de

classificação é obtida a partir de um limiar (tipicamente 0,5), que pode ser ajustado conforme o custo dos erros de classificação.

“A regressão logística combina simplicidade, interpretabilidade e eficiência computacional, sendo particularmente indicada para domínios regulados ou sensíveis à transparência da decisão” (Kuhn & Johnson, 2013).

3.5 TÉCNICAS PARA A AVALIAÇÃO DOS MODELOS

A avaliação rigorosa do desempenho dos modelos de classificação binária supervisionada é essencial para assegurar que as previsões geradas sejam não apenas precisas, mas também confiáveis e generalizáveis. Em problemas como o deste estudo — que busca prever o sucesso comercial de empreendimentos imobiliários — a escolha de métricas apropriadas e a adoção de estratégias de validação são fundamentais para orientar decisões baseadas em dados.

As métricas mais utilizadas para avaliar classificadores binários são derivadas da matriz de confusão, que resume as previsões do modelo em quatro categorias:

- Verdadeiros Positivos (VP): instâncias corretamente classificadas como pertencentes à classe positiva (sucesso);
- Falsos Positivos (FP): instâncias incorretamente classificadas como positivas;
- Verdadeiros Negativos (VN): instâncias corretamente classificadas como negativas (insucesso);
- Falsos Negativos (FN): instâncias incorretamente classificadas como negativas.

Com base nesses elementos, são calculadas as seguintes métricas (Han et al., 2012; Provost & Fawcett, 2013):

- Accuracy (Acurácia): proporção de classificações corretas (positivas e negativas) sobre o total de instâncias.

$$Accuracy = (VP + VN) / (VP + FP + VN + FN)$$

- Precision (Precisão): proporção de acertos entre as instâncias classificadas como positivas.

$$Precision = VP / (VP + FP)$$

- Recall (Sensibilidade): capacidade do modelo de identificar corretamente os casos positivos.

$$Recall = VP / (VP + FN)$$

- F1-Score: média harmônica entre *precision* e *recall*, utilizada quando há necessidade de balancear ambos os critérios.

$$F1\text{-}Score = 2 * (Precision * Recall) / (Precision + Recall)$$

- Área sob a Curva ROC (AUC-ROC): avalia a capacidade do modelo de distinguir entre as classes ao variar o limiar de decisão. É especialmente útil em casos de desbalanceamento entre as classes.

3.6 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (X-AI)

A crescente adoção de algoritmos de aprendizado de máquina em domínios sensíveis, como saúde, finanças e mercado imobiliário, trouxe à tona a necessidade de compreender, auditar e confiar nas decisões tomadas por modelos de inteligência artificial. Nesse contexto, surge o campo da Inteligência Artificial Explicável (X-AI – *Explainable Artificial Intelligence*), cujo objetivo é tornar transparentes os processos internos e os critérios de decisão de modelos preditivos, especialmente aqueles com alto poder de modelagem, mas baixa interpretabilidade.

Enquanto modelos lineares e árvores de decisão são naturalmente compreensíveis, algoritmos mais complexos — como *Random Forest*, *Gradient Boosting* ou Redes Neurais — tendem a se comportar como "caixas-pretas", dificultando a explicação direta dos resultados. A X-AI busca mitigar esse desafio por meio de técnicas que revelem quais atributos influenciaram cada predição e em que magnitude, promovendo confiança e responsabilidade nos sistemas automatizados.

3.6.1 Abordagens globais e locais

As técnicas de X-AI podem ser divididas em duas categorias principais (Guidotti et al., 2018):

- **Explicações globais**: revelam o comportamento geral do modelo em todo o conjunto de dados, como a importância média dos atributos ou as interações globais entre variáveis.

- **Explicações locais**: explicam predições específicas, apontando por que uma determinada instância foi classificada como sucesso ou insucesso, com base em seus atributos individuais.

3.6.2 Técnicas utilizadas

Neste trabalho, foram utilizadas abordagens amplamente reconhecidas como:

- **SHAP (*SHapley Additive ExPlanations*)**: baseia-se em teoria dos jogos cooperativos para calcular a contribuição de cada atributo para a predição de um modelo. Para cada instância, os valores SHAP indicam se um atributo aumentou ou diminuiu a probabilidade de pertencer à classe positiva. Além disso, os gráficos SHAP *summary* permitem avaliar a importância dos atributos em nível global.
- **PFI (Permutation Feature Importance)**: é um método de explicabilidade global que mede a importância de cada atributo com base na variação do desempenho do modelo quando seus valores são aleatoriamente embaralhados (permutados). A lógica é simples: se a permutação de um atributo provoca uma queda significativa na acurácia ou no coeficiente de determinação (R^2), isso indica que o atributo possui alta relevância para as predições. O PFI é amplamente utilizado por ser modelo-agnóstico, ou seja, pode ser aplicado a qualquer tipo de algoritmo, e fornece uma estimativa direta da dependência entre o desempenho do modelo e a informação contida em cada variável.
- **LIME (*Local Interpretable Model-Agnostic Explanations*)**: proposto por Ribeiro et al. (2016), é uma das abordagens mais populares para explicação local de modelos de aprendizado de máquina. O princípio fundamental do LIME é simples: para explicar a predição de um modelo complexo em uma instância específica, ele ajusta um modelo interpretável (como uma regressão linear ou árvore rasa) localmente ao redor daquela instância.

4 METODOLOGIA DO PROJETO

Este capítulo descreve a metodologia empregada na construção do modelo preditivo para avaliação do desempenho comercial de empreendimentos imobiliários na cidade do Recife. A abordagem adotada se fundamenta na aplicação sistemática do modelo de processo **CRISP-DM (*Cross Industry Standard Process for Data Mining*)**, reconhecido por sua flexibilidade, robustez e ampla aceitação tanto na academia quanto no setor produtivo (Chapman et al., 2000).

O processo metodológico percorreu todas as fases do CRISP-DM, desde o entendimento do domínio de negócio até a preparação final dos dados para os modelos de aprendizagem de máquina supervisionada, com foco na classificação binária. Foram utilizados dados reais do mercado imobiliário recifense, disponibilizados pela plataforma **RE.AI.s**, abrangendo características físicas e comerciais das unidades habitacionais, atributos dos empreendimentos, dados históricos de vendas e variáveis econômicas contextuais.

A definição das classes-alvo de sucesso comercial foi orientada por dois critérios complementares: **Velocidade de Vendas** (com foco no número de vendas de unidades nos três primeiros meses após o lançamento comercial) e **Resiliência de Vendas** (associada à sustentabilidade das vendas ao longo de 18 meses de comercialização). Essa modelagem dual permitiu uma análise mais abrangente e alinhada à dinâmica do ciclo de vida dos empreendimentos.

As subseções a seguir detalham cada uma das fases metodológicas do projeto, iniciando com a caracterização do domínio do problema, seguido pelo modelo conceitual da solução proposta, ambiente de experimentação e, por fim, a aplicação da metodologia CRISP-DM com suas respectivas fases.

4.1 CONTEXTO DO DOMÍNIO

O mercado imobiliário desempenha papel central na dinâmica econômica e social das cidades, sendo responsável por grandes volumes de investimento, geração de empregos e formação de patrimônio para famílias e empresas. No Brasil, esse setor se caracteriza por uma forte assimetria de informações, baixa padronização na coleta de dados e uma cultura decisória fortemente ancorada em

intuição, experiência empírica e análise retrospectiva pouco sistematizada. Em particular, o contexto de **Recife** — capital do estado de Pernambuco — apresenta desafios típicos de mercados urbanos intermediários: elevado grau de informalidade, flutuações cíclicas acentuadas e presença de múltiplos agentes atuando com graus variados de sofisticação analítica. A previsão do desempenho de vendas de empreendimentos, em particular, representa um desafio significativo devido à sua natureza cíclica, à heterogeneidade dos produtos e à influência de tendências locais e regionais.

Neste cenário, a incorporação de **metodologias de Inteligência Artificial e Mineração de Dados** oferece uma oportunidade concreta de transformação. Ao explorar dados históricos de empreendimentos residenciais lançados, atributos físicos e comerciais das unidades, e variáveis contextuais (econômicas e geográficas), torna-se possível antecipar padrões de comportamento do mercado, estimar o sucesso de vendas de novos empreendimentos e apoiar decisões estratégicas de incorporação e comercialização com maior robustez analítica.

O foco deste estudo está na **classificação binária supervisionada** de unidades habitacionais quanto ao sucesso de vendas, com base em dois critérios complementares: **Velocidade de Vendas** e **Resiliência de Vendas**. A granularidade da análise, conduzida no nível da unidade (e não apenas do empreendimento), permite captar variabilidades sutis e maximizar o poder preditivo dos modelos, além de aumentar a capacidade explicativa sobre os fatores determinantes do sucesso comercial.

A escolha do município de Recife como objeto de estudo deve-se à disponibilidade de uma base de dados consolidada, com cobertura histórica e riqueza de atributos, fornecida pela plataforma **RE.AI.s – Real Estate Artificial Intelligence Systems**, especializada em inteligência de mercado imobiliário. Essa plataforma reúne dados de mais de 300 empreendimentos e 28 mil unidades, com registros mensais de vendas no período de janeiro de 2021 a maio de 2025, permitindo a construção de análises com base empírica sólida e atualizada.

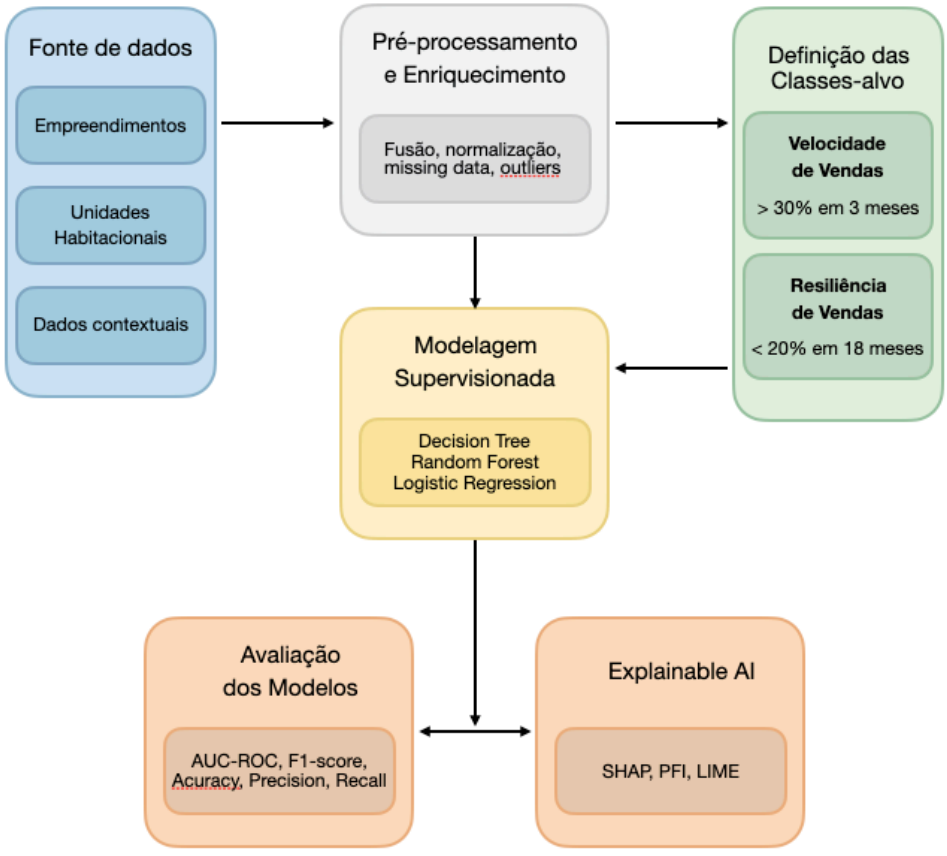
Com isso, este trabalho se insere em um esforço mais amplo de **aplicação da ciência de dados ao setor imobiliário brasileiro**, buscando contribuir com métodos replicáveis, baseados em dados reais, para a avaliação de projetos imobiliários residenciais em mercados urbanos semelhantes ao de Recife.

4.2 MODELO CONCEITUAL

O modelo conceitual deste estudo representa a arquitetura lógica que conecta os objetivos de negócio às etapas do processo de mineração de dados, detalhando como as fontes de informação, as variáveis relevantes e as operações de transformação se articulam para viabilizar a construção de modelos preditivos explicáveis no contexto do mercado imobiliário.

A figura a seguir ilustra a estrutura conceitual proposta, ancorada nos seguintes componentes principais:

Figura 4.1 - Modelo conceitual do estudo



Fonte: O autor (2025)

Fontes de Dados: Integra três categorias de dados:

Empreendimentos: atributos agregados dos projetos residenciais, como data de lançamento, número de unidades, construtora responsável e séries mensais de vendas.

Unidades Habitacionais: características físico-comerciais como área privativa, valor do m², padrão do imóvel, andar, posição e status de venda.

Dados Contextuais: indicadores econômicos (Selic, IPCA, Dólar, IGPM, INCC) e atributos espaciais (bairro, região geográfica, zoneamento).

Pré-processamento e Enriquecimento:

Fusão dos datasets via identificadores comuns (ex: id_empreendimento);

Tratamento de dados faltantes (inclusive MNAR), dados atípicos (*outliers)* e normalização;

Criação de atributos derivados como *região*, *padrão do imóvel*, *tempo de venda*, entre outros.

Definição das Classes-Alvo:

Velocidade de Vendas: classifica como “bem-sucedido” empreendimentos que venderam ≥30% das unidades nos três primeiros meses após o lançamento.

Resiliência de Vendas: considera como “bem-sucedido” empreendimentos com ≤20% de estoque remanescente após 18 meses.

Modelagem Supervisionada:

Aplicação de algoritmos de classificação binária, com ênfase em Decision Tree, Random Forest e Logistic Regression;

Adoção de estratégias como ajuste de hiperparâmetros via GridSearch, validação cruzada estratificada e tratamento de desbalanceamento de classes.

Explanação dos Modelos (X-AI):

Uso de técnicas como SHAP, LIME e PFI para garantir a interpretabilidade dos modelos, permitindo que especialistas compreendam e validem os fatores que mais influenciam o sucesso de vendas.

Avaliação e Aplicação:

Métricas como AUC-ROC, *F1-score*, *acuracy*, *precision*, *recall* são utilizadas para aferir o desempenho;

O modelo é proposto como ferramenta de apoio à tomada de decisão para incorporadoras e analistas de viabilidade imobiliária.

A estrutura conceitual conecta o conhecimento de domínio imobiliário com ferramentas avançadas de análise preditiva, buscando transformar dados históricos e contextuais em inteligência aplicável à avaliação de novos empreendimentos.

4.3 IMPLEMENTAÇÃO, FERRAMENTAS E AMBIENTE EXPERIMENTAL

A implementação deste projeto de mineração de dados foi concebida para garantir **rastreabilidade, reprodutibilidade e eficiência computacional** em todas as etapas, desde o tratamento dos dados brutos até a geração dos modelos preditivos explicáveis. Esta seção descreve os principais recursos computacionais utilizados, bem como a infraestrutura lógica e tecnológica que sustentou os experimentos desenvolvidos.

4.3.1 Linguagem e ambiente de programação

A linguagem de programação adotada foi o **Python**, amplamente reconhecida por sua flexibilidade e abrangência no ecossistema de ciência de dados. Todo o processamento, modelagem e explicação dos dados foram realizados utilizando o **Google Colab**, um ambiente de *notebooks* baseado em nuvem que permite experimentação interativa, reprodutibilidade e facilidade de compartilhamento.

4.3.2 Bibliotecas para manipulação e visualização de dados

- **Pandas**: Essencial para a manipulação de estruturas tabulares, limpeza de dados, tratamento de dados faltantes, criação de atributos e operações vetoriais.
- **NumPy**: Utilizada para operações numéricas de baixo nível e suporte a arrays multidimensionais.
- **Matplotlib** e **Seaborn**: Empregadas na geração de gráficos estatísticos, histogramas, boxplots e curvas ROC para análise exploratória e avaliação dos modelos.

4.3.3 Modelagem preditiva

- **Scikit-learn**: A principal biblioteca utilizada para modelagem de aprendizado supervisionado, incluindo algoritmos como *Decision Tree, Random Forest* e *Logistic Regression*, além de ferramentas para validação cruzada, métricas de avaliação e *pipelines* de transformação.

4.3.4 Fontes de dados

- **RE.AI.s – Real Estate Artificial Intelligence Systems**: Plataforma de inteligência de mercado imobiliário utilizada como principal fonte de dados. Forneceu os registros históricos de lançamentos, vendas e características físicas das unidades habitacionais na cidade do Recife entre janeiro de 2021 e maio de 2025.
- **Indicadores Econômicos**: Dados públicos extraídos de fontes como Banco Central do Brasil (Selic, Dólar), IBGE (IPCA) e FGV (IGPM e INCC), utilizados para contextualizar o cenário macroeconômico no momento da venda das unidades.
- **Prefeitura do Recife – Dados Geográficos**: Dados utilizados para o mapeamento de bairros em regiões administrativas, permitindo a criação do atributo derivado “região” com base na localização geográfica dos empreendimentos.

4.3.5 Organização e documentação

Todo o código-fonte, *pipelines* de transformação, notebooks e bases de dados foram organizados em repositórios versionados com o objetivo de garantir a **reprodutibilidade** da pesquisa. Além disso, os experimentos e análises foram conduzidos com **registro detalhado de etapas**, facilitando auditoria por terceiros e extensão futura dos estudos. O **Apêndice A** contém as informações sobre o repositório onde podem ser acessados os arquivos dessa dissertação.

4.4 PREPARAÇÃO DOS DADOS SEGUNDO O CRISP-DM

Este capítulo descreve como cada fase do CRISP-DM foi operacionalizada no contexto do mercado imobiliário recifense. Primeiro, em 4.4.1 Entendimento do Negócio, traduzimos os objetivos estratégicos da incorporação imobiliária em metas tangíveis de predição. Em seguida, 4.4.2 Entendimento dos Dados apresenta as fontes, a análise exploratória e as estatísticas descritivas que sustentam o diagnóstico inicial.

A seção 4.4.3 Preparação dos Dados — núcleo deste capítulo — detalha: (a) critérios de seleção de registros; (b) tratamento de dados faltantes e dados atípicos; (c) integração de fontes heterogêneas; (d) engenharia de atributos; e (e) transformações de setorização e normalização que antecedem a modelagem. Cada sub-etapa adota práticas recomendadas da literatura e know-how de domínio para maximizar completude e minimizar viés.

Por fim, 4.5 Métricas e Critérios de Avaliação fecha o ciclo, definindo indicadores quantitativos que guiarão a validação dos modelos nas fases posteriores de modelagem e implantação.

4.4.1 Entendimento do negócio

A fase de Entendimento do Negócio, conforme delineada pela metodologia CRISP-DM (Chapman et al., 2000), representa o ponto de partida fundamental em qualquer projeto de mineração de dados. Sua importância reside na capacidade de aprofundar a compreensão dos desafios e oportunidades de um determinado contexto de negócio, permitindo a identificação de problemas que mereçam ser explorados e que possam gerar valor significativo.

Para a presente dissertação, essa etapa é particularmente crucial. Isso se deve à complexidade inerente ao mercado imobiliário e às significativas deficiências em suas fontes de informação, que são caracterizadas pela fragmentação, heterogeneidade, falta de qualidade, pouca estruturação e, por vezes, indisponibilidade dos dados. Tal desafio, já reconhecido globalmente (Li & Fan, 2021), manifesta-se de forma ainda mais pronunciada e afeta diretamente o nosso mercado local.

Um dos maiores desafios enfrentados pelas incorporadoras é o risco de lançar um produto que não seja bem aceito pelo mercado, resultando em fracasso de vendas ou em um longo período de comercialização, o que já comprometeu a viabilidade de diversas empresas**.** A incerteza da demanda e a taxa de absorção de unidades representam um dos maiores riscos em projetos de desenvolvimento imobiliário, sendo um fator crítico que impacta diretamente a viabilidade financeira e a sobrevivência das incorporadoras no setor.

Neste contexto, o presente trabalho busca **analisar a influência dos atributos no resultado das vendas de apartamentos em empreendimentos no Recife através da utilização da mineração de dados e inteligência artificial, com a metodologia CRISP-DM, técnicas de aprendizagem de máquina e algoritmos de inteligência artificial explicável (X-AI)**. A partir de uma formulação clara das hipóteses e objetivos a serem testados, esta fase direciona todo o processo subsequente, desde a identificação das informações necessárias e sua origem até a preparação dos dados e a construção de modelos capazes de gerar percepções acionáveis e validáveis para o mercado imobiliário.

4.4.1.1 Objetivos de negócio

Em consonância com a metodologia CRISP-DM, projetos baseados em mineração de dados e inteligência artificial devem ser, primordialmente, guiados pelas necessidades concretas do negócio. Nesse sentido, esta seção visa explicitar os objetivos estratégicos deste estudo para o mercado imobiliário primário da cidade do Recife.

No contexto do mercado imobiliário de Recife, um dos desafios mais críticos enfrentados pelas incorporadoras reside na predominância de decisões empíricas e subjetivas para o desenvolvimento de novos empreendimentos. A escolha da localização e das características (atributos) do produto é frequentemente baseada na experiência individual e em análises históricas estáticas, desconsiderando a dinâmica e as premissas mutáveis do mercado. Embora pesquisas de mercado tradicionais sejam empregadas para mitigar riscos, estas são onerosas, geograficamente limitadas e capturam apenas uma intenção pontual do cliente,

tornando-se suscetíveis à desatualização em ciclos longos e complexos (Li & Fan, 2021).

A consequência direta dessa lacuna em ferramentas de apoio à decisão é o lançamento de produtos desalinhados com a demanda efetiva do mercado. Tal desalinhamento sobrecarrega o setor comercial, dificulta as estratégias de marketing e comunicação, acarreta atrasos nas vendas, necessidade de concessão de descontos não previstos, alongamento dos prazos médios de comercialização e aumento dos custos operacionais e financeiros que, em última instância, culmina em resultados financeiros inferiores ou negativos para os parceiros. Adicionalmente, mesmo para empreendimentos bem-sucedidos, a ausência de um acompanhamento dinâmico do mercado e da precificação dos concorrentes pode levar à subotimização de preços, resultando em receitas potenciais não realizadas.

Com base no problema de negócio central identificado, a presente pesquisa visou alcançar os seguintes objetivos específicos, que se traduzem em valor acionável para os *players* do mercado imobiliário:

(i) Capacitar a previsão da performance comercial de novos empreendimentos

Permitir que incorporadoras prevejam, por meio de uma classificação binária, o sucesso ou insucesso de um projeto imobiliário antes do seu lançamento sob critérios previamente estabelecidos e validados para o contexto do mercado imobiliário primário. Essa previsão será baseada em critérios objetivos de velocidade e consistência de vendas - como, por exemplo, atingir um percentual mínimo de vendas em determinado horizonte temporal ou reduzir o estoque disponível para um patamar inferior a um limite previamente definido para um intervalo de tempo - visando mitigar riscos associados ao lançamento de produtos desalinhados com a real demanda do mercado.

(ii) Subsidiar a otimização estratégica de produtos imobiliários com base em insights explicáveis

Oferecer recomendações estratégicas para o ajuste de atributos e características dos empreendimentos (como tipologia, mix de unidades, preço e localização), visando maximizar a chance de sucesso comercial e a resiliência de vendas. A aplicabilidade de técnicas de Inteligência Artificial Explicável (X-AI) é particularmente relevante no mercado imobiliário. Isso se deve ao fato de que as

decisões nesse setor envolvem investimentos de grande porte, horizontes temporais longos e elevada exposição ao risco financeiro. Proporcionando assim, maior confiança aos tomadores de decisão sobre as mudanças propostas e as razões por trás das previsões de sucesso ou insucesso.

(iii) Promover a tomada de decisão orientada por dados e estabelecer um framework analítico replicável e transparente

Incentivar uma cultura de decisão baseada em dados no desenvolvimento de novos produtos imobiliários. Além disso, estabelecer um *framework* de análise preditiva que, por meio da interpretabilidade oferecida pelo X-AI, possa ser replicado e adaptado para a avaliação de futuros projetos e mercados, contribuindo para a profissionalização e institucionalização de processos de análise preditiva e inteligência de dados no mercado imobiliário.

A articulação desses objetivos de negócio não apenas aborda as lacunas atualmente observadas no processo decisório de lançamento de empreendimentos residenciais na cidade do Recife, mas também estabelece as bases para um modelo analítico robusto, transparente e orientado a valor.

4.4.1.2 Avaliação da situação atual

(i) Recursos Disponíveis

Para a realização deste estudo, o recurso mais crítico e diferenciado é a base de dados provida pela plataforma **RE.AI.s – Real Estate Artificial Intelligence Systems**. Originada há mais de uma década como um CRM imobiliário, a RE.AI.s evoluiu para uma plataforma de inteligência de dados, consolidando um *datalake* robusto a partir de um vasto ecossistema de participantes. Com uma rede nacional que inclui mais de 760 construtoras, 1500 imobiliárias e 8.600 corretores – e, especificamente em Recife, 235 construtoras, 321 imobiliárias e 1913 corretores –, a plataforma capta informações sobre ofertas de empreendimentos, imóveis usados, bem como registra um volume massivo de interações e contatos de milhares de clientes em diversos canais digitais — incluindo portais imobiliários como OLX, Zap Imóveis, Viva Real, Expoimovel e Creci, além de plataformas como Google, YouTube, Facebook, Instagram e TikTok. Essa capilaridade a estabelece como uma

fonte inigualável de informações abrangentes e granulares para o mercado imobiliário primário.

É crucial destacar que todos os dados cadastrados, atualizados e utilizados pela plataforma, e consequentemente por este estudo, respeitam integralmente as diretrizes e regulamentações da LGPD – Lei Geral de Proteção de Dados (Lei nº 13.709/2018), garantindo a privacidade e a segurança das informações, em alinhamento com as melhores práticas de governança de dados estabelecidas pela LGPD brasileira. Todos os dados utilizados neste estudo são provenientes de informações públicas, autorreferenciadas pelas próprias construtoras ou disponibilizadas para fins comerciais nos canais de venda. Não há qualquer coleta de dados sensíveis de pessoas físicas, garantindo a aderência às melhores práticas de governança de dados e compliance ético-acadêmico (Tallon, 2013).

A qualidade, corretude e a completude dos dados da RE.AI.s são asseguradas por um processo híbrido e rigoroso, em consonância com práticas avançadas de governança de dados aplicadas a sistemas inteligentes urbanos. A plataforma se destaca por consolidar a quase totalidade dos empreendimentos lançados na região, com alto grau de detalhe, impulsionado por uma estratégia de validação colaborativa ('Wiki') pela própria comunidade de usuários. Complementarmente, para a finalidade deste estudo, a qualidade dos dados do mercado primário é garantida pela interação direta com as construtoras através de APIs e/ou pelo trabalho de equipes de engenharia de dados, permitindo o recebimento mensal de informações sobre novos empreendimentos lançados no mercado e atualizações precisas sobre unidades vendidas, disponíveis e preços dos empreendimentos em comercialização. Os dados, então, são submetidos a um sistema de duas camadas de revisão interligadas: primeiro, algoritmos de Inteligência Artificial leem e interpretam os dados para sua atualização; e, adicionalmente, uma equipe dedicada de engenheiros de dados da RE.AI.s realiza revisões e correções contínuas, assegurando a consistência e a confiabilidade das informações.

Para o propósito específico desta dissertação, foi curado um *dataset* exclusivo derivado do *datalake* da RE.AI.s, com foco no mercado imobiliário primário da cidade do Recife. **Este conjunto de dados, abrangendo o período de janeiro de 2022 a maio de 2025, é composto por informações de 235 construtoras, 271**

empreendimentos listados, totalizando mais de 27.000 unidades lançadas e um Valor Geral de Vendas (VGV) aproximado de R$ 17,3 bilhões.

Os dados foram organizados em duas granularidades:

1. Nível de empreendimento, contendo informações como data de lançamento, estágio da obra, quantidade total de unidades, histórico de vendas mensais, estoque disponível e características físicas e de localização do empreendimento;
2. Nível de unidade habitacional, com dados individuais de cada apartamento, incluindo atributos como status de venda, preço, área, pavimento, tipologia, características físicas específicas e posição no empreendimento.

O dataset contempla mais de 150 atributos por unidade, abrangendo desde aspectos físicos até informações de mercado e contexto econômico. Para a análise proposta, foram consideradas as vendas mensais desses empreendimentos no período compreendido entre janeiro de 2022 e maio de 2025.

Tabela 4.1 - Tabela resumo dos dados

ASPECTO	DETALHE
Período	Janeiro/2022 a Maio/2025
Construtoras	235
Empreendimentos	271
Unidades habitacionais	Mais de 27.000
Valor Global de Vendas (VGV) Lançado	Aproximadamente R$ 17,3 Bilhões
Granularidade	Empreendimento e Unidade
Número de atributos	Mais de 150 atributos

Fonte - O autor (2025)

Além da base de dados da R.E.A.I.s, foram incorporadas outras fontes de informação para enriquecer a análise preditiva. Para compreender em que grau a força da marca da construtora pode influenciar o desempenho comercial dos empreendimentos (Smit & van den Heuvel, 2010) – seja pela consolidação da marca, percepção da qualidade, segurança do investimento, solidez financeira, investimento em marketing ou outros fatores exógenos –, foram coletados dados como o ano de fundação da construtora, o número total de empreendimentos

lançados e entregues, e o volume de lançamentos realizados no período específico do estudo. Esses atributos, obtidos diretamente dos sites e canais digitais e sociais oficiais de cada construtora, servem como *proxies* para avaliar a presença e a reputação da marca no mercado. Essas variáveis permitem não apenas capturar a dimensão histórica e institucional da construtora, mas também mensurar sua atividade recente no mercado, oferecendo ao modelo preditivo uma visão mais contextualizada sobre sua presença e impacto no comportamento dos consumidores.

Adicionalmente, para capturar o impacto das condições econômicas mais amplas, foram coletados, para cada mês do período de estudo, os principais indicadores macroeconômicos e de mercado imobiliário. Estes incluem a taxa SELIC, o Índice Geral de Preços do Mercado (IGP-M), o Índice Nacional de Custo da Construção (INCC), a taxa de câmbio (Dólar), e o Índice Nacional de Preços ao Consumidor Amplo (IPCA). A inclusão desses indicadores, provenientes de fontes oficiais como o Instituto Brasileiro de Geografia e Estatística (IBGE), a Fundação Getúlio Vargas (FGV) e o Banco Central do Brasil (BACEN), permite controlar os efeitos de variações macroeconômicas, garantindo maior robustez e validade aos modelos preditivos propostos. Esses indicadores visam assegurar que os modelos analíticos consigam distinguir variações nas performances comerciais que sejam atribuíveis às condições macroeconômicas, evitando a atribuição indevida de variações sazonais ou conjunturais aos atributos físicos dos empreendimentos ou às características das construtoras.

A integração dessas múltiplas fontes de dados — tanto do mercado imobiliário primário quanto de variáveis macroeconômicas e institucionais — assegura que os modelos desenvolvidos estejam ancorados em uma visão holística, precisa e contextualizada da dinâmica de mercado, permitindo não apenas predições acuradas, mas também insights robustos para a tomada de decisão estratégica no setor.

(ii) Capacidade Computacional e Ferramentas

Para a execução deste estudo, foi utilizada uma infraestrutura tecnológica robusta suportada por uma arquitetura computacional híbrida, composta por recursos locais, ambientes em nuvem e infraestrutura da plataforma ***RE.AI.S – Real***

Estate Artificial Intelligence Systems, assegurando não apenas desempenho, mas também flexibilidade, escalabilidade e segurança.

Para a construção dos datasets específicos utilizados neste estudo, foi desenvolvido um pipeline de extração eficiente, alinhado às práticas modernas de engenharia de dados e manipulação de grandes volumes de informação diretamente a partir do datalake da plataforma RE.AI.S, por meio de APIs desenvolvidas na linguagem Java, responsáveis por gerar arquivos tabulares no formato .XLSX, não normalizados, otimizados para consumo analítico. Esses arquivos foram posteriormente integrados ao ambiente de Google Sheets, com o intuito de garantir portabilidade, acessibilidade multiplataforma (MacOS, Windows e Web) e facilidade na manipulação colaborativa dos dados.

O processamento dos dados e a execução dos modelos de mineração, aprendizado de máquina e inteligência artificial explicável foram realizados em um ambiente híbrido, utilizando tanto recursos locais — um equipamento MacBook Pro com processador Apple M3 PRO, 18 GB de memória RAM e SSD de 1TB — quanto ambientes em nuvem, especificamente o Google Colab, que oferece infraestrutura baseada em Google Cloud, permitindo acesso a GPUs e maior capacidade de processamento quando necessário.

A decisão pelo uso de ferramentas de código aberto reflete não apenas uma escolha técnica, mas também um compromisso com a construção de ciência aberta, garantindo o acesso universal, a transparência, a replicabilidade e a possibilidade de reuso dos resultados por outros pesquisadores.

Essa infraestrutura computacional demonstrou-se plenamente adequada tanto para o pré-processamento do volume dos dados — que incluíram a manipulação de mais de 27.000 unidades habitacionais, distribuídas em 271 empreendimentos, com histórico de vendas mensal, de janeiro de 2022 a maio de 2025 — quanto para o treinamento, validação e explicação de modelos de *machine learning*, garantindo robustez, desempenho e escalabilidade no desenvolvimento das soluções propostas.

(iii) Recursos Humanos/Conhecimento

A execução deste estudo e o desenvolvimento das soluções propostas foram viabilizados pela confluência de experiência setorial, conhecimento técnico

aprofundado em ciência de dados e inteligência artificial, e a sólida orientação acadêmica de um corpo docente altamente qualificado.

O pesquisador é mestrando em Ciências da Computação pela Universidade Federal de Pernambuco (UFPE), com foco em Inteligência Computacional, e possui vasta experiência profissional no mercado imobiliário e no desenvolvimento de *software* de larga escala. Desde 1999, atua na concepção e construção de ferramentas e aplicações para este setor, tendo participado ativamente da criação de um dos primeiros portais imobiliários do Brasil, o Expoimovel.com, ainda como estudante de graduação na disciplina de empreendedorismo do Prof. Dr. Hermano Perrelli de Moura. Sua trajetória inclui a co-fundação e a responsabilidade pela área de desenvolvimento e ciência de dados da plataforma RE.AI.s.

Para fortalecer o *background* acadêmico diretamente aplicável a este projeto, o pesquisador cursou disciplinas essenciais durante o mestrado, sob a orientação de renomados professores do Centro de Informática (CIn) da UFPE:

- Aprendizagem de Máquinas, com o Prof. Dr. Francisco de Assis Tenório de Carvalho;
- Mineração de Dados, com Prof. Dr. Paulo Jorge Leitão Adeodato (orientador);
- Inteligência Artificial Explicável, com o Prof. Dr. Ricardo Bastos Cavalcante Prudêncio;
- Princípios e Técnicas de Análise Estatística Experimental, com a Profa. Dra. Renata Maria Cardoso Rodrigues de Souza;
- Agentes Cognitivos e Adaptativos, com o Prof. Dr. Ricardo Bastos Cavalcante Prudêncio.

A orientação acadêmica é um pilar fundamental para este trabalho, sendo conduzida pelo Prof. Dr. Paulo Jorge Leitão Adeodato, como orientador principal, e pelo Prof. Dr. Bruno Campelo de Souza, como co-orientador.

O Prof. Dr. Paulo Jorge Leitão Adeodato possui um perfil único que integra mais de 35 anos de experiência profissional com uma sólida carreira de pesquisa. Professor no Centro de Informática da UFPE há mais de 20 anos, é uma das principais referências brasileiras em sistemas de apoio à decisão, mineração de dados, inteligência computacional aplicada e empreendedorismo. Doutor por King's College London, foi *Visiting Scholar* na *Stanford Graduate School of Education* e *Visiting Professor* na *Stanford School of Medicine*. No âmbito empresarial, foi

fundador e sócio da NeuroTech, com expertise em aplicações financeiras de mineração de dados.

O Prof. Dr. Bruno Campelo de Souza, como co-orientador, contribui com uma perspectiva valiosa, sendo Professor e Coordenador do Programa de Pós-Graduação em Administração da UFPE, com doutorado em Psicologia Cognitiva. Sua expertise complementa o rigor técnico do projeto, acrescentando uma perspectiva aprofundada sobre os aspectos comportamentais e cognitivos que podem influenciar as dinâmicas de mercado e a interpretação dos resultados, especialmente em um setor complexo como o imobiliário.

A combinação da vivência prática no mercado, domínio técnico-computacional e fundamentação acadêmica de excelência do pesquisador, aliada à expertise acadêmica e de pesquisa avançada dos professores orientadores, constituem um diferencial estratégico deste estudo, garantindo que as decisões metodológicas, os experimentos conduzidos e as interpretações realizadas estejam alinhadas tanto com as necessidades reais do setor quanto com o estado da arte da ciência de dados e da inteligência artificial.

(iv) Restrições do Estudo

Este estudo concentra-se no mercado imobiliário primário da cidade do Recife, Pernambuco, especificamente em empreendimentos residenciais verticais do tipo apartamento, lançados comercialmente no período compreendido entre janeiro de 2022 e maio de 2025. Essa delimitação temporal e geográfica foi imposta por diversos fatores, incluindo:

Disponibilidade Histórica de Dados: A principal restrição temporal foi a indisponibilidade de dados de vendas detalhados por empreendimento anteriores a janeiro de 2022. Embora a plataforma RE.AI.S possua uma vasta base de dados, a ausência de informações suficientemente granulares e consistentes antes de janeiro de 2022 impôs uma limitação metodológica relevante, o que se alinha com práticas que priorizam robustez e completude em modelagem preditiva.

Completude das Características dos Empreendimentos: Dados faltantes nas características dos empreendimentos iniciais da base exigiram um processo exaustivo de revisão minuciosa e curadoria. Para assegurar a representatividade, foi rigorosamente verificado que todas as construtoras cadastradas nos principais órgãos de classe (Ademi-PE e Sinduscon-PE) com empreendimentos válidos no

período tiveram suas informações devidamente coletadas, cadastradas e revisadas mensalmente.

Requisitos de Histórico de Vendas para Modelagem: Para a inclusão na modelagem preditiva, os empreendimentos necessitavam de, no mínimo, três meses consecutivos de histórico de vendas para a construção das métricas de velocidade de comercialização, e de dezoito meses para a métrica de resiliência de vendas. Essa exigência metodológica resultou na exclusão de empreendimentos com ciclos de vendas muito recentes ou com registros incompletos nesse intervalo.

Delimitação do Escopo por Prazo e Foco: O prazo disponível para a realização da pesquisa de mestrado inviabilizou a inclusão de fatores exógenos adicionais, como distância para parques, praças, ou equipamentos públicos. Esses aspectos são reconhecidos como relevantes e serão tratados em estudos futuros.

Não Inclusão de Dados Comportamentais de Clientes: Para manter o foco no produto e na estratégia de vendas, dados comportamentais dos clientes não foram utilizados para explicar as vendas, mesmo sabendo que essa é uma informação valiosa. Essa linha de investigação, embora de grande potencial, será explorada em futuras etapas de pesquisa no doutorado.

Aplicabilidade Regional: Embora o estudo seja concentrado em Recife, a generalização dos resultados para outras regiões deve ser feita com cautela. Apesar da crença de que padrões similares possam ser observados em mercados com características análogas, a validação de tal generalização exige comprovação. Para tanto, dados de capitais do Nordeste do Brasil já estão sendo coletados para um estudo comparativo mais aprofundado.

(v) Premissas do Estudo

As seguintes premissas foram adotadas para a viabilidade e a interpretação dos resultados deste estudo.

Representatividade e Qualidade dos Dados: Presume-se que os dados de vendas, as características dos empreendimentos e os atributos das construtoras, após o processo de curadoria e tratamento, são rigorosamente coletados e revisados mensalmente, e possuem a qualidade e a representatividade necessárias para a construção de modelos preditivos válidos para o mercado imobiliário de Recife.

Validade das Métricas de Desempenho:

- A velocidade de vendas (definida como, por exemplo, ≥30% de vendas em até 3 meses) é considerada um *proxy* válido para a atratividade comercial imediata do empreendimento e para sua capacidade de geração de caixa no início do ciclo de obras.
- A resiliência de vendas (definida como, por exemplo, ≤20% de unidades não vendidas após 18 meses) representa a solidez da estratégia comercial e a aderência do produto ao longo do tempo, mesmo após o período de lançamento.

Padrões Históricos como Preditor: Assume-se que os padrões históricos de desempenho comercial observados são representativos e preditivos para o comportamento de novos empreendimentos com atributos similares no mesmo mercado, conforme discutido em estudos de modelagem urbana.

Relevância dos Atributos das Construtoras: É premissa que o entendimento sobre os atributos das construtoras é relevante para explicar a força da marca e sua influência nas vendas dos empreendimentos.

Influência de Fatores Macroeconômicos: Assume-se que a influência de fatores macroeconômicos, como variações de inflação, custo da construção e taxa de juros, é relevante para explicar variações de vendas e, portanto, sua inclusão na modelagem é justificada e contribui para a robustez dos modelos (Himmelberg et al., 2005). Essa premissa implica que, exceto por mudanças abruptas no cenário macroeconômico (que poderiam afetar a validade dos modelos), o ambiente econômico subjacente permite a aplicação dos padrões identificados.

Adequação das Ferramentas e Linguagens: As ferramentas e linguagens de programação utilizadas (Python e suas bibliotecas, Google Colab) foram consideradas plenamente adequadas e suficientes para as análises, o processamento de dados e a execução dos modelos preditivos, não apresentando limitações técnicas que comprometessem o andamento metodológico do estudo.

Essas restrições e premissas foram cuidadosamente consideradas na definição da metodologia e na interpretação dos resultados, conferindo transparência, reprodutibilidade e contextualização científica ao presente estudo.

4.4.1.3 Objetivos de mineração de dados

Tabela 4.2 - Resumo das hipóteses e experimentos do estudo

Dimensão	Experimento 1: Velocidade de Vendas	Experimento 2: Resiliência de Vendas
Objetivo	Predizer se o empreendimento vendeu ≥30% das unidades nos 3 primeiros meses	Predizer se o empreendimento teve ≤20% de estoque remanescente após 18 meses
Classe Sucesso (1)	≥30% de vendas em até 3 meses	≤20% de estoque após 18 meses
Classe Insucesso (0)	<30% de vendas em até 3 meses	>20% de estoque após 18 meses
Horizonte de Observação	3 meses após o lançamento	18 meses após o lançamento
Aplicabilidade de Negócio	Liquidez inicial, planejamento de lançamento e marketing	Gestão de estoque, precificação de longo prazo
Tipo de Problema	Classificação supervisionada binária	Classificação supervisionada binária
Variável-Alvo	Velocidade de vendas	Resiliência de vendas
Interpretação Esperada	Identificação de empreendimentos com alta demanda imediata	Identificação de empreendimentos com bom desempenho ao longo do tempo

Fonte: O autor (2025)

Esta seção tem como finalidade traduzir os objetivos de negócio anteriormente definidos em metas operacionais e técnicas, que guiarão as fases subsequentes de preparação, modelagem e avaliação dos dados. Essa transposição conceitual é essencial para estabelecer com precisão o tipo de problema de aprendizado supervisionado envolvido, as tarefas computacionais associadas, bem como os critérios de avaliação e os requisitos metodológicos que guiarão o desenvolvimento analítico para que contribuam efetivamente para a tomada de decisões estratégicas. (HAN et al., 2011)

4.4.1.3.1 Formulação geral do problema de mineração de dados

O objetivo deste estudo é desenvolver modelos para classificar empreendimentos imobiliários como ‘bem-sucedidos’ ou ‘não bem-sucedidos’ com base em informações disponíveis dos atributos dos empreendimentos e de suas unidades habitacionais através de uma abordagem de classificação binária. A particularidade deste trabalho reside na definição de dois experimentos independentes, baseados em critérios objetivos e distintos de sucesso, com a construção de uma variável-alvo binária específica para cada um.

O objetivo principal deste estudo de mineração de dados desdobra-se em duas vertentes, cada uma correspondendo a um experimento de classificação binária distinto:

Experimento 1: Predição do Sucesso Baseado na Velocidade de Vendas.

- **Objetivo**: Desenvolver um modelo de classificação binária capaz de predizer o sucesso comercial de novos empreendimentos imobiliários em Recife, onde "sucesso" é definido pela velocidade de vendas. Onde um empreendimento é bem-sucedido (classe = 1) se atingir **≥30% das unidades vendidas nos primeiros 3 meses** após o lançamento; caso contrário, é classificado como mal-sucedido (classe = 0)
- **Conexão com o Negócio**: Este modelo visa apoiar construtoras e incorporadoras na identificação precoce de empreendimentos com alto potencial de liquidez inicial, otimizando estratégias de lançamento e alocação de recursos.

Experimento 2: Predição do Sucesso Baseado na Resiliência de Vendas.

- **Objetivo**: Desenvolver um modelo de classificação binária capaz de predizer o sucesso comercial de novos empreendimentos imobiliários em Recife, onde "sucesso" é definido pela resiliência de vendas. Onde um empreendimento é bem-sucedido (classe = 1) se possuir **≤20% de estoque remanescente após 18 meses** do lançamento; caso contrário, é considerado mal-sucedido (classe = 0).
- **Conexão com o Negócio**: Este modelo busca auxiliar na avaliação da aderência do produto ao longo do tempo e na solidez da estratégia comercial, permitindo decisões mais assertivas sobre precificação e gestão de estoque a médio e longo prazo.

Todo o processo de análise, incluindo preparação de dados, modelagem e avaliação, será conduzido separadamente para cada um desses experimentos, permitindo uma análise comparativa posterior dos resultados obtidos sob cada critério de sucesso.

4.4.1.3.2 Tipo de tarefa de mineração de dados

O problema de modelagem em ambos os experimentos é do tipo **classificação supervisionada binária** (ALPAYDIN, 2020). A tarefa consiste em aprender, a partir de dados históricos rotulados, um modelo capaz de prever a probabilidade de sucesso comercial de um novo empreendimento, dado um conjunto de atributos conhecidos.

A abordagem adotada é adequada, pois:

- O atributo-alvo é binário (sucesso ou insucesso);
- As variáveis explicativas são majoritariamente categóricas e numéricas, incluindo atributos de empreendimento, construtora, localização e indicadores macroeconômicos;
- O foco está na predição acurada e interpretável para tomada de decisão estratégica.

4.4.1.3.3 Objetivos secundários de mineração de dados

Além dos objetivos principais de predição, este estudo busca alcançar os seguintes objetivos secundários, que são comuns a ambos os experimentos e visam enriquecer a compreensão do mercado:

[O1] Desenvolver modelos de classificação binária para prever o sucesso comercial de empreendimentos com base em:

- Atributos do empreendimento (tipologia, metragem, preço médio, estágio da obra);
- Atributos da construtora (experiência, volume de lançamentos, histórico de entregas);
- Condições de mercado (mês do lançamento, indicadores econômicos como IGP-M, INCC, IPCA, SELIC etc.).

[O2] Comparar os dois critérios de sucesso, avaliando o impacto de cada definição sobre a performance dos modelos e a seleção de variáveis explicativas.

[O3] Identificar os atributos mais relevantes para a classificação correta das instâncias por meio de técnicas de interpretabilidade e importância de variáveis como SHAP, feature importance, árvores de decisão explicativas.

[O4] Avaliar o desempenho dos modelos preditivos com base em métricas de classificação como:

- AUC-ROC;
- F1-Score;
- Accuracy;
- - Precision;
- - Recall.

[O5] Fornecer insights explicáveis e acionáveis para profissionais do setor imobiliário, alinhados com o conceito de Inteligência Artificial Explicável (X-AI), garantindo transparência, confiabilidade e aplicabilidade prática dos resultados preditivos (Guidotti et al., 2018).

4.4.1.3.4 Considerações finais

A decisão de estruturar dois experimentos distintos permite analisar o comportamento do mercado sob diferentes perspectivas temporais. Enquanto a **velocidade de vendas** reflete o apelo inicial do produto, a **resiliência de vendas** está associada à capacidade sustentada de comercialização ao longo do tempo. Essa abordagem comparativa permitirá uma análise mais rica e aplicável dos fatores que explicam o sucesso comercial de empreendimentos, servindo como base para futuras pesquisas mais generalistas ou em outras regiões geográficas.

4.4.1.4 Plano do projeto - Resumo

O presente estudo adota a metodologia **CRISP-DM (Cross-Industry Standard Process for Data Mining)** como estrutura guia para a execução das atividades de pesquisa e desenvolvimento.

Este projeto tem como objetivo geral desenvolver modelos de classificação binária para prever o sucesso comercial de empreendimentos imobiliários na cidade do Recife. A iniciativa parte da constatação, evidenciada na seção de objetivos de

negócio, de que decisões críticas de lançamento ainda são majoritariamente empíricas no setor imobiliário, e visa contribuir com inteligência preditiva orientada por dados.

Do ponto de vista técnico, a tarefa de mineração de dados foi formalizada como um problema de **classificação supervisionada binária**, estruturado em dois experimentos independentes:

Experimento 1: predizer se um empreendimento será bem-sucedido com base na **velocidade de vendas** (≥30% das unidades vendidas nos três primeiros meses);

Experimento 2: predizer se um empreendimento será bem-sucedido com base na **resiliência de vendas** (≤20% de estoque remanescente após 18 meses).

Cada experimento será conduzido de forma autônoma, contemplando as fases de preparação dos dados, construção da variável-alvo, modelagem, avaliação e interpretação dos resultados, permitindo comparações estruturadas entre os dois critérios.

O escopo do projeto está restrito a empreendimentos **residenciais verticais do tipo apartamento**, lançados no **mercado primário de Recife** entre **janeiro de 2022 e maio de 2025**, e que possuam registros completos de atributos e histórico de vendas. Dados de comportamento do cliente, bem como fatores exógenos espaciais (como distância a praças ou escolas), serão deliberadamente excluídos deste escopo, conforme critérios definidos em Restrições e Premissas, e poderão ser abordados em um posterior doutorado.

4.4.1.4.1 Visão geral das fases do projeto

O plano de projeto para a obtenção dos objetivos de mineração de dados previamente definidos compreende as seguintes fases principais:

(i) Compreensão dos Dados:

(ii) Preparação dos Dados:

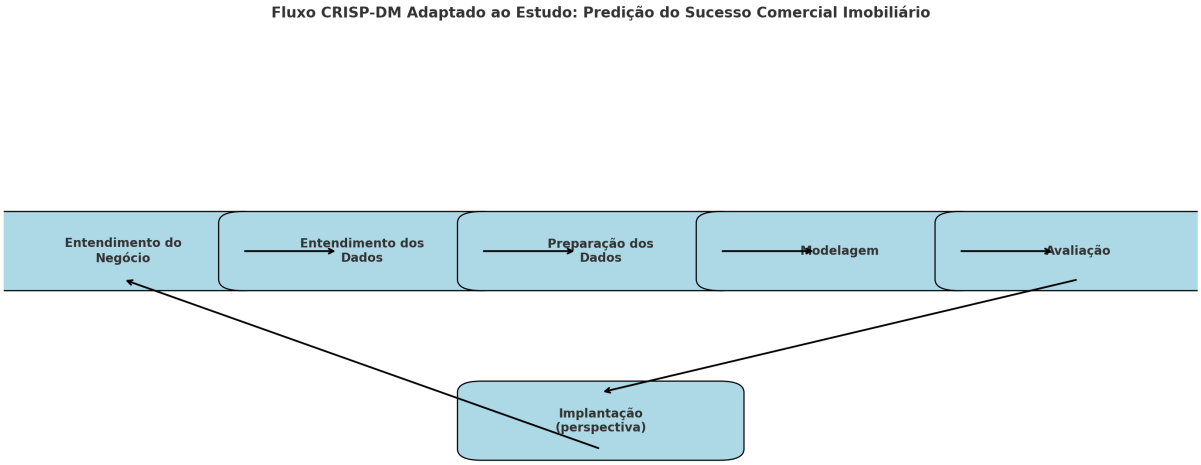
(iii) Modelagem:

(iv) Avaliação:

(v) Implantação (Perspectiva de Aplicação):

A figura 4.2 apresenta o fluxograma adaptado da metodologia CRISP-DM que resume o plano do projeto aplicado nesta dissertação. O diagrama explicita as principais fases, entregas e a abordagem analítica adotada em cada experimento de classificação binária, reforçando a conexão entre os objetivos de negócio e as decisões técnicas implementadas.

Figura 4.2 - Fluxo do CRISP-DM



Fonte: O autor

4.4.1.4.2 Resultados esperados

Ao final deste estudo, espera-se entregar:

- **Modelos Preditivos:** Modelos de classificação binária validados para prever o sucesso comercial de empreendimentos imobiliários com base em velocidade e resiliência de vendas.
- ***Insights* Acionáveis:** Identificação dos atributos mais relevantes que influenciam o desempenho de vendas, oferecendo subsídios para o desenvolvimento de produtos e estratégias comerciais.
- **Conhecimento Científico:** Contribuição para a literatura sobre a aplicação de Ciência de Dados e IA no mercado imobiliário, com foco na predição de sucesso comercial e na interpretabilidade dos modelos.

O sucesso do projeto será avaliado com base em dois eixos complementares:

1. **Eixo técnico:** desempenho dos modelos em termos de capacidade preditiva, robustez estatística e interpretabilidade;
2. **Eixo de negócio:** geração de valor para o setor por meio de *insights* claros e acionáveis sobre os fatores que influenciam o sucesso de um empreendimento.

A estruturação deste plano proporciona clareza quanto ao escopo, objetivos, abordagem metodológica e critérios de sucesso do projeto, sendo fundamental para garantir a consistência e a rastreabilidade de todas as etapas seguintes da metodologia CRISP-DM.

Tabela 4.3 - Estrutura de entregas por fase do CRISP-DM

Fase	Entregas Principais
Entendimento do Negócio	Problema definido; critérios de sucesso formulados
Entendimento dos Dados	Descrição da base da RE.AI.s; análise exploratória; verificação de consistência e completude
Preparação dos Dados	Criação das variáveis-alvo binárias; seleção de atributos; codificação categórica; padronização
Modelagem	Treinamento de modelos supervisivos (árvore, random forest, XGBoost); tuning de hiperparâmetros
Avaliação	Métricas de desempenho (AUC, F1, precisão, revocação); comparação entre experimentos
Interpretação e Ação	Aplicação de X-AI; insights para decisões de produto e estratégia; recomendações ao setor
Documentação Final	Dissertação final com metodologia, experimentos, análises e conclusões publicadas

Fonte: O autor (2025)

Tabela 4.4 - Quadro de planejamento resumido

Dimensão	Detalhamento
Problema	Predição binária do sucesso comercial de empreendimentos (velocidade ou resiliência de vendas)
Tipo de Tarefa	Classificação supervisionada binária
Experimentos	1. Velocidade de vendas (≥30% em 3 meses); 2. Resiliência (≤20% não vendidas após 18 meses)
Dados	271 empreendimentos (2022–2025), 27.000+ unidades, dados mercadológicos, econômicos e institucionais
Ferramentas	Python, Pandas, Scikit-learn, XGBoost, SHAP, Google Colab, RE.AI.s APIs
Avaliação dos Modelos	Acurácia, Precisão, Recall, F1-Score, AUC-ROC
X-AI	SHAP, feature importance, árvores explicativas
Resultados Esperados	Modelos preditivos robustos e interpretáveis; recomendações para lançamentos e gestão de produtos

Fonte: O autor (2025)

4.4.2 Entendimento dos dados

A fase de Entendimento dos Dados, conforme estabelecido pela metodologia CRISP-DM, tem como objetivo realizar uma imersão sistemática no universo informacional do problema, examinando a origem, estrutura, qualidade e relevância dos dados disponíveis para a tarefa de mineração. Esta etapa é fundamental para assegurar a compatibilidade entre os dados e os objetivos de negócio previamente definidos, além de permitir a antecipação de possíveis desafios relacionados à completude, consistência e viabilidade analítica das variáveis. A partir desse diagnóstico inicial, torna-se possível fundamentar decisões metodológicas mais robustas nas fases subsequentes de preparação, modelagem e avaliação.

4.4.2.1 Descrição das bases de dados utilizadas - *Datasets*

A base informacional deste estudo é predominantemente oriunda da plataforma **RE.AI.S – Real Estate Artificial Intelligence Systems**, uma solução proprietária de inteligência de dados desenvolvida com foco no monitoramento e análise dos mercados imobiliários primário e secundário e no comportamento do clientes. Esses dados incluem:

- Informações dos empreendimentos comercializados, lançamentos ou em estoque nas incorporadoras, disponibilidade das unidades em comercialização, tabelas de preços dos empreendimentos coletadas por APIs ou diretamente pela equipe de engenharia de dados com construtoras, incorporadoras, imobiliárias e corretores. Compondo um repositório relacional capaz de fornecer tanto informações agregadas de empreendimentos quanto dados no nível de cada unidade habitacional;
- Dados de clientes interessados através de anúncios em portais imobiliários (OLX, Zap Imóveis, Viva Real, Expoimóvel, etc.) e canais digitais e sociais - Estes dados não fazem parte deste estudo atual, mas vão ser parte crucial na continuação deste estudo no doutorado.

A arquitetura da plataforma compreende os seguintes componentes:

1. **Módulo de Captura**: Cadastro direto de informações de oferta e demanda pelos parceiros. APIs de integração com as principais plataformas do mercado, equipe de engenharia de dados para coleta de dados do mercado.
2. **ETL**: Processos de extração, transformação e carga (ETL) estruturados com base em boas práticas (Kimball & Caserta, 2004).
3. **Banco de Dados Relacional**: Armazenamento estruturado e normalizado por empreendimento, unidade e mês.
4. **Camada Analítica**: Fornece saídas de dados e inteligência com painéis, dashboards, relatórios automatizados e datasets analíticos.

O mercado primário reflete diretamente a dinâmica de lançamentos, estratégias comerciais, precificação inicial e absorção pelo mercado — fatores que são cruciais para tomada de decisão por parte das construtoras, investidores e incorporadoras. Diferentemente do mercado secundário, onde os imóveis são usados e sujeitos a fatores como liquidez de revenda, negociações individuais, reformas e financiamento entre pessoas físicas, o mercado primário representa a interação direta entre incorporadoras e consumidores finais na aquisição de unidades novas, além de refletir as decisões estratégicas de desenvolvimento urbano.

Portanto, a utilização do mercado primário proporciona dados alinhados com os objetivos deste trabalho, permitindo a criação de modelos preditivos que podem ser aplicados para apoiar decisões comerciais e de lançamento de novos empreendimentos com foco na velocidade de vendas e na resiliência comercial ao longo do ciclo de oferta.

4.4.2.1.1 Estrutura dos datasets principais utilizados

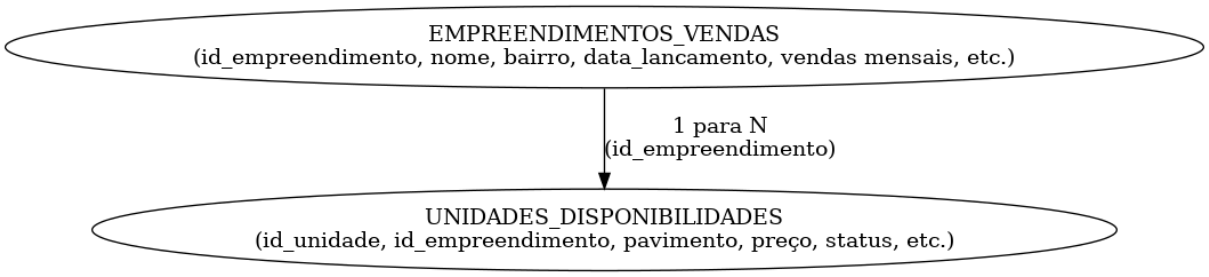
Neste estudo, os dados coletados do mercado primário foram divididos em dois *datasets* principais

Dataset 1 – Empreendimentos e Vendas: contém informações no nível do empreendimento, como bairro, construtora, data de lançamento, número de unidades, status da obra, vendas mensais (jan/2022 a mai/2025), e estoque.

Dataset 2 – Unidades e Disponibilidade: inclui, no nível individual de unidade habitacional, variáveis como preço, área, pavimento, posição, status (vendido ou disponível), tipo da unidade, e dados adicionais de características físicas e locacionais.

Os dois conjuntos possuem granularidades distintas e complementares, sendo o primeiro voltado à caracterização e desempenho das vendas agregadas de cada empreendimento listado, e o segundo às características individuais de cada venda de cada unidade habitacional. A integração desses conjuntos de dados visa a formação de uma base de informações robusta para a análise da velocidade de vendas e resiliência no mercado imobiliário.

Figura 4.4 - Datasets Utilizados no Estudo



Fonte: O autor (2025)

Além disso, dois **conjuntos de dados auxiliares** foram construídos e os dados foram incorporados ao Dataset Unidades Disponibilidades:

- Variáveis macroeconômicas agregadas por mês (IGP-M, INCC, IPCA, SELIC, dólar), utilizadas para controle de contexto de mercado nas análises.
- Atributos relacionados a marca da construtora/incorporadora para analisar o impacto da força e confiança da marca nas vendas.

A seguir, são apresentados os detalhes de cada *dataset*, incluindo os atributos, seus tipos, definições e a motivação para sua utilização.

4.4.2.1.2 Dataset Empreendimentos e Vendas

Este *dataset* compreende informações consolidadas das vendas para os empreendimentos imobiliários listados. Cada linha representa um empreendimento

único com atributos descritivos e um histórico de vendas mensais de janeiro de 2022 até março de 2025 a partir da data do lançamento comercial do empreendimento.

Ele serve como a base principal para a identificação e acompanhamento da performance de vendas dos empreendimentos no tempo, e será fundamental para o posterior cálculo das variáveis-alvo relacionadas à velocidade e resiliência de vendas.

No **Apêndice B** podem ser visualizadas e analisadas em detalhe as informações de cada atributo do dataset Empreendimentos e Vendas.

4.4.2.1.3 Dataset Unidades e Disponibilidades

Este dataset contém informações individuais sobre cada unidade habitacional (apartamento), incluindo características físicas, preço, localização, entre outras. Um atributo muito importante para o nosso estudo é o Status de Venda, que indica se a unidade foi vendida ou ainda está disponível para compra.

No **Apêndice C** podem ser visualizadas e analisadas em detalhe as informações de cada atributo do dataset Unidades e Disponibilidades.

4.4.2.1.4 Observações finais sobre os datasets brutos

A descrição dos dois *datasets* evidencia a complementaridade entre a perspectiva agregada dos empreendimentos e a análise granular das unidades habitacionais.

O *dataset* Empreendimento e Vendas permite a definição da variável-alvo binária a partir do desempenho agregado de vendas no tempo, enquanto o *dataset* Unidades e Disponibilidades oferece uma visão granular das características de cada unidade habitacional, enriquecida por informações de mercado e da construtora.

A próxima etapa consiste na análise exploratória e descritiva dessas variáveis, como apresentado nas seções 4.4.2.2 e 4.4.2.3.

4.4.2.2 Análise exploratória dos dados (AED)

A fase de Análise Exploratória dos Dados (AED) é um componente crítico do processo de mineração de dados, focada na compreensão das características

principais do conjunto de dados por meio de técnicas visuais e estatísticas resumidas. Seu propósito é identificar padrões, tendências, anomalias e relações entre as variáveis, fornecendo *insights* valiosos que guiam as etapas subsequentes de preparação e modelagem dos dados (TUKEY, 1977).

O processo para este estudo envolveu a inspeção de distribuições de variáveis, a detecção de valores ausentes e *outliers*, e a análise de relações bivariadas, especialmente com as variáveis-alvo de sucesso.

4.4.2.2.1 Análise exploratória do dataset Empreendimentos e Vendas

Este *dataset* é central para a definição e avaliação das classes de sucesso em velocidade e resiliência de vendas. A análise exploratória revelou as seguintes características e *insights*:

(i) Estrutura e Composição dos Dados

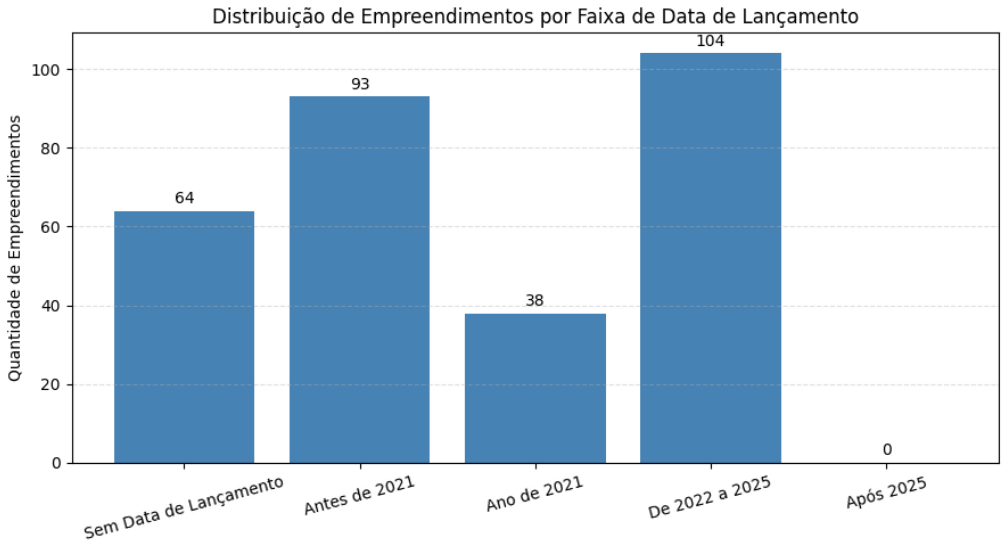
O dataset "Empreendimentos e Vendas" compreende um total de 278 registros, representando empreendimentos imobiliários únicos. É composto por 55 colunas, categorizadas majoritariamente como:

- Datas: 4 atributos, indicando marcos temporais importantes do empreendimento.
- Texto: 6 atributos, para informações descritivas.
- Numéricas: 45 atributos, que incluem características intrínsecas dos empreendimentos e, predominantemente, o histórico mensal de vendas.

Para o presente estudo, a data de lançamento (*empreendimento_data_lancamento*) dos empreendimentos é um atributo crucial, dada sua relevância direta para a definição das classes-alvo de Velocidade de Vendas e Resiliência de Vendas.

A distribuição dos empreendimentos em relação à sua data de lançamento é ilustrada no histograma da Figura 4.5:

Figura 4.5 - Histograma dos empreendimentos listados no *dataset* em relação à data de lançamento



Fonte: O autor (2025)

A distribuição temporal dos 299 empreendimentos é composta por:

- 64 empreendimentos não possuem informação sobre a data de lançamento.
- 93 empreendimentos foram lançados antes de 2021.
- 38 empreendimentos foram lançados entre Janeiro e Dezembro de 2021.
- 104 empreendimentos foram lançados entre Janeiro de 2022 e Maio de 2025.

Esta segmentação temporal é fundamental para a definição das bases de análise de cada experimento:

Para a definição da classe-alvo de **Velocidade de Vendas**, serão considerados os **104 empreendimentos** lançados entre Janeiro de 2022 e Maio de 2025, representando **34,8%** do total de empreendimentos inicialmente listados. Este recorte temporal garante que haja um período adequado de observação para a métrica de vendas nos três primeiros meses.

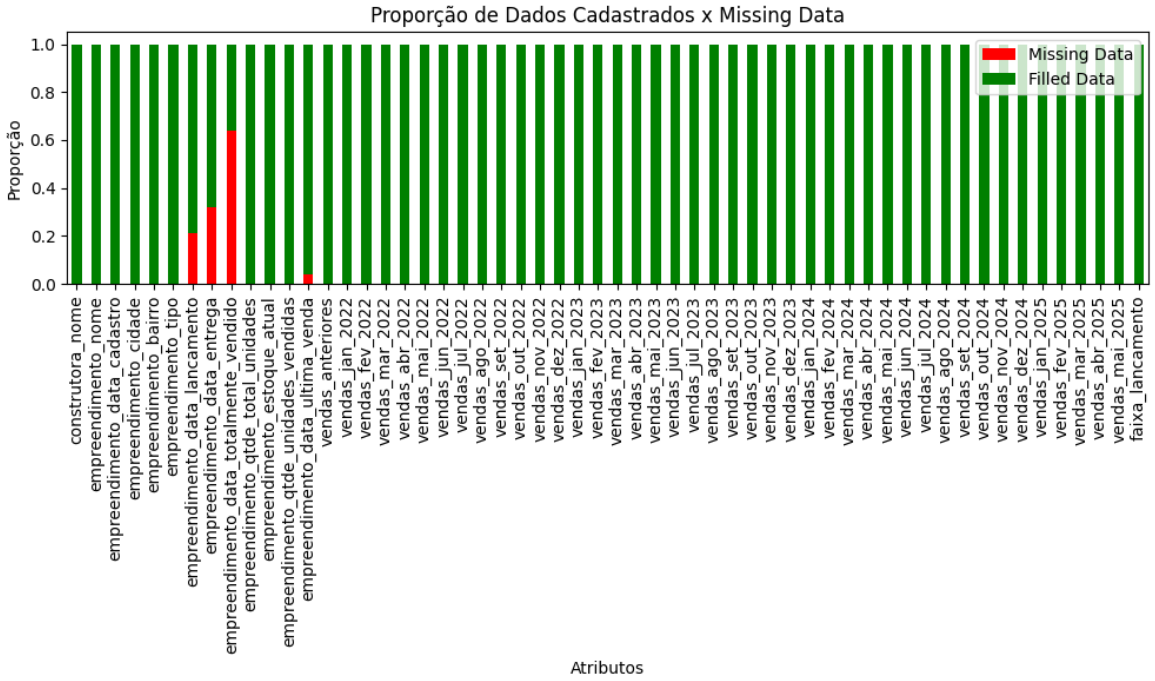
Para a definição da classe-alvo de **Resiliência de Vendas**, a análise abrangerá os empreendimentos lançados a partir de 2021. Isso inclui os 38 empreendimentos lançados em 2021 somados aos 104 empreendimentos lançados entre Janeiro de 2022 e Maio de 2025, totalizando **142 empreendimentos**. Este conjunto corresponde a **47,5%** dos empreendimentos listados, permitindo um período de observação suficiente para a métrica de 18 meses de resiliência.

Os **64 empreendimentos** que não possuem das de lançamento e os **93 empreendimentos** lançados antes de 2021 serão excluídos das análises preditivas por inviabilizarem o cálculo das métricas temporais de desempenho.

(ii) Análise de Dados Faltantes (Missing Data)

A análise da completude do *dataset* revelou a presença de valores ausentes em apenas três atributos, conforme ilustrado na Figura 4.6 e na tabela 4.5 a seguir:

Figura 4.6 - Proporção de Dados Cadastrados X Dados Faltantes por atributo



Fonte: O autor (2025)

Tabela 4.5 - Percentual de Dados Faltantes por atributo

Atributo	Faltantes	%
empreendimento_data_totalmente_vendido	191	63.9%
empreendimento_data_entrega	95	31.8%
empreendimento_data_lancamento	64	21.4%
empreendimento_data_ultima_venda	12	4.0%

Fonte: O autor (2025)

Os atributos com dados faltantes e suas respectivas interpretações são:

- **empreendimento_data_totalmente_vendido**: Apresentou 63,9% de valores ausentes. Esta ocorrência é esperada e não representa um

problema de qualidade de dados, visto que estes registros correspondem a empreendimentos que ainda se encontram em fase ativa de comercialização no período analisado, não tendo atingido 100% de suas vendas.

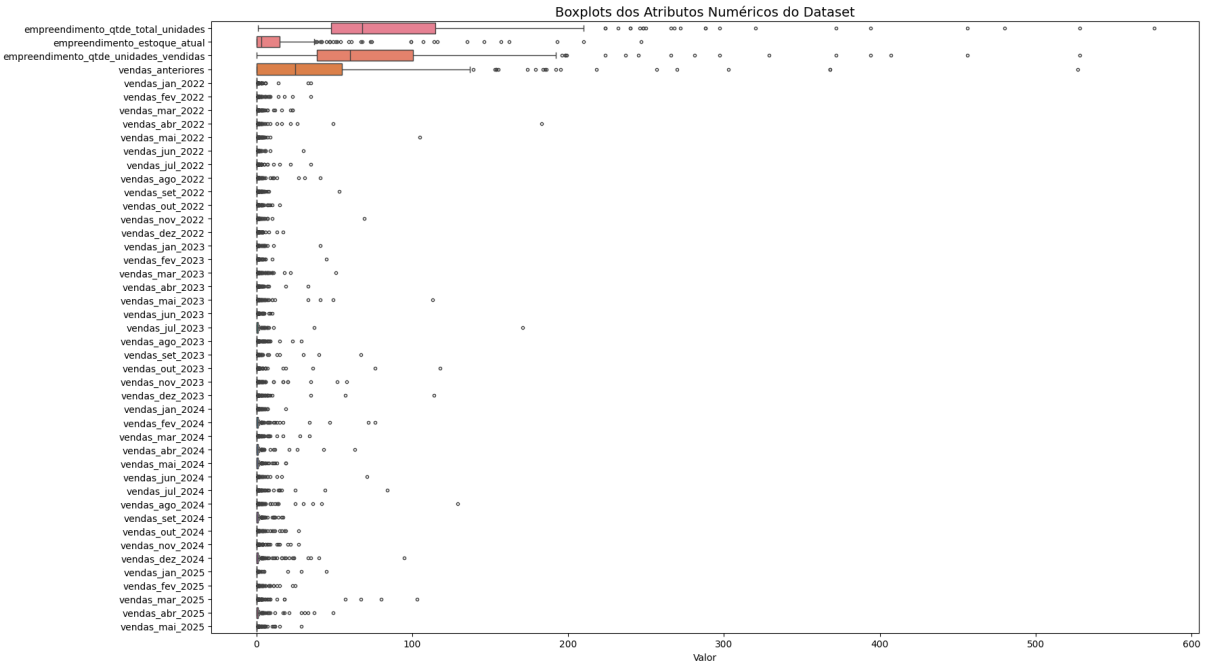
- **empreendimento_data_entrega**: Com 31,8% de valores ausentes. Similarmente, a ausência nesta variável é interpretada como indicativo de empreendimentos que, na data da coleta dos dados, ainda estavam nas fases iniciais de lançamento comercial ou em construção, e portanto, não haviam sido formalmente entregues.
- **empreendimento_data_lancamento**: Identificou-se 21,4% de valores ausentes. A ausência nesta variável é crítica, pois a data de lançamento é um atributo fundamental para a construção das variáveis-alvo de sucesso, tanto para a métrica de velocidade de vendas quanto para a de resiliência. Para garantir a integridade da análise preditiva, os 45 registros nos quais esta informação estava ausente serão excluídos do conjunto de dados, um processo a ser detalhado na fase de Preparação dos Dados.

A exclusão dos 64 empreendimentos representa 21,4% da amostra, porém é necessária para garantir a construção correta das variáveis-alvo.

(iii) Análise de Dados Atípicos (*Outliers*)

A metodologia adotada segue a definição clássica de Tukey (1977) para detecção de *outliers* univariados baseada no IQR (*Interquartile Range*), que define valores discrepantes como aqueles situados além de 1,5 vezes o Intervalo Interquartil (IQR = Q3 - Q1) abaixo do primeiro quartil (Q1) ou acima do terceiro quartil (Q3).

Figura 4.7 - Bloxplots dos atributo numéricos do *dataset* Empreendimentos e Vendas



Fonte: O autor (2025)

A Tabela 4.6, mostrada na página a seguir, sumariza as métricas descritivas e os limites para detecção de *outliers* via regra de Tukey para os principais atributos numéricos, incluindo contagem, mínimo, máximo, média, desvio padrão, mediana, Q1, Q3 e o número de *outliers* identificados.

Tabela 4.6 - Métricas descritivas do *dataset* Empreendimentos e Vendas

Atributo	Mín	Máx	Média	Desvio Padrão	Q2 Mediana	Q1 (25%)	Q3 (75%)	IQR	Limite Inferior	Limite Superior	Outlier
empreendimento_qtde_total_unidades	1	576	94,6	82,2	68,0	48,0	115,0	67,0	-52,5	215,5	22
empreendimento_estoque_atual	0	247	14,9	32,3	3,0	0,0	15,0	15,0	-22,5	37,5	32
empreendimento_qtde_unidades_vendi	0	528	79,7	72,2	60,0	39,0	100,5	61,5	-53,3	192,8	16
vendas_anteriores	0	527	41,0	63,9	25,0	0,0	55,0	55,0	-82,5	137,5	19
vendas_jan_2022	0	35	0,6	3,0	0,0	0,0	0,0	0,0	0,0	0,0	49
vendas_fev_2022	0	35	0,8	3,0	0,0	0,0	0,0	0,0	0,0	0,0	60
vendas_mar_2022	0	23	0,7	2,8	0,0	0,0	0,0	0,0	0,0	0,0	65
vendas_abr_2022	0	183	1,4	11,2	0,0	0,0	0,0	0,0	0,0	0,0	61
vendas_mai_2022	0	105	0,8	6,2	0,0	0,0	0,0	0,0	0,0	0,0	62
vendas_jun_2022	0	30	0,5	2,0	0,0	0,0	0,0	0,0	0,0	0,0	65
vendas_jul_2022	0	35	0,6	2,8	0,0	0,0	0,0	0,0	0,0	0,0	57

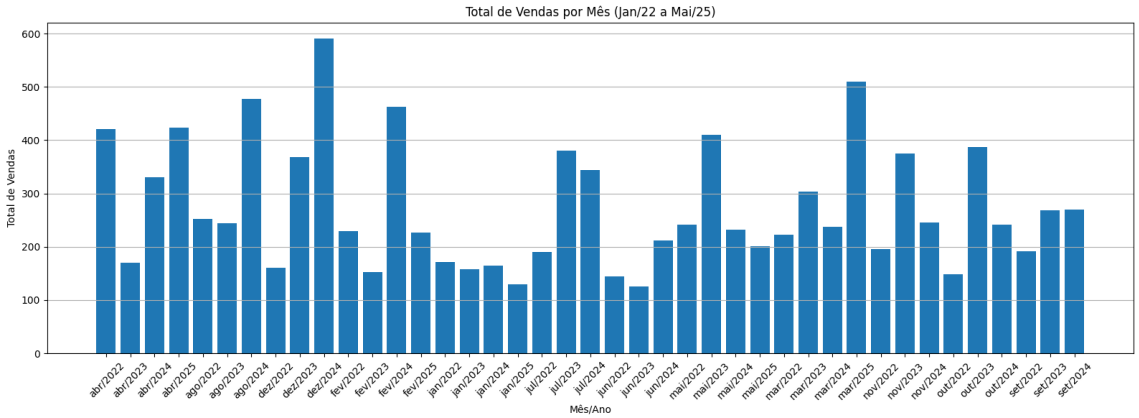
Atributo	Mín	Máx	Média	Desvio Padrão	Q2 Mediana	Q1 (25%)	Q3 (75%)	IQR	Limite Inferior	Limite Superior	Outlier
vendas_ago_2022	0	41	0,8	3,7	0,0	0,0	0,0	0,0	0,0	0,0	63
vendas_set_2022	0	53	0,6	3,3	0,0	0,0	0,0	0,0	0,0	0,0	63
vendas_out_2022	0	15	0,5	1,6	0,0	0,0	0,0	0,0	0,0	0,0	58
vendas_nov_2022	0	69	0,7	4,1	0,0	0,0	0,0	0,0	0,0	0,0	56
vendas_dez_2022	0	17	0,5	1,6	0,0	0,0	0,0	0,0	0,0	0,0	72
vendas_jan_2023	0	41	0,5	2,6	0,0	0,0	0,0	0,0	0,0	0,0	55
vendas_fev_2023	0	45	0,5	2,8	0,0	0,0	0,0	0,0	0,0	0,0	56
vendas_mar_2023	0	51	1,0	3,8	0,0	0,0	0,0	0,0	0,0	0,0	72
vendas_abr_2023	0	33	0,6	2,4	0,0	0,0	0,0	0,0	0,0	0,0	55
vendas_mai_2023	0	113	1,4	7,8	0,0	0,0	0,0	0,0	0,0	0,0	65
vendas_jun_2023	0	10	0,4	1,2	0,0	0,0	0,0	0,0	0,0	0,0	58
vendas_jul_2023	0	171	1,3	10,2	0,0	0,0	1,0	1,0	-1,5	2,5	20
vendas_ago_2023	0	29	0,8	2,7	0,0	0,0	0,0	0,0	0,0	0,0	69
vendas_set_2023	0	67	0,9	5,0	0,0	0,0	0,0	0,0	0,0	0,0	63
vendas_out_2023	0	118	1,3	8,5	0,0	0,0	0,0	0,0	0,0	0,0	63
vendas_nov_2023	0	58	1,3	5,5	0,0	0,0	0,0	0,0	0,0	0,0	65
vendas_dez_2023	0	114	1,2	7,7	0,0	0,0	0,0	0,0	0,0	0,0	65
vendas_jan_2024	0	19	0,5	1,6	0,0	0,0	0,0	0,0	0,0	0,0	61
vendas_fev_2024	0	76	1,5	7,1	0,0	0,0	1,0	1,0	-1,5	2,5	29
vendas_mar_2024	0	34	0,8	3,1	0,0	0,0	0,0	0,0	0,0	0,0	66
vendas_abr_2024	0	63	1,1	4,9	0,0	0,0	1,0	1,0	-1,5	2,5	24
vendas_mai_2024	0	19	0,8	2,3	0,0	0,0	1,0	1,0	-1,5	2,5	25
vendas_jun_2024	0	71	0,7	4,4	0,0	0,0	0,0	0,0	0,0	0,0	61
vendas_jul_2024	0	84	1,2	6,0	0,0	0,0	0,0	0,0	0,0	0,0	60
vendas_ago_2024	0	129	1,6	8,6	0,0	0,0	0,0	0,0	0,0	0,0	72
vendas_set_2024	0	17	0,9	2,4	0,0	0,0	1,0	1,0	-1,5	2,5	33
vendas_out_2024	0	27	0,8	2,9	0,0	0,0	0,0	0,0	0,0	0,0	59
vendas_nov_2024	0	27	0,8	2,9	0,0	0,0	0,0	0,0	0,0	0,0	62
vendas_dez_2024	0	95	2,0	7,4	0,0	0,0	1,0	1,0	-1,5	2,5	42
vendas_jan_2025	0	45	0,4	3,3	0,0	0,0	0,0	0,0	0,0	0,0	21
vendas_fev_2025	0	25	0,8	2,7	0,0	0,0	0,0	0,0	0,0	0,0	55
vendas_mar_2025	0	103	1,7	9,2	0,0	0,0	0,0	0,0	0,0	0,0	68
vendas_abr_2025	0	49	1,4	5,2	0,0	0,0	1,0	1,0	-1,5	2,5	30
vendas_mai_2025	0	29	0,7	2,5	0,0	0,0	0,0	0,0	0,0	0,0	50

Fonte: O autor (2025)

Nos casos onde Q1 = Q3 = 0, o IQR é nulo, e portanto a regra de Tukey não é aplicável — todo valor diferente de zero seria considerado outlier. Assim, optou-se por manter esses valores, interpretando-os como picos naturais de vendas.

Para contextualizar os principais atributos do Dataset Empreendimentos Vendas, apresentamos o histograma das vendas mensais:

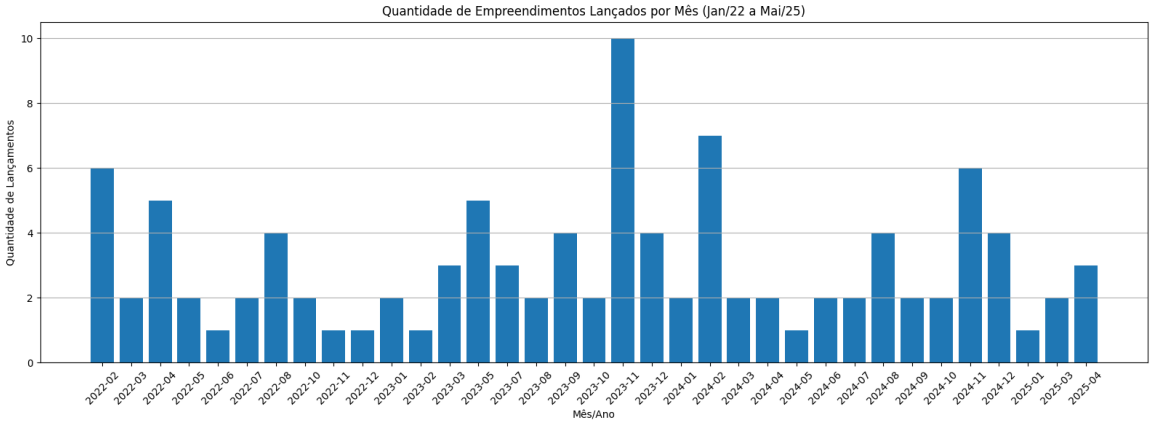
Figura 4.8 - Histograma do total de vendas mensais (Janeiro de 2022 a Maio de 2025)



Fonte: O autor (2025)

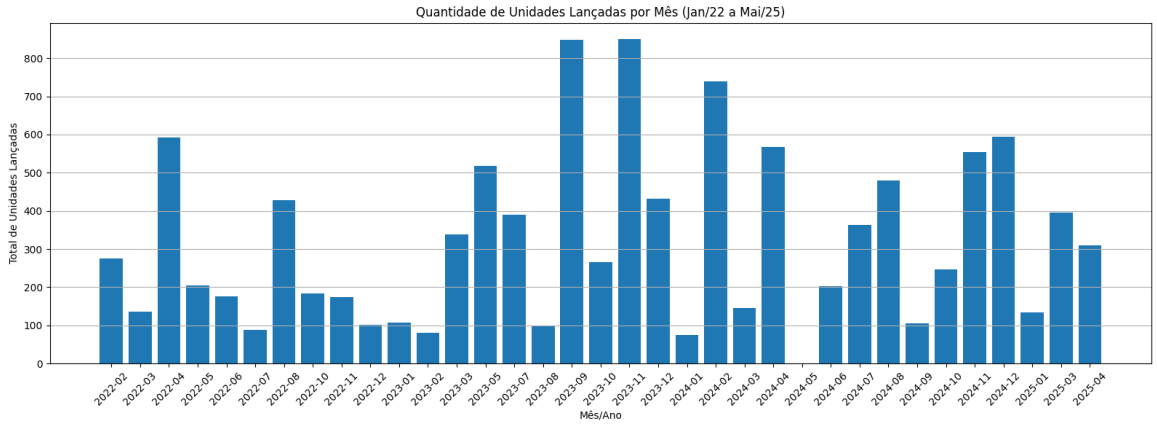
A Figura 4.8 demonstra que a distribuição das vendas totais mensais apresenta picos em determinados períodos, o que se alinha com a natureza sazonal e concentrada dos lançamentos imobiliários. Essa variabilidade é reforçada pelas distribuições da quantidade de empreendimentos lançados (Figura 4.9) e pela quantidade de unidades habitacionais lançadas (Figura 4.10):

Figura 4.9 - Quantidade de empreendimentos lançados por mês (Janeiro/22 a Maio/25)



Fonte: O autor (2025)

Figura 4.10 - Quantidade de unidades habitacionais lançadas por mês (Janeiro/22 a Maio/25)

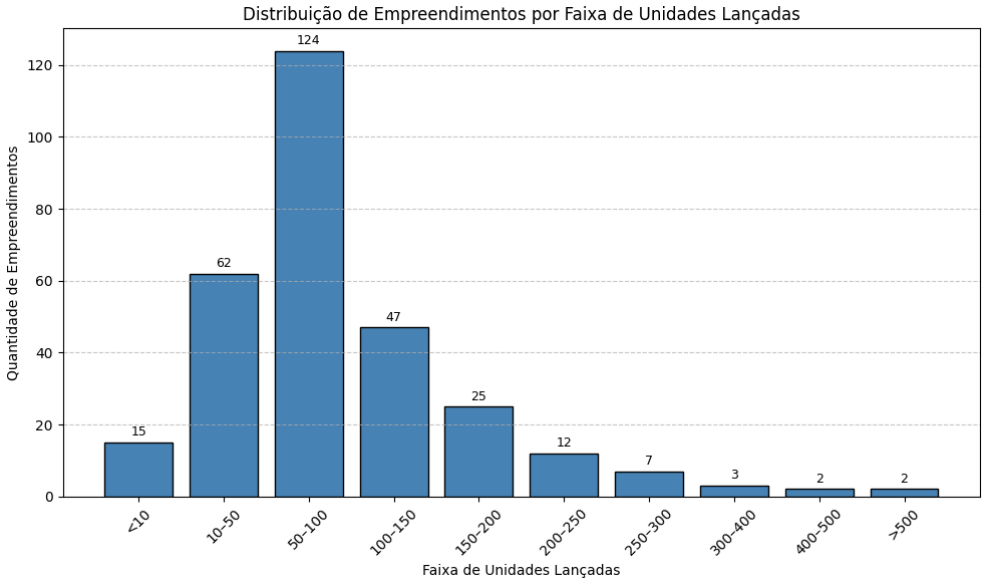


Fonte: O autor (2025)

As figuras acima ilustram que meses com maior volume de lançamentos (tanto em número de empreendimentos quanto em quantidade de unidades) tendem a correlacionar-se com maiores volumes de vendas.

Em uma análise mais aprofundada para o tratamento de *outliers* que poderiam impactar a representatividade do modelo, foi estabelecido um critério adicional: empreendimentos com menos de 10 unidades cadastradas foram considerados exceções. A Figura 4.11 ilustra a distribuição dos empreendimentos por faixas de quantidade de unidades lançadas.

Figura 4.11 - Distribuição de empreendimentos por faixa de quantidade de unidades lançadas



Fonte: O autor (2025)

Empreendimentos com menos de 10 unidades representam outliers estruturais e fogem ao padrão de produto predominante no mercado vertical de Recife. Foram removidos 15 empreendimentos, correspondendo cerca de 5% da amostra.

(iv) Conclusão da Análise Exploratória do Dataset Empreendimentos e Vendas

As evidências encontradas na análise exploratória orientaram decisões cruciais na etapa de preparação dos dados, especialmente na definição de critérios de exclusão, tratamento de *outliers* e construção das variáveis-alvo para os experimentos de classificação binária descritos na seção 4.4.3.

4.4.2.2.2 Análise exploratória do dataset Unidades e Disponibilidades

A Análise Exploratória dos Dados (AED) para o *dataset* "Unidades e Disponibilidades" é um passo fundamental que complementa a análise realizada no nível do empreendimento. Este conjunto de dados oferece uma visão granular das características individuais das unidades habitacionais e de seu status de disponibilidade, o que é essencial para compreender as dinâmicas de comercialização em um nível micro. Tal granularidade é crucial para investigar os fatores que impulsionam o sucesso ou insucesso de unidades específicas dentro de um empreendimento, contribuindo diretamente para a análise da Velocidade de Vendas e da Resiliência de Vendas.

Esta análise aprofundada permite uma compreensão abrangente de como atributos específicos de um empreendimento (localização, características, equipamentos, entre outros) e das unidades (como preço, área, número de quartos, pavimento, entre outros) interagem com seu status de venda ao longo do tempo. Esta granularidade é fundamental para:

- **Identificar perfis de unidades com alta ou baixa demanda:** Compreender quais características de unidades se correlacionam com maior velocidade de venda ou com permanência prolongada em estoque.
- **Analisar a absorção de mercado por tipologia:** Avaliar como diferentes tipos de unidades (ex: 2 quartos, 3 quartos, coberturas) são absorvidas pelo mercado.

- **Fornecer *insights* para precificação e *mix* de produtos:** Oferecer subsídios para construtoras e incorporadoras otimizarem o planejamento de futuros lançamentos, precificação e o *mix* de unidades.

(i) Estrutura e Composição dos Dados

O *dataset* Unidades Disponibilidade compreende um total de 28.728 registros, cada um representando uma unidade única de empreendimento imobiliário. A estrutura do *dataset* compreende **106 colunas**, que foram categorizadas da seguinte forma:

- Datas: 4 atributos, indicando marcos temporais importantes das unidades dos empreendimentos.
- Texto: 14 atributos, para informações descritivas.
- Numéricas: 87 atributos, que incluem características intrínsecas das unidades dos empreendimentos.
- Booleano: 1 Atributo com característica da unidade do empreendimento.

A predominância de atributos numéricos reflete a modelagem orientada à análise quantitativa detalhada do comportamento das unidades.

Nesta análise inicial, foi identificado que vários atributos destinados a serem do tipo booleano não foram corretamente classificados ou interpretados no carregamento dos dados. Esta questão será abordada na fase de Preparação dos Dados, onde a correção desses tipos de dados será realizada para garantir a clareza semântica e facilitar análises posteriores.

Para uma compreensão mais estruturada do *dataset* Unidades e Disponibilidades, os atributos foram categorizados de acordo com sua relevância e tipo de informação que representam no contexto do mercado imobiliário.

Características do Empreendimento

Estes atributos descrevem o empreendimento como um todo, englobando sua localização, porte e os diferenciais estruturais e de lazer oferecidos.

- Identificação e Contexto Geral;
- Linha do Tempo e Comercialização;
- Estrutura e Porte;
- Localização e Acessibilidade (Diferenciais Externos);
- Lazer e Amenidades (Áreas Comuns/Qualidade de Vida).

Características da Unidade

Estes atributos detalham a unidade habitacional individual, seus aspectos físicos, acabamentos e diferenciais.

- Identificação e Tipo;
- Dimensão e Valor;
- Layout e Tipologia Interna;
- Cômodos e Funcionalidades Internas;
- Acabamento e Mobiliário (Diferenciais de Entrega).

Dinâmica de Vendas e Status da Unidade

Atributos que registram o status de comercialização e o contexto da venda da unidade.

Dados de Mercado e Contexto Macroeconômico

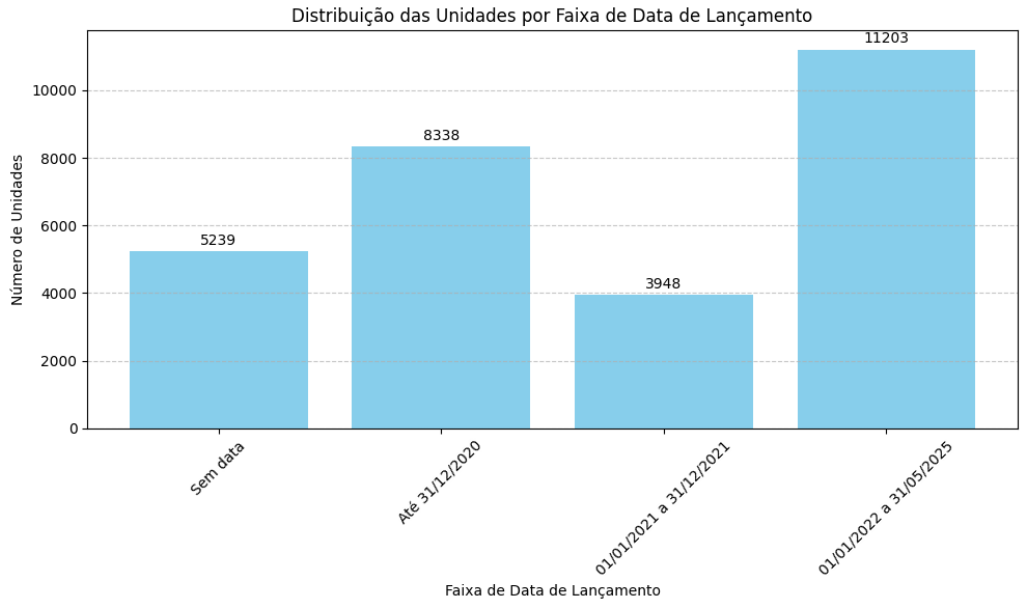
Atributos que refletem as condições econômicas gerais e específicas do mercado imobiliário no momento da venda ou análise.

Reputação da Construtora (Marca e Experiência)

Atributos que medem a experiência e o reconhecimento da construtora no mercado.

A **data de lançamento do empreendimento** (*empreendimento_data_lancamento*) é um atributo crítico neste *dataset*, pois sua relevância é direta para a definição das classes-alvo de **Velocidade de Vendas** e **Resiliência de Vendas**. A distribuição das instâncias das unidades dos empreendimentos em relação à sua data de lançamento é visualizada no histograma da Figura 4.12:

Figura 4.12 - Histograma da distribuição por faixa de data de lançamento



Fonte: O autor (2025)

A análise temporal das 28.728 instâncias das unidades revela a seguinte composição:

- 5.239 unidades que não possuem informação sobre a data de lançamento.
- 8.338 unidades de empreendimentos lançados antes de 2021.
- 3.948 unidades de empreendimentos lançados entre Janeiro e Dezembro de 2021.
- 11.203 unidades de empreendimentos lançados entre Janeiro de 2022 e Maio de 2025.

Esta segmentação temporal é fundamental para a definição das bases de análise de cada experimento:

- Para a **definição da classe-alvo de Velocidade de Vendas**, serão consideradas as **11.203 instâncias de unidades** de empreendimentos lançados entre Janeiro de 2022 e Maio de 2025. Este subconjunto representa **39,0%** do total das instâncias listadas, e seu recorte temporal garante um período adequado de observação para a métrica de vendas nos três primeiros meses, conforme os critérios estabelecidos.

- Para a **definição da classe-alvo de Resiliência de Vendas**, a análise abrangerá as instâncias das unidades de empreendimentos lançados a partir de 2021. Este grupo inclui as **3.948 instâncias** de unidades de

empreendimentos lançados em 2021, somadas às **11.203 instâncias** de empreendimentos lançados entre Janeiro de 2022 e Maio de 2025, totalizando **15.151 instâncias de unidades**. Este conjunto corresponde a **52,7%** das instâncias listadas, proporcionando um período de observação suficiente para a métrica de 18 meses de resiliência.

Por fim, as **5.239 instâncias de unidades de empreendimentos** que não possuem data de lançamento e as **8.338 instâncias de unidades de empreendimentos** lançados antes de 2021 serão excluídas das análises preditivas. Esta exclusão é justificada pela inviabilidade do cálculo preciso das métricas temporais de desempenho para essas unidades, garantindo a integridade do modelo.

(ii) Análise de Dados Faltantes (Missing Data)

A análise da completude do *dataset* revelou a presença de valores ausentes em alguns atributos. Para uma compreensão detalhada da extensão dos dados faltantes, a Tabela 4.7 apresenta os percentuais de dados preenchidos e faltantes para todos os atributos afetados:

Tabela 4.7 - Proporção de Dados Preenchidos X Dados Faltantes no *dataset*

Atributo	% dados preenchidos	% dados faltantes
construtora_ano_fundação	0	100
construtora_empreendimentos_entregues	0	100
construtora_empreendimentos_entregues_3_anos	0	100
empreendimento_beira_rio	0	100
unidade_vendas_mes_economia_igpm	0	100
unidade_vendas_mes_economia_incc	0	100
unidade_vendas_mes_economia_taxa_selic	0	100
unidade_vendas_mes_economia_valor_dolar	0	100
empreendimento_rooftop	9,45	90,55
empreendimento_estagio_obra_atual	14,94	85,06
empreendimento_situação_atual	15,52	84,48
empreendimento_qtde_estoque	15,52	84,48

Atributo	% dados preenchidos	% dados faltantes
empreendimento_data_totalmente_vendido	36,43	63,57
unidade_venda_mes_relacao_dataentrega	57,78	42,22
unidade_venda_semestre_dataentrega	57,78	42,22
emprendimento_situacao_unidade_mes_venda	61,93	38,07
unidade_venda_semestre_datalancamento	68,3	31,7
unidade_venda_mes_datalancamento	68,3	31,7
empreendimento_data_entrega	70,53	29,47
unidade_suites	75,75	24,25
unidade_sub_tipo	76,72	23,28
unidade_banheiros	80,36	19,64
empreendimento_data_lancamento	81,76	18,24
empreendimento_numero_meses_comercializacao	82,45	17,55
unidade_venda_valor	83,3	16,7
empreendimento_percentual_estoque_unidade_mes_venda	84,3	15,7
unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia	84,3	15,7
unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia	84,3	15,7
unidade_venda_data	84,3	15,7
unidade_venda_estoque_empreendimento	84,3	15,7
unidade_garagem	86,63	13,37
unidade_salas	89,88	10,12
empreendimento_elevador	90,51	9,49
unidade_valor_m2_imovel	98,09	1,91
padrao_valor_m2_imovel	98,09	1,91
unidade_area	98,14	1,86
empreendimento_pavimentos	98,37	1,63
empreendimento_unidades_por_andar	98,59	1,41
unidade_valor_imovel	99,68	0,32
unidade_quartos	99,69	0,31

Atributo	% dados preenchidos	% dados faltantes
unidade_andar	99,82	0,18
unidade_pretensao	99,92	0,08

Fonte: O autor (2025)

A presença de dados faltantes, no entanto, é compreendida e justificada para diversos grupos de atributos, refletindo a natureza do processo de comercialização imobiliária:

Atributos da Construtora:

Os campos *construtora_ano_fundacao*, *construtora_empreendimentos_entregues*, e *construtora_empreendimentos_entregues_3_anos* serão preenchidos em uma etapa posterior, durante a fase de Preparação dos Dados, por meio da integração com um dataset externo específico de construtoras.

Atributos de Índices Econômicos:

Similarmente, os atributos *unidade_vendas_mes_economia_igpm*, *unidade_vendas_mes_economia_incc*, *unidade_vendas_mes_economia_taxa_selic*, e *unidade_vendas_mes_economia_valor_dolar* serão complementados na fase de Preparação dos Dados, com a aplicação dos valores correspondentes ao mês da venda da unidade ou ao mês atual para unidades ainda disponíveis.

Atributos Condicionais de Disponibilidade:

Os atributos *empreendimento_estagio_obra_atual*, *empreendimento_situação_atual*, e *empreendimento_qtde_estoque* terão valores preenchidos apenas para as unidades que ainda se encontram disponíveis para venda, sendo nulos para unidades já comercializadas.

Atributos Condicionais de Unidades Vendidas:

Inversamente, os atributos *unidade_venda_mes_relacao_dataentrega*, *unidade_venda_semestre_dataentrega*, *empreendimento_situacao_unidade_mes_venda*, *unidade_venda_semestre_datalancamento*, *unidade_venda_mes_datalancamento*, *unidade_venda_valor*, *empreendimento_percentual_estoque_unidade_mes_venda*, *unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia*, *unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia*,

unidade_venda_data, e *unidade_venda_estoque_empreendimento* estarão preenchidos exclusivamente para as unidades que já foram vendidas.

Atributos Condicionais de Empreendimento Entregue/Vendido:

Por fim, os atributos *empreendimento_data_entrega*, *empreendimento_data_totalmente_vendido*, e *empreendimento_numero_meses_comercializacao* só apresentarão valores para empreendimentos que foram entregues ou tiveram todas as suas unidades comercializadas.

Conforme mencionado na análise da estrutura dos dados, as instâncias de unidades com dados faltantes na data de lançamento (*empreendimento_data_lancamento*) serão removidas da base para as análises preditivas, dada a criticidade dessa informação para o cálculo das variáveis-alvo temporais.

(iii) Análise de Dados Atípicos (Outliers)

Para a análise de dados fora da série (*outliers*) no *dataset* Unidades e Disponibilidades, foi adotada uma metodologia de seleção criteriosa dos atributos numéricos a serem inspecionados.

As seguintes considerações guiaram a escolha das variáveis (Han et al., 2011; Rousseeuw & Hubert, 2011):

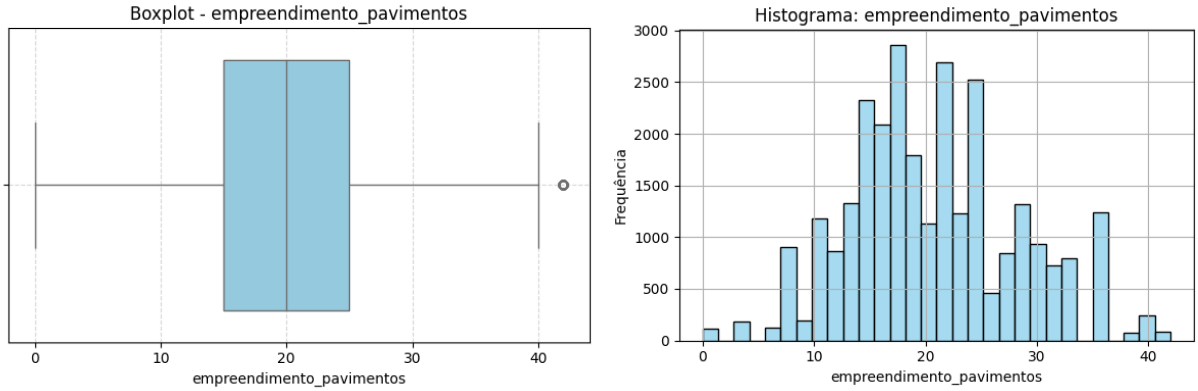
- Exclusão de Atributos Binários/Booleanos: Variáveis que representam categorias ou flags (com valores 0 ou 1) foram excluídas, pois não possuem a variabilidade contínua ou discreta necessária para a aplicação das técnicas de detecção de outliers baseadas em distribuição.
- Seleção de Variáveis Contínuas ou Discretas com Variabilidade Real: O foco foi em atributos que pudessem apresentar uma gama de valores numéricos significativos, refletindo características mensuráveis do empreendimento ou da unidade.
- Verificação de Ausência de Alta Correlação Redundante: Evitou-se a inclusão de atributos altamente correlacionados entre si que poderiam levar a análises redundantes de outliers, priorizando variáveis que oferecem informações únicas sobre diferentes dimensões dos dados.
- Verificação de Suficiência de Dados: Somente atributos com uma proporção de dados preenchidos superior a 70% (inferior a 30% de dados

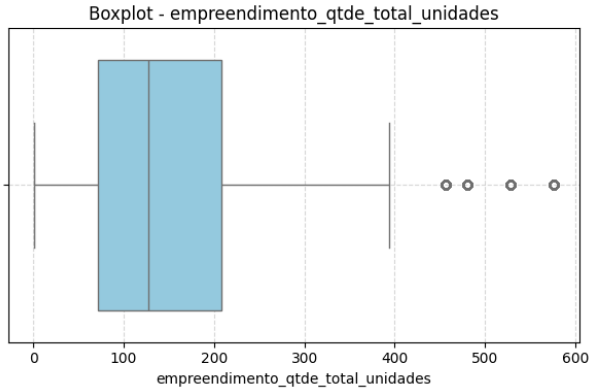
faltantes) foram considerados, garantindo que a análise de *outliers* fosse baseada em uma amostra representativa e robusta de informações.

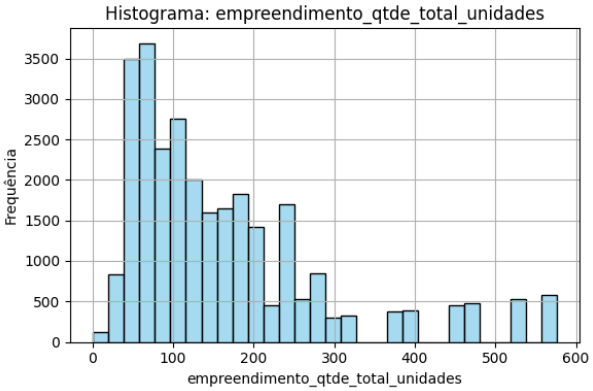
Com base nesses critérios, os atributos numéricos selecionados para a análise de outliers incluíram: *empreendimento_qtde_total_unidades*, *empreendimento_pavimentos*, *empreendimento_unidades_por_andar*, *empreendimento_numero_meses_comercializacao*, *unidade_quartos*, *unidade_valor_imovel*, *unidade_valor_m2_imovel*, *unidade_area*, *unidade_suites*, *unidade_garagem*, *unidade_andar*, *unidade_venda_valor*, *unidade_venda_semestre_datalancamento*, *empreendimento_percentual_estoque_unidade_mes_venda*, *unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia*, e *unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia*.

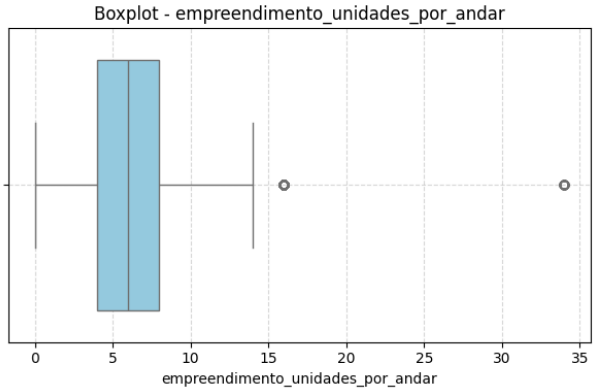
A exploração foi feita através da geração de boxplots e histogramas para cada um desses atributos conforme ilustrado na figura abaixo:

Figura 4.13 - Listagem de Bloxplot e Histograma dos atributos analisados

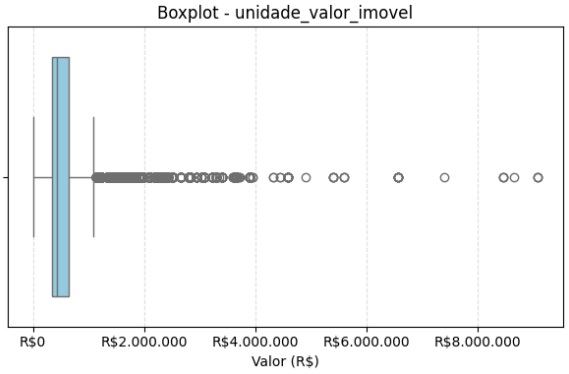


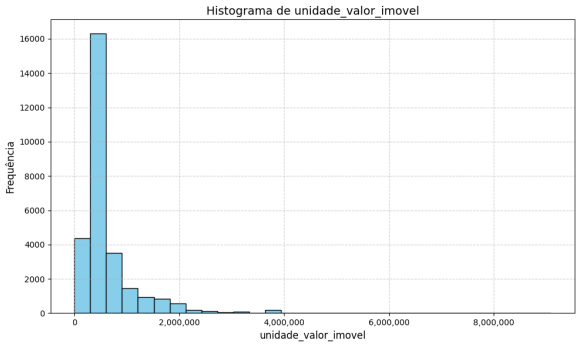

Boxplot - empreendimento_qtde_total_unidades
0
100
200
300
400
500
600
empreendimento_qtde_total_unidades

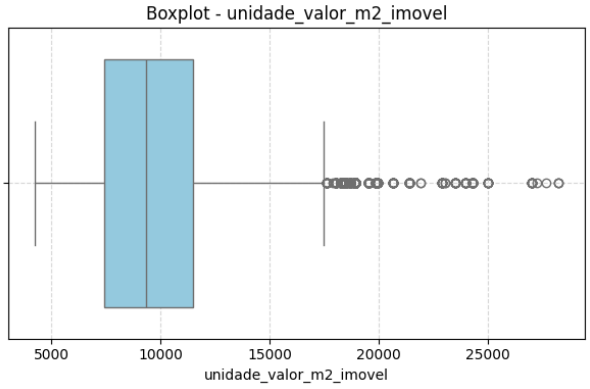

Histograma: empreendimento_qtde_total_unidades
Frequência
0
500
1000
1500
2000
2500
3000
3500
0
100
200
300
400
500
600
empreendimento_qtde_total_unidades

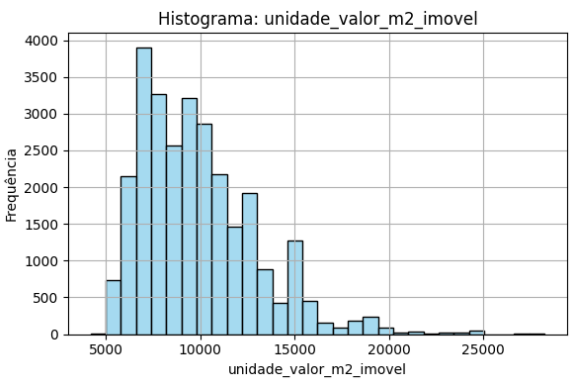

Boxplot - empreendimento_unidades_por_andar
0
5
10
15
20
25
30
35
empreendimento_unidades_por_andar

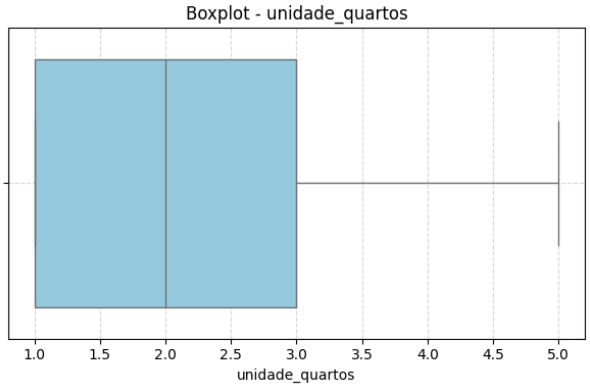

Histograma: empreendimento_unidades_por_andar
Frequência
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25
30
35
empreendimento_unidades_por_andar

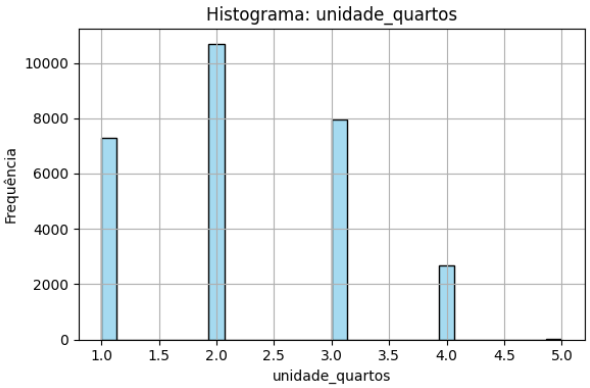

Boxplot - unidade_valor_imovel
R$0
R$2.000.000
R$4.000.000
R$6.000.000
R$8.000.000
Valor (R$)

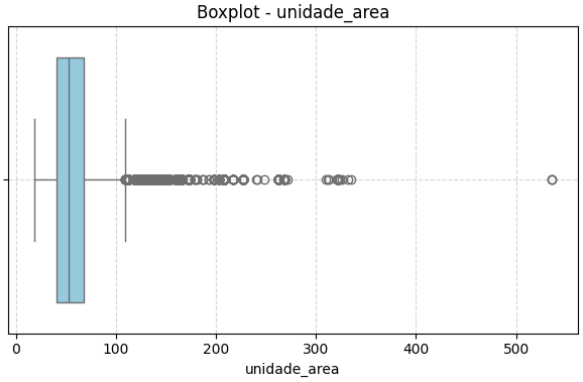

Histograma de unidade_valor_imovel
Frequência
0
2000
4000
6000
8000
10000
12000
14000
16000
0
2,000,000
4,000,000
6,000,000
8,000,000
unidade_valor_imovel

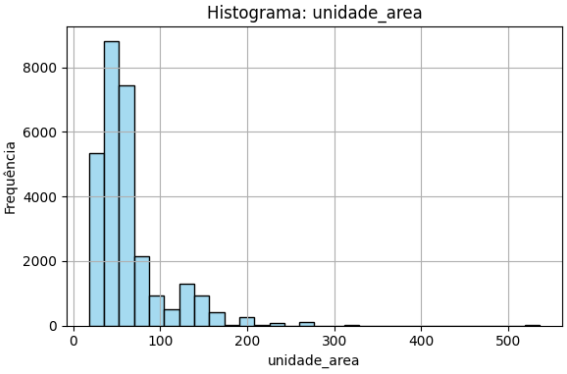

Boxplot - unidade_valor_m2_imovel
5000
10000
15000
20000
25000
unidade_valor_m2_imovel

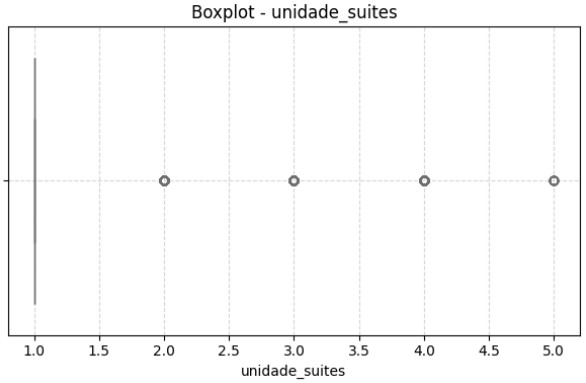

Histograma: unidade_valor_m2_imovel
Frequência
0
500
1000
1500
2000
2500
3000
3500
4000
5000
10000
15000
20000
25000
unidade_valor_m2_imovel

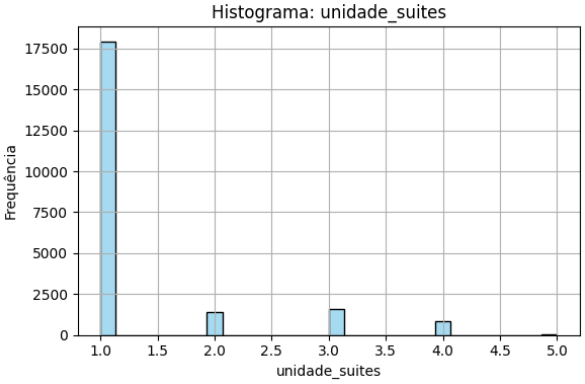

Boxplot - unidade_quartos
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_quartos

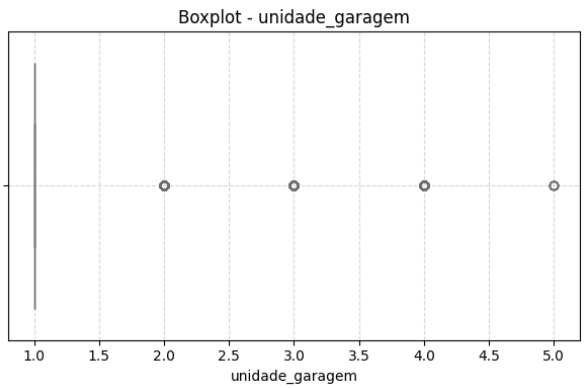

Histograma: unidade_quartos
Frequência
0
2000
4000
6000
8000
10000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_quartos

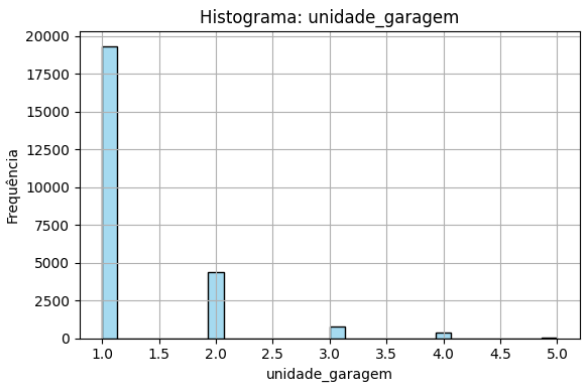

Boxplot - unidade_area
0
100
200
300
400
500
unidade_area

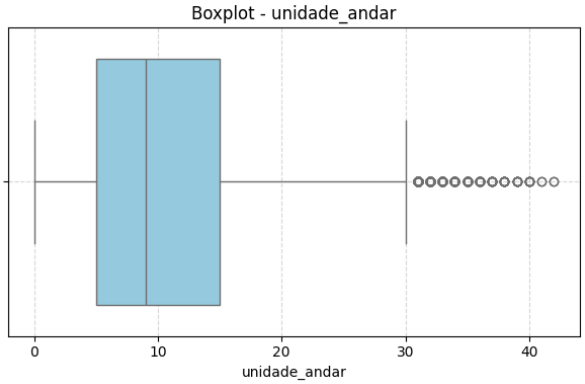

Histograma: unidade_area
Frequência
0
2000
4000
6000
8000
0
100
200
300
400
500
unidade_area

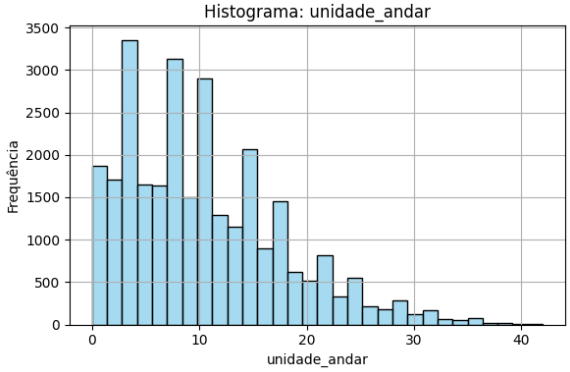

Boxplot - unidade_suites
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_suites

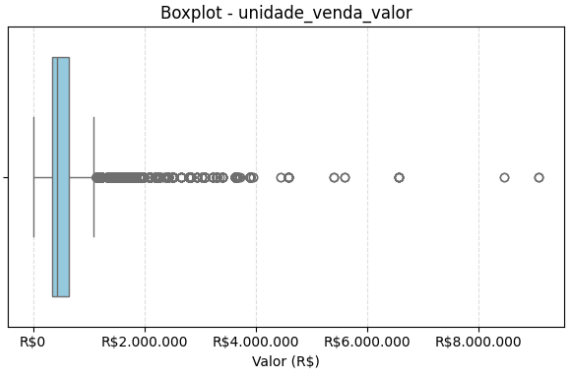

Histograma: unidade_suites
Frequência
0
2500
5000
7500
10000
12500
15000
17500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_suites

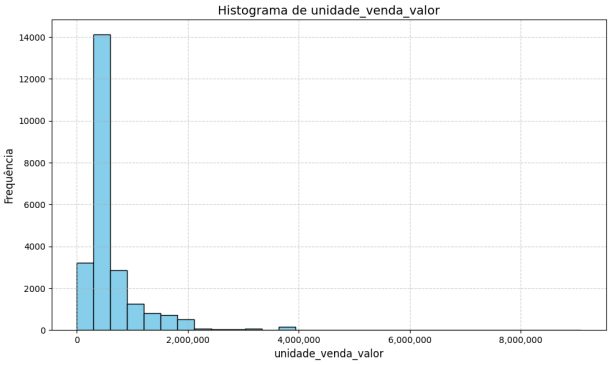

Boxplot - unidade_garagem
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_garagem

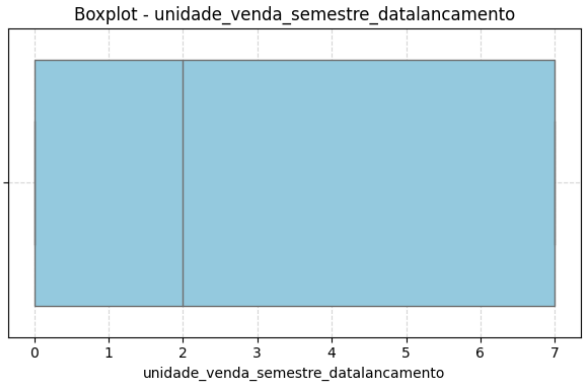

Histograma: unidade_garagem
Frequência
0
2500
5000
7500
10000
12500
15000
17500
20000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
unidade_garagem

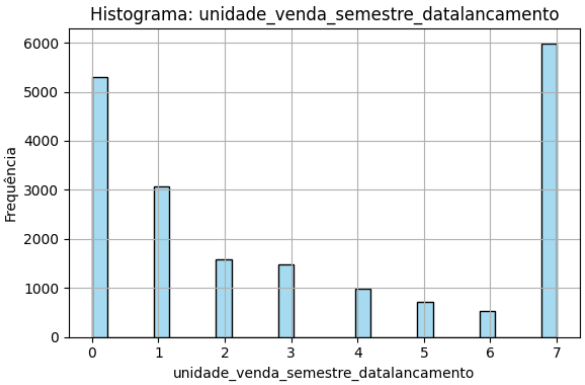

Boxplot - unidade_andar
0
10
20
30
40
unidade_andar

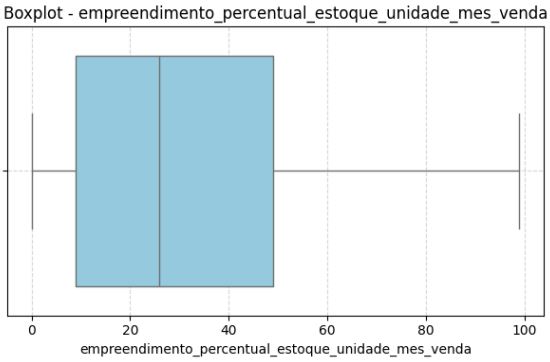

Histograma: unidade_andar
Frequência
0
500
1000
1500
2000
2500
3000
3500
0
10
20
30
40
unidade_andar

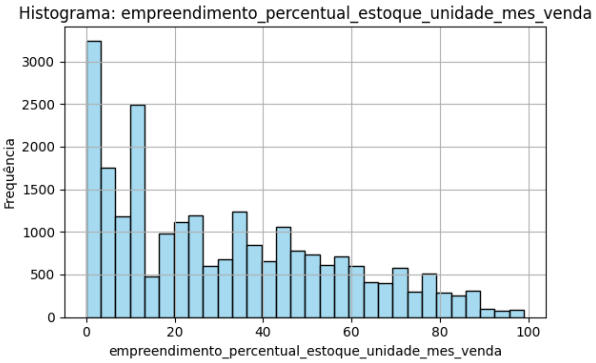

Boxplot - unidade_venda_valor
R$0
R$2.000.000
R$4.000.000
R$6.000.000
R$8.000.000
Valor (R$)

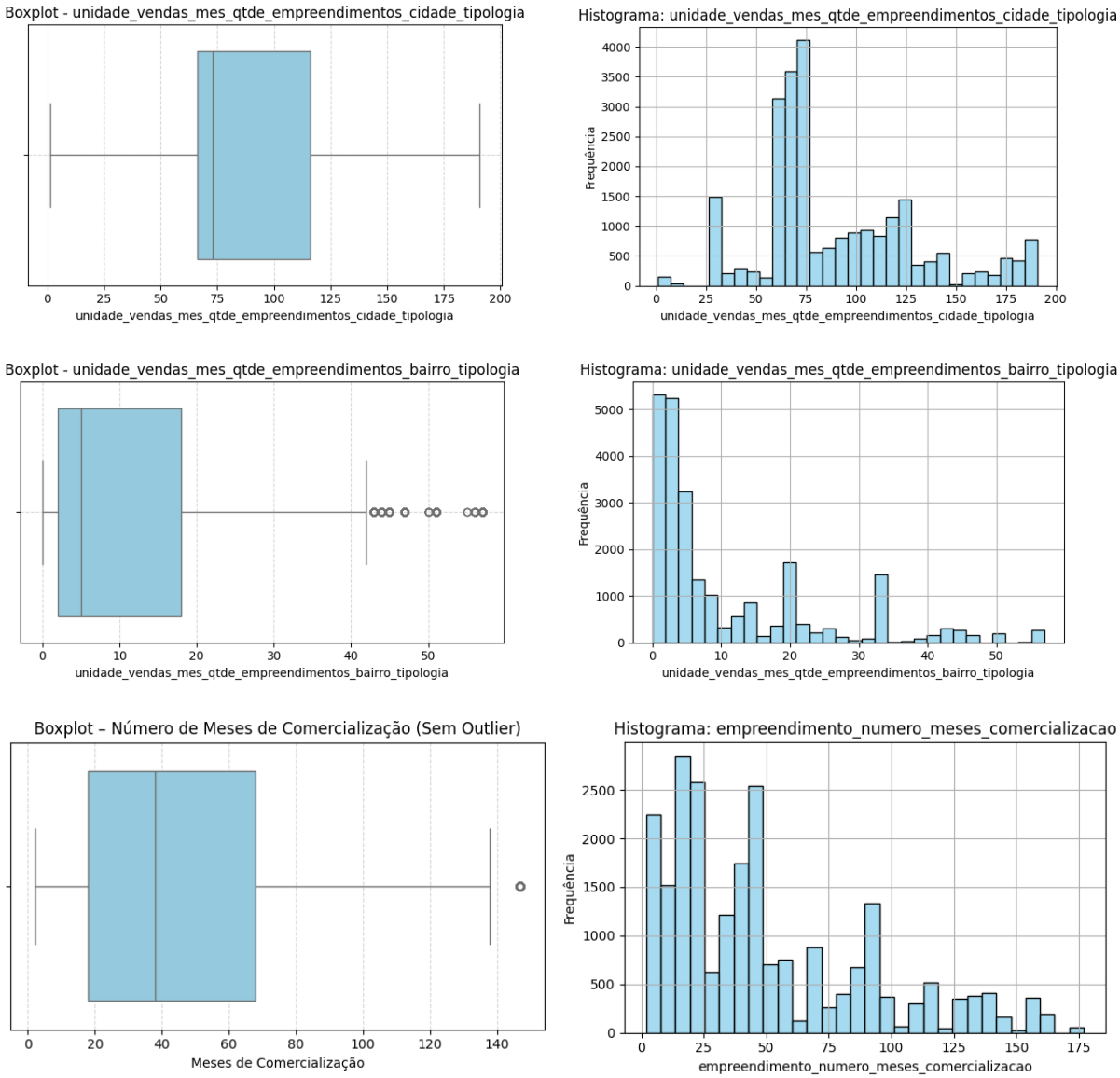

Histograma de unidade_venda_valor
Frequência
0
2000
4000
6000
8000
10000
12000
14000
0
2,000,000
4,000,000
6,000,000
8,000,000
unidade_venda_valor


Boxplot - unidade_venda_semestre_datalancamento
0
1
2
3
4
5
6
7
unidade_venda_semestre_datalancamento


Histograma: unidade_venda_semestre_datalancamento
Frequência
0
1000
2000
3000
4000
5000
6000
0
1
2
3
4
5
6
7
unidade_venda_semestre_datalancamento


Boxplot - empreendimento_percentual_estoque_unidade_mes_venda
0
20
40
60
80
100
empreendimento_percentual_estoque_unidade_mes_venda


Histograma: empreendimento_percentual_estoque_unidade_mes_venda
Frequência
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100
empreendimento_percentual_estoque_unidade_mes_venda



Fonte: O autor (2025)

Apesar da ampla variabilidade inerente ao mercado imobiliário da cidade do Recife — que abrange desde empreendimentos com grande número de unidades até os mais compactos, preços que variam do segmento econômico ao alto luxo, apartamentos de diferentes tamanhos, e ritmos de vendas distintos (rápidos ou mais lentos) —, não foram identificados valores que pudessem ser categorizados como *outliers* problemáticos ou anômalos que necessitassem de tratamento específico ou remoção. A qualidade dos dados é refletida no momento da captação por um time de ciências de dados exclusivo para este fim.

Ainda, segundo Aggarwal (2017), em contextos de alta variabilidade como o mercado imobiliário, valores extremos podem representar nichos de mercado legítimos, e não necessariamente ruído nos dados.

Isso sugere que a dispersão observada nos dados reflete a realidade multifacetada do mercado imobiliário primário, e não falhas ou erros na coleta das informações. A diversidade de valores, mesmo em extremos, representa nichos e características legítimas de empreendimentos e unidades, o que é um achado positivo para a robustez da base de dados e para a capacidade de generalização dos futuros modelos preditivos.

4.4.2.2.3 Conclusão da análise exploratória do dataset Unidades e Disponibilidades

A análise exploratória do *dataset* Unidades e Disponibilidades foi fundamental para obter uma compreensão granular das características que moldam o mercado imobiliário de Recife no nível da unidade. Os *insights* obtidos sobre a estrutura dos dados, a distribuição dos atributos e a ausência de *outliers* problemáticos são de grande valor para as etapas subsequentes do projeto.

Especificamente, a clareza na categorização dos atributos, a identificação precisa da volumetria de dados para cada classe-alvo (Velocidade e Resiliência de Vendas) e a confirmação de que a variabilidade dos dados reflete a realidade do mercado — em vez de anomalias — fortalecem a base para a modelagem preditiva. Embora alguns atributos booleanos e de contexto (construtoras, índices econômicos) demandem tratamento na fase de Preparação dos Dados, suas ausências foram compreendidas e mapeadas para futuras imputações ou integrações.

Esta fase de AED solidificou o entendimento sobre as variáveis que potencialmente influenciam a venda e a resiliência das unidades, pavimentando o caminho para a engenharia de *features* e a construção de modelos preditivos robustos e representativos do mercado imobiliário estudado.

4.4.2.3 Estatística descritiva

A estatística descritiva consiste na aplicação de medidas numéricas para resumir, organizar e interpretar as principais características de um conjunto de dados. Neste estudo, essa abordagem foi empregada com o objetivo de fornecer uma visão quantitativa detalhada dos atributos numéricos e categóricos dos *datasets* Empreendimentos e Vendas e Unidades e Disponibilidade. Conforme proposto por Tukey (1977), a análise exploratória de dados (AED) fornece uma estrutura robusta para sintetizar e visualizar padrões, detectar anomalias e compreender a estrutura subjacente dos dados antes da modelagem preditiva.

4.4.2.3.1 Estatística descritiva do dataset Empreendimentos e Vendas

Esta subseção detalha as características quantitativas do *dataset* Empreendimentos e Vendas, que é fundamental para a compreensão da dinâmica dos lançamentos e vendas no mercado imobiliário.

(i) Atributos Numéricos Selecionados e Medidas Estatísticas Calculadas

Para a análise descritiva, foram selecionados os seguintes atributos numéricos considerados mais relevantes para o problema de modelagem preditiva do sucesso comercial:

- empreendimento_qtde_total_unidades;
- empreendimento_qtde_unidades_vendidas;
- empreendimento_estoque_atual;
- vendas_jan_2022 até vendas_mar_2025.

Para cada um desses atributos, foram calculadas as seguintes medidas estatísticas, fornecendo um panorama completo de sua distribuição e dispersão:

- **Média (μ)** e **mediana (Q2)**: medidas de tendência central;
- **Desvio padrão (σ)**: medida de dispersão;
- **Primeiro (Q1) e terceiro quartil (Q3)**: Para análise da distribuição e identificação da faixa central dos dados.
- **Intervalo interquartil (IQR = Q3 - Q1)**: Indicador da dispersão dos 50% centrais dos dados e base para a detecção de *outliers*.
- **Limites inferior e superior para outliers (LI e LS) para outliers**: Definidos pela regra de Tukey como LI = Q1 – 1,5×IQR e LS = Q3 + 1,5×IQR;

(ii) Interpretação e Insights Quantitativos

Dispersão e Heterogeneidade: Observa-se uma dispersão significativa em diversos atributos numéricos, como *empreendimento_qtde_total_unidades* e as colunas de vendas mensais. Essa característica reflete a heterogeneidade inerente ao mercado imobiliário de Recife, onde coexistem empreendimentos de portes variados e as dinâmicas de vendas diferem consideravelmente entre si e ao longo do tempo.

Casos Específicos de *Outliers* (Reiteração e Justificativa): Conforme abordado na seção 4.4.2.2, é crucial reiterar que, em situações onde Q1 =Q3 =0 para determinadas variáveis de vendas, o IQR se torna nulo, inviabilizando a aplicação direta da regra de Tukey. Nestes casos, valores não-zero seriam tecnicamente classificados como *outliers*. Contudo, optou-se por manter tais valores, interpretando-os como picos naturais de vendas ou transações em meses específicos, que são representativos da dinâmica real do mercado e não como anomalias a serem removidas.

- *empreendimento_qtde_total_unidades*: Este atributo demonstra uma média de 99,7 unidades, porém com uma acentuada assimetria positiva e a presença de *outliers* significativos acima do limite superior. Este perfil indica a coexistência de um grande número de empreendimentos de pequeno e médio porte com alguns poucos empreendimentos de grande escala, refletindo a diversidade da oferta imobiliária.
- *empreendimento_estoque_atual* e *empreendimento_qtde_unidades_vendidas*: Ambos os atributos exibiram elevada dispersão e assimetria, o que é esperado e reflete os diferentes estágios de comercialização dos empreendimentos no momento da coleta dos dados. Valores baixos em estoque atual e altos em unidades vendidas são indicativos de empreendimentos com boa absorção, enquanto o oposto aponta para desafios na comercialização.
- Colunas de Vendas Mensais (*vendas_jan_2022* a *vendas_mar_2025*): A análise estatística dessas séries temporais de vendas evidenciou uma alta variabilidade, com uma concentração notável de valores zero para períodos anteriores ao lançamento de um empreendimento ou após o esgotamento total de suas unidades. Essa característica reforça a

premissa de que a consideração da data de lançamento e da data de venda total do empreendimento é fundamental para uma análise precisa do seu desempenho e para a construção das variáveis-alvo.

No **Apêndice D**, é possível analisar todas as métricas dos atributos do dataset Empreendimentos e Vendas.

4.4.2.3.2 Estatística descritiva do dataset Unidades e Disponibilidade

A análise estatística descritiva deste *dataset* busca sumarizar numericamente as principais características das unidades habitacionais comercializadas no mercado primário de Recife entre janeiro de 2022 e maio de 2025. Por meio de medidas de tendência central, dispersão e forma, obtém-se uma visão mais precisa sobre o comportamento dos atributos quantitativos e seu papel na construção dos modelos preditivos deste estudo.

(i) Atributos Numéricos Selecionados e Medidas Estatísticas Calculadas

A seleção dos atributos numéricos para esta análise estatística detalhada foi um processo criterioso, fundamentado em sua **relevância intrínseca para a caracterização do produto imobiliário** e sua **potencial influência nas decisões de compra e no desempenho de vendas**. Além de atenderem aos critérios de qualidade de dados (variabilidade real, baixa redundância e suficiência de dados) já estabelecidos na seção 4.4.2.2.2 sobre Análise de Dados Fora da Série (*Outliers*), esses atributos foram escolhidos por representarem as dimensões mais cruciais para a compreensão da **Velocidade de Vendas** e da **Resiliência de Vendas**.

Os atributos selecionados permitem uma investigação aprofundada sobre:

- Características físicas e de dimensionamento das unidades: unidade_quartos, unidade_area, unidade_suites, unidade_garagem, unidade_andar. Esses fatores são primordiais na adequação da unidade às necessidades e preferências dos compradores, impactando diretamente seu valor e liquidez.
- Aspectos financeiros e de mercado da unidade: unidade_valor_imovel, unidade_valor_m2_imovel, unidade_venda_valor. Essas variáveis são indicadores diretos da precificação e do posicionamento da unidade no mercado.

- Dimensões do empreendimento e seu contexto de mercado: empreendimento_qtde_total_unidades, empreendimento_pavimentos, empreendimento_unidades_por_andar. Esses atributos fornecem um panorama do porte e da densidade do projeto, elementos que influenciam a percepção de exclusividade e a dinâmica de oferta.
- Dinâmica temporal e de estoque: empreendimento_numero_meses_comercializacao, unidade_venda_semestre_datalancamento, empreendimento_percentual_estoque_unidade_mes_venda, unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia, unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia. Estes atributos são fundamentais para medir a evolução das vendas e a pressão do estoque, aspectos-chave para as métricas de velocidade e resiliência.

Em suma, a escolha desses atributos reflete uma abordagem focada em capturar os elementos mais explicativos do comportamento de vendas no mercado imobiliário, fornecendo a base quantitativa necessária para os modelos preditivos subsequentes.

A análise das medidas de tendência central (média e mediana) e de dispersão (desvio padrão, quartis e *IQR - Interquartile Range*) para os atributos numéricos selecionados forneceu uma compreensão aprofundada da distribuição e variabilidade dos dados.

(ii) Interpretação e Insights Quantitativos

- *empreendimento_pavimentos*: Empreendimentos em Recife possuem, em média, 20.6 pavimentos, com metade concentrada entre 15 e 25 andares. A presença de empreendimentos com 0 pavimentos reflete a inclusão de tipologias horizontais.
- *empreendimento_qtde_total_unidades*: A distribuição de unidades por empreendimento é assimétrica positiva (média de 165.6 vs. mediana de 127.0), indicando predominância de projetos de médio porte, mas com a existência de grandes empreendimentos (até 576 unidades).
- *empreendimento_unidades_por_andar*: A maioria dos andares (mediana de 6.0) possui entre 4 e 8 unidades, com uma distribuição simétrica e

baixa dispersão, mas com alguns andares com alta densidade (até 34 unidades).

- *empreendimento_numero_meses_comercializacao*: O ciclo de comercialização apresenta assimetria positiva (média de 49.4 vs. mediana de 40.0), com metade dos empreendimentos vendida em até 40 meses. A presença de empreendimentos com até 177 meses de comercialização (quase 15 anos) pode indicar relançamentos, estoque residual ou desaceleração extrema na absorção do produto.
- *unidade_andar*: A maioria das unidades está concentrada entre o 5º e o 15º andar (mediana de 9.0), com leve assimetria positiva devido à presença de andares mais altos (até 42º andar).
- *unidade_valor_imovel*: Os valores de imóveis exibem distribuição altamente assimétrica à direita, com média de R$ 630.615,90 e mediana de R$ 441.000,00, sugerindo concentração de preços entre R$ 340.000 e R$ 650.000. No entanto, a presença de unidades com valores superiores a R$ 9 milhões reflete o impacto de empreendimentos voltados ao segmento de alto padrão, criando uma cauda longa e elevando a média da distribuição. Essa heterogeneidade do mercado imobiliário já foi destacada por Zhu e Ferreira (2014), que demonstraram como a dinâmica urbana pode gerar clusters com valores significativamente distintos dentro de uma mesma cidade, refletindo diferentes perfis de absorção e liquidez.
- *unidade_area*: A área das unidades também é assimétrica positiva (média de 63.1 m² vs. mediana de 52.5 m²), predominando apartamentos compactos a médio-porte (40.8 m² a 68.0 m²), mas com unidades maiores (até 536 m²) representando a diversidade do mercado.
- *unidade_valor_m2_imovel*: O valor por m² tem leve assimetria positiva (média de R$ 9.882,30/m² vs. mediana de R$ 9.348,20/m²). A faixa de R$ 7.432,30/m² a R$ 11.491,60/m² concentra a maioria, mas a amplitude (R$ 4.210/m² a R$ 28.226/m²) destaca a vasta segmentação de mercado.
- *unidade_quartos*: A tipologia de 2 quartos apresenta a maior frequência, indicando sua predominância no mercado local (média de 2.2, mediana de 2.0), com a maioria variando entre 1 e 3 quartos, refletindo a demanda por imóveis de menor e médio porte.

- *unidade_suites*: Observa-se baixa variabilidade na quantidade de suítes, com predomínio absoluto de 1 suíte por unidade (Q1 = Q2 = Q3 = 1).
- *unidade_garagem*: De forma similar, a grande maioria das unidades (mediana e Q1/Q3 em 1.0) dispõe de 1 vaga de garagem, sendo esse o padrão de mercado.
- *empreendimento_qtde_estoque*: O volume de estoque remanescente por empreendimento é assimétrico positivo (média de 80.9 vs. mediana de 52.0), indicando que, embora muitos tenham estoque reduzido, há empreendimentos com grande volume de unidades disponíveis.
- *unidade_venda_estoque_empreendimento*: Fortemente assimétrico positivo (média de 52.6 vs. mediana de 28.0), mostrando que muitas vendas ocorrem com baixo estoque restante, mas uma parcela significativa acontece mesmo com o empreendimento tendo um grande volume de unidades.
- *unidade_venda_valor*: As estatísticas para o valor de venda (média de R$ 632.234,60 vs. mediana de R$ 440.000,00) são consistentes com o valor inicial do imóvel, confirmando a forte assimetria positiva impulsionada pelo segmento de luxo.
- *unidade_venda_semestre_datalancamento*: A distribuição é assimétrica, com muitas vendas nos primeiros semestres (Q1 em 0.0, mediana em 2.0), mas com um ciclo de vendas que pode se estender por até 7 semestres (quase 3,5 anos) para algumas unidades.
- *unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia*: No momento da venda, a quantidade de empreendimentos concorrentes similares na cidade varia, concentrando-se entre 66 e 116, com uma média de 89.1 empreendimentos por tipologia na cidade.
- *unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia*: A concorrência por bairro e tipologia é altamente assimétrica (média de 10.8 vs. mediana de 5.0), indicando que a maioria das vendas ocorre em contextos de baixa concorrência local, mas com picos em bairros mais saturados (até 57 empreendimentos).
- *empreendimento_percentual_estoque_unidade_mes_venda*: As vendas ocorrem frequentemente com o estoque em níveis médios a baixos

(mediana de 26%), mas também com percentuais de estoque maiores (até 99%), refletindo a variabilidade nas estratégias e momentos de venda.

No **Apêndice E**, é possível analisar todas as métricas dos atributos do dataset Unidades e Disponibilidades.

4.4.2.3.3 Considerações finais

A análise estatística descritiva realizada nas subseções anteriores fornece fundamentos robustos para a etapa de modelagem preditiva ao oferecer uma compreensão aprofundada das características dos dados. Por meio do cálculo de métricas como média, mediana, desvio padrão, quartis, IQR (Intervalo Interquartil) e limites para detecção de *outliers*, foi possível caracterizar a dispersão, a assimetria e a variabilidade dos atributos mais relevantes dos dois *datasets* utilizados neste estudo.

No *dataset* **Empreendimentos e Vendas**, observou-se grande heterogeneidade entre os projetos imobiliários, tanto em termos de porte quanto em dinâmica comercial. Essa variabilidade foi refletida nas diferenças de volume de unidades lançadas, de estoque atual e no padrão de vendas mensais, reforçando a importância de tratar temporalmente cada empreendimento em relação à sua data de lançamento. O comportamento das variáveis temporais, com muitos valores nulos, está intrinsecamente associado à ausência de comercialização em determinados períodos, sendo, portanto, representativo da realidade mercadológica e não de inconsistências.

No *dataset* **Unidades e Disponibilidades**, os atributos analisados revelaram estruturas assimétricas em várias dimensões — como valor dos imóveis, área útil e pavimentos — com caudas longas associadas à diversidade de produtos (de unidades econômicas a alto padrão). Essa distribuição heterogênea confirma achados de estudos como Zhu e Ferreira (2014), que demonstram como a dinâmica urbana e a segmentação do mercado imobiliário se refletem em padrões estatísticos dispersos.

Além disso, conforme defendido por Tukey (1977), a aplicação da análise exploratória de dados permite interpretar valores extremos não necessariamente

como anomalias, mas como sinais de subgrupos ou segmentos distintos do mercado. Assim, optou-se por manter a totalidade das observações, mesmo aquelas que extrapolam os limites definidos por regras clássicas de *outlier*, desde que coerentes com o domínio do problema.

Essas análises subsidiam importantes decisões para as próximas etapas, como:

- **Engenharia de atributos**: seleção de variáveis com alto poder explicativo e menor redundância;
- **Tratamento de valores extremos**: interpretação contextualizada ao invés de remoção automática;
- **Filtragem de registros**: refinamento de amostras com base na completude e temporalidade dos dados;
- **Ajuste de granularidade**: diferenciação entre análise em nível de empreendimento versus unidade, conforme apropriado ao objetivo de cada experimento.

Com isso, esta etapa conclui a fase de **Entendimento dos Dados** no ciclo CRISP-DM, estabelecendo um diagnóstico sólido e direcionado, que guiará os processos subsequentes de preparação dos dados, modelagem e avaliação.

4.4.3 Preparação dos dados

A fase de **Preparação dos Dados** é considerada uma das mais críticas no ciclo de desenvolvimento de projetos de mineração de dados, conforme definido pela metodologia CRISP-DM. Esta etapa compreende o conjunto de atividades necessárias para transformar dados brutos em dados analiticamente preparados e adequados à modelagem preditiva, garantindo qualidade, consistência, relevância e adequação estatística dos dados ao problema de negócio.

Segundo Han, Kamber e Pei (2011), a qualidade dos dados envolve múltiplas dimensões, como **acurácia, completude, consistência, pontualidade, credibilidade e interpretabilidade**, sendo essencial que tais atributos sejam rigorosamente tratados antes da aplicação de algoritmos de *machine learning*.

Além disso, a literatura destaca que etapas como seleção, integração, transformação e tratamento de dados ausentes ou atípicos não apenas reduzem o

ruído, mas aumentam substancialmente a eficácia dos modelos preditivos. A integração de fontes diversas, por sua vez, impõe desafios técnicos e semânticos que exigem rigor no mapeamento e na fusão das informações (Rahm & Do, 2000), especialmente em contextos complexos como o mercado imobiliário.

Dada a complexidade e a granularidade dos dados utilizados neste estudo — oriundos de diferentes fontes e com níveis variados de estruturação semântica e temporal —, esta etapa foi cuidadosamente estruturada em **sete subtópicos** que descrevem, de forma detalhada, o processo de preparação aplicado aos dois datasets principais (Empreendimentos e Vendas e Unidades e Disponibilidades):

- 4.4.3.1 Seleção dos Dados;
- 4.4.3.2 Redução dos Dados;
- 4.4.3.3 Tratamento de *Missing Data* e *Outliers*;
- 4.4.3.4 Integração e Consolidação dos Dados;
- 4.4.3.5 Criação de Novos Atributos;
- 4.4.3.6 Definição da Classe-alvo;
- 4.4.3.7 Transformação dos Dados.

Esta estrutura visa assegurar a transparência e a reprodutibilidade do processo de preparação, reforçando a integridade metodológica da pesquisa e estabelecendo bases sólidas para a etapa subsequente de modelagem preditiva.

4.4.3.1 Seleção dos dados

A etapa de seleção dos dados, conforme previsto na metodologia **CRISP-DM**, é responsável por determinar os subconjuntos mais adequados de dados a serem utilizados nas análises. Essa fase é essencial para garantir que os dados alimentem corretamente os modelos preditivos, estejam alinhados aos objetivos do projeto e possuam qualidade suficiente para sustentar inferências estatísticas confiáveis.

A seleção criteriosa dos dados contribui para maior eficiência nas fases posteriores do projeto, além de reduzir ruído, redundâncias e vieses que podem comprometer os resultados analíticos.

Neste estudo, foram utilizados dois conjuntos principais de dados:

- **Empreendimentos e Vendas**, contendo informações consolidadas no nível de empreendimento, incluindo dados cadastrais, datas importantes, atributos estruturais e série histórica mensal de vendas;
- **Unidades e Disponibilidades**, estruturado no nível da unidade habitacional, com atributos físicos, econômicos e comerciais detalhados.

Para viabilizar a integração entre esses dois *datasets*, foi criado um identificador único, denominado *id_empreendimento*, que garante a consistência relacional e a rastreabilidade das relações entre os níveis de granularidade distintos.

4.4.3.1.1 Critérios de seleção aplicados ao dataset Empreendimentos e Vendas

A filtragem dos registros seguiu os seguintes critérios:

Período de Lançamento: Todos os empreendimentos com data de lançamento registrada no atributo empreendimento_data_lancamento foram inicialmente mantidos, desde que válidos e completos. Nesta etapa, 64 empreendimentos foram removidos devido à ausência ou inconsistência na data de lançamento, impactando a remoção de 5.398 instâncias correspondentes no *dataset* Unidades e Disponibilidades. Os critérios específicos de corte temporal utilizados nos dois experimentos (Velocidade de Vendas e Resiliência de Vendas) serão detalhados na subseção 4.4.3.8, que trata da definição das classes-alvo para a modelagem supervisionada.

Tipo de Empreendimento: Apenas empreendimentos classificados como **“Apartamento”** foram incluídos, de forma a garantir **homogeneidade tipológica** e reduzir a variabilidade indesejada decorrente de outros tipos construtivos, como casas ou condomínios horizontais. Todos os registros dos datasets já estavam filtrados para este critério, não sendo necessário remover nenhuma informação.

Localização Geográfica: O escopo da pesquisa foi delimitado à cidade do **Recife**, conforme o campo *empreendimento_cidade*, em consonância com o foco geográfico da pesquisa. Todos os registros dos datasets já estavam filtrados para este critério, não sendo necessário remover nenhuma informação.

4.4.3.1.2 Critérios de seleção aplicados ao dataset Unidades e Disponibilidades

Após a aplicação dos filtros no *dataset* de empreendimentos, o *dataset* de unidades foi submetido a um processo de filtragem correspondente. Apenas as unidades associadas aos empreendimentos previamente selecionados foram mantidas, garantindo integridade referencial por meio do campo *id_empreendimento*.

Adicionalmente, foram eliminados registros de unidades que apresentavam inconsistências estruturais ou ausência de dados essenciais, tais como:

- **Preço total igual a zero ou nulo**;
- **Área privativa nula ou negativa**;
- **Registros sem vínculo identificável com empreendimento válido**.

Com o objetivo de **focar a análise nos imóveis residenciais destinados à venda**, também foram excluídas unidades cuja categoria não fosse classificada como “Residencial”, bem como aquelas cuja pretensão de comercialização não estivesse definida como “Vender”. Isso garante que o *dataset* represente o segmento específico de **mercado de apartamentos residenciais para venda** de interesse da pesquisa.

Após a aplicação rigorosa desses critérios de seleção, o *dataset* Empreendimentos e Vendas foi consolidado para um total de **227 empreendimentos**, enquanto o *dataset* Unidades e Disponibilidade resultou em **22.417 instâncias** de unidades habitacionais prontas para análise.

Este processo de filtragem assegurou a integridade e a confiabilidade dos dados analisados, assegurando que as instâncias incluídas na modelagem representem **observações reais, economicamente relevantes e tecnicamente robustas e confiáveis**.

A aplicação sistemática desses filtros resultou em um subconjunto de dados coeso, limpo, consistente e alinhado com os objetivos de modelagem do projeto. Essa base selecionada constitui a fundação para as etapas de preparação, modelagem e extração de conhecimento descritas nas próximas seções desta dissertação.

4.4.3.2.Redução dos dados

A fase de redução dos dados tem como objetivo eliminar atributos redundantes, invariantes ou irrelevantes para a tarefa preditiva, contribuindo para a

simplificação do espaço amostral, redução de ruído e melhoria da eficiência computacional dos modelos (Han et al., 2022). A exclusão de atributos foi conduzida com base em dois critérios principais: (i) variância nula, e (ii) natureza identificadora (*surrogate keys* ou atributos operacionais).

(i)Atributos com variância nula

A primeira categoria de atributos removidos diz respeito àqueles que apresentaram **variância igual a zero** ao longo de todo o conjunto de dados. Por possuírem **o mesmo valor para todas as instâncias**, tais variáveis não oferecem qualquer poder discriminativo para os modelos de aprendizagem de máquina. Os atributos eliminados por esse critério foram:

- unidade_tipo = "Apartamento"
- unidade_pretensao = "Vender"
- empreendimento_categoria = "Residencial"
- empreendimento_cidade = "Recife"

Tais atributos representam características invariantes em todo o dataset analisado. Sua manutenção não apenas seria redundante como também potencialmente prejudicial ao desempenho dos algoritmos, por aumentar desnecessariamente a dimensionalidade do problema.

(ii)Atributos identificadores (*surrogate keys*)

A segunda categoria de variáveis excluídas compreende os chamados **atributos identificadores ou técnicos**, que não carregam informação preditiva, não são úteis para segmentação, e tampouco se relacionam com o comportamento da variável-alvo. Esses atributos têm como única função o controle operacional do sistema, sendo normalmente usados para indexação, rastreabilidade ou auditoria. Foram removidos por esse critério:

- unidade_id
- empreendimento_data_cadastro

Um conjunto específico de atributos foi excluído da base de variáveis explicativas devido à sua **relação de endogeneidade com a classe-alvo**, configurando risco de *data leakage*. Tais atributos são derivados diretamente de métricas de desempenho comercial — em particular, de indicadores de vendas e

estoque que **compõem os próprios critérios de definição das classes-alvo** de velocidade e resiliência de vendas. Os atributos removidos por esse critério foram:

- *empreendimento_percentual_estoque_unidade_mes_venda*;
- *unidade_venda_semestre_datalancamento*;
- *unidade_venda_mes_datalancamento*
- *empreendimento_qtde_estoque*;
- *unidade_venda_estoque_empreendimento*.

Atributos categóricos com alta cardinalidade e baixa frequência relativa foram avaliados quanto ao seu potencial de induzir sobreajuste nos modelos. Em particular, os atributos *empreendimento_nome*, *construtora_nome* foram removidos por sua natureza identificadora, apresentando um valor único por instância e, portanto, **não contribuindo para a generalização do modelo**.

Conforme recomendado por Guyon & Elisseeff (2003), tais atributos devem ser excluídos na fase de preparação, uma vez que sua inclusão pode levar à sobreajuste (*overfitting*) em modelos supervisionados — especialmente quando possuem alta cardinalidade ou são únicos por instância. Além disso, atributos como empreendimento_data_cadastro carregam informações administrativas, mas não necessariamente refletem aspectos relevantes ao comportamento de mercado, principalmente após a normalização temporal dos dados de vendas.

4.4.3.3.Tratamento de dados faltantes (*missing data*) e valores atípicos(*outliers*)

O tratamento de dados faltantes e valores atípicos constitui uma etapa fundamental no processo de preparação dos dados, conforme delineado na metodologia CRISP-DM. A presença de valores ausentes ou extremos pode comprometer significativamente a qualidade estatística do *dataset*, afetando a performance dos algoritmos de aprendizado de máquina e introduzindo vieses ou ruído nos modelos preditivos.

4.4.3.3.1 Tratamento de dados faltantes (missing data)

Parte do tratamento de dados ausentes foi realizada ainda na etapa de **Seleção dos Dados** (Seção 4.4.3.1), onde foram removidas:

- Empreendimentos sem data de lançamento registrada (*empreendimento_data_lancamento*);
- Unidades com valor de imóvel (*unidade_valor_imovel*) nulo ou igual a zero;
- Unidades com área privativa (*unidade_area*) nula ou negativa;
- Registros sem vínculo identificável com empreendimentos válidos.

Além disso, atributos com proporções superiores a 40% de valores ausentes foram analisados quanto à sua relevância para os objetivos analíticos. Optou-se por mantê-los quando os valores ausentes não representavam falhas de coleta, mas sim regras de negócio explícitas, conforme discutido nas Seções 4.4.2.2.2 e 4.4.2.2.3 da Análise Exploratória dos Dados.

Destaca-se ainda a presença de atributos cuja ausência é esperada e derivada de regras de negócio.

No dataset Empreendimentos e Vendas, determinados campos são preenchidos de acordo com a sua data de lançamento e sua data de finalização de comercialização.

4.4.3.3.2 Tratamento de valores ausentes estruturais (Non-Applicable Values)

Em projetos com atributos temporais vinculados a marcos do ciclo de vida dos produtos, como datas de lançamento e comercialização, é comum a ocorrência de **valores ausentes estruturais**, ou seja, atributos cujo preenchimento não é aplicável a determinadas instâncias. Em vez de representarem dados faltantes acidentais, esses casos refletem **restrições lógicas ou temporais do negócio**.

No *dataset* Empreendimentos e Vendas, os atributos de vendas mensais (*vendas_jan_2022* a *vendas_mai_2025*) foram analisados com base nas seguintes datas de referência:

- *empreendimento_data_lancamento*: indica o início da comercialização;
- *empreendimento_data_totalmente_vendido*: indica a finalização do processo de vendas.

Da mesma forma, o atributo *empreendimento_data_totalmente_vendido* foi avaliado em relação ao campo *empreendimento_estoque_atual*. Para os empreendimentos ainda em comercialização (estoque atual > 0), a ausência desse

campo não representa perda de informação, mas sim uma **não aplicabilidade** derivada do estágio atual da venda.

Dessa forma:

- Para os meses **anteriores ao lançamento**, não era possível realizar vendas, portanto os valores foram tratados como NaN, indicando **não aplicabilidade** e **não ausência acidental**;
- Para os meses **após a finalização das vendas**, as unidades não estavam mais disponíveis, portanto, os valores também foram substituídos por NaN;
- Apenas para os meses **entre essas duas datas**, os valores numéricos foram mantidos (inclusive zeros reais), pois representam o efetivo comportamento de vendas no período de interesse.

No caso do atributo *empreendimento_data_totalmente_vendido*, quando o empreendimento apresentava unidades em estoque, o campo foi preenchido com NaT (Not a Time), indicando que o dado **não é aplicável** naquele momento.

No *dataset* Unidades e Disponibilidades, o atributo empreendimento_beira_rio foi excluído por ter 100% dos dados faltantes.

Para este *dataset*, existem situações especiais onde determinados campos são preenchidos apenas quando a unidade está vendida. Entre eles:

- *unidade_venda_valor,*
- *unidade_venda_data,*
- *unidade_venda_estoque_empreendimento,*
- *unidade_venda_mes_datalancamento,*
- *unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia,*
- *unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia,*
- *emprendimento_situacao_unidade_mes_venda,*
- *empreendimento_percentual_estoque_unidade_mes_venda.*

Em contrapartida também existem atributos que só fazem sentido estarem preenchidos se a unidade não estiver vendida:

- *empreendimento_qtde_estoque*
- *empreendimento_situação_atual*
- *empreendimento_estagio_obra_atual*

Adicionalmente, variáveis como *unidade_venda_semestre_dataentrega* e *unidade_venda_mes_relacao_dataentrega* somente são preenchidas quando a unidade está vendida e o empreendimento encontra-se entregue.

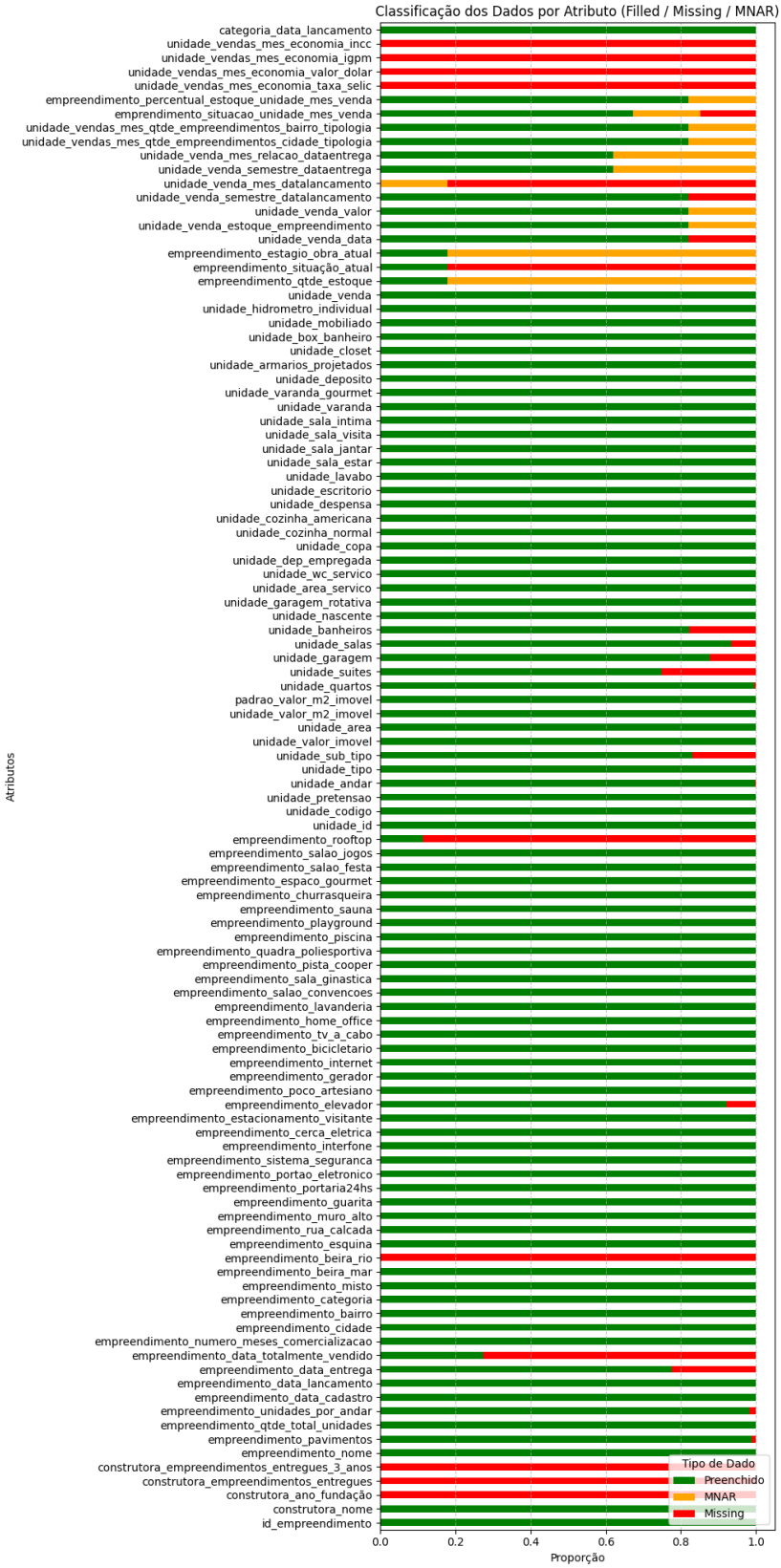
Para todos os atributos acima, os valores ausentes foram representados por NaN ou NaT, conforme seu tipo de dado (numérico ou temporal).

Essas ausências são consideradas ***ausências condicionais,*** cuja ocorrência depende diretamente do status da unidade e do estágio de comercialização do empreendimento, tal comportamento enquadra-se na categoria de ***Missing Not At Random*** **(MNAR)**, uma vez que a ausência está sistematicamente relacionada a variáveis observáveis. Por essa razão, esses valores foram mantidos no dataset como *missing não imputáveis*, sendo devidamente documentados e tratados como categorias especiais durante a modelagem.

Figura 4.14 - Proporção de Dados Cadastrados X Missing Data X MNAR depois do tratamento no *dataset* Empreendimentos e Vendas

Fonte: O autor (2025)

Figura 4.15 - Proporção de dados cadastrados X dados faltantes X MNAR depois do tratamento no *dataset* Unidades e Disponibilidade



Fonte: O autor (2025)

4.4.3.3.3 Tratamento de dados atípicos (Outliers)

Após a imputação e remoção de dados faltantes, foi conduzida uma análise detalhada de outliers com base em critérios estatísticos (boxplots e cálculo de limites via IQR) e critérios empíricos fundamentados no conhecimento do domínio do mercado imobiliário. Conforme argumentado por Tukey (1977), a análise exploratória permite identificar padrões e exceções relevantes, mas deve ser complementada com conhecimento contextual, a fim de evitar a exclusão inadvertida de observações legítimas, sobretudo em domínios com ampla variabilidade, como o mercado imobiliário.

Neste trabalho, adotou-se uma abordagem híbrida, combinando regras estatísticas com limites baseados na experiência prática do mercado local (Recife/ PE). Foram excluídas as instâncias que se enquadram nas seguintes condições:

- **Empreendimentos com menos de 10 unidades** (*empreendimento_qtde_total_unidades*), pois representam casos atípicos com baixa escala de comercialização. Nesta etapa, 5 empreendimentos foram removidos, impactando a remoção de 14 instâncias correspondentes no *dataset* "Unidades e Disponibilidades";
- **Área útil da unidade** (*unidade_area*) menor que 10 m² ou superior a 1.000 m², por se tratar de valores considerados fora do padrão de construções residenciais urbanas. Nenhuma instância com essas características foi identificada;
- **Valor por metro quadrado (***unidade_valor_m2_imovel***)** inferior a R$ 1.000, por ser economicamente inviável frente aos custos mínimos de construção e incorporação. Nenhuma instância foi identificada;
- **Empreendimentos com mais de 50 pavimentos** (*empreendimento_pavimentos*), limite acima da média dos edifícios residenciais da cidade e frequentemente associado a inconsistências cadastrais. Nenhuma instância foi identificada;
- **Unidades com mais de 6 quartos** (*unidade_quartos*), representando casos extremamente raros e fora do perfil padrão da maioria das tipologias comercializadas no mercado primário. Nenhuma instância foi identificada.

A adoção sistemática dessas regras de exclusão, fundamentadas em conhecimento técnico e estatístico, resultou em um conjunto de dados mais homogêneo, robusto e aderente à realidade do mercado estudado. Essa base limpa e confiável fornece maior estabilidade aos modelos preditivos e favorece sua interpretabilidade.

4.4.3.4 Integração e consolidação dos dados

A fase de integração e consolidação dos dados constitui uma etapa crítica no processo de preparação de dados, conforme estabelecido na metodologia CRISP-DM. Em projetos de *Data Mining*, a capacidade preditiva dos modelos frequentemente depende da incorporação de dados externos e contextuais que complementem a base primária, enriquecendo-a com variáveis latentes que refletem o ambiente real de negócio. Esse processo de enriquecimento semântico dos atributos é essencial para capturar nuances relevantes e construir modelos mais robustos e explicativos (Rahm & Do, 2000).

Embora os *datasets* Empreendimentos e Vendas e Unidades e Disponibilidades mantenham uma relação estrutural por meio do identificador comum id_empreendimento, não foi realizada uma junção completa entre eles nesta etapa. Em vez disso, a presente fase concentrou-se na integração de informações externas ao *dataset* Unidades e Disponibilidades, a fim de fortalecer sua capacidade analítica e preditiva.

O enriquecimento foi realizado por meio da incorporação de dois *datasets* auxiliares: **Construtoras Indicadores** e **Economia Indicadores**.

Adicionalmente, incorporou-se ao pipeline o *dataset* auxiliar **Regiões Administrativas do Recife**, com o objetivo de permitir a criação posterior do atributo categórico região. Embora não tenha ocorrido fusão direta nesse momento, a vinculação conceitual entre os bairros e suas respectivas macrozonas foi pré-processada para uso posterior na engenharia de atributos geográficos.

A seguir, detalha-se a lógica de integração adotada em cada caso:

(i) Integração com o *dataset* Construtoras Indicadores

Para adicionar ao *dataset* Unidades e Disponibilidades variáveis representativas do histórico e da reputação institucional das empresas construtoras,

foi utilizado o *dataset* **Construtoras Indicadores**. Este *dataset* foi compilado a partir de informações secundárias consultadas nos *sites* oficiais e redes sociais de cada construtora presente na amostra, contendo atributos importantes sobre o histórico e a performance das empresas construtoras no mercado.

Atributos Integrados:

Os atributos adicionados ao *dataset* Unidades e Disponibilidades foram: construtora_ano_fundação, construtora_empreendimentos_entregues, e construtora_empreendimentos_entregues_3_anos.

Chave de Integração:

A junção foi realizada por meio da correspondência dos nomes das construtoras nos dois *datasets*. Para garantir a consistência e a integridade da integração, os nomes foram previamente padronizados e normalizados, e técnicas de validação automática e manual foram empregadas para assegurar a unicidade e a correlação correta entre registros.

Processo:

Para cada unidade no *dataset* Unidades e Disponibilidades, realizou-se uma operação de junção (*left merge usando a biblioteca Pandas*) com o *dataset* Construtoras Indicadores com base na chave de integração, permitindo o preenchimento dos atributos designados em cada registro de unidade.

Após o processo de integração com o *dataset* Construtoras Indicadores, as proporções finais de completude e ausência para cada atributo foram as seguintes: apresentadas a seguir:

Ano de Fundação (*construtora_ano_fundacao***):**

Completude: 98,7%

Ausência: 1,3%

Portfólio de Empreendimentos Entregues (*construtora_empreendimentos_entregues***):**

Completude: 97,0%

Ausência: 3,0%

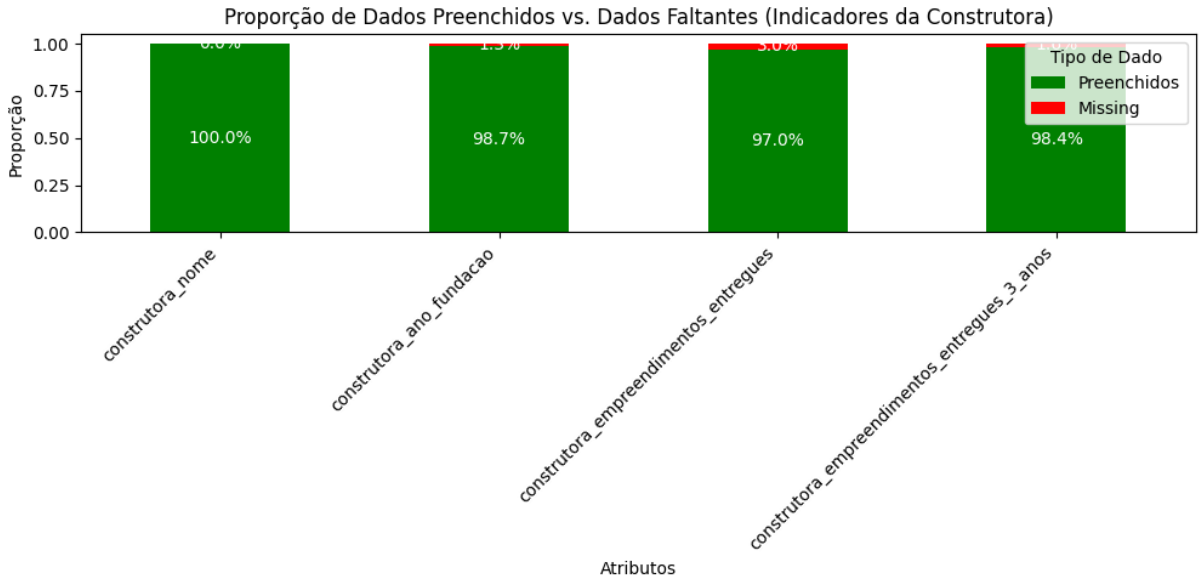
Empreendimentos Entregues nos Últimos 3 Anos (*construtora_empreendimentos_entregues_3_anos***):**

Completude: 98,4%

Ausência: 1,6%

A Figura 4.18 ilustra graficamente a distribuição entre dados preenchidos e ausentes para esses indicadores de construtoras após a etapa de integração, evidenciando o elevado grau de completude alcançado.

Figura 4.16 - Indicadores das construtoras - Dados cadastrados X dados faltantes depois da integração



Fonte: O autor (2025)

(ii) Integração com o *dataset* Economia Indicadores

A literatura reconhece que o desempenho do mercado imobiliário é fortemente sensível a fatores macroeconômicos, como taxa de juros, inflação e variação cambial (Zhu & Ferreira, 2014).

Com base nessa premissa, o *dataset* Unidades e Disponibilidades foi enriquecido com variáveis econômicas históricas extraídas do *dataset* **Economia Indicadores**. Os dados para este *dataset* foram extraídos de *sites* oficiais de instituições reconhecidas pela sua credibilidade em estatísticas econômicas, incluindo o Banco Central do Brasil (BC), a Fundação Getúlio Vargas (FGV) e o Instituto Brasileiro de Geografia e Estatística (IBGE). Este *dataset* contém séries históricas de índices relevantes.

Atributos Integrados:

Foram adicionados os atributos:

- unidade_vendas_mes_economia_taxa_selic;
- unidade_vendas_mes_economia_valor_dolar;

- unidade_vendas_mes_economia_ipca;
- unidade_vendas_mes_economia_igpm;
- unidade_vendas_mes_economia_incc.

Regra de Integração Condicional:

A atribuição dos valores econômicos seguiu uma lógica condicional baseada no *status* de comercialização da unidade, garantindo a relevância temporal dos indicadores:

Para Unidades Vendidas (unidade_venda = TRUE): Os atributos econômicos foram preenchidos com os valores correspondentes à data da venda da unidade (unidade_venda_data). Esta abordagem captura o cenário macroeconômico vigente no momento exato da transação, refletindo as condições de mercado sob as quais a venda ocorreu.

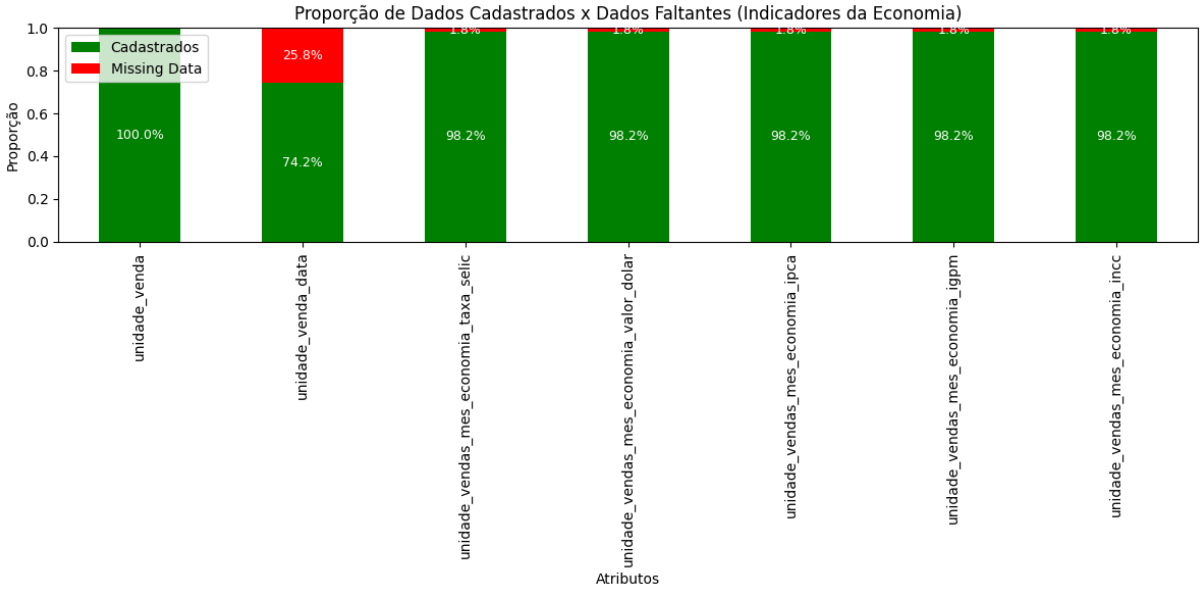
Para Unidades Em Comercialização (unidade_venda = FALSE): Para essas instâncias, que representam o estoque atual ou futuro de unidades, os atributos econômicos foram preenchidos com os valores referentes a **Maio de 2025**. Esta padronização fornece uma base comparável e atualizada dos indicadores para unidades ainda ativas no mercado.

A integração do *dataset* Unidades e Disponibilidades com o *dataset* Economia Indicadores resultou em um preenchimento quase completo dos atributos econômicos, que anteriormente apresentavam valores 100% ausentes. Esta alta taxa de completude é um reflexo direto da estratégia de integração condicional empregada, que considerou a data de venda para unidades já transacionadas e uma data de referência futura (Março de 2025) para unidades ainda em comercialização.

Após este processo, os atributos unidade_vendas_mes_economia_taxa_selic, unidade_vendas_mes_economia_valor_dolar, unidade_vendas_mes_economia_ipca, unidade_vendas_mes_economia_igpm e unidade_vendas_mes_economia_incc apresentaram uma **taxa de completude de 98,2%**, com apenas 1,8% de valores ausentes em cada um. Este resultado otimiza significativamente a base de dados para análises que incorporam fatores macroeconômicos.

A Figura 4.17 ilustra graficamente a proporção de completude alcançada para os indicadores econômicos após a integração.

Figura 4.17 - Indicadores Econômicos - Dados Cadastrados X Dados Faltantes depois da integração



Fonte: O autor (2025)

Impactos Analíticos da Consolidação:

A integração dessas variáveis contextuais no *dataset* Unidades e Disponibilidades ampliou substancialmente o escopo e a profundidade da base analítica. Seus principais benefícios são:

- **Aumento do poder preditivo**: A introdução de atributos sobre reputação da construtora e variáveis econômicas permite aos modelos detectar interações mais complexas, aumentando sua capacidade explicativa e generalizável.
- **Contextualização das vendas**: Ao associar o status da unidade com o contexto macroeconômico e institucional vigente, torna-se possível identificar padrões de comportamento e sazonalidade ligados ao ambiente externo.
- **Suporte à interpretabilidade**: Variáveis integradas favorecem a construção de modelos mais interpretáveis, alinhados com abordagens de X-AI (Explainable Artificial Intelligence), permitindo compreender o motivo pelos quais determinadas unidades têm desempenho superior em vendas.

Em síntese, a integração e consolidação de dados auxiliares ao *dataset* Unidades e Disponibilidades elevou significativamente sua qualidade e complexidade informacional. Esta etapa representa uma ponte crítica entre a

estruturação bruta dos dados e a geração de conhecimento preditivo, sustentando análises mais sofisticadas nas etapas subsequentes deste estudo.

4.4.3.5 Criação de atributos

A criação de atributos (*feature engineering*) é considerada uma das etapas mais determinantes no sucesso de projetos de aprendizado de máquina, conforme evidenciado na metodologia CRISP-DM. Esta fase extrapola a simples manipulação de dados ao incorporar conhecimento de domínio, criatividade e visão estratégica para derivar novas variáveis mais informativas e potencialmente mais preditivas (Zheng & Casari, 2018). Quando bem conduzida, a criação de atributos pode tanto elevar a qualidade dos indicadores dos modelos quanto aprimorar sua interpretabilidade.

Neste estudo, a criação de atributos foi conduzida a partir dos *dataframes* previamente consolidados, com foco em: (i) reforçar conectividade relacional entre os níveis de granularidade; (ii) agregar contexto geográfico; (iii) criar representações mais abstratas de variáveis contínuas; e (iv) incorporar atributos derivados relevantes à realidade do mercado imobiliário.

Atributo id_empreendimento: Reforço da conectividade relacional

Embora inicialmente introduzido na Seção 4.4.3.1 - Seleção dos dados, o atributo id_empreendimento adquire importância estratégica nesta etapa por sua função como chave de ligação entre os *datasets* Empreendimentos e Vendas e Unidades e Disponibilidades. Essa variável atua como elo relacional entre os diferentes níveis de granularidade – empreendimento e unidade –, permitindo:

- rastreamento e consistência entre registros,
- projeção de atributos do empreendimento para o nível da unidade, e
- consolidação de atributos de vendas para definição posterior da classe-alvo supervisionada.

Além disso, sua presença viabiliza análises cruzadas, enriquecimento semântico e modelagem hierárquica com maior robustez.

Atributo Região: Segmentação geográfica estratégica

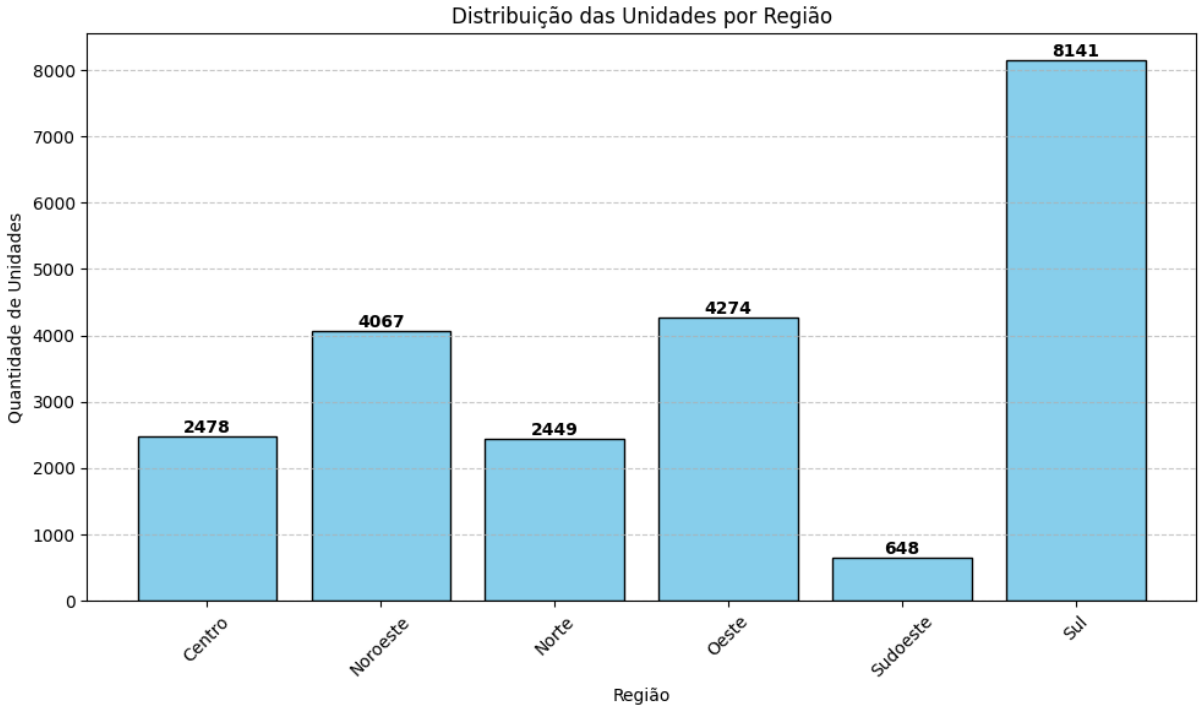
Atributos de localização muito específicos, como bairro, podem comprometer a generalização dos modelos devido ao alto número de categorias. Para mitigar esse risco e capturar dinâmicas territoriais mais amplas, foi criado o atributo região.

Justificativa: a generalização espacial dos bairros em regiões permite explorar padrões latentes influenciados por infraestrutura urbana, acessibilidade, centralidade, zonas de expansão e fatores socioeconômicos. A criação desse atributo alinha-se a estratégias de regularização e redução de dimensionalidade categórica, promovendo maior capacidade de generalização dos modelos.

Procedimento: o novo atributo foi derivado de um mapeamento entre o campo unidade_bairro e a divisão oficial de regiões político administrativas (RPAs) da cidade do Recife, conforme definido pela Prefeitura- h*ttps:// www2.recife.pe.gov.br/servico/perfil-dos-bairros* - e amplamente utilizado por construtoras e associações do setor. As categorias das regiões estabelecidas foram: Centro (RPA1), Norte (RPA2, Noroeste (RPA3), Oeste (RPA4), Sudeste (RPA5) e Sul (RPA6).

Impacto Esperado: espera-se que o atributo região atue como variável proxy para disparidades estruturais entre macrozonas urbanas, refletindo diferenças relevantes nos padrões de precificação, velocidade de vendas e atratividade comercial.

Figura 4.18 - Histograma com a distribuição das unidades por região



Fonte: O autor (2025)

Atributo padrao_imovel: Classificação por faixa de valor

O valor do metro quadrado (*unidade_valor_m2_imovel*) é uma métrica contínua amplamente usada como indicador de padrão de acabamento, perfil de público-alvo e valor agregado do imóvel. Contudo, variáveis contínuas com alta variabilidade podem reduzir a interpretabilidade dos modelos e amplificar ruídos.

Justificativa: A categorização do preço por metro quadrado permite segmentar o mercado em grupos de valor mais homogêneos (ex: econômico, médio, alto padrão), o que pode ser mais intuitivo para a interpretabilidade dos modelos e refletir diferentes nichos de compradores ou estratégias de venda. Além disso, muitos algoritmos de *Machine Learning* podem beneficiar-se de variáveis categóricas bem definidas.

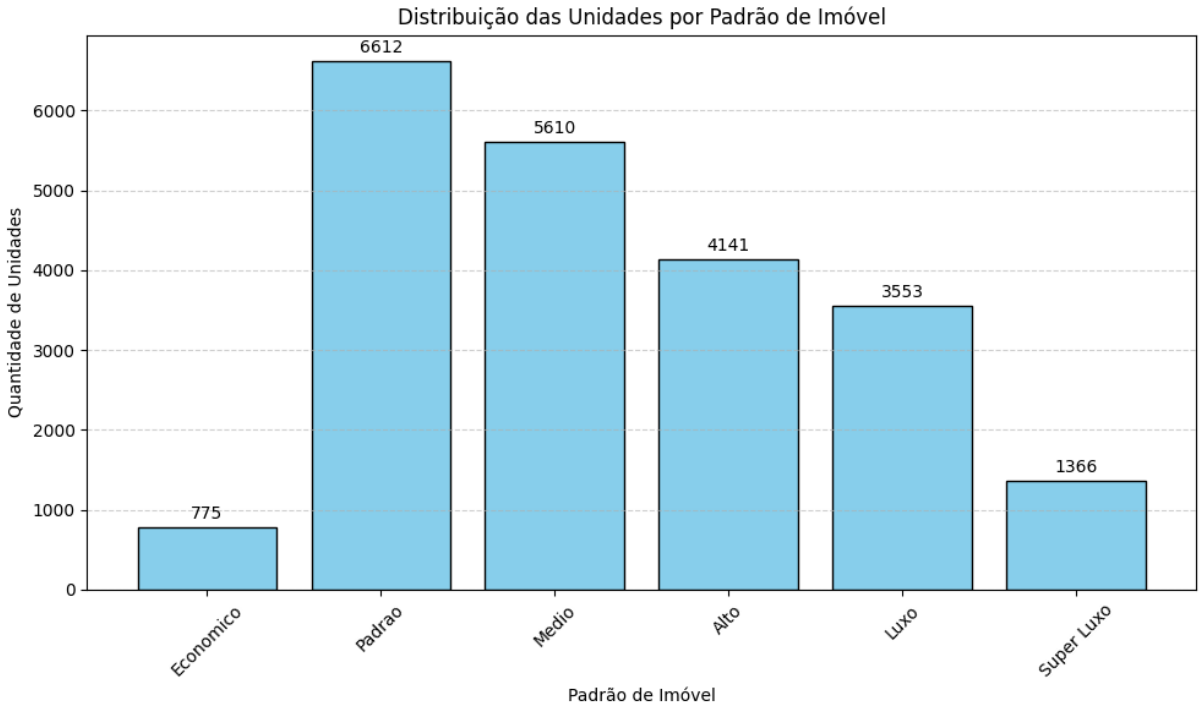
Processo de Criação: O atributo *padrao_imovel* foi derivado do *unidade_valor_m2_imovel* através de um processo de discretização. Os limiares para a definição das categorias foram estabelecidos com base em conhecimento empírico do mercado e padrões de classificação utilizados por empresas de pesquisa de mercado. As categorias resultantes são:

- **Econômico:** R$/m² até R$ 6.000

- **Padrão:** R$/m² entre R$ 6.001 e R$ 8.000
- **Médio:** R$/m² entre R$ 8.001 e R$ 10.000
- **Alto:** R$/m² entre R$ 10.001 e R$ 12.000
- **Luxo:** R$/m² entre R$ 12.001 e R$ 15.000
- **Super Luxo:** R$/m² acima de R$ 15.001

Impacto Esperado: o atributo deve permitir que os modelos identifiquem comportamentos de consumo e estratégias de precificação alinhadas a nichos específicos do mercado.

Figura 4.19 - Histograma com a distribuição das unidades por faixa de valor



Fonte: O autor (2025)

Atributo tempo_venda: Proximidade temporal ao lançamento

Justificativa: Em estudos de mercado (Sirmans et al., 2006; Rosenthal, 2020), o tempo decorrido entre o lançamento do empreendimento e a venda da unidade tem se mostrado uma variável altamente explicativa da atratividade e desempenho comercial de um imóvel.

Procedimento: o atributo foi criado como a diferença (em meses) entre a data de lançamento (empreendimento_data_lancamento) e a data de venda da unidade (unidade_venda_data), utilizando a seguinte fórmula: **Δ meses = (ano_2-**

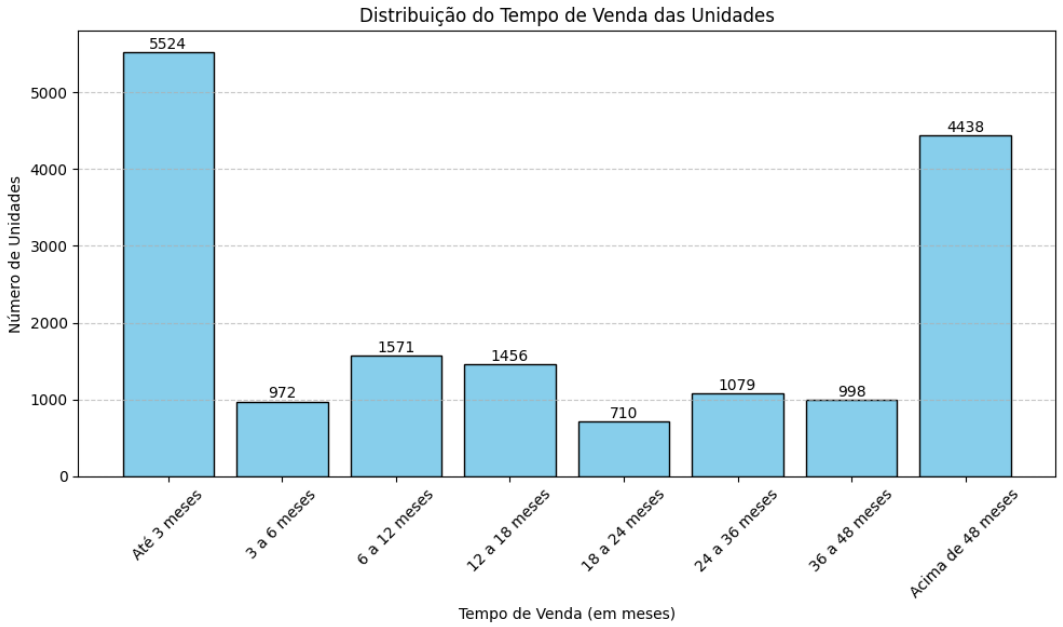
ano$_1$)*12 + (mês$_2$-mês). Para unidades ainda não vendidas, o valor foi imputado como NaN.

Impacto Esperado: esse atributo pode revelar padrões temporais de absorção de estoque, capturando o efeito da “curva de vendas” ao longo do ciclo de vida do empreendimento.

A criação estratégica desses atributos transformou o *dataset* em um formato mais propício para a análise preditiva. Essas novas variáveis, enraizadas no conhecimento de domínio, não apenas enriquecem a base informacional, mas também fornecem uma representação mais significativa das características do mercado imobiliário.

Espera-se que esses atributos ampliem a capacidade dos modelos de capturar relações não-lineares e dependências contextuais, além de fornecer maior transparência e valor analítico à tomada de decisão no mercado imobiliário.

Figura 4.20 - Histograma com a distribuição das unidades por proximidade temporal ao lançamento



Fonte: O autor (2025)

A etapa de **Criação de Atributos** concluiu a fase de enriquecimento semântico dos dados, convertendo variáveis brutas em descritores alinhados ao domínio imobiliário e adequados à interpretação dos modelos X-AI. O conjunto resultante de atributos cobre dimensões *espaciais* (região), *econômicas* (padrão de imóvel, indicadores macroeconômicos), *temporais* (idade comercial, tempo de

venda) e *organizacionais* (solidez da construtora), atendendo aos critérios de relevância, granularidade e completude estabelecidos no início da preparação.

4.4.3.6 Definição da Classe-alvo

A definição precisa da classe-alvo (ou variável dependente) é uma etapa fundamental no processo de mineração de dados, conectando diretamente os objetivos de negócio com a formulação do problema preditivo para os algoritmos de *Machine Learning* (Han et al., 2012).

No complexo cenário do mercado imobiliário, o conceito de "sucesso de vendas" pode ser multifacetado, englobando tanto a agilidade inicial na comercialização quanto a sustentabilidade da demanda ao longo do tempo. Para capturar essas distintas dimensões do desempenho, este estudo propõe a definição de duas classes-alvo complementares: **Velocidade de Vendas** e **Resiliência de Vendas**.

A estratégia de definição dessas classes envolve a classificação inicial no nível do empreendimento, utilizando o *dataset* Empreendimentos e Vendas, que agrega os dados de vendas mensais. Posteriormente, a classe-alvo determinada para cada empreendimento é transferida para as instâncias de unidade correspondentes, utilizando o atributo id_empreendimento para o *dataset* Unidades e Disponibilidades. É crucial ressaltar que, para cada classe-alvo, será gerado um *dataframe* de unidades distinto, garantindo que o subconjunto de observações e o contexto temporal sejam adequados à métrica de interesse.

Abaixo a tabela com a visão geral das regras e métricas de rotulagem:

Tabela 4.8 - Visão geral das regras e métricas de rotulagem

Métrica	Janela de observação	Regra de rotulagem (classe = 1)	Justificativa
Velocidade de Vendas	0–3 meses pós-lançamento	pct_vendido_3m $\geq 30\,\%$	Atratividade imediata e geração de caixa
Resiliência de Vendas	0–18 meses pós-lançamento	pct_remanescente_18m $\leq 20\,\%$	Aderência de produto no médio prazo

Fonte: O autor (2025)

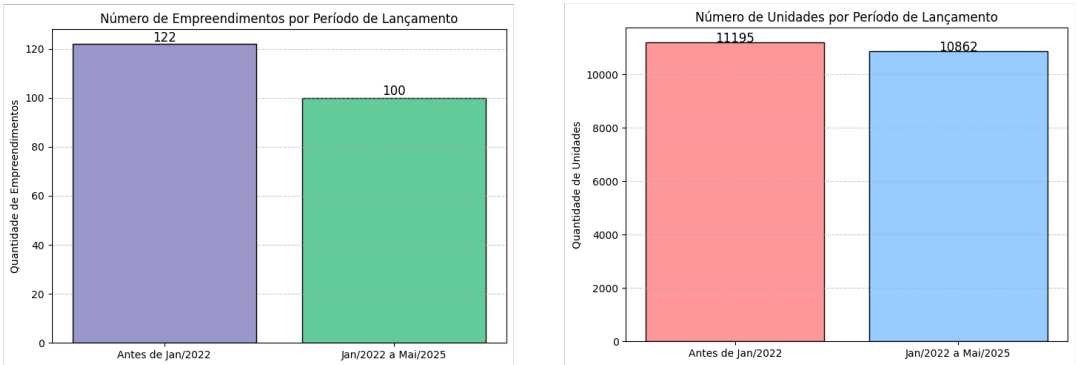
4.4.3.6.1 Classe-alvo: Velocidade de Vendas

A classe-alvo **Velocidade de Vendas** visa predizer a capacidade de um empreendimento de gerar atratividade comercial imediata e assegurar um fluxo de caixa rápido nos estágios iniciais de seu ciclo de obras.

Definição Operacional: Um empreendimento é classificado como de **Alta Velocidade de Vendas** se atingir ou superar 30% de suas unidades vendidas nos primeiros três meses após o lançamento. Caso contrário, é classificado como de Baixa Velocidade de Vendas. Essa métrica é considerada um *proxy* válido para a atratividade comercial imediata do empreendimento.

Escopo dos Dados: Para a definição desta classe-alvo, foram consideradas 100 empreendimentos e 10.862 instâncias de unidades habitacionais pertencentes a estes empreendimentos lançados entre Janeiro de 2022 e Maio de 2025. Este subconjunto representa 49,2% do total das instâncias listadas no *dataset* Unidades e Disponibilidades depois da fase anterior de Preparação de Dados. O recorte temporal específico garante um período de observação completo e adequado para a mensuração da métrica de vendas nos três primeiros meses pós-lançamento.

Figura 4.21 - Número de empreendimentos e unidades habitacionais por período após o lançamento



Fonte: O autor (2025)

Processo de Derivação: A classificação inicial é iniciada no *dataset* Empreendimentos e Vendas, , por meio do cálculo do percentual de unidades vendidas em relação ao total de unidades do empreendimento nos três primeiros meses após a **data de lançamento**. Com base nesse critério, cada empreendimento

é classificado segundo sua performance inicial de vendas. Em seguida, essa classificação é propagada para o dataset **Unidades e Disponibilidades**, replicando a mesma classe-alvo para todas as unidades vinculadas ao empreendimento, utilizando o atributo id_empreendimento como chave de relacionamento. A partir dessa junção, é gerado um novo dataframe contendo apenas as unidades pertinentes à análise de *Velocidade de Vendas*.

Durante a etapa de rotulagem supervisionada, observou-se que a classe inicialmente definida como “bem-sucedida” — ou seja, empreendimentos que venderam mais de 30% de suas unidades nos três primeiros meses — resultou como **classe majoritária** no conjunto de dados.

Segundo Chawla et al. (2010), muitos conjuntos de dados do mundo real apresentam distribuição assimétrica entre as classes, sendo compostos predominantemente por exemplos “normais”, enquanto apenas uma pequena fração corresponde a casos “anormais” ou “interessantes” — justamente aqueles que despertam maior interesse analítico. Em contextos de classificação binária, essa assimetria torna a **classe minoritária particularmente relevante**, pois ela frequentemente representa o evento crítico que se busca identificar. Por essa razão, é prática comum na literatura considerar essa classe menos frequente como **classe positiva**, principalmente em domínios nos quais a detecção de eventos raros (como fraude, falha ou doença) é prioritária.

Assim, optou-se por adotar a **classe com menor frequência** como a classe positiva nesta análise, por três motivos principais:

- Representa o evento de maior interesse — neste caso, empreendimentos com desempenho de vendas insatisfatório;
- Permite a utilização de métricas mais sensíveis ao comportamento da classe positiva, como sensibilidade (*recall*), F1-score, PR-AUC e Brier score;
- Está alinhada com práticas consolidadas em problemas de aprendizado supervisionado com desbalanceamento de classes.

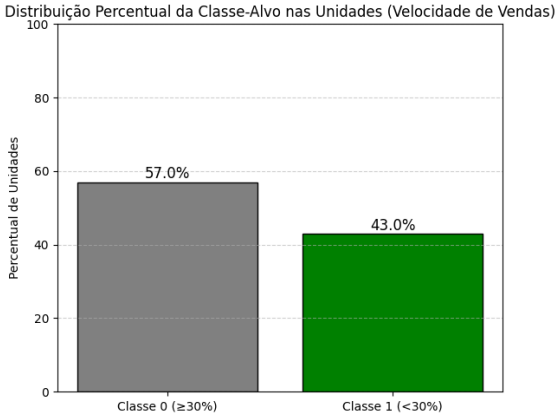
Abaixo, a tabela com a nova regra de rótulos após a classificação:

Tabela 4.9 - Nova regra de rotulagem para a classe-alvo positiva de Velocidade de Vendas

Métrica	Janela de observação	Regra de rotulagem (classe = 1)	Justificativa
Falha nas Vendas	0–3 meses pós-lançamento	pct_vendido_3m < 30 %	Risco comercial e de exposição de caixa.

Fonte: O autor (2025)

Figura 4.22 - Distribuição percentual da classe-alvo Velocidade de Vendas no dataset Unidades e Disponibilidade



Fonte: O autor (2025)

Dessa forma, a investigação volta-se para a identificação de fatores associados ao **insucesso comercial** de um empreendimento, isto é, à **falha em atingir 30% de vendas nos três primeiros meses após o lançamento** — um sinal de baixa velocidade de escoamento e, portanto, de alto risco comercial.

4.4.3.6.2 Classe-alvo: Resiliência de Vendas

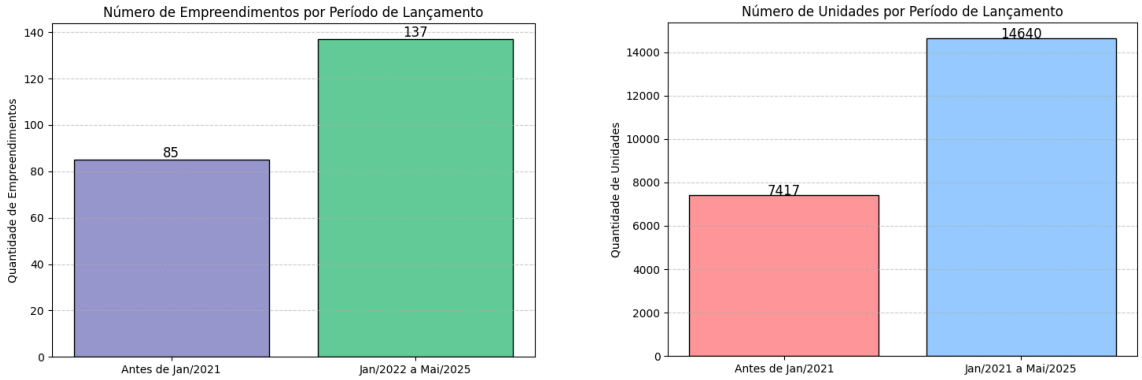
A classe-alvo **Resiliência de Vendas** foca na capacidade do empreendimento de manter a aderência do produto e a solidez da estratégia comercial ao longo do tempo, mesmo após o período de pico de lançamento.

Definição Operacional: Um empreendimento é classificado como de **Alta Resiliência de Vendas** se apenas 20% ou menos de suas unidades permanecerem não vendidas após 18 meses do lançamento. Empreendimentos que não atendem a este critério são classificados como de Baixa Resiliência de Vendas. Esta métrica reflete a longevidade e a capacidade do produto de sustentar a demanda no médio prazo.

Escopo dos Dados: Para a análise da Resiliência de Vendas, o escopo de dados é expandido para abranger instâncias de unidades de empreendimentos

lançados a partir de 2021. Este grupo compreende 137 empreendimentos e 14.640 instâncias de unidades habitacionais pertencentes a estes empreendimentos lançados entre Janeiro de 2022 e Maio de 2025, correspondendo a 66,2% das instâncias listadas, proporcionando um período de observação suficiente para a métrica de 18 meses de resiliência.

Figura 4.23 - Número de empreendimentos e unidades habitacionais por período após o lançamento



Fonte: O autor (2025)

Processo de Derivação: Similarmente à Velocidade de Vendas, a classificação inicial ocorre no *dataset* Empreendimentos e Vendas. A proporção de unidades vendidas é calculada 18 meses após a data_lançamento. A regra para rotulagem é a seguinte:

1. Empreendimento tem 20% ou menos de unidades à venda em 18 meses ou menos, classe positiva.

2. Empreendimento tem mais de 18 meses de lançado e ainda tem mais de 20% de unidades à venda, classe negativa.

3. Empreendimento tem menos de 18 meses de lançado e ainda tem mais de 20% de unidades à venda, não é rotulado e é excluído do conjunto.

Uma vez definida a classe binária para cada empreendimento, ela é transferida para as unidades correspondentes via id_empreendimento para o dataset Unidades e Disponibilidade. Um *dataframe* de unidades específico para a Resiliência de Vendas é então criado, contendo as observações adequadas a esta classe-alvo.

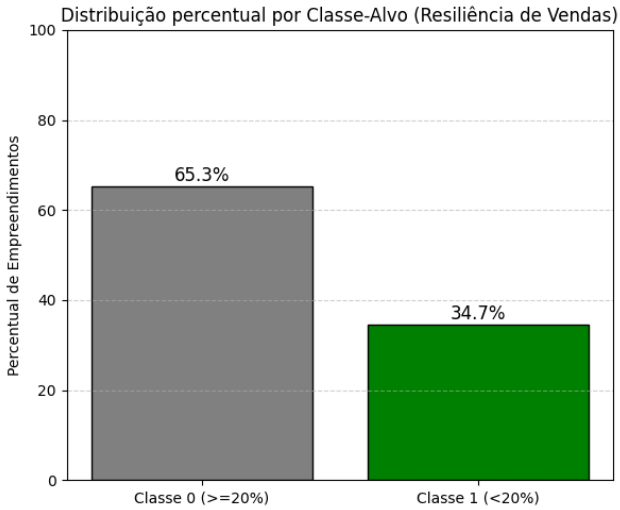
Abaixo, a tabela com a regra de rótulos após a classificação:

Tabela 4.10 - Regra de rotulagem para a classe-alvo positiva de Resiliência de Vendas

Métrica	Janela de observação	Regra de rotulagem (classe = 1)	Justificativa
Resiliência de Vendas	0–18 meses pós-lançamento	pct_remanescente_18m ≤ 20 %	Aderência de produto no médio prazo

Fonte: O autor (2025)

Figura 4.24 - Distribuição percentual da classe-alvo Resiliência de Vendas no dataset Unidades e Disponibilidade



Fonte: O autor (2025)

4.4.3.6.3 Resumo do processo de rotulagem (Nível Empreendimento → Nível Unidade)

A rotulagem é inicialmente calculada no *dataset* de **Empreendimentos e Vendas** e depois propagada ao *dataset* **Unidades e Disponibilidade** por meio de *id_empreendimento*, garantindo consistência entre granularidades.

Tabela 4.11 - Resumo das etapas do processo de rotulagem da classe-alvo

Etapa	Ação	Saída / Objeto Gerado
1	**Filtrar** df_empreendimentos_original_novos_atr: • VV → lançamentos 01/2022-05/2025 • RV → lançamentos 01/2021-05/2025	Subconjunto de empreendimentos elegíveis para cada métrica
2	**Calcular métricas** em nível empreendimento: • pct_vendido_3m = (vendas_0_3m / estoque_total) • pct_remanescente_18m = (1 - vendas_0_18m / estoque_total)	Colunas de proporção adicionadas ao dataframe de empreendimentos
3	**Gerar labels binários**: • classe_VV = 1 se pct_vendido_3m <0,30 • classe_RV = 1 se pct_remanescente_18m ≤ 0,20	df_alvo_VV e df_alvo_RV contendo id_empreendimento + label

4	**Mesclar** labels ao dataframe de unidades via id_empreendimento	df_unidades_VV e df_unidades_RV com granularidade unidade + classe-alvo
5	**Validar janelas** e **excluir** empreendimentos sem período completo	Dataframes finais prontos para transformação e modelagem

Fonte: O autor (2025)

4.4.3.6.4 Considerações finais

A utilização dessas duas classes-alvo distintas permite uma análise mais abrangente do desempenho de vendas no mercado imobiliário, possibilitando o desenvolvimento de modelos preditivos que enderecem tanto a atratividade inicial quanto a capacidade de sustentação do produto ao longo do tempo, fornecendo *insights* estratégicos para diferentes etapas do ciclo de vida dos empreendimentos. A criação de *dataframes* separados para cada classe-alvo assegura a integridade dos dados e a validade dos experimentos de modelagem subsequentes.

4.4.3.7 Transformações de dados

A etapa de transformação dos dados consiste na aplicação de operações que modificam atributos existentes ou derivam novos, com o objetivo de adaptar o conjunto de dados às exigências dos algoritmos de modelagem e melhorar a qualidade das representações disponíveis. No contexto da metodologia CRISP-DM, essa fase é fundamental para assegurar a **consistência, compatibilidade e interpretabilidade dos dados**, além de viabilizar comparações entre diferentes instâncias.

Esta fase também marca a transição conceitual entre a **engenharia de atributos** — voltada à construção de variáveis com significado interpretável e relevância para o negócio — e a **transformação numérica**, que busca adequar os dados às **exigências matemáticas dos algoritmos preditivos**, como escalonamento, codificação e representação vetorial. Ao integrar essas duas dimensões: semântica e computacional, essa fase finaliza o processo de preparação dos dados conforme o roteiro metodológico proposto pelo CRISP-DM, resultando em

um conjunto de dados adequado tanto à análise estatística quanto à aplicação de técnicas de aprendizado de máquina.

Cada uma das transformações realizadas será descrita nos subtópicos a seguir, incluindo sua motivação, a metodologia empregada e o impacto esperado na modelagem preditiva, com atenção especial à **manutenção da interpretabilidade** dos modelos e da **coerência semântica** com o problema de negócio.

As transformações foram agrupadas por tipo de atributo, respeitando suas características estruturais e operacionais, conforme recomendações de Chapman et al. (2000) e Han et al. (2012). Quando necessário, destacam-se abordagens específicas aplicadas a atributos individuais.

Transformações nos atributos de data

A transformação de atributos temporais teve como objetivo decompor informações de data em componentes mais simples, interpretáveis e compatíveis com algoritmos de modelagem supervisionada. Foram consideradas quatro variáveis principais com representação temporal:

- *empreendimento_data_lancamento*: data de lançamento comercial do empreendimento;
- *empreendimento_data_totalmente_vendido*: data em que todas as unidades foram vendidas;
- *empreendimento_data_entrega*: data de entrega do empreendimento para os moradores;
- *unidade_venda_data*: data de venda individual da unidade habitacional.

Esses atributos originais foram convertidos para o tipo *datetime*, permitindo a extração de variáveis derivadas que representam diferentes dimensões temporais. Entre as variáveis derivadas, destacam-se:

- **Ano (ano_*)**: representa a tendência temporal no horizonte do estudo, capturando efeitos macroeconômicos ou regulatórios por ano;
- **Mês (mes_*)**: permite capturar padrões de sazonalidade intra-anual no comportamento de compra;
- **Trimestre (trimestre_*)**: usado para capturar ciclos trimestrais de mercado, alinhados a práticas contábeis e estratégias de lançamento.

Essa decomposição segue uma prática amplamente adotada na literatura de Ciência de Dados, conhecida como *calendrical feature extraction*, e enquadra-se na fase de Transformação de Dados da metodologia CRISP-DM.

A opção por não aplicar *cyclical encoding* (por exemplo, transformações seno e cosseno para variáveis sazonais como mês ou dia da semana) foi deliberada, considerando a ênfase deste estudo em explicabilidade. Embora esse tipo de codificação possa representar a natureza circular de atributos temporais (ex: mês 12 é próximo de mês 1), ele reduz significativamente a interpretabilidade dos modelos, especialmente em técnicas baseadas em árvore e em análises orientadas ao negócio.

Após a extração dos componentes temporais, os atributos originais de data foram removidos do conjunto de dados final por três razões principais:

1. **Incompatibilidade com algoritmos de modelagem**: a maioria dos modelos supervisionados não lida bem diretamente com atributos do tipo datetime;
2. **Eliminação de redundância**: todas as informações relevantes foram decompostas em variáveis derivadas mais úteis e explicáveis;
3. **Prevenção de colinearidade e overfitting**: a permanência da variável original poderia introduzir correlação indesejada com os atributos derivados, distorcendo a aprendizagem do modelo.

Essa abordagem fortalece a interpretabilidade do modelo e facilita a análise de padrões temporais no mercado imobiliário.

Transformações nos atributos booleanos

A base de dados analisada contém 54 atributos booleanos, representando a presença ou ausência de características específicas das unidades habitacionais e dos empreendimentos. Embora atributos booleanos sejam naturalmente compatíveis com algoritmos de classificação, foi necessário aplicar uma série de procedimentos para avaliar sua variabilidade, relevância preditiva e associação com a variável de interesse (classe_alvo).

(i) Etapas da análise

A avaliação e transformação desses atributos seguiu um processo estruturado em quatro etapas principais:

- **Análise de Associação Estatística**

  Foi aplicado o **teste do Qui-Quadrado de independência** entre cada atributo booleano e a variável-alvo. Esse teste permite verificar se a proporção de casos positivos e negativos da classe difere significativamente entre os grupos True e False de cada atributo. A hipótese nula considera ausência de associação. Foram considerados estatisticamente relevantes os atributos com valor-p < 0,05.

- **Importância no Modelo Preditivo**

  A importância de cada atributo foi avaliada a partir do método **Permutation Feature Importance (PFI)** aplicado a um modelo de Árvore de Decisão treinado com validação cruzada. O PFI estima o impacto de um atributo na acurácia do modelo ao embaralhá-lo aleatoriamente, permitindo medir sua contribuição marginal à predição (Breiman, 2001).

- **Análise de Variância e Desbalanceamento**

  A distribuição dos valores True e False foi examinada para identificar **atributos com baixa variância**, ou seja, características extremamente raras ou quase sempre presentes (com proporção ≥95% de um dos valores). Em geral, atributos severamente desbalanceados não contribuem significativamente para modelos supervisionados e podem introduzir ruído.

- **Critério de Seleção Combinado**

  A seleção final dos atributos booleanos relevantes baseou-se em um critério combinado:

•O atributo foi **mantido** se:

a) apresentou associação estatística significativa com a classe-alvo (valor-p < 0,05), **ou**

b) apresentou **importância relevante** no modelo (PFI), **e** não foi considerado severamente desbalanceado.

Esse critério híbrido buscou equilibrar **evidência estatística**, **contribuição prática para os modelos** e **adequação estatística** (variabilidade suficiente), garantindo que os atributos finais fossem não apenas informativos, mas também estáveis e interpretáveis.

Atributos removidos

Com base nesse processo de análise combinada, **20 atributos booleanos** foram removidos do conjunto de dados. Esses atributos apresentaram **baixa variância** e **ausência de associação estatística com a classe-alvo**, ou ainda, importância nula ou irrelevante nos modelos preditivos. Abaixo estão listados:

- *unidade_dep_empregada*;
- *empreendimento_sistema_seguranca*;
- *empreendimento_portaria24hs*;
- *unidade_wc_servico*;
- *unidade_sala_estar*;
- *empreendimento_beira_mar*;
- *empreendimento_cerca_eletrica*;
- *empreendimento_esquina*;
- *empreendimento_misto*;
- *empreendimento_rooftop*;
- *empreendimento_tv_a_cabo*;
- *unidade_armarios_projetados*;
- *unidade_box_banheiro*;
- *unidade_despensa*;
- *unidade_escritorio*;
- *unidade_mobiliado*;
- *unidade_nascente*;
- *unidade_sala_intima*;
- *unidade_sala_visita*.

A remoção desses atributos reduz a dimensionalidade do conjunto de dados, mitigando riscos de *overfitting*, melhora a eficiência computacional e contribui para modelos mais interpretáveis, sem perda significativa de desempenho preditivo. A tabela com os valores de proporção da classe-alvo (classe_alvo) para cada atributo booleano (proporcao_classe1_true, proporcao_classe1_false), valor-p do teste Qui-quadrado, e as frequências de 0 e 1 (% Valor 0 e % Valor 1) pode ser consultada no **Apêndice F**.

Transformações em atributos categóricos

Os atributos categóricos foram transformados com o objetivo de viabilizar sua utilização em algoritmos supervisionados de aprendizado de máquina, ao mesmo tempo em que se preserva a interpretabilidade do modelo e a rastreabilidade semântica das variáveis originais. A escolha da técnica de codificação levou em consideração o nível de cardinalidade, a distribuição de frequência das categorias e a compatibilidade com abordagens explicáveis como SHAP, LIME e coeficientes lineares.

(i) Atributo “empreendimento_bairro” – Agrupamento de categorias e One-Hot Encoding

O atributo empreendimento_bairro, de alta cardinalidade, apresentou 30 categorias distintas com distribuição altamente assimétrica. Para reduzir a dimensionalidade, foi adotada uma estratégia de agrupamento baseada em frequência acumulada: os bairros que, em conjunto, representavam até 90% das observações foram mantidos como categorias individuais, enquanto os demais, com baixa representatividade estatística (≤10%), foram agrupados sob a nova categoria OUTROS_BAIRROS.

Essa abordagem, com suporte na literatura de pré-processamento para modelagem supervisionada (Han et al., 2022), preserva a rastreabilidade semântica das principais regiões da cidade, melhora a performance de modelos lineares e de árvore, e facilita a interpretação dos resultados por meio de variáveis *dummies* interpretáveis.

Após o agrupamento, foi aplicado *One-Hot Encoding* e o atributo original empreendimento_bairro foi removido do conjunto de dados, visto que seu conteúdo informacional encontra-se representado de forma distribuída nas variáveis derivadas.

Tabela 4.12 - Benefícios técnicos e explicativos dessa estratégia

Benefício técnico	Benefício explicável
Reduz número de colunas no One-Hot	Fica claro para o leitor quais bairros foram tratados individualmente
Evita sparsidade	Permite que modelos lineares e de árvore lidem melhor com os dados
Mantém representação de todos os dados	Nenhum bairro é descartado, apenas agrupado
Excelente para XAI (SHAP, LIME, coef.)	Atributos binários são intuitivos e rastreáveis

Fonte: O autor (2025)

(ii) Atributos de baixa cardinalidade – One-Hot Encoding direto

Para os atributos categóricos com baixa cardinalidade, o One-Hot Encoding foi aplicado diretamente, sem necessidade de agrupamento prévio, conforme prática recomendada para variáveis nominais com poucas categorias. Esses atributos são:

- *empreendimento_regiao*;
- *unidade_sub_tipo*;
- *padrao_imovel*;
- *emprendimento_situacao_unidade_mes_venda*;
- *empreendimento_estagio_obra_atual*;
- *empreendimento_situação_atual*.

Todos apresentavam até 5 categorias, com distribuição de frequência razoavelmente balanceada, o que garante boa representatividade estatística e compatibilidade com técnicas preditivas e explicáveis.

Transformações em atributos numéricos

Nesta etapa, atributos contínuos ou discretos foram tratados com diferentes estratégias conforme sua natureza estatística, distribuição, cardinalidade e relação com a variável-alvo.

(i) Atributos Inteiros de Baixa Cardinalidade

Atributos inteiros com baixo número de categorias distintas (geralmente até 6) foram transformados preferencialmente por *One-Hot Encoding*, preservando sua natureza semântica e maximizando a interpretabilidade dos modelos.

- ***unidade_quartos***: variando entre 1 e 5 quartos, apresentou associação significativa com a classe-alvo ($p < 0.01$). Como o número de quartos representa uma variável arquitetônica fundamental, foi aplicada codificação binária com uma coluna para cada valor.
- ***unidade_suites***: discretamente distribuído entre 0 e 3 suítes, também mostrou relação estatisticamente significativa com o desfecho. Optou-se por *One-Hot Encoding* para manter a granularidade e facilitar a análise de impacto por quantidade de suítes.

- ***unidade_banheiros***: embora originalmente representado como *float*, trata-se de um atributo naturalmente inteiro e foi convertido para *int*. Recebeu codificação categórica com base nos valores observados, dada sua contribuição estatística e semântica.
- ***unidade_garagem***: a associação com a classe-alvo foi fraca, mas, por razões de explicabilidade e sua possível interação com atributos de padrão, o atributo foi mantido discretizado em faixas (0, 1, 2+ vagas).
- ***unidade_salas***: com baixa variabilidade (0 a 3), e sem associação estatística clara, foi simplificado em dois grupos: 0–1 salas e 2 ou mais salas, visando reduzir complexidade e ruído.

Tabela 4.13 - Transformações e justificativas para atributos numéricos de baixa cardinalidade

Atributo	Transformação Sugerida	Justificativa Principal
unidade_quartos	OneHotEncoding	Relação não linear clara e alta explicabilidade
unidade_suites	OneHotEncoding	Boa distinção por quantidade de suítes
unidade_banheiros	iOneHot ou Ordinal	Forte associação e fácil interpretação
unidade_garagem	Int ou binarização (≥2)	Relevância baixa, mas pode ajudar
unidade_salas	Int + possível agrupamento	Pouco informativo isoladamente

Fonte: O autor

A análise estatística completa dos atributos numéricos de baixa cardinalidade pode ser encontrada no **Apêndice G**.

(ii) Atributos Inteiros de Alta Cardinalidade

Atributos com maior amplitude e distribuição assimétrica exigiram transformações mais cuidadosas para evitar distorções nos modelos lineares e, ao mesmo tempo, manter sua capacidade explicativa.

- ***construtora_ano_fundacao***: transformado em idade_construtora = 2025 - ano_fundacao, posteriormente discretizado em faixas (0–10, 11–20, >20 anos). Essa transformação torna o atributo mais semântico e aplicável à análise de maturidade empresarial.
- ***construtora_empreendimentos_entregues***: dada sua dispersão e assimetria, foi discretizado em três faixas (0–5, 6–15, >15). Essa

abordagem favorece a explicabilidade, mesmo sendo possível aplicar log(1+x) em outros contextos.

- ***unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia*** e **cidade_tipologia**: ambos relacionados ao grau de concorrência ou saturação da tipologia em determinada localidade. Foram discretizados em tercis, criando categorias de baixa, média e alta competitividade.
- ***unidade_venda_semestre_dataentrega***: mantido como variável ordinal, já que representa o intervalo de tempo entre venda e entrega em semestres (0 a 7). Sua escala interpretável justifica o uso direto.
- ***construtora_empreendimentos_entregues_3_anos***: pelo acúmulo de zeros e poucos valores altos, foi discretizado em faixas (0, 1–2, >2 entregas recentes). Alternativamente, pode ser binarizado (entregou ou não).
- ***unidade_venda_mes_relacao_dataentrega***: transformado em faixas temporais com base em conhecimento de negócio: até 6 meses (antecipada), 7–12 (proximidade), 13+ (pós-entrega). Essa discretização facilita insights sobre timing de vendas.
- ***empreendimento_unidades_por_andar***: discretizado em categorias estruturais: 1, 2, 3, 4, 5+ unidades por andar. A escolha está alinhada com a semântica da arquitetura de torres residenciais.
- ***unidade_andar***: discretizado em faixas de altura com significado prático: baixo (≤4), médio (5–10), alto (11–20), muito alto (>20), refletindo percepção de valor por localização vertical.
- ***empreendimento_qtde_total_unidades***: agrupado com base em porte do empreendimento: pequeno (≤40), médio (41–120), grande (>120). A decisão foi orientada por práticas comuns do setor.
- ***empreendimento_pavimentos***: discretizado conforme critérios arquitetônicos: baixo (≤4), médio (5–10), alto (11–20), muito alto (>20). Isso favorece análises com modelos explicáveis como Árvores de Decisão.
- ***empreendimento_numero_meses_comercializacao***: variável de tempo decorrido para esgotamento das unidades. Discretizado em rápida (≤12

meses), média (13–24), lenta (25–36), muito lenta (>36). Permite associação direta com velocidade de vendas.

Tabela 4.14 - Transformações e justificativas para atributos numéricos de alta cardinalidade

Atributo	Transformação sugerida	Justificativa
construtora_ano_fundacao	Derivar idade e discretizar em 0–10, 11–20, 20+ anos. Aplicar OHE (nova, estabelecida, tradicional).	Ano absoluto não é interpretável. Idade da construtora é mais explicável e relevante.
construtora_empreendimentos_entregues	Discretizar em 0–5, 6–15, 16+ e aplicar OHE (pequena, média, grande).	Reflete porte da construtora. Evita influência de escala nos modelos.
unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia	Discretizar em tercis (baixa, média, alta). Aplicar OHE.	Representa competição local por tipologia. Alto valor explicativo.
unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia	Discretizar em tercis (baixa, média, alta). Aplicar OHE.	Similar ao atributo anterior, mas em escala mais ampla.
unidade_venda_semestre_dataentrega	Manter como está (ordinal de 0 a 7).	Já está discretizado com boa interpretabilidade.
construtora_empreendimentos_entregues_3_anos	Discretizar em tercis e aplicar OHE (pouco, média, muito ativa).	Capta atividade recente da construtora, mais relevante que histórico amplo.
unidade_venda_mes_relacao_dataentrega	Discretizar em faixas: até 6 meses, 7–12, 13+.	Facilita interpretação da antecedência ou estoque.
empreendimento_unidades_por_andar	Discretizar em 1, 2, 3, 4, 5+ unidades.	Captura densidade de planta por pavimento, com efeito direto na atratividade.
unidade_andar	Discretizar em (0–4), (5–10), (11–20), (20+). Aplicar OHE (baixo, médio, alto, muito alto).	Relaciona-se com percepção de vista, ventilação e segurança.

Fonte: O autor (2025)

A análise estatística completa dos atributos numéricos de alta cardinalidade pode ser encontrada no **Apêndice H**.

(iii)Atributos Contínuos

- ***unidade_venda_valor***: mantido como contínuo. A variável apresenta assimetria à direita, mas seu valor absoluto é relevante. Modelos baseados em árvore lidam bem com sua forma original. Em regressão, pode ser normalizado.

- ***unidade_valor_m2_imovel***: removido da modelagem, pois seu conteúdo já se encontra representado pela variável categórica padrao_imovel, discretizada em categorias explicáveis (Econômico, Médio, Luxo, etc.).
- ***unidade_area***: discretizada em faixas de área útil com significado de mercado:
    - até 20 m²: super compacto
    - 21–40 m²: compacto
    - 41–70 m²: médio compacto
    - 71–100 m²: médio
    - 101–150 m²: médio grande
    - 151–200 m²: grande
    - 200 m²: muito grande

Essas faixas foram definidas com base em critérios práticos adotados no setor imobiliário e aplicadas com codificação *One-Hot Encoding*. Tal abordagem facilita a comparação entre perfis de unidades e aumenta a interpretabilidade dos modelos.

5. DESENVOLVIMENTO E AVALIAÇÃO DOS MODELOS PREDITIVOS

Esta seção apresenta a execução prática da fase de modelagem preditiva, conforme definida na metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*). Embora a modelagem seja formalmente identificada como a Fase 5 do processo CRISP-DM, optou-se por organizá-la nesta **Seção 5 da dissertação** com o objetivo de separar conceitualmente a preparação dos dados (Seção 4) da aplicação efetiva dos algoritmos de aprendizagem de máquina, seus resultados e a posterior extração de conhecimento.

Essa decisão tem como finalidade facilitar a compreensão do leitor, permitindo que a Seção 4 se concentre exclusivamente na engenharia, tratamento e transformação dos dados, enquanto esta Seção 5 concentra-se na **execução dos modelos**, **avaliação de desempenho** e **interpretação dos resultados** por meio de técnicas de inteligência artificial explicável (X-AI).

Foram consideradas duas hipóteses principais:

(i) a **Velocidade de Vendas**, como indicador de tração inicial do empreendimento; e

(ii) a **Resiliência de Vendas**, como medida de estabilidade e liquidez ao longo do ciclo comercial.

Essas hipóteses foram modeladas como problemas de classificação supervisionada, com classes binárias previamente definidas com base em critérios de negócios. Para abordar tais problemas, foram implementados três modelos com diferentes características de explicabilidade: D***ecision Tree, Random Forest* e *Logistic Regression***.

Estão seção está organizada em quatro subseções. Primeiramente, a **Arquitetura da solução** descreve o ambiente computacional e as ferramentas utilizadas. Em seguida, a **Redução da dimensionalidade e seleção dos atributos para modelagem** documenta os últimos ajustes antes do treinamento. A terceira subseção trata da **Seleção dos modelos**, com base em sua adequação ao problema e à exigência de interpretabilidade.

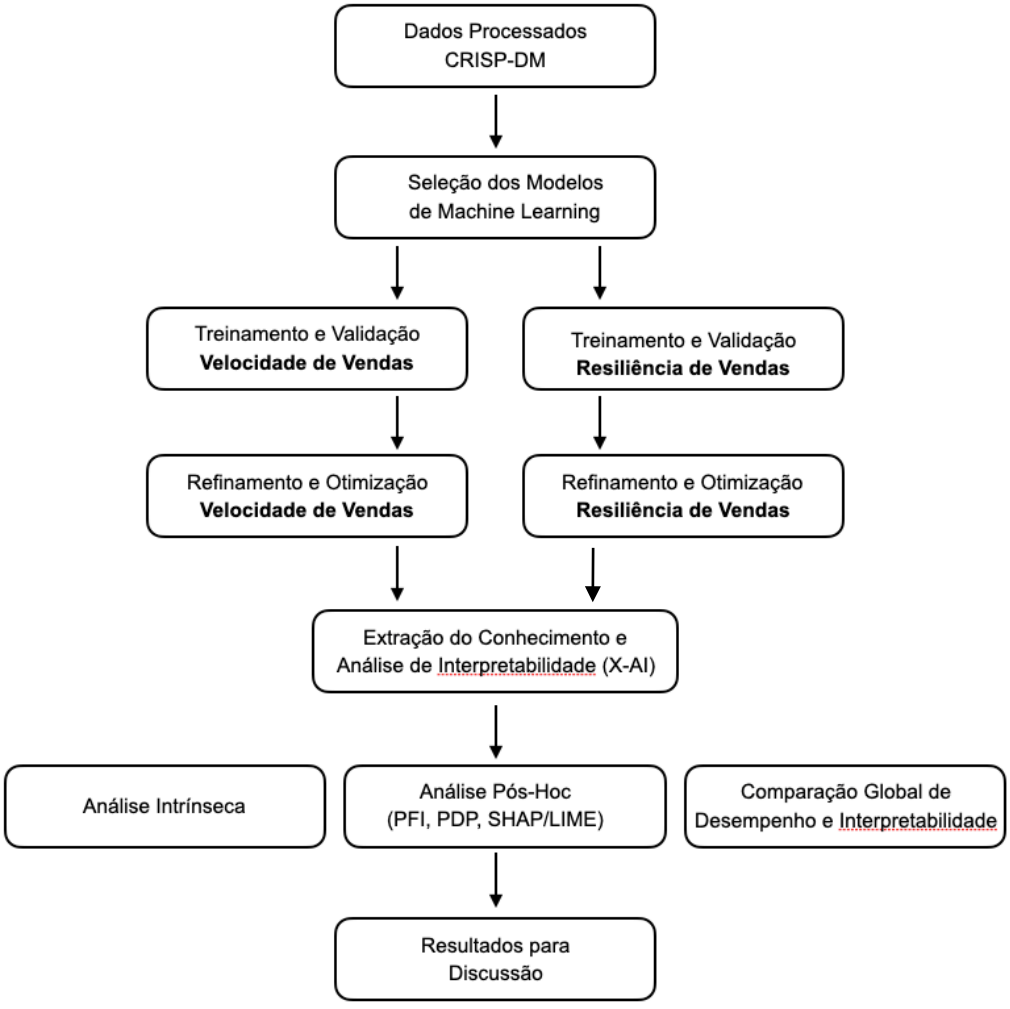
Na quarta parte, **Refinamento, execução e avaliação dos modelos**, o refinamento detalha os processos de ajuste fino dos modelos, incluindo tuning de hiperparâmetros e engenharia adicional de atributos, a execução utiliza os modelos

com os melhores hiperparâmetros encontrados e a avaliação extrai os indicadores encontrados.

5.1 ARQUITETURA DA SOLUÇÃO

A implementação computacional da solução preditiva proposta nesta dissertação foi conduzida por meio de uma arquitetura modular, composta por ferramentas e bibliotecas amplamente utilizadas em projetos de ciência de dados e aprendizado de máquina. A Figura 5.1 apresenta o fluxograma geral da solução desenvolvida, destacando o fluxo de trabalho desde a preparação dos dados até a extração de conhecimento explicável por meio de técnicas de X-AI.

Figura 5.1 – Arquitetura da Solução para Modelagem Preditiva e Extração de Conhecimento



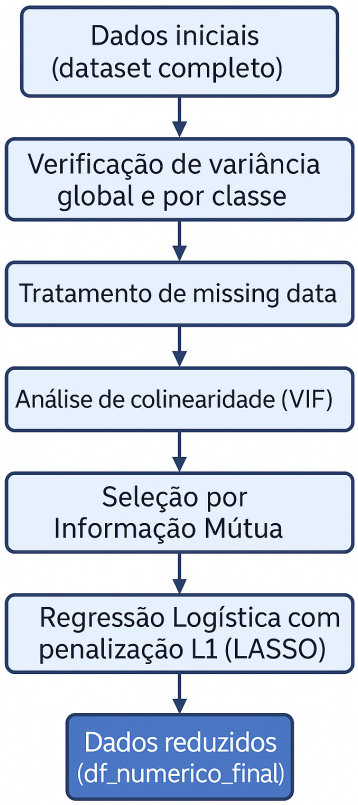
Fonte: O autor (2025)

Essa arquitetura foi essencial para garantir um fluxo de trabalho reprodutível, robusto e compatível com os princípios de interpretabilidade e transparência exigidos para a aplicação prática dos modelos no setor imobiliário.

5.2 REDUÇÃO DE DIMENSIONALIDADE E SELEÇÃO DE ATRIBUTOS PARA MODELAGEM

A fim de reduzir a complexidade dos dados e aumentar a robustez dos modelos preditivos, foi realizado um processo sistemático de redução de dimensionalidade e seleção de atributos. Esse processo visou eliminar variáveis redundantes, pouco informativas ou potencialmente ruidosas, resultando em um conjunto de atributos mais consistente com o objetivo de previsão. A Figura 5.2 apresenta as principais etapas do pipeline executado.

Figura 5.2 - Etapas aplicadas para redução de dimensionalidade dos dados.



Fonte: O autor

(i) Verificação de variância global e por classe

Inicialmente, avaliou-se a variabilidade de cada atributo, considerando tanto a variância global quanto a variância intra-classe. A análise permitiu identificar variáveis sem variação informativa, isto é, atributos com **variância igual a zero**, que não contribuem para a discriminação entre classes. Na base Velocidade de Vendas, apenas o atributo empreendimento_salao_convencoes apresentou variância global nula, sendo removido do conjunto de dados, e na base Resiliência de Vendas foram removidos os atributos empreendimento_salao_convencoes e o atributo empreendimento_estacionamento_visitante.

(ii) Tratamento de dados ausentes (*missing data*)

Na etapa seguinte, tratou-se a incompletude dos dados. Adotou-se a regra de exclusão progressiva:

- Atributos com **mais de 10% de valores ausentes** foram completamente removidos;
- Atributos com **menos de 10% de valores nulos** foram mantidos, mas as instâncias correspondentes foram excluídas.

Como resultado, no *dataframe* Velocidade de Vendas 14 atributos foram eliminados e 248 instâncias foram descartadas, e no *dataframe* Resiliência de Vendas foram eliminados 20 atributos e 707 instâncias, garantindo um balanceamento entre completude e preservação do tamanho amostral.

(iii) Análise de colinearidade

Para mitigar o impacto de correlações excessivas entre variáveis independentes — que podem prejudicar a interpretabilidade e estabilidade dos modelos —, foi realizada uma análise de correlação linear de Pearson. Foram removidos os atributos com **correlação acima de 0,90**, o que resultou na exclusão de 14 atributos (Velocidade de Vendas) e 13 atributos (Resiliência de Vendas) altamente correlacionadas. Essa etapa reduziu significativamente a multicolinearidade estrutural entre os preditores.

(iv) Verificação do Fator de Inflação da Variância (VIF)

A etapa seguinte aplicou a métrica *Variance Inflation Factor* (VIF) para verificar redundâncias multivariadas. **Valores de VIF > 10** indicam dependência linear significativa entre variáveis, podendo distorcer estimativas dos modelos lineares. Após essa verificação, 1 atributo foi removido em Velocidade de Vendas, completando o controle de colinearidade.

(v) Seleção por Informação Mútua

Posteriormente, foi aplicada uma abordagem baseada em informação mútua, visando identificar variáveis com maior relevância preditiva em relação à classe-alvo. Para tornar a análise mais robusta, foram empregados três modelos distintos — *Random Forest, Decision Tree e Logistic Regression* —, e avaliou-se o impacto de cada atributo sobre a métrica AUC-ROC.

Foram eliminadas variáveis cuja exclusão não alterava a AUC em mais de **5% do limiar de estabilidade**, considerando múltiplas iterações. Para Velocidade de Vendas *Logistic Regression* apresentou o desempenho mais consistente, e ao longo de 8 interações promoveu a exclusão de 84 atributos. Para Resiliência de Vendas *Decision Tree* apresentou o desempenho mais consistente, e ao longo de 7 interações promoveu a exclusão de 76 atributos.

(Vi) *Logistic Regression* com penalização L1 (LASSO)

Para refinar ainda mais o conjunto de variáveis, aplicou-se uma *Logistic Regression* com penalização L1 (LASSO), técnica amplamente utilizada para seleção automática de atributos. A penalização L1 força os coeficientes menos relevantes a zero, eliminando atributos que não contribuem significativamente para a explicação da variável-alvo. Nesse estágio, no conjunto de Velocidade de Vendas 6 atributos foram removidos e no conjunto Resiliência de Vendas 17 atributos foram removidos, todos por apresentarem coeficientes nulos.

(vii) Controle de atributos com alto potencial de vazamento de informação

Por fim, foram excluídas variáveis **com AUC individual superior a 0,95**, consideradas como de alto potencial de vazamento da classe-alvo (data leakage). Essa etapa preventiva resultou na exclusão de 3 atributos adicionais na base de Velocidade de Vendas e 4 atributos na base de Resiliência de Vendas, assegurando que os modelos mantivessem caráter genuinamente preditivo e não retrospectivo.

Resultado Final

Ao término das etapas descritas, obteve-se o conjunto final de **Velocidade de Vendas ficou com 49 atributos relevantes e 10.614 instâncias válidas**, e o conjunto final de **Resiliência de Vendas ficou comm 43 atributos e 9.567 instâncias válidas**. Representando assim, uma base de dados equilibrada,

informativa e estatisticamente consistente. Esse processo garantiu a redução de ruído, colinearidade e sobreajuste, contribuindo diretamente para a melhoria do desempenho e da interpretabilidade dos modelos preditivos subsequentes.

5.3 SELEÇÃO DOS MODELOS DE APRENDIZAGEM DE MÁQUINAS

A seleção dos modelos preditivos foi orientada por três critérios centrais: (i) capacidade de realizar classificação binária com desempenho robusto, (ii) viabilidade de aplicação sobre dados estruturados com múltiplos tipos de atributos e (iii) alto grau de interpretabilidade dos resultados, em consonância com os objetivos de extração de conhecimento e apoio à decisão no domínio imobiliário. Neste estudo, foram selecionados três modelos amplamente consolidados na literatura de Ciência de Dados e Inteligência Artificial: ***Decision Tree***, ***Random Forest*** e ***Logistic Regression***.

O modelo ***Decision Tree*** foi selecionado por sua capacidade de capturar relações não-lineares e hierárquicas entre os atributos, além de gerar regras de decisão facilmente interpretáveis. As árvores particionam recursivamente o espaço de atributos de forma que cada ramo represente uma regra lógica "if-then", fornecendo uma estrutura de decisão transparente e intuitiva (Tan, Steinbach & Kumar, 2018). Essa abordagem tem sido especialmente útil em domínios onde a compreensão das condições que levam ao sucesso ou insucesso de um empreendimento é crítica.

Logistic Regression foi escolhido como baseline estatístico clássico para problemas de classificação binária. Por sua natureza linear, permite inferir diretamente o impacto (sinal e magnitude) de cada variável explicativa sobre a probabilidade de ocorrência da classe positiva. Essa característica a torna especialmente adequada para cenários em que a explicabilidade e a comunicação dos resultados com stakeholders não técnicos são fundamentais (Kuhn & Johnson, 2019). Ainda que limitada na modelagem de relações não-lineares e interações complexas entre atributos, sua transparência a torna uma referência importante na análise comparativa.

O modelo ***Random Forest*** foi incorporado ao estudo como um método de *ensemble learning* capaz de reduzir a variância e mitigar o sobreajuste característico

de classificadores baseados em uma única árvore de decisão. Proposto por Breiman (2001), o algoritmo combina múltiplas árvores independentes, construídas a partir de amostras aleatórias dos dados e subconjuntos de atributos, agregando suas previsões por meio de votação majoritária. Essa estratégia de *bagging* (*bootstrap aggregating*) resulta em modelos mais estáveis e robustos a ruídos nos dados, preservando, contudo, parte da interpretabilidade herdada das árvores individuais.

A escolha desses três modelos representa um equilíbrio entre **precisão preditiva**, **transparência do modelo** e **potencial de explicação de resultados** – elementos fundamentais para o uso de modelos de IA em processos decisórios no setor imobiliário. Além disso, sua complementaridade metodológica permitirá análises comparativas que destacam não apenas o desempenho quantitativo, mas também o valor qualitativo da explicabilidade na adoção de soluções baseadas em dados.

5.4 REFINAMENTO, EXECUÇÃO E AVALIAÇÃO DOS MODELOS

Esta seção descreve as etapas de refinamento, execução e avaliação dos modelos preditivos — Decision Tree, Random Forest e Logistic Regression — aplicados às duas hipóteses formuladas neste trabalho: Velocidade de Vendas e Resiliência de Vendas.

O objetivo foi garantir que cada modelo fosse avaliado de forma justa e estatisticamente robusta, assegurando comparabilidade e interpretabilidade entre os resultados obtidos.

Divisão das dados para treinamento e teste

Todas as divisões entre os conjuntos de treino e teste foram estratificadas com base na classe-alvo, assegurando que a proporção entre as classes (“sucesso” e “insucesso”) fosse mantida em cada subconjunto. Essa abordagem é essencial em cenários de desbalanceamento da variável-alvo, pois reduz vieses de amostragem e melhora a estabilidade das métricas de desempenho.

Com isso, cada modelo foi treinado e avaliado em bases equivalentes, reforçando o rigor metodológico e a comparabilidade entre hipóteses e algoritmos.

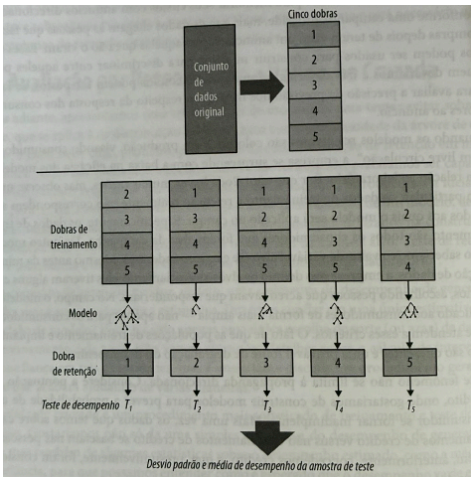
Refinamento dos hiperparâmetros
GridSearch com Validação Cruzada

Na fase inicial de treinamento, foi conduzido o refinamento dos hiperparâmetros por meio da técnica de *Grid Search* combinada à validação cruzada estratificada em 10 dobras (*folds*).

- Esse procedimento consistiu em:
- Dividir o conjunto de treinamento em dez subconjuntos (folds);
- Treinar o modelo em nove folds e validar no fold restante, repetindo o processo até que todos tivessem sido usados como validação;
- Calcular a média da métrica AUC-ROC obtida em todas as dobras;
- Selecionar automaticamente a combinação de hiperparâmetros que maximizou o desempenho médio.

Importante ressaltar que apenas o conjunto de treino foi utilizado neste processo — o conjunto de teste permaneceu isolado até a etapa final de avaliação, evitando qualquer contaminação de dados (data leakage) e garantindo validade estatística às comparações.

Figura 5.3 - Exemplo de ilustração do funcionamento da validação cruzada com 5 folds



Fonte: PROVOST; FAWCETT, 2016. p. 126–128.

Execução dos modelos com os melhores hiperparâmetros

Após a identificação dos melhores hiperparâmetros para cada algoritmo, procedeu-se à execução definitiva dos modelos utilizando todo o conjunto de treino

(com os parâmetros otimizados) e posterior avaliação no conjunto de teste independente.

- Essa fase teve como objetivos principais:
- Estimar o desempenho real do modelo em dados não vistos;
- Verificar a generalização das soluções otimizadas;
- Registrar métricas comparativas de desempenho (AUC-ROC, *F1-Score*, *Acuracy*, *Precision* e *Recall*).

A execução final foi realizada separadamente para cada hipótese investigada (Velocidade e Resiliência de Vendas), permitindo comparar não apenas os modelos, mas também as diferenças estruturais de cada problema preditivo.

Os resultados obtidos serviram de base para as etapas seguintes de interpretação dos modelos (*Explainable AI*), onde técnicas como *Permutation Feature Importance* (PFI), SHAP e LIME fossem aplicadas para explicar o impacto de cada atributo nas previsões.

Nas seções 5.4.1 e 5.4.2 a seguir, são apresentados os resultados obtidos nas etapas de refinamento, execução e avaliação dos modelos preditivos, considerando cada uma das hipóteses analisadas.

5.4.1 Hipótese 1: Velocidade de Vendas

5.4.1.1 Refinamento dos Hiperparâmetros - GridSearch com Validação Cruzada

Decision Tree

O conjunto de hiperparâmetros considerados incluiu:

- max_depth $\in$ {3, 5, 7} - Limite da profundidade da árvore.
- min_samples_split $\in$ {60, 100, 150, 200} - Amostras para dividir o nó.
- min_samples_leaf $\in$ {30, 50, 75, 100} - Amostras na folha.
- criterion $\in$ {“gini”, “entropy”} - Critério de pureza.
- class_weight $\in$ {None, “balanced”} - Ponderação para desbalanceamento.

Na execução foram analisados 192 candidatos, totalizando 1.920 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- max_depth: 7
- min_samples_split: 60

- min_samples_leaf: 30
- criterion: entropy
- class_weight: balanced

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.950.**

Random Forest

O conjunto de hiperparâmetros considerados incluiu:

- n_estimators ∈ {100, 200} - Número de árvores independentes.
- max_depth ∈ {5, 7} - Limite da profundidade da árvore,
- min_samples_split ∈ {60, 100, 150} - Amostras para dividir o nó.
- min_samples_leaf ∈ {30, 50, 75} - Amostras na folha.
- max_features ∈ {“sqrt”, “log2”} - Num. de atributos em cada divisão.
- class_weight ∈ {None, “balanced”} - Ponderação para desbalanceamento.

Na execução foram analisados 144 candidatos, totalizando 1.440 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- n_estimators: 200
- max_depth: 7
- min_samples_split: 60
- min_samples_leaf: 30
- max_features: sqrt
- class_weight: balanced

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.989.**

Logistic Regression

O conjunto de hiperparâmetros considerados incluiu:

- penalty: {l1, l2}
- C: {0.01, 0.1, 1, 10}
- solver: {liblinear}
- class_weight: {None, balanced}

Na execução foram analisados 16 candidatos, totalizando 160 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- penalty: l1
- C: 10
- solver: liblinear
- class_weight: balanced

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.976.**

5.4.1.2 Execução dos modelos com o melhor conjunto de parâmetros

Abaixo são apresentadas as métricas de execução dos modelos para a hipótese de Velocidade de Vendas com a execução definitiva dos modelos utilizando todo o conjunto de treino (com os parâmetros otimizados) e avaliação no conjunto de teste independente.

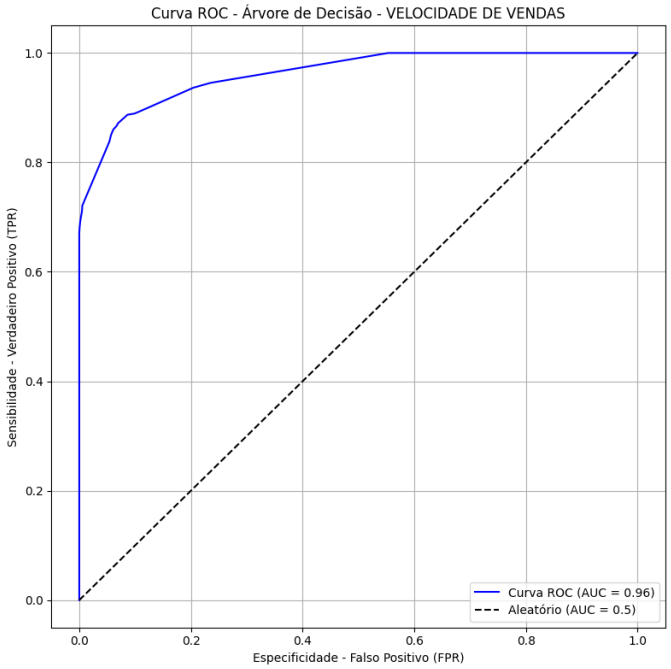
Decision Tree

Tabela 5.1 - Métricas de avaliação do modelo Decision Tree com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Decision Tree	0.962	0.885	0.902	0.883	0.887

Fonte: O autor (2025)

Figura 5.4 - AUC-ROC do modelo Decision Tree com os melhores hiperparâmetros



Fonte: O autor (2025)

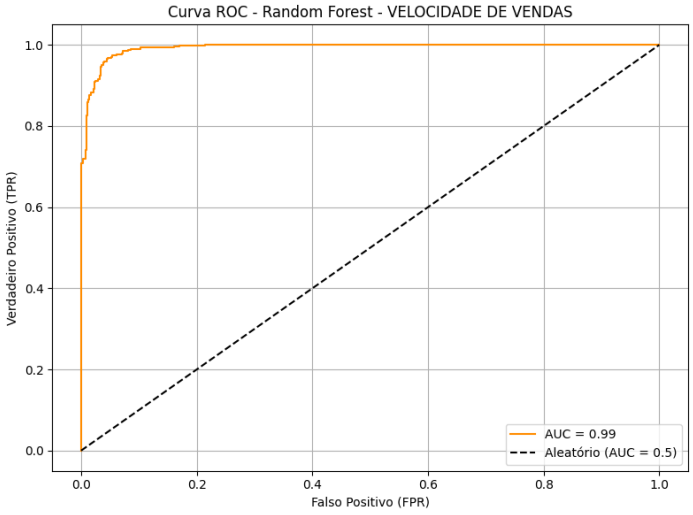
Random Forest

Tabela 5.2 - Métricas de avaliação do modelo Random Forest com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Random Forest	0.992	0.950	0.956	0.933	0.967

Fonte: O autor (2025)

Figura 5.5 - AUC-ROC do modelo Random Forest com os melhores hiperparâmetros



Fonte: O autor (2025)

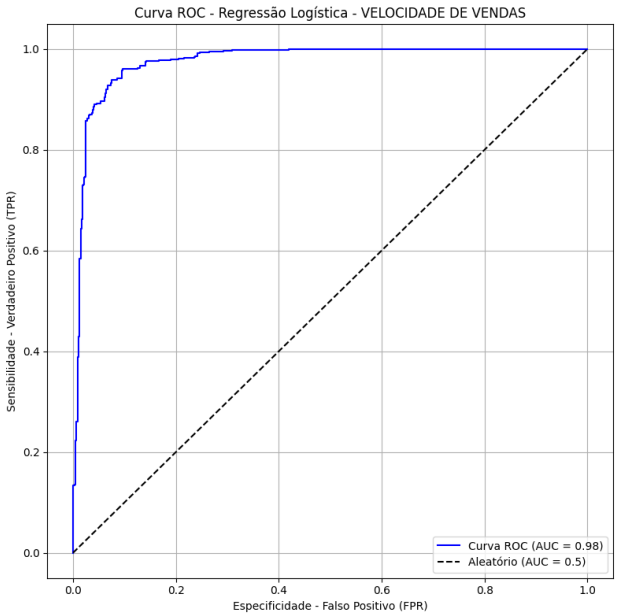
Logistic Regression

Tabela 5.3 - Métricas de avaliação do modelo Logistic Regression com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Logistic Regression	0.975	0.919	0.930	0.903	0.937

Fonte: O autor (2025)

Figura 5.6 - AUC-ROC do modelo Logistic Regression com os melhores hiperparâmetros



Fonte: O autor (2025)

5.4.2 Hipótese 2: Resiliência de Vendas

5.4.2.1 Refinamento dos Hiperparâmetros - GridSearch com Validação Cruzada

Decision Tree

O conjunto de hiperparâmetros considerados incluiu:

- max_depth ∈ {3, 5, 7} - Limite da profundidade da árvore.
- min_samples_split ∈ {60, 100, 150, 200} - Amostras para dividir o nó.
- min_samples_leaf ∈ {30, 50, 75, 100} - Amostras na folha.
- criterion ∈ {“gini”, “entropy”} - Critério de pureza.

- class_weight ∈ {None, “balanced”} - Ponderação para desbalanceamento.

Na execução foram analisados 192 candidatos, totalizando 1.920 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- max_depth: 7
- min_samples_split: 60
- min_samples_leaf: 100
- criterion: entropy
- class_weight: None

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.975.**

Random Forest

O conjunto de hiperparâmetros considerados incluiu:

- n_estimators ∈ {100, 200} - Número de árvores independentes.
- max_depth ∈ {5, 7} - Limite da profundidade da árvore,
- min_samples_split ∈ {60, 100, 150} - Amostras para dividir o nó.
- min_samples_leaf ∈ {30, 50, 75} - Amostras na folha.
- max_features ∈ {“sqrt”, “log2”} - Num. de atributos em cada divisão.
- class_weight ∈ {None, “balanced”} - Ponderação para desbalanceamento.

Na execução foram analisados 144 candidatos, totalizando 1.440 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- n_estimators: 200
- max_depth: 7
- min_samples_split: 60
- min_samples_leaf: 30
- max_features: sqrt
- class_weight: None

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.998.**

Logistic Regression

O conjunto de hiperparâmetros considerados incluiu:

- penalty: {l1, l2}

- C: {0.01, 0.1, 1, 10}
- solver: {liblinear}
- class_weight: {None, balanced}

Na execução foram analisados 16 candidatos, totalizando 160 conjuntos.

Para este grid o melhor conjunto de hiperparâmetros encontrado foi:

- penalty: l2
- C: 0.1
- solver: liblinear
- class_weight: balanced

Este conjunto obteve a melhor média na validação cruzada com **AUC-ROC de 0.991.**

5.4.2.2 Execução dos modelos com o melhor conjunto de parâmetros

Abaixo são apresentadas as métricas de execução dos modelos para a hipótese de Velocidade de Vendas com a execução definitiva dos modelos utilizando todo o conjunto de treino (com os parâmetros otimizados) e avaliação no conjunto de teste independente.

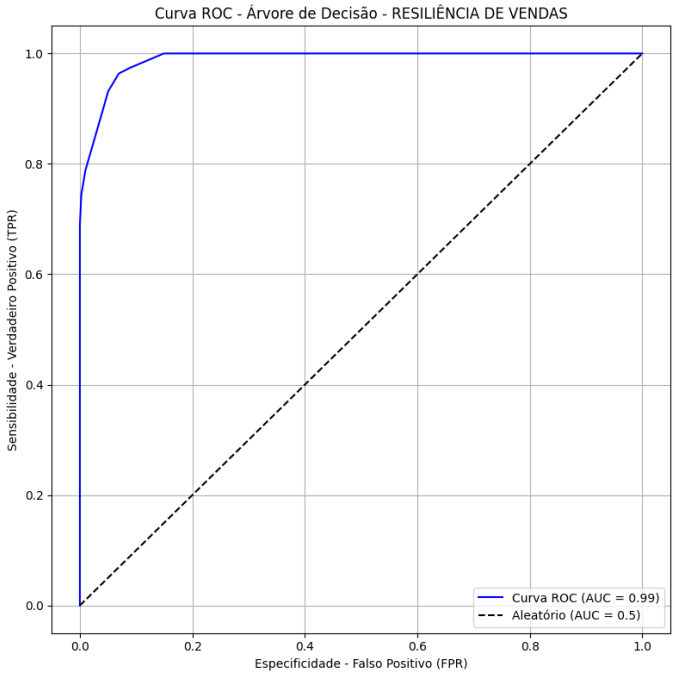
Decision Tree

Tabela 5.4 - Métricas de avaliação do modelo Decision Tree com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Decision Tree	0.989	0.913	0.943	0.895	0.931

Fonte: O autor (2025)

Figura 5.7 - AUC-ROC do modelo Decision Tree com os melhores hiperparâmetros



Fonte: O autor (2025)

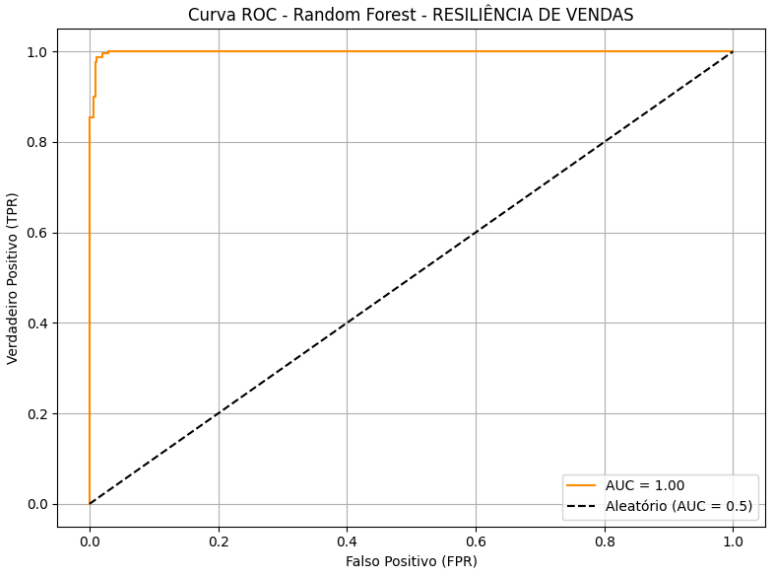
Random Forest

Tabela 5.5 - Métricas de avaliação do modelo Random Forest com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Random Forest	0.998	0.904	0.944	1.000	0.826

Fonte: O autor (2025)

Figura 5.8 - AUC-ROC do modelo Random Forest com os melhores hiperparâmetros



Fonte: O autor (2025)

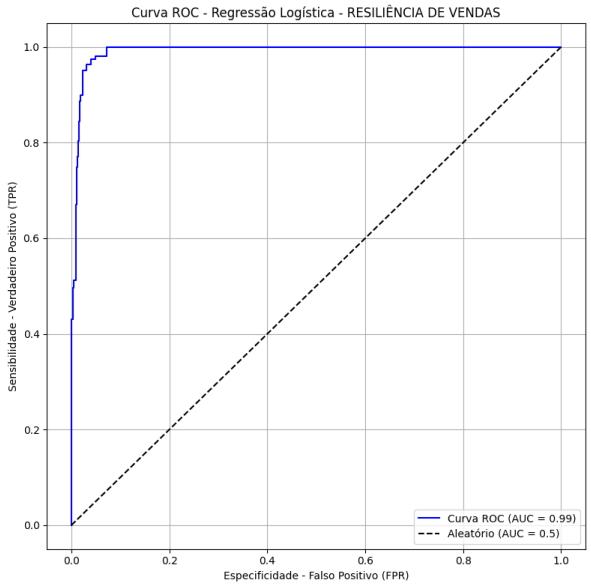
Logistic Regression

Tabela 5.6 - Métricas de avaliação do modelo Logistic Regression com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acuracy	Precision	Recall
Logistic Regression	0.991	0.946	0.964	0.919	0.974

Fonte: O autor (2025)

Figura 5.9 - AUC-ROC do modelo Logistic Regression com os melhores hiperparâmetros



Fonte: O autor (2025)

As métricas completas das avaliações e os gráficos informativos gerados para cada modelo podem ser acessados no **Apêndice I** (*Decision Tree*) e **Apêndice J** (*Logistic Regression*).

6 RESULTADOS E DISCUSSÕES

6.1 APRESENTAÇÃO DOS RESULTADOS

Nesta subseção são apresentados os principais resultados dos modelos preditivos aplicados às hipóteses de Velocidade de Vendas e Resiliência de Vendas, considerando tanto as métricas de desempenho quanto as análises de interpretabilidade geradas por técnicas *pós-hoc*. A apresentação está organizada por hipótese, facilitando a leitura comparativa e a análise dos fatores preditivos mais relevantes.

6.1.1 Hipótese 1: Velocidade de Vendas

Após a execução dos modelos com os hiperparâmetros otimizados, procedeu-se à avaliação quantitativa de desempenho, a fim de verificar a capacidade preditiva e a robustez de cada algoritmo frente aos conjuntos de dados de teste.

As métricas consideradas foram a Área sob a Curva ROC (AUC-ROC), *F1-score, acuracy, precision* e *recall,* escolhidas por oferecerem uma visão abrangente do equilíbrio entre acertos e erros, especialmente em cenários com possível desbalanceamento entre as classes.

Os resultados apresentados na Tabela 6.1 sintetizam o desempenho alcançado por cada modelo — Decision Tree, Random Forest e Logistic Regression — possibilitando uma comparação direta entre suas capacidades de generalização e adequação à hipótese analisada.

Tabela 6.1 - Métricas de avaliação dos modelos executados com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acurácia	Precisão	Revocação
Decision Tree	0.962	0.885	0.902	0.883	0.887
Random Forest	**0.992**	**0.950**	**0.956**	**0.933**	**0.967**
Logistic Regression	0.975	0.919	0.930	0.903	0.937

Fonte: O autor (2025)

Com base nas métricas apresentadas, o modelo **Random Forest** demonstrou desempenho superior em relação aos demais, apresentando o maior valor de AUC-ROC. Diante disto, o modelo Random Forest foi selecionado como referência para as análises de interpretabilidade dos seus resultados.

A etapa de interpretabilidade pós-hoc visa ampliar a transparência dos modelos preditivos por meio de técnicas que explicam seu funcionamento **a posteriori**, ou seja, com base nas predições já realizadas. Diferentemente dos modelos intrinsicamente interpretáveis, que possuem estrutura naturalmente explicável (como Árvores de Decisão e Regressão Logística), as abordagens pós-hoc permitem **investigar modelos mais complexos ou complementar a explicação dos modelos já analisados**.

Neste trabalho, foram aplicadas três técnicas amplamente reconhecidas na literatura de **XAI (eXplainable Artificial Intelligence)**:

- **PFI (Permutation Feature Importance)**: avalia a importância de cada atributo com base na perda de desempenho do modelo quando seus valores são embaralhados;
- **Accumulated Local Effects (ALE)**: permite observar o efeito marginal de cada variável sobre a predição;
- **SHAP (SHapley Additive exPlanations)**: quantifica o impacto de cada atributo em cada predição, com base em valores de Shapley da Teoria dos Jogos;
- **LIME (Local Interpretable Model-agnostic Explanations)**: explica decisões específicas do modelo a partir de aproximações locais, utilizando modelos lineares simples.

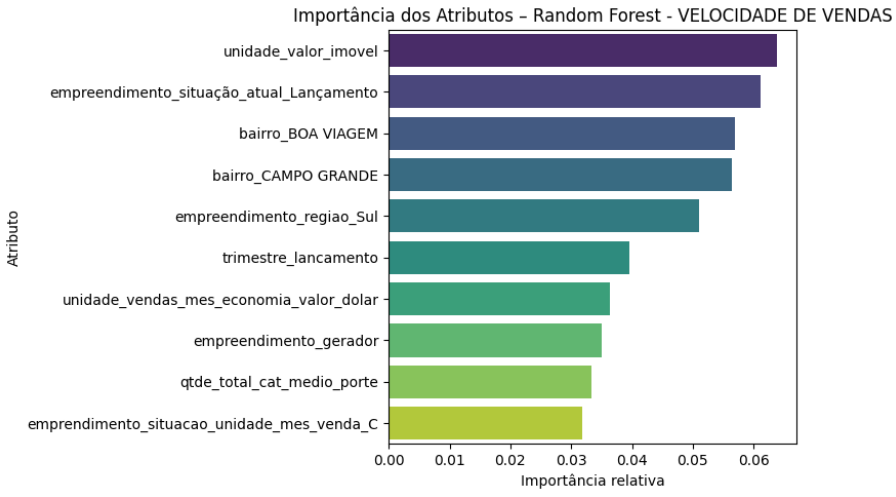
Essas abordagens complementares foram aplicadas de forma integrada, de modo a promover uma compreensão abrangente e transparente das decisões do modelo. As análises equivalentes para os modelos Decision Tree e Logistic Regression são apresentadas de forma completa no **Apêndice I** e **Apêndice J**, respectivamente, permitindo comparação metodológica entre os diferentes algoritmos.

6.1.1.1 Interpretabilidade intrínseca dos resultados do modelo Random Forest

A etapa de modelagem preditiva não se restringiu à busca por desempenho estatístico, mas também priorizou a geração de explicações compreensíveis para os agentes decisórios no mercado imobiliário. A adoção de modelos intrinsecamente interpretáveis teve como objetivo central ampliar a transparência analítica e facilitar a tradução dos resultados para ações práticas no domínio de negócios.

Além de sua capacidade preditiva, o modelo do Random Forest aplicado ao problema de classificação binária permitiu a geração de uma representação gráfica que facilita a compreensão dos mecanismos internos da predição. Por ser um modelo intrinsecamente interpretável, cada decisão pode ser descrita como uma sequência lógica de condições sobre os atributos do empreendimento ou da unidade habitacional.

Figura 6.1 – Importância dos atributos - Velocidade de vendas



Fonte: O autor (2025)

A Figura 6.1 apresenta a importância relativa dos atributos calculada pelo modelo Random Forest na tarefa de classificação da hipótese de Velocidade de Vendas. Os valores exibidos representam a contribuição média de cada variável para a redução da impureza (*Mean Decrease in Impurity*) nas árvores do *ensemble*, servindo como medida de relevância estrutural dentro do processo de decisão do modelo.

Observa-se que o atributo valor da unidade apresenta a maior importância relativa, seguido de empreendimento em lançamento e bairro Boa Viagem. Essas variáveis são as mais utilizadas pelo modelo para divisão de nós e definição de regras preditivas.

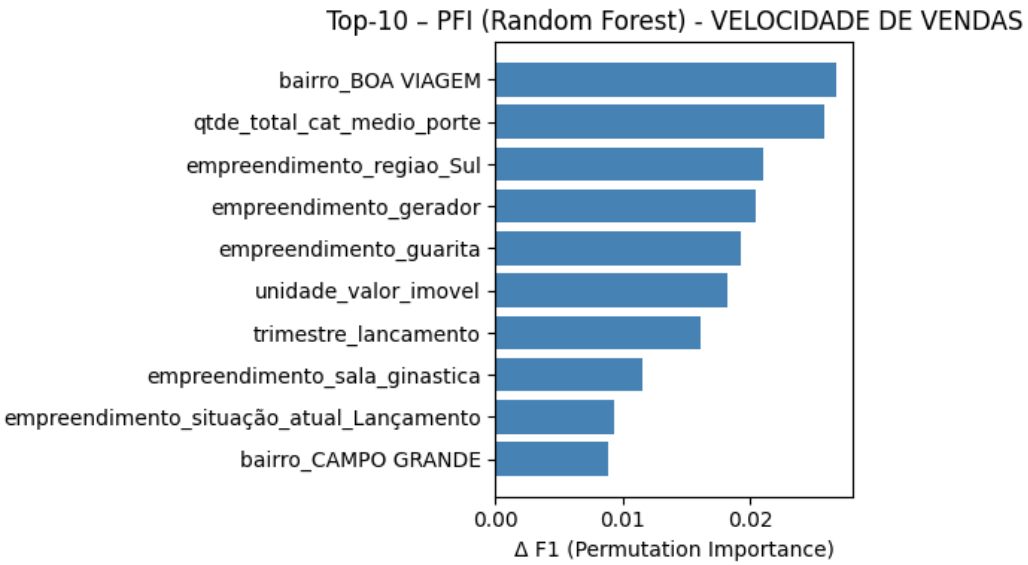
A distribuição dos valores indica uma concentração de importância entre variáveis quantitativas, espaciais e temporais, sem, contudo, permitir inferências sobre direção de influência ou relação causal. A métrica de importância apresentada limita-se a quantificar o impacto relativo de cada atributo na estrutura interna de decisão do *Random Forest*, sem implicar em interpretação sobre o comportamento do mercado ou sobre o sentido do efeito observado.

Em síntese, a Figura 6.1 documenta o comportamento interno do classificador, mostrando quais variáveis foram mais frequentemente selecionadas nas divisões das árvores e, portanto, exerceram maior influência no processo de classificação da variável-alvo relacionada à velocidade de vendas.

6.1.1.2 Interpretabilidade global dos resultados do modelo Random Forest

Importância Global – PFI (Δ F1)

Figura 6.2 – PFI Global (AUC) – Velocidade de vendas



Fonte: O autor (2025)

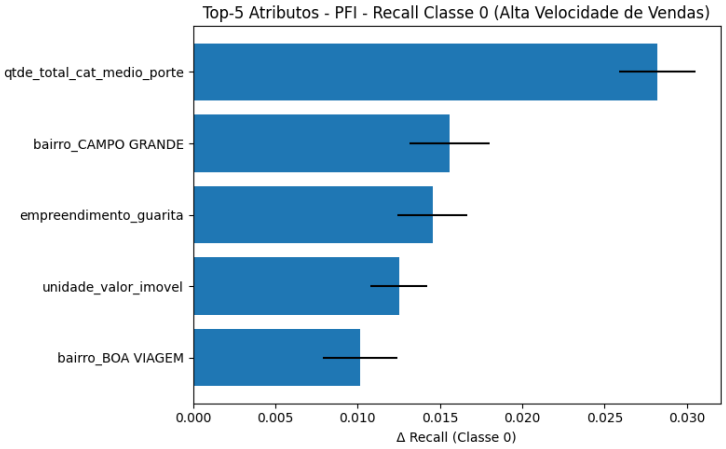
O gráfico de ***PFI - Permutation Feature Importante*** - evidencia que os atributos **Bairro Boa Viagem** e **Empreendimento de Porte Médio** são os mais sensíveis à permutação, provocando as maiores quedas no desempenho do modelo. Isso significa que o bairro onde o empreendimento está localizado e o porte total do empreendimento (número de unidades) são determinantes para a velocidade de vendas. Em seguida aparecem atributos que combinam aspectos espaciais, estruturais e temporais.

A presença de variáveis como **Empreendimento em Situação de Lançamento** e o **Bairro Campo Grande** nas últimas posições do top-10 indica que, apesar de relevantes na estrutura, perderam importância marginal quando avaliadas de forma independente, sugerindo redundância informacional — isto é, já capturadas por outros atributos correlacionados.

A PFI global revela que a localização (bairro/região) e o porte do empreendimento são vetores dominantes na explicação da velocidade de vendas, reforçando o peso do contexto urbano e da escala de produção como preditores da absorção imobiliária.

Importância Global Específica – PFI Classe 0 (Alta Velocidade de Vendas)

Figura 6.3 – Top-5 atributos com maior impacto no Recall (Classe 0) – Alta velocidade de vendas

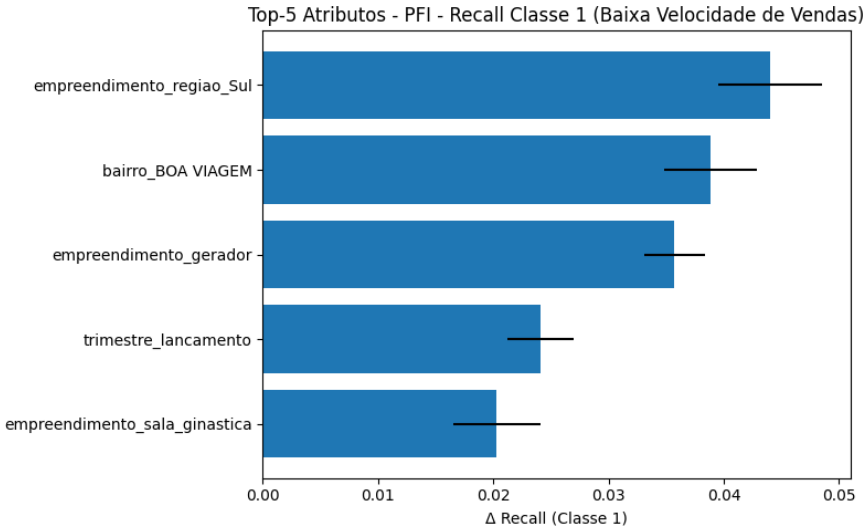


Fonte: O autor (2025)

Os resultados da Figura 6.3 evidenciam as variáveis que o modelo considerou mais relevantes para **reconhecer empreendimentos de alta velocidade de vendas**. Entre os principais fatores estão o **porte médio do empreendimento**, a **localização em bairros como Campo Grande e Boa Viagem**, e variáveis relacionadas ao **valor do imóvel** e à **presença de guarita**. Esses atributos demonstram ser determinantes na capacidade preditiva do modelo para identificar corretamente os casos de rápida absorção, ainda que o gráfico de PFI não indique a direção do efeito de cada variável.

Importância Global Específica – PFI Classe 1 (Baixa Velocidade de Vendas)

Figura 6.4 – Top-5 atributos com maior impacto no Recall (Classe 1) – Baixa velocidade de vendas

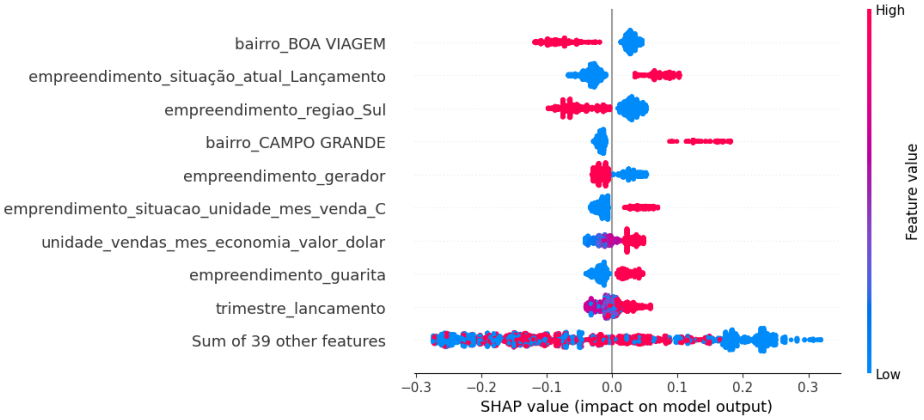


Fonte: O autor (2025)

Os resultados da Figura 6.4 indicam as variáveis que o modelo Random Forest considerou mais relevantes para distinguir empreendimentos com baixa velocidade de vendas. Entre elas, destacam-se fatores locacionais — como estar na região Sul e no bairro de Boa Viagem — e características estruturais e de conforto, como presença de gerador e sala de ginástica. Essas variáveis contribuem de forma decisiva para o modelo reconhecer os padrões associados à classe 1, embora o gráfico de PFI não indique a direção de seu efeito. A interpretação detalhada de seu impacto (positivo ou negativo) sobre as previsões é aprofundada na etapa de explicabilidade local, por meio das técnicas SHAP e LIME.

Importância Global – SHAP

Figura 6.5 – SHAP Summary Plot – Impacto global dos atributos no modelo



Fonte: O Autor (2025)

O gráfico do tipo beeswarm (Figura 6.5) apresenta os principais atributos de acordo com os valores SHAP médios do modelo, onde valores positivos indicam contribuição para o aumento da probabilidade de baixa velocidade de vendas (classe 1), enquanto valores negativos indicam influência na alta velocidade de vendas (classe 0). Observa-se que os fatores espaciais e geográficos exercem maior influência sobre o comportamento global do modelo. O atributo **Bairro Boa Viagem** apresenta o impacto mais expressivo, com valores altos (em vermelho) concentrando-se no lado negativo do eixo, o que significa que a presença de empreendimentos nessa localidade tende a reduzir a probabilidade de baixa velocidade de vendas e, consequentemente, aumentar a probabilidade de vendas aceleradas.

Por outro lado, o atributo **Bairro Campo Grande** apresenta efeito oposto, com valores elevados associados a maior probabilidade de baixa velocidade de vendas. O atributo **Empreendimento na Região Sul** reforça o padrão observado, associando regiões consolidadas a melhor desempenho comercial. Já atributos como **Empreendimento com Gerador** e **Empreendimento com Guarita** exibem dispersão mais heterogênea, refletindo características contextuais relacionadas ao padrão construtivo e ao público-alvo de cada empreendimento.

Comparação entre PFI e SHAP - Interpretabilidade Global

A comparação entre os resultados obtidos pelas técnicas de interpretabilidade global — Permutation Feature Importance (PFI) e SHAP — evidencia forte convergência na identificação dos atributos mais determinantes para o modelo Random Forest. Em ambas as abordagens, os fatores espaciais e geográficos se destacam como principais vetores explicativos da velocidade de vendas, com o atributo **Bairro Boa Viagem** apresentando o maior impacto sobre o desempenho preditivo. O **PFI**, ao quantificar a perda de desempenho (ΔF1) causada pela permutação de cada variável, apontou também a relevância de **Empreendimentos de Médio Porte** e **Empreendimento na Região Sul**, reforçando a influência de características regionais e de porte do empreendimento. Já o método **SHAP**

complementa essa análise ao revelar a direção e o sentido do impacto de cada variável, mostrando que empreendimentos localizados em **Boa Viagem** e na **Região Sul** aumentam a probabilidade de alta velocidade de vendas, enquanto aqueles situados em Campo Grande tendem a apresentar vendas mais baixas.

Além disso, ambos os métodos identificam a importância de variáveis temporais e estruturais, como **Empreendimento Situação de Lançamento**, **Empreendimento com Guarita** e **Empreendimento com Gerador**, sugerindo que o estágio do ciclo de vendas e o padrão construtivo exercem influência relevante sobre o ritmo de absorção do estoque. Essa coerência entre os resultados do PFI e do SHAP fortalece a robustez interpretativa do modelo, demonstrando que o *Random Forest* aprendeu relações consistentes entre atributos geográficos, temporais e de porte, e que essas relações são explicáveis, estáveis e alinhadas com a dinâmica observada no mercado imobiliário do Recife.

6.1.1.3 Interpretabilidade local dos resultados do modelo Random Forest

A interpretabilidade local busca compreender como um modelo de aprendizado de máquina toma decisões em nível individual, ou seja, quais atributos mais influenciam a previsão para uma instância específica. Enquanto o SHAP (SHapley Additive exPlanations) quantifica a contribuição de cada variável com base em princípios da teoria dos jogos, oferecendo uma decomposição consistente e aditiva da previsão, o LIME (Local Interpretable Model-agnostic Explanations) cria modelos lineares simplificados ao redor de uma observação, aproximando o comportamento local do modelo complexo. A aplicação conjunta dessas técnicas permite identificar, de forma transparente, quais características do empreendimento ou unidade mais pesaram para a classificação de uma determinada amostra como “vendas altas” ou “vendas baixas”.

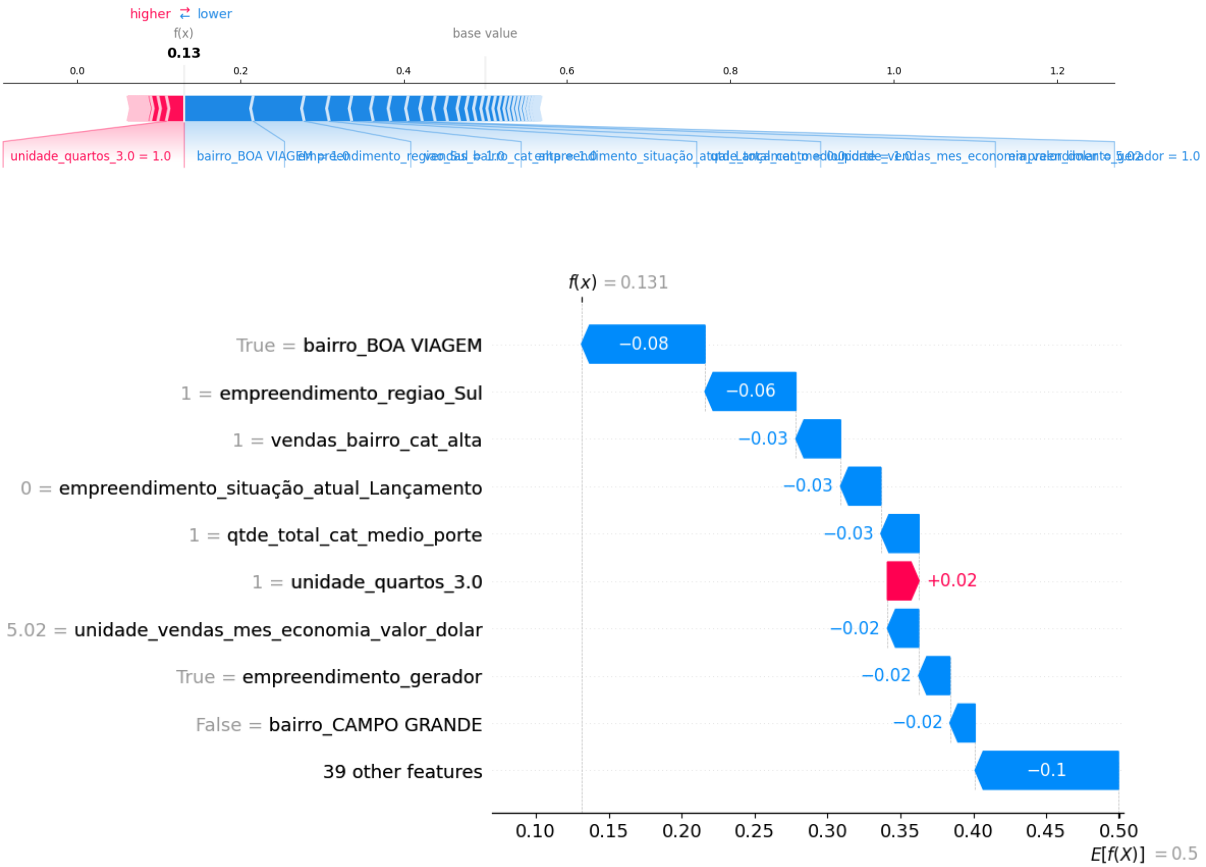
Importância Local – SHAP

O figura 6.6 abaixo parte de uma probabilidade base de aproximadamente 0,50 (valor médio esperado das previsões) e, após somar os efeitos locais dos atributos dessa instância, chega a uma probabilidade final de cerca de 0,13,

indicando que o modelo classificou o empreendimento como provável de apresentar alta velocidade de vendas.

No gráfico waterfall (inferior), as barras azuis indicam fatores que reduzem a probabilidade de alta velocidade (favorecendo vendas rápidas), enquanto as vermelhas indicam os que aumentam essa probabilidade (ou seja, dificultam as vendas).

Figura 6.6 – SHAP Force Plot – Instância com alta probabilidade de sucesso



Fonte: O autor (2025)

O modelo SHAP revela que esta instância representa um empreendimento localizado em Boa Viagem, na Zona Sul, de porte médio, recém-lançado e com boas condições de mercado, o que justifica sua classificação como empreendimento de vendas rápidas. O único fator de leve contrapeso foi a tipologia de três quartos, que elevou marginalmente a chance de menor velocidade, mas não suficiente para inverter a decisão.

Em resumo, o SHAP confirma que localização, porte e estágio de lançamento foram determinantes para o bom desempenho previsto.

Importância Local – LIME

A Figura 6.7 apresenta a explicação local gerada pelo método Local Interpretable Model-Agnostic Explanations (LIME) para a mesma instância analisada na seção anterior pelo SHAP. O modelo de Random Forest previu, para essa observação, uma probabilidade de 87% de pertencer à classe “Alta Velocidade de Vendas”, demonstrando elevada confiança na classificação. O LIME decompõe essa decisão em contribuições lineares das variáveis mais relevantes para o resultado predito, permitindo compreender o comportamento do modelo de forma individualizada e transparente.

Figura 6.7 – LIME – Instância com 87% de probabilidade de alta velocidade de vendas

Fonte: O autor (2025)

Comparação entre PFI e SHAP - Interpretabilidade Global

De modo geral, a explicação local do LIME confirma a coerência das relações observadas nas análises globais com SHAP e PFI, evidenciando que o modelo aprendeu padrões consistentes com o comportamento real do mercado. Assim, empreendimentos situados especialmente em Boa Viagem e na Região Sul do Recife, tendem a apresentar maior velocidade de vendas, validando a capacidade interpretativa do modelo e a utilidade do LIME como ferramenta de explicabilidade em nível individual.

6.1.2 Hipótese 2: Resiliência de Vendas

Após a execução dos modelos com os hiperparâmetros otimizados, procedeu-se à avaliação quantitativa de desempenho, a fim de verificar a capacidade preditiva e a robustez de cada algoritmo frente aos conjuntos de dados de teste.

As métricas consideradas foram a Área sob a Curva ROC (AUC-ROC), F1-Score, Acurácia, Precisão e Revocação, escolhidas por oferecerem uma visão abrangente do equilíbrio entre acertos e erros, especialmente em cenários com possível desbalanceamento entre as classes.

Os resultados apresentados na Tabela 6.2 sintetizam o desempenho alcançado por cada modelo — Decision Tree, Random Forest e Logistic Regression — possibilitando uma comparação direta entre suas capacidades de generalização e adequação à hipótese analisada.

Tabela 6.2 - Métricas de avaliação dos modelos executados com os melhores hiperparâmetros

Modelo	AUC-ROC	F1-Score	Acurácia	Precisão	Revocação
Decision Tree	0.989	0.913	0.943	0.895	0.931
Random Forest	**0.998**	**0.904**	**0.944**	**1.000**	**0.826**
Logistic Regression	0.991	0.946	0.964	0.919	0.974

Fonte: O autor (2025)

Com base nas métricas apresentadas, o modelo **Random Forest** demonstrou desempenho superior em relação aos demais, apresentando o maior valor de AUC-ROC. Diante disto, o modelo Random Forest foi selecionado como referência para as análises de interpretabilidade dos seus resultados.

A etapa de interpretabilidade pós-hoc visa ampliar a transparência dos modelos preditivos por meio de técnicas que explicam seu funcionamento **a posteriori**, ou seja, com base nas predições já realizadas. Diferentemente dos

modelos intrinsicamente interpretáveis, que possuem estrutura naturalmente explicável (como Árvores de Decisão e Regressão Logística), as abordagens pós-hoc permitem **investigar modelos mais complexos ou complementar a explicação dos modelos já analisados**.

Neste trabalho, foram aplicadas três técnicas amplamente reconhecidas na literatura de **XAI (eXplainable Artificial Intelligence)**:

- **PFI (Permutation Feature Importance)**: avalia a importância de cada atributo com base na perda de desempenho do modelo quando seus valores são embaralhados;
- **Accumulated Local Effects (ALE)**: permite observar o efeito marginal de cada variável sobre a predição;
- **SHAP (SHapley Additive exPlanations)**: quantifica o impacto de cada atributo em cada predição, com base em valores de Shapley da Teoria dos Jogos;
- **LIME (Local Interpretable Model-agnostic Explanations)**: explica decisões específicas do modelo a partir de aproximações locais, utilizando modelos lineares simples.

Essas abordagens complementares foram aplicadas de forma integrada, de modo a promover uma compreensão abrangente e transparente das decisões do modelo. As análises equivalentes para os modelos *Decision Tree, Random Forest* e *Logistic Regression* são apresentadas de forma completa no Apêndice I e Apêndice J, respectivamente, permitindo comparação metodológica entre os diferentes algoritmos.

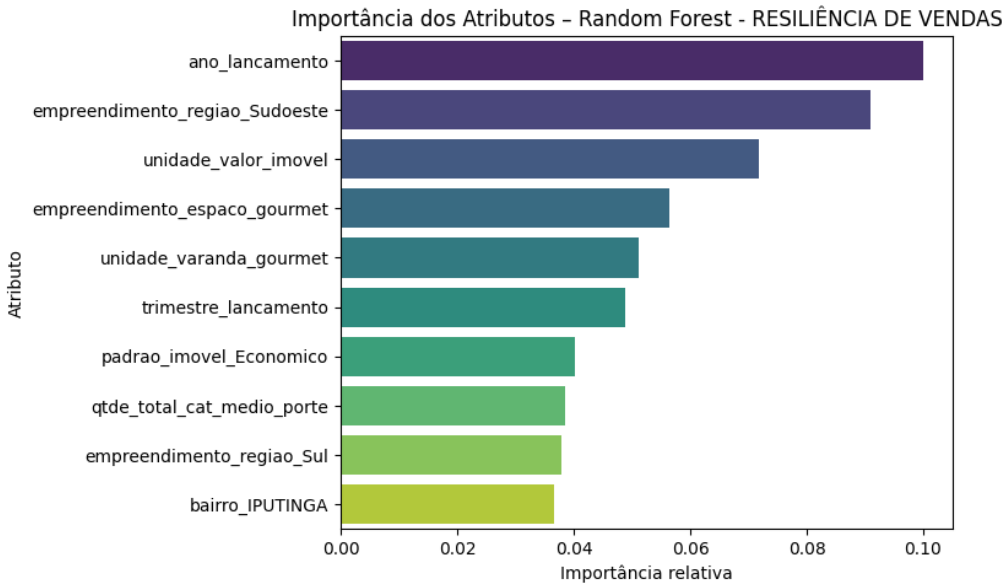
6.1.2.1 Interpretabilidade intrínseca dos resultados do modelo Random Forest

A etapa de modelagem preditiva não se restringiu à busca por desempenho estatístico, mas também priorizou a geração de explicações compreensíveis para os agentes decisórios no mercado imobiliário. A adoção de modelos intrinsecamente interpretáveis teve como objetivo central ampliar a transparência analítica e facilitar a tradução dos resultados para ações práticas no domínio de negócios.

Além de sua capacidade preditiva, o modelo do Random Forest aplicado ao problema de classificação binária permitiu a geração de uma representação gráfica

que facilita a compreensão dos mecanismos internos da predição. Por ser um modelo intrinsecamente interpretável, cada decisão pode ser descrita como uma sequência lógica de condições sobre os atributos do empreendimento ou da unidade habitacional.

Figura 6.8 – Importância dos atributos - Resiliência de vendas



Fonte: O autor (2025)

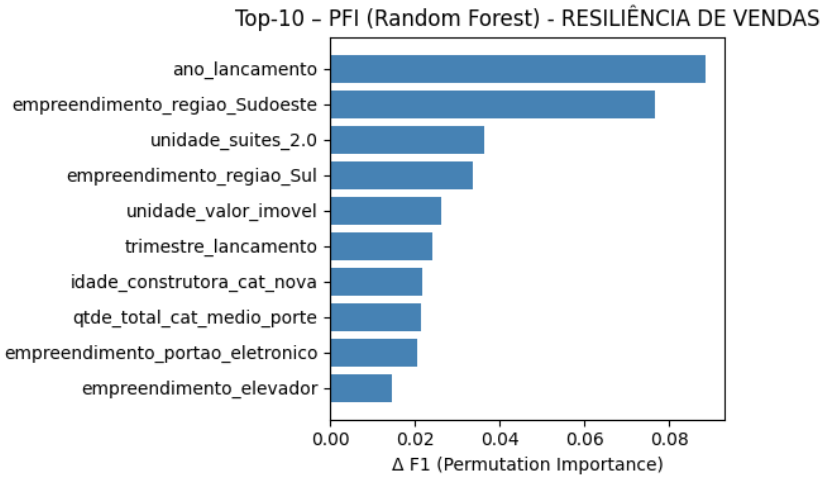
A Figura 6.8 apresenta o ranking dos dez atributos com maior importância relativa no modelo Random Forest para a hipótese de Resiliência de Vendas. Os valores exibidos correspondem às pontuações médias de importância calculadas a partir da redução da impureza (*Mean Decrease in Impurity*) ao longo das árvores que compõem o ensemble.

Observa-se que o atributo ano lançamento apresenta a maior contribuição relativa para o modelo, seguido por região sudoeste e valor da unidade. Esses três atributos concentram as maiores importâncias, indicando que o classificador utiliza fortemente essas variáveis para a construção de suas regras de decisão.

6.1.2.2 Interpretabilidade global dos resultados do modelo Random Forest

Importância Global – PFI (Δ F1)

Figura 6.9 – PFI Global (AUC) - Resiliência de vendas



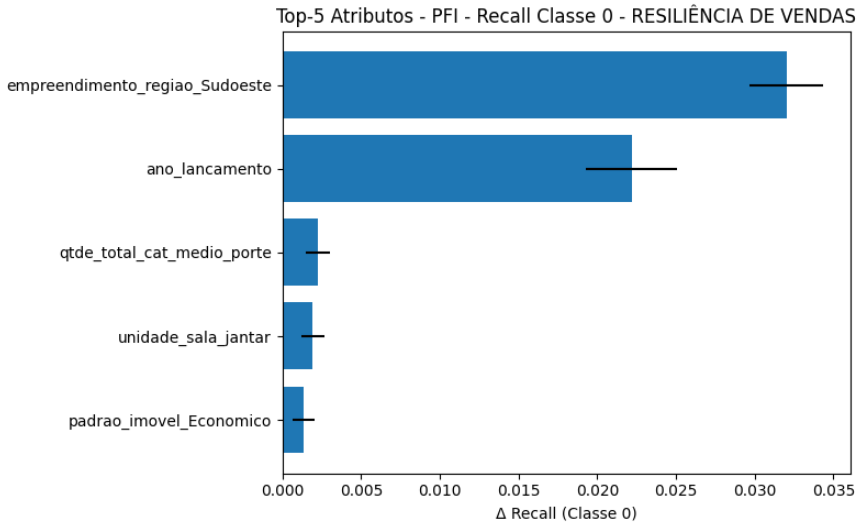
Fonte: O autor (2025)

O gráfico de **Permutation Feature Importance (PFI)** evidencia que os atributos **Ano de Lançamento** e **Região Sudoeste** são os mais sensíveis à permutação, provocando as maiores reduções no desempenho do modelo. Isso indica que o momento do lançamento e a localização geográfica são fatores decisivos para a resiliência de vendas, isto é, a capacidade de um empreendimento manter ritmo de absorção após o pico inicial de comercialização. O peso dessas variáveis demonstra que tanto o período de entrada no mercado quanto o posicionamento espacial determinam fortemente a estabilidade das vendas ao longo do tempo.

Na sequência, destacam-se atributos estruturais, como **Quantidade de Suítes**, **Valor do Imóvel** e **Trimestre de Lançamento**, que reforçam a influência conjunta de características físicas e temporais na manutenção do desempenho comercial. De forma geral, a análise de PFI confirma que a resiliência de vendas é moldada por uma combinação de fatores temporais (ano e trimestre de lançamento), espaciais (região) e estruturais (suítes, porte e padrão do empreendimento). O modelo de *Random Forest* capturou, assim, um comportamento coerente com a dinâmica real do mercado imobiliário, no qual localização, momento de entrada e qualidade construtiva são vetores fundamentais para a sustentabilidade das vendas ao longo do ciclo de vida do projeto.

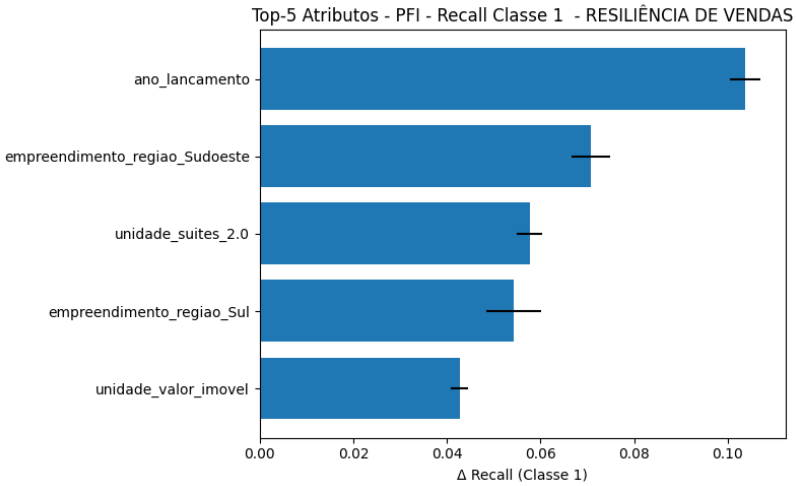
Importância Global Específica

Figura 6.10 – Top-5 atributos com maior impacto no Recall (Classe 0) – Baixa Resiliência de Vendas



Fonte: O autor (2025)

Figura 6.11 – Top-5 atributos com maior impacto no Recall (Classe 1) – Alta velocidade de Vendas



Fonte: O autor (2025)

As Figuras 6.10 e 6.11 apresentam os resultados da ***Permutation Feature Importance (PFI)*** calculada separadamente para as classes 0 (baixa resiliência de vendas) e 1 (alta resiliência de vendas). A análise comparativa permite observar quais atributos provocam maior variação no desempenho do modelo quando seus

valores são permutados dentro de cada classe, indicando, portanto, sua relevância relativa para o processo de classificação.

Para a Classe 1 (alta resiliência), o modelo destacou os atributos ano de lancamento, região sudoeste, suítes, como os de maior impacto sobre a métrica de *recall*. Esses resultados indicam que tais variáveis apresentam maior sensibilidade à permutação e, portanto, contribuem de forma mais expressiva para o reconhecimento dos casos classificados como de alta resiliência.

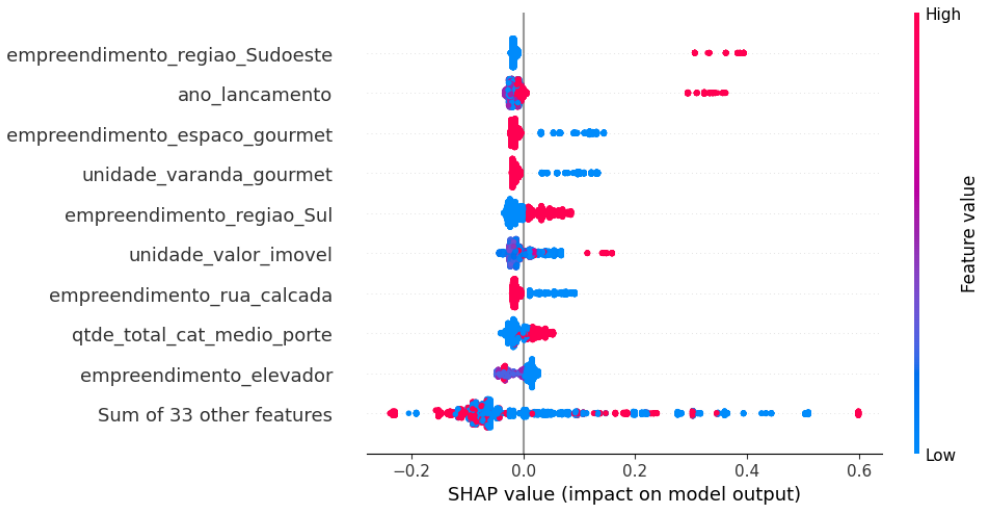
Na Classe 0 (baixa resiliência), os atributos de maior importância relativa foram região sudoeste, ano lançamento, empreendimento de médio porte,. Esses atributos foram os que mais afetaram o desempenho do modelo na identificação correta dos casos pertencentes a essa classe, sendo, portanto, os principais componentes utilizados pelo classificador para distinguir instâncias de baixa resiliência.

De forma geral, observa-se que ano lançamento e região sudoeste aparecem entre os atributos mais relevantes em ambas as classes, sugerindo que o modelo utiliza essas variáveis como elementos estruturais de separação. As demais variáveis variam entre as classes, refletindo diferentes combinações de atributos relevantes para a predição de cada tipo de comportamento de vendas.

Essas variáveis contribuem de forma decisiva para o modelo reconhecer os padrões associados às classes, embora o gráfico de PFI não indique a direção de seu efeito. A interpretação detalhada de seu impacto (positivo ou negativo) sobre as previsões é aprofundada na etapa de explicabilidade local, por meio das técnicas SHAP e LIME.

Importância Global – SHAP

Figura 6.12 – SHAP Summary Plot – Impacto global dos atributos no modelo



Fonte: O Autor (2025)

A Figura 6.12 apresenta o gráfico de valores SHAP (SHapley Additive exPlanations) para o modelo Random Forest aplicado à hipótese de Resiliência de Vendas. O gráfico resume a contribuição média e a dispersão dos efeitos de cada atributo sobre as predições do modelo, permitindo avaliar a importância global e o impacto direcional de suas variações sem inferir causalidade externa.

No eixo vertical, estão listados os principais atributos ordenados por importância média absoluta. Cada ponto no gráfico representa uma observação individual do conjunto de dados, e sua posição horizontal indica o efeito marginal estimado (valor SHAP) sobre a saída do modelo. Valores positivos deslocam a predição em direção à classe positiva (classe 1, alta resiliência), enquanto valores negativos deslocam em direção à classe negativa (classe 0, baixa resiliência).

Entre os atributos mais relevantes para a explicação do modelo, destacam-se região sudoeste, ano lançamento, empreendimento com espaço gourmet, os quais apresentam maior dispersão horizontal e amplitude de valores SHAP, indicando maior impacto no resultado preditivo.

Comparação entre PFI e SHAP - Interpretabilidade Global

De forma geral, há consistência entre as variáveis identificadas como mais relevantes pelos dois métodos. Embora ambos os métodos mensurem a relevância das variáveis para o comportamento do modelo, eles o fazem por **abordagens conceitualmente distintas**: o PFI avalia a **queda de desempenho global** quando um atributo é permutado aleatoriamente, enquanto o SHAP quantifica o **impacto aditivo e local** de cada variável sobre as predições do modelo.

Os atributos **ano lançamento** e **região sudoeste** aparecem nas primeiras posições em ambos os gráficos, evidenciando que são utilizados de maneira recorrente e influente pelo modelo.

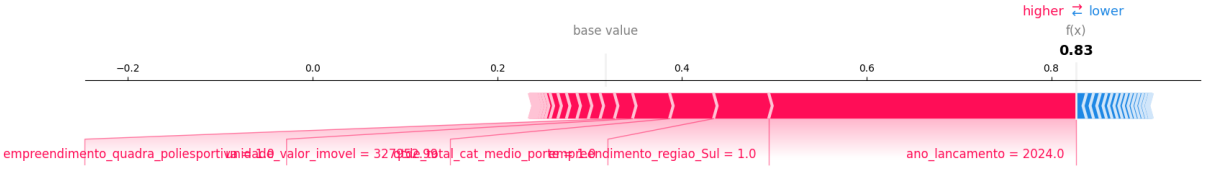
Também há convergência parcial em variáveis intermediárias como **valor da unidade**, **empreendimento de médio porte** e **região sul**, que aparecem entre os conjuntos de maior relevância em ambos os métodos, embora em ordens diferentes. Essa coincidência entre as técnicas reforça a estabilidade estrutural do modelo quanto à seleção de atributos relevantes para o processo de decisão.

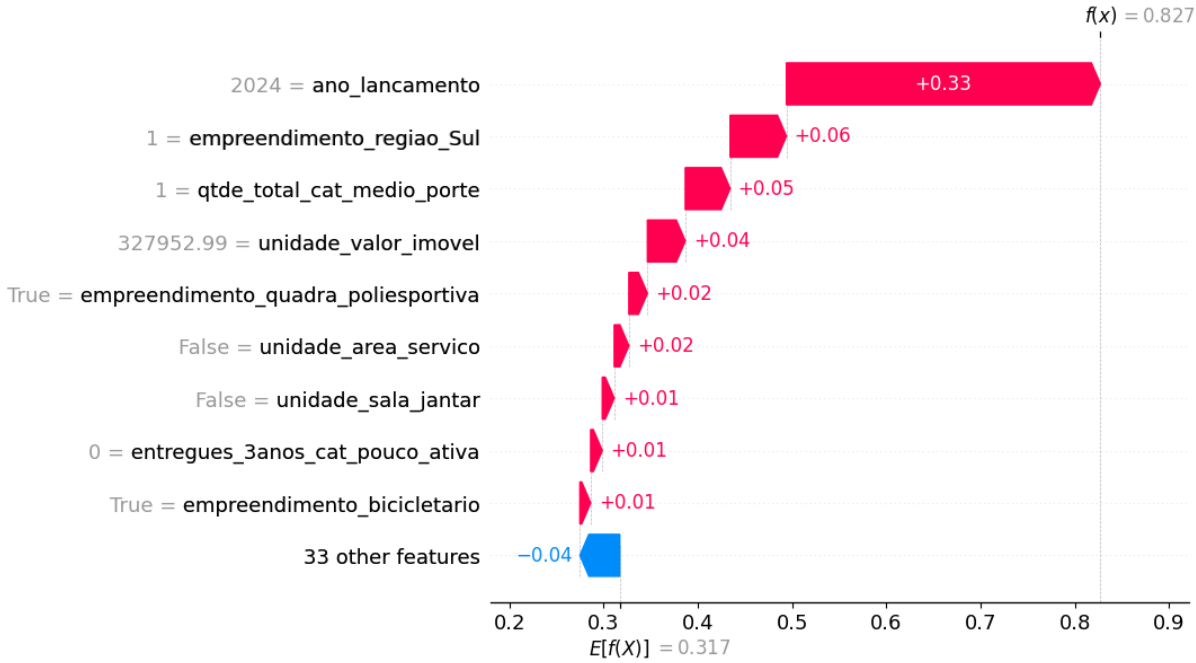
6.1.2.3 Interpretabilidade local dos resultados do modelo Random Forest

A interpretabilidade local busca compreender como um modelo de aprendizado de máquina toma decisões em nível individual, para isto, vamos continuar utilizando o SHAP e o LIME. A aplicação conjunta dessas técnicas permite identificar, de forma transparente, quais características do empreendimento ou unidade mais pesaram para a classificação de uma determinada amostra.

Importância Local – SHAP

Figura 6.13 – SHAP Force Plot – Instância com alta probabilidade de sucesso





Fonte: O autor (2025)

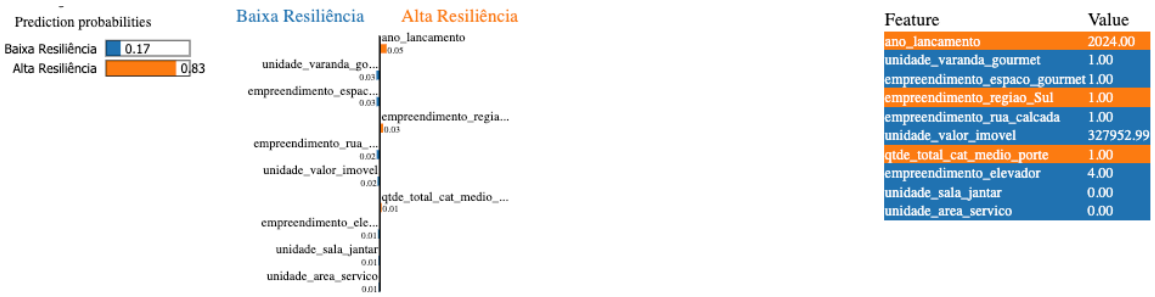
O modelo SHAP revela que esta instância representa um empreendimento localizado na região sul, com porte médio e ano de lançamento em 2024 e com boas condições de mercado, o que justifica sua classificação como empreendimento de alta resiliência. Os outros atributos atribuíram peso para baixa resiliência, mas não o suficiente para inverter a decisão.

Em resumo, o SHAP confirma que localização, porte e estágio de lançamento foram determinantes para o bom desempenho previsto.

Importância Local – LIME

A Figura 6.14 apresenta a explicação local gerada pelo método Local Interpretable Model-Agnostic Explanations (LIME) para a mesma instância analisada na seção anterior pelo SHAP. O modelo de Random Forest previu, para essa observação, uma probabilidade de 83% de pertencer à classe “Alta Resiliência de Vendas”, demonstrando elevada confiança na classificação. O LIME decompõe essa decisão em contribuições lineares das variáveis mais relevantes para o resultado predito, permitindo compreender o comportamento do modelo de forma individualizada e transparente.

Figura 6.14 – LIME – Instância com 83% de probabilidade de alta resiliência



Fonte: O autor (2025)

Comparação entre PFI e SHAP - Interpretabilidade Global

De modo geral, a explicação local do LIME confirma a coerência das relações observadas nas análises globais com SHAP e PFI, evidenciando que o modelo aprendeu padrões consistentes com o comportamento real do mercado.

6.2 OTIMIZAÇÃO DO PONTO DE OPERAÇÃO DA CURVA ROC

6.2.1 Contexto e problema da seleção do limiar

Modelos de classificação binária, como os utilizados neste estudo, geralmente produzem uma probabilidade de uma instância pertencer à classe positiva. A transformação dessa probabilidade em uma classificação binária (e.g., "alta velocidade de vendas" ou "baixa velocidade de vendas"/ "alta resiliência" ou "baixa resiliência") requer a definição de um **limiar (threshold)**. A escolha desse limiar é uma decisão crítica, pois impacta diretamente o equilíbrio entre as diferentes métricas de desempenho do modelo, como Precisão, Recall, Taxa de Falsos Positivos (FPR) e Taxa de Verdadeiros Positivos (TPR) e, consequentemente, impacta diretamente os resultados financeiros e operacionais de uma decisão de negócio.

A Curva Característica de Operação do Receptor (ROC) é uma ferramenta visual poderosa que ilustra todas as combinações possíveis de TPR e FPR que um modelo pode alcançar em diferentes limiares. No entanto, a Curva ROC por si só não indica qual limiar é o "melhor" para um cenário de negócio específico. A seleção de um ponto de operação ideal demanda a consideração de objetivos estratégicos e dos custos assimétricos associados a Falsos Positivos e Falsos Negativos, e também pode envolver a ponderação dos **custos e**

recompensas assimétricas associados a cada tipo de classificação correta ou incorreta (Verdadeiro Positivo, Falso Positivo, Verdadeiro Negativo, Falso Negativo).

6.2.2 Definições fundamentais e métrica de otimização

Para a compreensão do algoritmo de otimização, é fundamental revisar as definições das métricas envolvidas:

- **Verdadeiro Positivo (TP - True Positive):** Instâncias da classe positiva corretamente classificadas como positivas.
- **Falso Positivo (FP - False Positive):** Instâncias da classe negativa incorretamente classificadas como positivas.
- **Verdadeiro Negativo (TN - True Negative):** Instâncias da classe negativa corretamente classificadas como negativas.
- **Falso Negativo (FN - False Negative):** Instâncias da classe positiva incorretamente classificadas como negativas.

Com base nessas definições, as métricas-chave para a Curva ROC são:

- **Taxa de Verdadeiros Positivos (TPR - True Positive Rate):** Proporção de instâncias positivas que foram corretamente identificadas.

$$TPR=TP/TP+FNTP$$

- **Taxa de Falsos Positivos (FPR - False Positive Rate):** Proporção de instâncias negativas que foram incorretamente identificadas como positivas.

$$FPR=FP/FP+TNFP$$

6.2.3 O Algoritmo de Seleção do Ponto de Operação

Para automatizar e objetivar a seleção do limiar de classificação com base em requisitos de negócio, foi desenvolvido um algoritmo que permite ao usuário especificar um par de valores-alvo desejados para a Taxa de Falsos Positivos (FPRalvo) e a Taxa de Verdadeiros Positivos (TPRalvo). O objetivo do algoritmo é encontrar o limiar de probabilidade que minimize a "distância" entre o ponto

desejado (FPRalvo ,TPRalvo) e os pontos reais (FPR,TPR) existentes na Curva ROC do modelo.

As etapas do algoritmo são as seguintes:

1. **Geração de Pares (FPR, TPR):** Para cada limiar de probabilidade único gerado pelas previsões do modelo no conjunto de dados de validação, são calculados os valores correspondentes de FPR e TPR. Isso efetivamente constrói todos os pontos da Curva ROC.
2. **Definição dos Valores-Alvo:** O usuário insere os valores desejados para FPRalvo e TPRalvo com base em metas ou restrições de negócio.
3. **Cálculo da Distância:** Para cada par (FPR,TPR) da Curva ROC gerada na etapa 1, é calculada a distância em relação ao ponto-alvo (FPRalvo ,TPRalvo). A métrica de distância utilizada é a **distância Euclidiana**, conforme a fórmula:

$$\text{Distância} = \sqrt{(\text{FPR}_{\text{alvo}} - \text{FPR})^2 + (\text{TPR}_{\text{alvo}} - \text{TPR})^2}$$

4. **Seleção do Limiar Ótimo:** O algoritmo identifica o ponto na Curva ROC (e seu respectivo limiar de probabilidade) para o qual a distância calculada na etapa 3 é mínima. Este limiar corresponde ao ponto de operação que melhor se aproxima dos objetivos de FPRalvo e TPRalvo .

O algoritmo retorna o limiar de probabilidade selecionado e as métricas de desempenho (TPR, FPR, Precisão, Recall, F1-Score) que seriam obtidas ao operar o modelo com esse limiar.

Por exemplo, se no caso de Velocidade de Vendas, quisermos um algoritmo conservador que aceita apenas 1% dos falsos negativos. Logo:
FNR = 00.1 -> TPR = 1 - FNR = 0.99

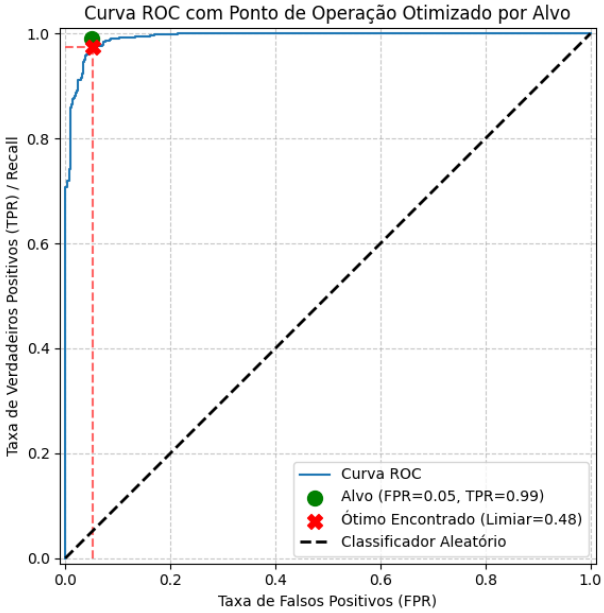
E podemos aceitar alguns falsos positivos (podemos deixar escapar alguns bons empreendimentos), podemos deixar a taxa de FPR = 0.05.

Esse raciocínio nos leva a seguinte situação:

- O custo de deixar um empreendimento ruim passar (FN) é muito alto — ele pode gerar meses de estoque encalhado, gasto com marketing e prejuízo financeiro.

- O custo de classificar erroneamente um bom empreendimento como lento (FP) é bem menor — você apenas analisa mais a fundo ou revisa as condições antes de lançar.

Figura 6.15 – Ponto de Operação



Fonte: O autor (2025)

--- Resultados do Algoritmo ---

Limiar Ótimo Selecionado: 0.4822

FPR no Limiar Ótimo : 0.0530

TPR (Recall) no Limiar Ótimo: 0.9742

Precisão no Limiar Ótimo : 0.9315

F1-Score no Limiar Ótimo : 0.9524

Menor Distância Euclidiana ao Alvo: 0.0161

6.2.4 O Algoritmo de Seleção do Lucro Esperado

Além do propósito mostrado na seção anterior, o algoritmo pode operar com base nos seguintes parâmetros de negócio, a serem definidos pelo usuário:

- mathbfR: Recompensa (ganho) associada a cada **Verdadeiro Positivo (TP)**.
- mathbfC: Custo (perda) associado a cada **Falso Negativo (FN)**, ou seja, uma oportunidade perdida.

- mathbfC_FP: Custo (perda) associado a cada **Falso Positivo (FP)**, ou seja, um alarme falso.

A métrica de otimização central é o **Lucro Esperado (Expected Value - EV)** por instância, calculado para cada limiar de probabilidade. Ele representa o retorno médio esperado ao aplicar o modelo em uma nova instância, considerando as probabilidades de ocorrência de cada classe e os custos/recompensas definidos:

$$EV=(Ppos \times(R\times TPR-C\times(1-TPR)))-(Pneg \times CFP \times FPR)$$

Onde:

- P_textpos: Probabilidade *a priori* de uma instância pertencer à classe positiva (proporção de positivos no dataset).
- P_textneg: Probabilidade *a priori* de uma instância pertencer à classe negativa (proporção de negativos no dataset).
- textTPR: Taxa de Verdadeiros Positivos no limiar atual.
- textFPR: Taxa de Falsos Positivos no limiar atual.
- (1−textTPR): Taxa de Falsos Negativos (também conhecida como textFNR - False Negative Rate).

6.2.5 O Algoritmo de Maximização do Lucro Esperado

O algoritmo desenvolvido busca identificar o limiar de probabilidade que **maximiza o Lucro Esperado (EV)** para o modelo, dadas as recompensas e custos de negócio definidos pelo usuário.

As etapas de funcionamento são as seguintes:

1. **Geração da Curva ROC:** Utilizam-se os rótulos verdadeiros e as probabilidades preditas do modelo no conjunto de dados de validação para gerar todos os pares (textFPR,textTPR) e seus respectivos limiares (*thresholds*) da Curva ROC.
2. **Cálculo das Probabilidades Preditivas:** As proporções de classes (textP∗textpos e textP∗textneg) são determinadas a partir dos dados de validação.

3. **Cálculo do Lucro Esperado por Limiar:** Para cada limiar (*threshold*) e seus correspondentes valores de textFPR e textTPR (obtidos na etapa 1), o Lucro Esperado (EV) é calculado utilizando a fórmula apresentada em na seção 6.2.4 e os valores de mathbfR, mathbfC e mathbfC_FP fornecidos pelo usuário.
4. **Seleção do Limiar Ótimo:** O algoritmo identifica o limiar de probabilidade para o qual o Lucro Esperado (EV) é máximo. Este é o ponto de operação que, do ponto de vista econômico/operacional, é o mais vantajoso.

O algoritmo retorna o limiar de probabilidade selecionado e as métricas de desempenho (TPR, FPR) e o Lucro Esperado (EV) obtidos nesse ponto de operação.

Figura 6.16 - Algoritmo de maximização do lucro esperado

```
# ------------------------------------------------------------------
# 1. Insira aqui seu vetor de rótulos (0/1) e as probabilidades
# ------------------------------------------------------------------
y_true  = y_test_dt
y_score = y_proba_rf  # ex.: modelo.predict_proba(X)[:, 1]

# ------------------------------------------------------------------
# 2. Defina recompensa (R) e custo (C)
# ------------------------------------------------------------------
R =  50   # ganho por acerto (TP)
C =  5   # custo por perder oportunidade (FN)
C_FP = 50

# ------------------------------------------------------------------
# 3. Calcula ROC e lucro esperado em cada limiar
# ------------------------------------------------------------------
fpr, tpr, thr = roc_curve(y_true, y_score)

P_pos = np.mean(y_true == 1)
P_neg = 1 - P_pos
ev    = P_pos * (R * tpr - C * (1 - tpr)) - P_neg * C_FP * fpr

idx_best = np.argmax(ev)
t_best   = thr[idx_best]
ev_best  = ev[idx_best]
tpr_best = tpr[idx_best]
fpr_best = fpr[idx_best]

# ------------------------------------------------------------------
# 4. Plot
# ------------------------------------------------------------------
fig, ax = plt.subplots(figsize=(6, 5))
RocCurveDisplay(fpr=fpr, tpr=tpr).plot(ax=ax, label="ROC")
ax.scatter(fpr_best, tpr_best, color="red", zorder=5, label=f"Limiar = {t_best:.2f}")
ax.plot([0, 1], [0, 1], linestyle="--", color="grey")
ax.set_title("Curva ROC com ponto ótimo de lucro")
ax.legend()

# Anotação de texto
txt = (f"TPR = {tpr_best:.3f}\n"
       f"FPR = {fpr_best:.3f}\n"
       f"Lucro esperado = {ev_best:.2f}")
ax.annotate(txt, xy=(fpr_best, tpr_best),
            xytext=(fpr_best+0.05, tpr_best-0.1),
            arrowprops=dict(arrowstyle="->"))

plt.tight_layout()
plt.show()

# ------------------------------------------------------------------
# 5. Imprimir métricas no console
# ------------------------------------------------------------------
print(f"--- Limiar ótimo ---")
print(f"   Threshold   : {t_best:.4f}")
print(f"   TPR (Recall): {tpr_best:.4f}")
print(f"   FPR         : {fpr_best:.4f}")
print(f"   EV (Retorno,Custo)    : {ev_best:.2f}")
```

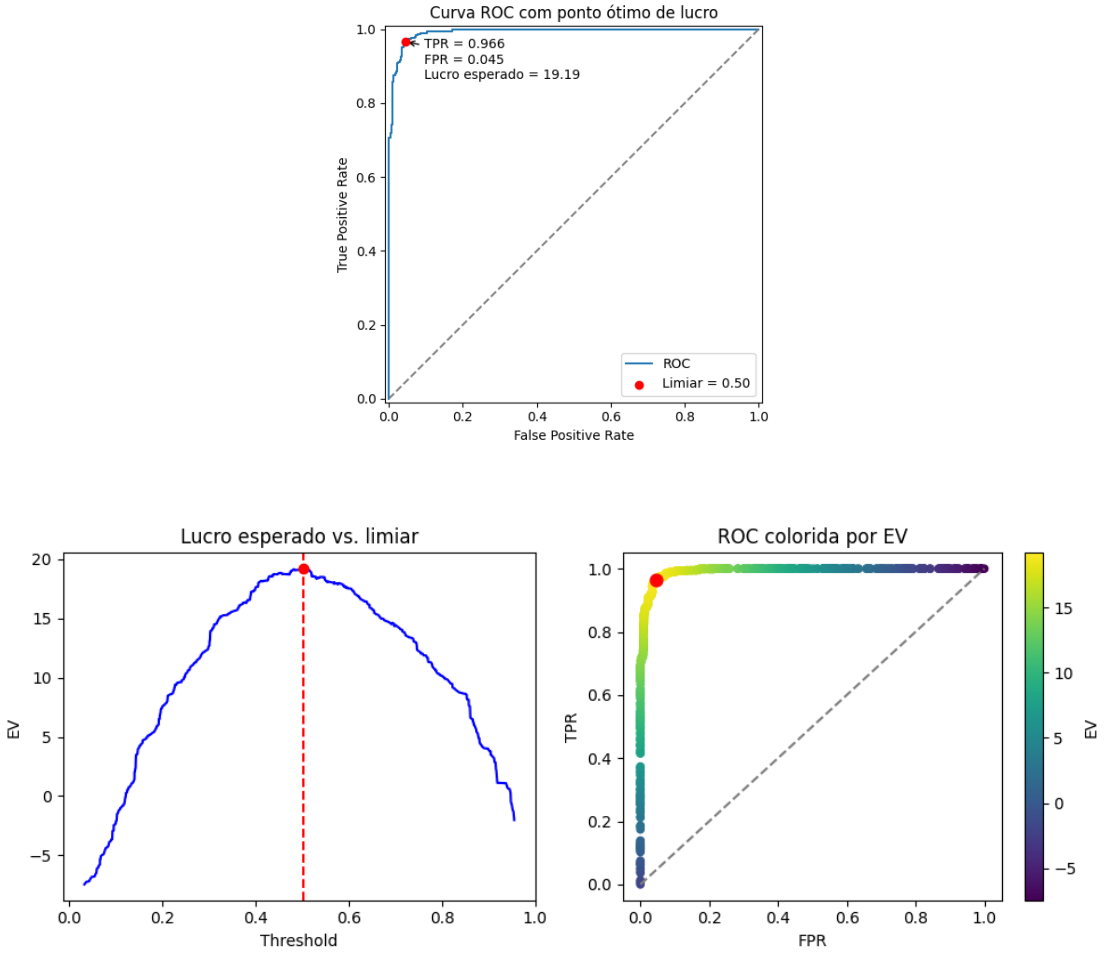
Fonte: O autor (2025)

Figura 6.17 - Definição do ponto penalizando o Falso-Positivo
(Exemplo hipotético - mercado imobiliário)

```
# ------------------------------------------------------------------
# 2. Defina recompensa (R) e custo (C)
# ------------------------------------------------------------------
R =  50   # ganho por acerto (TP)
C =  5   # custo por perder oportunidade (FN)
C_FP = 50
```

Fonte: O autor

Figura 6.18 - Definição do ponto do lucro máximo esperado



Fonte: O autor

6.2.6 Propósito e Implicações Práticas

A capacidade de selecionar o ponto de operação de forma programática, baseando-se na maximização do Lucro Esperado, tem implicações significativas para a aplicabilidade dos modelos no mercado imobiliário:

- **Otimização Direta do Resultado de Negócio:** Permite que as construtoras e incorporadoras calibrem o modelo não apenas com base em métricas estatísticas, mas diretamente na otimização de variáveis financeiras ou operacionais (lucro, custo, etc.). Por exemplo, o modelo pode ser otimizado para minimizar perdas por empreendimentos com baixa resiliência (Falsos Negativos) ou para maximizar o retorno em campanhas de marketing (Verdadeiros Positivos), considerando seus respectivos pesos econômicos.
- **Decisão Estratégica e Risco Mitigado:** Transforma a avaliação do modelo em uma discussão baseada em risco e recompensa financeiros, facilitando a

comunicação com *stakeholders* não técnicos e a tomada de decisão em cenários de alto custo do erro, como é o caso do mercado imobiliário.

- **Alocação Otimizada de Recursos:** Ajuda na alocação mais eficiente de recursos (financeiros, de pessoal, de marketing) ao indicar o ponto onde o retorno esperado é maximizado, considerando os custos associados a diferentes ações.
- **Padronização e Reprodução de Decisões:** Fornece um método objetivo e reproduzível para a definição do limiar, eliminando a subjetividade e garantindo consistência nas operações.

Este algoritmo, portanto, atua como uma ponte vital entre a performance preditiva dos modelos de Machine Learning e as necessidades operacionais e financeiras do setor imobiliário, capacitando os tomadores de decisão a extrair o máximo valor das análises de dados com uma visão clara do lucro esperado.

6.3 ANÁLISE CRÍTICA

A análise dos resultados obtidos revela aspectos importantes sobre o desempenho dos modelos e os desafios inerentes ao domínio imobiliário. Para além das métricas quantitativas, esta subseção aprofunda-se na problemática dos dados no setor, na forma como as decisões são tradicionalmente tomadas e na importância da inteligência artificial como vetor de transformação.

O desafio de tratar com 2 hipóteses diferentes e ter que dobrar todo o trabalho executado foi um ponto significativo para a complexidade deste estudo. Contudo, essa abordagem permitiu uma compreensão multifacetada dos fatores que influenciam o sucesso de um empreendimento no mercado imobiliário, destacando a capacidade dos modelos em endereçar diferentes facetas do fenômeno de vendas.

Para esse estudo foi um desafio trabalhar com todos esses modelos, que exigiram um cuidado e tratamento meticuloso dos dados para a execução. Reconhecemos que não conseguimos extrair todo o potencial dos modelos no 'estado da arte' em Machine Learning. No entanto, o objetivo principal não era esse, mas sim o de colocar a pedra fundamental para um fluxo contínuo de estudos e pesquisas na área, incentivando outros pesquisadores a explorar este campo que, diferentemente do setor financeiro ou de saúde, ainda carece de maior atenção.

6.3.1 Problemática dos dados e a decisão no mercado imobiliário

Um dos maiores desafios e, ao mesmo tempo, uma das principais justificativas para este trabalho, reside na **problemática dos dados** no setor imobiliário. Atualmente, a informação crítica para a tomada de decisões está dispersa entre dezenas de milhares de construtoras, incorporadoras, imobiliária e corretores, sem um padrão consistente de coleta, tratamento e armazenamento. Isso impede a construção de uma visão holística e baseada em evidências.

Nesse contexto, a importância de um sistema, plataforma ou modelo que sirva de **sistema de suporte à decisão (SSD)** torna-se inegável. A decisão de conceituar um empreendimento, com todas as suas características e a necessidade de ser aderente às necessidades de um mercado em constante mutação, é de alto risco e custo. Modelos avançados de recomendação baseados em Machine Learning, por mais sofisticados que sejam, não terão utilidade sem uma **base de dados confiável e consistente**.

Para mitigar essa lacuna, é fundamental ter um ***data warehouse* imobiliário** robusto, consistente e constantemente atualizado para dar suporte a esses sistemas. Iniciativas futuras deverão focar na criação de formas automáticas de atualização das bases de dados, por meio de APIs, LLMs (*Large Language Models*), Agentes Inteligentes e outros meios automatizados ou semi-automatizados. A própria captura de dados pode ser impulsionada pela criação de ecossistemas autogeridos, nos quais sistemas de informação como os CRMs (*Customer Relationship Management*) possam gerar uma quantidade massiva de dados, a exemplo do que já ocorre com as redes sociais.

A forma como as incorporadoras decidem seus projetos atualmente é um reflexo direto dessa escassez de dados qualificados. Quase sempre, as decisões são tomadas considerando a **função do custo/perda**. Por falta de informações precisas ou sistemas de recomendação seguros – comparáveis, por exemplo, aos sistemas de detecção de fraudes usados por bancos –, as construtoras dependem predominantemente da experiência e conhecimento empírico dos seus gestores, ou de uma oportunidade que o custo de risco percebido não ultrapasse uma barreira subjetiva. Isso ressalta a urgência de uma mudança para calcular o **resultado/**

ganho de forma proativa. Tal transição só será possível com sistemas robustos de suporte à decisão que utilizem uma quantidade e qualidade de dados consideráveis, capazes de prover o nível de precisão necessário em um mercado onde o custo do erro é notoriamente alto.

6.3.2 A Importância transformadora da Inteligência Artificial

Neste cenário, a **Inteligência Artificial (IA)** emerge como um agente transformador. A IA já está revolucionando diversos setores da sociedade e, assim como em outros mercados, tem um potencial imenso para impactar o mercado imobiliário.

É nesse contexto que o nosso trabalho, mesmo em sua fase embrionária, se mostra necessário. Ele representa um primeiro passo fundamental para apresentar e educar o mercado sobre a existência da Ciência de Dados e da Inteligência Artificial – muito além das LLMs (*Large Language Models*) –, focadas em problemas de predição, classificação e suporte à decisão. A adoção de uma metodologia científica rigorosa e o uso integrado de X-AI (Inteligência Artificial Explicável) e CRISP-DM (*Cross-Industry Standard Process for Data Mining*) não apenas permitiram a reprodutibilidade científica, mas também garantiram a aplicabilidade prática, fortalecendo a confiança nos modelos.

Trabalhamos neste projeto em diferentes níveis de granularidade – empreendimento e unidade habitacional –, adicionando dados de indicadores financeiros e de força de marca das construtoras, algo raro em estudos acadêmicos ou comerciais. Este rigor na criação de atributos, tratamento de dados ausentes e redução de viés fortalece a confiança nos modelos desenvolvidos.

Mesmo diante das limitações de tempo e experiência inerentes ao pesquisador, a capacidade de coletar, tratar e transformar atributos de empreendimentos, unidades habitacionais, mercado e construtoras – informações que, muitas vezes, eram informais – foi uma conquista. Isso demonstra que, em um certo ponto de maturidade, é possível criar sistemas robustos de apoio à tomada de decisão para este mercado.

Este estudo, ademais, nos deu o vislumbre de que é possível, com dados qualificados e uma equipe experiente de cientistas de dados e profissionais da área

de inteligência artificial, criar modelos extremamente eficientes, gerando uma proposta de valor imediata para o mercado e seus gestores.

6.4 VALIDAÇÃO DA HIPÓTESE

A validação das hipóteses de pesquisa formuladas neste estudo é crucial para consolidar as contribuições e direções futuras da pesquisa no campo da Ciência de Dados aplicada ao mercado imobiliário. Abaixo, cada hipótese é revisitada à luz dos resultados obtidos.

As hipóteses inicialmente propostas para este trabalho foram:

- **H1:** É possível construir um modelo preditivo com qualidade, bom desempenho e transparência para classificar o sucesso de vendas de empreendimentos residenciais com base em dados históricos e atributos estruturais.
- **H2:** O desempenho comercial de um empreendimento (velocidade e resiliência de vendas) está fortemente associado à sua capacidade de atender ao perfil de demanda do mercado, o que pode ser inferido a partir do comportamento de compra dos consumidores.
- **H3:** Modelos explicáveis de classificação, ao fornecerem interpretações acessíveis de suas decisões, favorecem a adoção prática de soluções baseadas em IA no setor imobiliário e fortalecem a confiança dos tomadores de decisão.

6.4.1 Análise da H1: Viabilidade e desempenho dos modelos preditivos

A **Hipótese 1**, focada na viabilidade de construção de modelos preditivos com bom desempenho e transparência, foi analisada sob uma perspectiva multifacetada. Do ponto de vista técnico, os modelos construídos para classificar a Velocidade e a Resiliência de Vendas (Decision Tree, Random Forest e Logistic Regression), conforme detalhado na Seção 6.1, apresentaram um desempenho estatístico **satisfatório e promissor**, especialmente considerando a natureza complexa e desafiadora dos dados imobiliários e o caráter pioneiro do estudo. Esse

desempenho já indica uma propensão para a viabilidade de classificação para as duas métricas de sucesso de vendas investigadas (Velocidade e Resiliência).

A transparência mencionada na hipótese também foi abordada através da escolha e aplicação de modelos intrinsecamente interpretáveis e por meio de técnicas de Inteligência Artificial Explicável (X-AI), como LIME, SHAP e PFI. Essa abordagem garante que o desempenho não seja uma 'caixa preta', permitindo a compreensão dos fatores que impulsionam as previsões e, assim, validando a dimensão da transparência da hipótese.

No entanto, a validação experimental permanece parcial, por três motivos principais:

1. **Ausência de aplicação prática em casos reais**: os modelos ainda não foram usados para prever o desempenho de empreendimentos futuros e acompanhar sua evolução no tempo;
2. **Restrições de generalização**: os dados utilizados estão restritos ao município do Recife e a um subconjunto de tipologias habitacionais verticais;
3. **Falta de avaliação econômica associada**: Apesar de não ser foco principal deste estudo, a avaliação econômica é vital para a extrapolação da academia para o mercado. Os modelos classificam empreendimentos quanto ao sucesso ou insucesso, mas ainda não incorporam uma análise de custo-benefício associado às decisões classificatórias.

Apesar dessas limitações, os resultados obtidos **reforçam a plausibilidade e aplicabilidade preliminar** da hipótese H1, tanto no plano estatístico quanto interpretativo. A consistência dos atributos explicativos identificados (como padrão do imóvel, localização, número de unidades, data de entrega, entre outros) reforça a plausibilidade de que as variáveis selecionadas capturam elementos determinantes do desempenho comercial.

6.4.2 Análise da H2: Associação entre desempenho comercial e perfil de demanda

A **Hipótese 2** é **parcialmente corroborada e fundamentada** pelos achados do estudo.

Neste estudo, foram avaliadas duas situações distintas e complementares para o mercado imobiliário:

- **Velocidade de Vendas**: empreendimentos com dificuldade de atingir 30% de vendas acumuladas nos três primeiros meses;
- **Resiliência de Vendas**: empreendimentos com mais de 20% de estoque ainda disponível no 18º mês após o lançamento.

Esses indicadores de Velocidade e Resiliência foram formuladas a partir de padrões de mercado historicamente observados, mas também com base em conhecimento tácito de especialistas do setor, refletindo uma combinação entre critérios operacionais e senso prático do mercado. Ainda que esses *thresholds* tenham fundamento empírico, reconhece-se que sua **validação definitiva requer diálogo estruturado com *stakeholders*** — como construtoras, incorporadoras e entidades do setor — para alinhamento entre modelagem acadêmica e aceitabilidade prática.

Embora o trabalho não tenha diretamente inferido o perfil de demanda a partir de um **comportamento de compra explícito e individualizado de consumidores** (o que exigiria dados transacionais e comportamentais mais granulares), a capacidade preditiva dos modelos para Velocidade e Resiliência de Vendas sugere que os atributos utilizados – que indiretamente refletem a adequação do empreendimento ao mercado (ex: localização, tipo de unidade, indicadores de mercado) – são representativos da demanda.

Os *insights* extraídos por meio das técnicas de X-AI (Seção 6) reforçam essa associação. Ao identificar quais características são mais importantes para classificar o sucesso de vendas, os modelos implicitamente revelam os aspectos dos empreendimentos que ressoam com a demanda do mercado. Por exemplo, se a área privativa ou número de vagas de garagem se mostram como atributos altamente preditivos, isso indica que o mercado (ou seja, o perfil de demanda) atribui grande valor a essas características. Assim, o estudo estabelece uma correlação indireta, mas robusta, entre o atendimento a esses atributos e o desempenho comercial, validando a premissa de que o sucesso está ligado à adequação à demanda inferida dos dados.

Para uma validação mais direta da inferência do perfil de demanda a partir do comportamento de compra dos consumidores, seriam necessários dados mais

detalhados sobre as interações dos consumidores com os empreendimentos (visitas, leads, etc.), o que é proposto para trabalhos futuros.

6.4.3 Análise da H3: Modelos explicáveis e a adoção da IA no setor imobiliário

A **Hipótese 3** conseguiu ser **evidenciada** pelos resultados e pela própria metodologia empregada. O presente trabalho não apenas construiu modelos com desempenho preditivo satisfatório, mas também integrou explicitamente a dimensão da explicabilidade (X-AI) desde o seu planejamento.

A capacidade de entender e explicar para participantes ativos do mercado o significado do que estávamos fazendo e os resultados e insights obtidos, evidenciam que modelos explicáveis criam uma ponte concreta entre a complexidade da IA e a necessidade de compreensão e confiança dos tomadores de decisão do setor imobiliário. Ao explicitar as regras de negócio e a importância das características de forma acessível, a pesquisa demonstra como a X-AI pode transformar a percepção dos dados de desperdício de tempo para um ativo valioso.

Embora a adoção prática em larga escala ainda seja um trabalho futuro – pois não houve um teste da solução no mercado –, o estudo estabelece as condições e a proposta de valor para que essa adoção ocorra. A criação de uma expectativa de até onde isso pode ser levado e alavancar o mercado por meio da capacidade de interpretar os modelos, indica que a transparência é, de fato, um fator crítico para a confiança e a absorção de soluções de IA em um setor tradicionalmente avesso a inovações.

6.5 ESTUDOS COMPARATIVOS

Nesta seção, comparamos os resultados e abordagens adotadas nesta dissertação com trabalhos acadêmicos e aplicados previamente realizados na interseção entre Ciência de Dados, Modelagem Preditiva e Mercado Imobiliário. O objetivo é destacar convergências, inovações e oportunidades de avanço a partir do comparações com produções acadêmicas nacionais e internacionais.

6.5.1 Comparativo com dissertações e pesquisas acadêmicas

Analisamos diversas dissertações e artigos que aplicam técnicas de mineração de dados ao mercado imobiliário, com destaque para estudos voltados à precificação de imóveis ou à análise de fatores de valorização. No entanto, poucos têm como foco a classificação de sucesso comercial de empreendimentos residenciais com base em atributos estruturais e comportamento de vendas.

Entre os estudos nacionais mais recentes, destaca-se o trabalho de Brito, Felzemburgh e Jardim (2025), que desenvolveram um modelo de aprendizado de máquina baseado em redes neurais artificiais (RNA) para precificação imobiliária na cidade de Salvador (BA). Os autores coletaram dados reais de anúncios por meio de *web scraping* e aplicaram técnicas de pré-processamento, normalização e remoção de *outliers* semelhantes às adotadas nesta pesquisa. O modelo final, uma rede neural densa e *feedforward*, apresentou $R^2 = 0,87$, MAE = R$ 915,44/m² e MAPE = 18,06%, superando as abordagens tradicionais de regressão hedônica em termos de desempenho preditivo. Entretanto, a natureza intrinsecamente opaca da RNA empregada impediu uma análise interpretativa aprofundada sobre a influência dos atributos no valor final dos imóveis, restringindo a compreensão do comportamento subjacente do modelo.

No presente estudo, embora o problema-alvo seja distinto — classificação binária de sucesso de vendas de empreendimentos residenciais em vez de regressão de preços —, a arquitetura metodológica e o domínio de aplicação apresentam notável convergência. Ambos os trabalhos partem de bases regionais do mercado imobiliário nordestino e exploram atributos estruturais, espaciais e contextuais para predição de desempenho comercial. Contudo, diferentemente de Brito et al. (2025), esta dissertação prioriza a transparência e interpretabilidade dos modelos, utilizando algoritmos intrinsecamente explicáveis e métodos pós-hoc como PFI, SHAP e LIME. Assim, o presente estudo amplia a contribuição científica ao demonstrar que é possível manter níveis comparáveis de qualidade das métricas dos modelos sem abrir mão da explicabilidade das decisões, evidenciando que a integração entre desempenho e transparência é essencial para a adoção prática da inteligência artificial no setor imobiliário brasileiro — reforçando, portanto, a Hipótese H3 desta pesquisa.

6.5.2 Inovações e contribuições originais

As contribuições centrais deste estudo, quando comparadas ao estado da arte, estão concentradas em três eixos:

1. **Formulação de Hipóteses de Sucesso Comercial com Base em Regras de Negócio:** diferentemente de modelos preditivos baseados em variáveis-contínuas (como preço), este estudo propõe e valida duas hipóteses operacionais binárias ligadas à velocidade e à resiliência de vendas, com base em patamares reconhecidos pelo mercado (30% em 3 meses; 80% em 18 meses).
2. **Integração de Fontes e Criação de Atributos Explicativos Multi-nível:** o modelo combina atributos de unidades, empreendimentos, grau de maturidade das construtoras e indicadores econômicos do mercado, criando uma estrutura explicativa mais rica e fundamentada em teorias de oferta e demanda habitacional.
3. **Aplicabilidade de Modelos Explicáveis (X-AI) com Potencial de Transferência Imediata ao Mercado:** ao empregar modelos como Decision Tree, Random Forest, Logistic Regression e técnicas como SHAP, LIME e PFI, o estudo torna a interpretação acessível para decisores não técnicos, algo raramente encontrado em estudos da área.

6.5.3 Limites e oportunidades para generalização

Ao comparar com trabalhos internacionais destaca-se que a maior parte da literatura utiliza dados de portais ou dados abertos com foco em preço. Este estudo inova ao aplicar *data mining* com foco em **desempenho comercial prospectivo**, e com base em **dados estruturados e proprietários** de alta granularidade.

Apesar de limitado a um contexto geográfico (Recife) e tipológico (empreendimentos verticais), o modelo e a metodologia podem ser expandidos para outros territórios e tipos de produtos residenciais, desde que observadas as especificidades locais e a disponibilidade de dados. Estudos comparativos futuros

podem explorar aplicações semelhantes em cidades com diferentes padrões de verticalização, renda, regulação e comportamento de compra.

6.5.4 Consideração Final

Ao se posicionar entre trabalhos acadêmicos e aplicações de mercado, este estudo contribui para consolidar a emergente linha de pesquisa em **inteligência explicável aplicada ao mercado imobiliário** no Brasil. Ao combinar rigor metodológico com orientação prática, abre caminho para a criação de sistemas de apoio à decisão com aplicabilidade real, confiável e escalável em ambientes de negócio complexos e tradicionalmente resistentes à inovação.

7 CONCLUSÃO E TRABALHOS FUTUROS

7.1 PRINCIPAIS ACHADOS

Esta dissertação demonstrou a viabilidade de construir modelos preditivos com qualidade, bom desempenho e interpretabilidade para classificar o sucesso comercial de empreendimentos residenciais, tanto sob a ótica da Velocidade de Vendas quanto da Resiliência de Vendas. A hipótese H1 foi corroborada por meio de modelos como *Decision Tree, Random Forest* e *Logistic Regression*, que apresentaram desempenho satisfatório e, sobretudo, coerente com o domínio de aplicação.

Um dos achados centrais desta pesquisa reside na estruturação de um processo robusto de coleta, tratamento e organização de dados em um setor historicamente desestruturado e avesso a práticas analíticas formais. O trabalho consolidou uma base multi-nível integrada — combinando atributos de unidades, empreendimentos, construtoras e contexto de mercado — articulada sob a metodologia CRISP-DM. Esta fundação foi essencial para o desenvolvimento analítico e representa, por si só, uma contribuição relevante para o avanço da ciência aplicada ao setor imobiliário.

Destaca-se, também, a adição de atributos relacionados à força de marca das construtoras, inferida de forma indireta a partir do tempo de atuação e do número de empreendimentos efetivamente entregues. Esse enriquecimento da base de dados permitiu capturar variáveis qualitativas com forte valor explicativo para o desempenho comercial, frequentemente negligenciadas em estudos tradicionais. Da mesma forma, a integração de indicadores econômicos — como taxa Selic, IPCA, câmbio e indicadores de inflação da construção civil — agregou uma nova dimensão analítica ao estudo, permitindo conexões entre o comportamento de vendas e o contexto macroeconômico, e abrindo possibilidades futuras para integração com outras bases socioeconômicas.

A aplicação de técnicas de Inteligência Artificial Explicável (X-AI), como SHAP, LIME e PFI, permitiu não apenas classificar empreendimentos com sucesso ou dificuldade de vendas, mas também compreender os fatores determinantes dessas classificações. A presença de atributos como padrão construtivo, região geográfica e

momento de lançamento entre os mais influentes reforça a hipótese H2, ao evidenciar a ligação entre sucesso comercial e o alinhamento do produto ao perfil de demanda latente. Mais do que performance, os modelos forneceram explicações transparentes — validando a hipótese H3 — e facilitaram o diálogo entre os dados e os decisores do setor.

Entre os principais achados deste trabalho, destaca-se a capacidade de operar com diferentes níveis de granularidade de dados — tanto no nível do empreendimento quanto no nível das unidades habitacionais. Trabalhar com múltiplas dimensões de granularidade se mostrou um diferencial técnico e metodológico importante, por representar: (i) um desafio superado em termos de integração e preparação de dados; (ii) uma inovação no domínio, já que muitas análises limitam-se a um único nível; (iii) um ganho substancial na riqueza informacional, possibilitando inferências mais detalhadas e relevantes para diferentes estágios do ciclo imobiliário; e (iv) uma base robusta para modelos mais precisos e explicativos. Adicionalmente, uma contribuição prática e metodológica significativa foi o desenvolvimento de um algoritmo para a seleção automatizada do **ponto de operação na curva ROC**. Esta ferramenta permite que os *stakeholders* alinhem a performance do modelo à metas de negócio específicas, definindo os balanços desejados entre a Taxa de Verdadeiros Positivos (TPR) e a Taxa de Falsos Positivos (FPR), com pesos de recompensas e custos, o que é essencial para uma tomada de decisão otimizada e sensível ao contexto de negócio.

Como resultado prático, esta pesquisa oferece uma estrutura replicável para suporte à decisão em planejamento de empreendimentos, com potencial de impactar significativamente o processo de concepção, lançamento e comercialização no mercado imobiliário. O trabalho inaugura um caminho para a adoção de abordagens *data-driven* em um setor tradicionalmente orientado por experiências individuais e intuição empírica, ao fornecer ferramentas analíticas alinhadas às reais necessidades de compreensão e previsibilidade dos agentes do mercado.

Por fim, este estudo representa um marco pioneiro no contexto brasileiro ao aplicar um arcabouço rigoroso de Ciência de Dados e Inteligência Artificial Explicável à análise do desempenho comercial de empreendimentos imobiliários. Ao estabelecer uma base metodológica clara e demonstrar sua aplicabilidade prática,

inaugura-se uma linha de pesquisa promissora com potencial de transformação estrutural no setor.

7.2 LIMITAÇÕES DO ESTUDO

Embora esta pesquisa tenha obtido avanços significativos, é essencial reconhecer as limitações que delimitam o escopo de generalização e aplicação dos resultados obtidos. A primeira limitação refere-se à cobertura geográfica: os dados utilizados concentram-se exclusivamente na cidade do Recife, o que implica em restrições quanto à extrapolação dos modelos para outras realidades urbanas com características socioeconômicas, demográficas e regulatórias distintas.

Além disso, o estudo concentrou-se em empreendimentos residenciais verticais, o que restringe suas conclusões à tipologia mais comum no mercado recifense, sem considerar produtos como casas, loteamentos horizontais ou empreendimentos comerciais. Essa delimitação, embora metodologicamente necessária, convida a estudos complementares para avaliar o comportamento preditivo em outros nichos do setor imobiliário.

Outro ponto de limitação está relacionado à ausência de dados comportamentais dos consumidores. A inferência do perfil de demanda foi feita a partir da análise do comportamento de compra agregado e atributos dos produtos, sem o uso de variáveis como visitas, intenção de compra, registros de CRM ou interações digitais — elementos que poderiam enriquecer significativamente os modelos com *insights* mais individualizados.

Por fim, cabe ressaltar que, embora os modelos tenham alcançado desempenho satisfatório, não foram otimizados para atingir o estado da arte em termos de performance ou sofisticação algorítmica. Essa escolha não decorre de limitação técnica, mas sim de uma estratégia metodológica deliberada: garantir a construção de uma base sólida, reprodutível e interpretável era o maior desafio e prioridade neste projeto. Ao consolidar um *pipeline* robusto — desde a coleta até a modelagem explicável — estabeleceu-se uma infraestrutura analítica que, futuramente, poderá ser refinada em cada etapa: com aprimoramentos nos modelos, na qualidade e variedade dos dados, e nas ferramentas de explicação. O foco,

portanto, foi lançar a pedra fundamental de um processo sustentável e escalável de inteligência aplicada ao mercado imobiliário.

A ausência de uma aplicação prática dos modelos em ambiente real também configura uma limitação relevante. Os resultados obtidos até aqui são experimentais, ainda que baseados em dados reais. Avaliar o impacto real da adoção dessas soluções em ciclos de decisão do mercado imobiliário dependerá de colaborações futuras com empresas do setor e validações em contexto operacional.

Mesmo com essas restrições, as limitações aqui descritas não comprometem a validade dos achados, mas reforçam a importância de continuidade e expansão da linha de pesquisa, ampliando sua abrangência territorial, tipológica e comportamental.

7.3 TRABALHOS FUTUROS

A continuidade deste trabalho pode se desdobrar em várias direções complementares, tanto no campo científico quanto na aplicação prática no mercado imobiliário. A primeira vertente refere-se ao fortalecimento da metodologia já estabelecida, com base no CRISP-DM, algoritmos de decisão binária e técnicas de Inteligência Artificial Explicável (X-AI). Esses pilares demonstraram grande aderência ao problema investigado e oferecem um caminho sólido para pesquisas futuras.

Embora o avanço tecnológico continue a fornecer novos algoritmos e abordagens sofisticadas — como modelos baseados em *deep learning* ou *ensemble* de modelos —, seu valor real só pode ser alcançado se sustentado por dados corretos, completos e de qualidade. Grandes modelos não substituem a necessidade de bases estruturadas; assim como uma Ferrari não pode rodar sem uma boa estrada, a inteligência de um sistema analítico depende da robustez da infraestrutura de dados que o alimenta.

Nesse sentido, a ênfase futura deve recair sobre a ampliação e qualificação das fontes de dados, o refinamento da engenharia de atributos e a integração com bases públicas e comportamentais. Estudos que expandam o escopo geográfico e tipológico — indo além de Recife e dos empreendimentos verticais — permitirão testar a generalização dos modelos e abrir novas hipóteses analíticas.

Do ponto de vista da aplicação prática, destaca-se a necessidade de realizar o aplicação dos modelos em ambiente real, colaborando com agentes do mercado na validação contínua dos resultados e na retroalimentação dos sistemas preditivos. Essa interação entre pesquisa e prática permitirá consolidar um ecossistema de inteligência imobiliária com impacto direto no planejamento urbano, na concepção de empreendimentos e na eficiência da alocação de recursos.

Além disso, uma das direções mais promissoras diz respeito à dimensão econômica da análise. A construção de funções de avaliação baseadas em ganhos e custos — como o impacto financeiro de uma decisão preditiva correta ou equivocada — permitirá transformar o sistema de classificação em uma ferramenta mais sofisticada de gestão de risco e projeção de resultados. No estado da arte, essa abordagem poderá até mesmo estimar margens de lucratividade, auxiliando as construtoras na precificação de produtos, alocação de recursos e avaliação de viabilidade de projetos.

Em resumo, os trabalhos futuros deverão se concentrar não apenas no refinamento técnico dos modelos, mas, sobretudo, na expansão da infraestrutura de dados e no fortalecimento das metodologias que garantam transparência, replicabilidade e valor prático para o setor.

Considerações Finais

Concluir esta pesquisa representa mais do que encerrar um ciclo acadêmico — é consolidar um marco inicial para uma transformação gradual e necessária no mercado imobiliário brasileiro. Este trabalho reafirma que a aplicação de ciência de dados, associada à Inteligência Artificial Explicável e à modelagem preditiva, é não apenas viável, mas desejável para a construção de um setor mais analítico, transparente e orientado por evidências.

Ao propor um arcabouço técnico robusto, com base metodológica sólida e modelos interpretáveis, esta dissertação lança os alicerces para que o mercado passe a enxergar os dados como ativos estratégicos. Mais do que gerar previsões, os modelos aqui desenvolvidos ajudaram a compreender o "porquê" do sucesso ou fracasso nas vendas de empreendimentos — um passo fundamental para a mudança de paradigma na tomada de decisão.

Reconhecendo seus limites, o estudo não pretende ser definitivo, mas sim inaugural. Ele aponta caminhos, propõe alternativas e desafia práticas estabelecidas, demonstrando que é possível aplicar o rigor científico sem perder de vista a aplicabilidade prática. Ao trabalhar com múltiplas granularidades, atributos derivados, fontes econômicas e explicações acessíveis, o projeto transcende a simples modelagem estatística e entrega uma plataforma conceitual para a inovação contínua.

A trajetória aqui percorrida, marcada por desafios de dados, integração metodológica e validação empírica, serve como ponto de partida para pesquisadores, profissionais e agentes públicos que desejem construir um mercado imobiliário mais inteligente, sustentável e humano. Em tempos de cidades complexas e demandas sociais crescentes, transformar informação em inteligência é mais do que uma vantagem competitiva — é uma responsabilidade coletiva.

Este trabalho, assim, retorna ao mercado com a maturidade conquistada na academia e oferece uma contribuição concreta para que decisões futuras sejam mais informadas, justas e eficazes.

REFERÊNCIAS

AGGARWAL, Charu C. *Outlier Analysis*. 2. ed. Cham: Springer, 2017. (Springer Series in Statistics). DOI: https://doi.org/10.1007/978-3-319-47578-3. Acesso em: 29 jul. 2025.

ALMEIDA, S. M.; AMANO, A. G.; TUPY, A. F. V. Fatores de sucesso em empreendimentos residenciais: uma análise empírica no Brasil. *Revista de Gestão Imobiliária*, v. 8, n. 2, p. 123–142, 2022.

ALPAYDIN, Ethem. *Introduction to Machine Learning*. 4. ed. Cambridge, MA: MIT Press, 2020.

BALDOMINOS, A. et al. Identifying real estate opportunities using machine learning. *Applied Sciences*, v. 8, n. 12, p. 2321, 2018.

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324

BRITO, Bruno Leão de; FELZEMBURGH, Maurício; JARDIM, Rei Carlos. Desenvolvimento de modelo de aprendizado de máquina para precificação imobiliária. In: 5º Simpósio Brasileiro de Tecnologia da Informação e Comunicação na Construção (SBTIC), 2025, Florianópolis. Anais [...]. Porto Alegre: ANTAC, 2025.

CHAPMAN, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium, 2000.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM Computing Surveys*, v. 51, n. 5, p. 1–42, 2018.

GUYON, I.; ELISSEEFF, A. *An introduction to variable and feature selection*. Journal of Machine Learning Research, v. 3, p. 1157–1182, 2003.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining: Concepts and Techniques*. 3. ed. Waltham: Morgan Kaufmann, 2011.

HIMMELBERG, Charles; MAYER, Christopher; SINAI, Todd. Assessing high house prices: Bubbles, fundamentals and misperceptions. *Journal of Economic Perspectives*, v. 19, n. 4, p. 67–92, 2005. Disponível em: https://doi.org/10.1257/089533005775196769. Acesso em: 29 jul. 2025.

JHA, S. et al. Predicting real estate prices using machine learning. *International Journal of Advance Research, Ideas and Innovations in Technology*, v. 6, n. 3, p. 217–224, 2020.

KIMBALL, Ralph; CASERTA, Joe. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Indianapolis: Wiley Publishing, 2004.

LI, X.; FAN, C. Real estate price prediction with heterogeneous data: A review and new perspectives. *Cities*, v. 114, p. 103207, 2021. DOI: https://doi.org/10.1016/j.cities.2021.103207.

LOCATELLI, M. M. et al. O mercado imobiliário no Brasil e seus ciclos econômicos. *Revista de Economia Contemporânea*, v. 21, n. 2, p. 1–28, 2017.

MORENO-FORONDA, I.; SÁNCHEZ-MARTÍNEZ, J.; PAREJA-EASTAWAY, M. Machine learning in real estate market analysis: A systematic literature review. *Cities*, v. 139, 2023.

PROVOST, Foster; FAWCETT, Tom. Data Science para Negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. 1. ed. Rio de Janeiro: Alta Books, 2016.

RAHM, Erhard; DO, Hong Hai. *Data cleaning: Problems and current approaches*. IEEE Data Engineering Bulletin, v. 23, n. 4, p. 3-13, 2000.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. “Why should I trust you?”: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 1135–1144.

ROSENTHAL, Stuart S. Are private markets and filtering a viable source of low-income housing? Estimates from a “repeat income” model. *American Economic Journal: Economic Policy*, v. 12, n. 2, p. 225–252, 2020.

ROUSSEEUW, Peter J.; HUBERT, Mia. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 1, n. 1, p. 73–79, 2011. Disponível em: https://doi.org/10.1002/widm.2. Acesso em: 29 jul. 2025.

SHEARER, Colin. *The CRISP-DM model: The new blueprint for data mining*. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000.

SILVA, Bruno Henrique de Pádua; SANTANA, Rogério Alves; ROCHA, Honovan Paz. Aplicação de algoritmos de aprendizado de máquina à previsão do valor de imóveis no Norte de Minas Gerais. In: Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware 2023). Foz do Iguaçu: Sociedade Brasileira de Computação, 2023. Disponível em: https://sol.sbc.org.br/index.php/latinoware/article/view/26091.

SIRMANS, G. Stacy; MACPHERSON, David A.; ZIETZ, Emily N. The composition of hedonic pricing models. *Journal of Real Estate Literature*, v. 13, n. 1, p. 3–43, 2005.

SMIT, Jan; VAN DEN HEUVEL, Martijn. Building brand equity through real estate marketing. *Journal of Real Estate Literature*, v. 18, n. 2, p. 253–270, 2010.

SOUZA, Genival Evangelista de. *Fatores que influenciam a decisão de compra de imóveis residenciais do tipo apartamento na cidade de São Paulo*. 2010. Dissertação

(Mestrado em Administração) — Universidade Municipal de São Caetano do Sul, São Caetano do Sul, 30 jun. 2010. Disponível em: https://repositorio.uscs.edu.br/handle/123456789/1222.

TALLON, Paul P. Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer*, v. 46, n. 6, p. 32–38, 2013. DOI: 10.1109/MC.2013.155

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. *Introdução à mineração de dados*. 2. ed. São Paulo: Pearson, 2018.

TUKEY, John W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

XU, Y.; NGUYEN, H. Analyzing the impact of COVID-19 on housing prices using SHAP. *Real Estate Economics*, v. 50, n. 3, p. 661–684, 2022.

ZHENG, Alice; CASARI, Amanda. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol: O'Reilly Media, 2018.

ZHU, Y.; FERREIRA, J. Knowledge-based urban data mining. *Urban Informatics*, v. 1, p. 89–104, 2014.

APÊNDICE A – REPOSITÓRIO DE ARQUIVOS E DADOS DA DISSERTAÇÃO

Todos os arquivos, bases de dados, scripts e materiais utilizados no desenvolvimento desta dissertação estão disponibilizados em acesso público no repositório digital mantido pelo autor. O repositório reúne os datasets originais e tratados, os códigos de modelagem e análise, bem como as figuras, tabelas e documentos complementares que sustentam as etapas metodológicas descritas no trabalho.

O repositório tem como objetivo garantir a transparência, reprodutibilidade e validação científica dos resultados apresentados, em conformidade com as boas práticas de ciência aberta e pesquisa replicável recomendadas pela comunidade acadêmica.

O conteúdo pode ser acessado por meio do seguinte link: https://drive.google.com/drive/folders/1Qwqd5yQQSv4JHbPWrA94w4OtCn6rxdyX?usp=drive_link

APÊNDICE B – TABELA COM A DESCRIÇÃO DOS ATRIBUTOS DO DATASET EMPREENDIMENTOS E VENDAS

Tabela B.1 - Descrição dos Atributos do *dataset* Empreendimentos e Vendas

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo
construtora_nome	Texto	Nome da construtora principal responsável pelo empreendimento	String de texto de no máximo 100 caracteres.	Identifica a empresa responsável pelo lançamento, essencial para análises por marca.	Construtora Alpha
empreendimento_nome	Texto	Nome comercial do empreendimento	String de texto de no máximo 100 caracteres.	Permite a identificação e diferenciação de empreendimentos únicos.	Residencial Alameda
empreendimento_cidade	Texto	Cidade onde se localiza o empreendimento	String de texto de no máximo 100 caracteres.	Usado para delimitar o território de análise e controle geográfico.	Recife
empreendimento_bairro	Texto	Bairro onde se localiza o empreendimento	String de texto de no máximo 100 caracteres.	Permite a análise por micro-localização e comparação entre bairros.	Boa Viagem
empreendimento_tipo	Texto	Indica o tipo das unidades principais do empreendimento	Apartamento, Casa, etc….	Permite diferenciar segmentos de mercado e produtos distintos.	Apartamento
empreendimento_data_cadastro	Data	Data de cadastro do empreendimento no sistema	Formato DD/MM/AAAA	Usado para analisar os empreendimentos que farão ou não parte do estudo	20/02/2023
empreendimento_data_lancamento	Data	Data de lançamento comercial para o mercado	Formato DD/MM/AAAA	Usado para cálculo de métricas temporais como velocidade e resiliência de vendas.	01/03/2023
empreendimento_data_entrega	Data	Data de entrega do empreendimento, caracterizando o empreendimento como pronto.	Formato DD/MM/AAAA. O campo pode estar vazio se o empreendimento não foi entregue.	Usado para cálculo de métricas temporais como velocidade e resiliência de vendas.	01/12/2026
empreendimento_data_totalmente_vendido	Data	Data final de quando todas as unidades do empreendimento foram vendidas.	Formato DD/MM/AAAA. O campo pode estar vazio se o empreendimento não foi totalmente vendido.	Permite identificar ciclo completo de vendas.	01/04/2025
empreendimento_qtde_total_unidades	Numérico	Número total de unidades/imóveis lançadas pelo empreendimento no mercado	Formato númerico inteiro, maior que zero. Vazio para dados faltantes	Permite verificar o tamanho do empreendimento e o total de unidades lançadas.	96

empreendimento_estoque_atual	Numérico	Número total de unidades a venda no estoque no mês atual	Formato númerico. Vazio para dados faltantes	Indica volume disponível atual; usado para cálculo de resiliência.	0
empreendimento_qtde_unidades_vendidas	Numérico	Número total de unidades vendida saté o mês atual	Formato númerico. Vazio para dados faltantes	Permite analisar o número total de unidades vendidas e a velocidade de vendas	98
empreendimento_data_ultima_venda	Data	Data que foi realizada a última venda do empreendimento.	Formato DD/MM/AAAA	Permite estimar a dinâmica de vendas e a estagnação do estoque.	30/11/2023
vendas_anteriores	Numérico	Número de unidades vendidas antes do mês de Janeiro de 2022. Início da série histórica de vendas.	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	0
vendas_jan_2022	Numérico	Vendas mensal do empreendimento em Janeiro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	20
vendas_fev_2022	Numérico	Vendas mensal do empreendimento em fevereiro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	12
vendas_mar_2022	Numérico	Vendas mensal do empreendimento em março de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	10
vendas_abr_2022	Numérico	Vendas mensal do empreendimento em abril de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	8
vendas_mai_2022	Numérico	Vendas mensal do empreendimento em maio de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	6
vendas_jun_2022	Numérico	Vendas mensal do empreendimento em junho de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	6
vendas_jul_2022	Numérico	Vendas mensal do empreendimento em julho de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	6
vendas_ago_2022	Numérico	Vendas mensal do empreendimento em agosto de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	5

vendas_set_2022	Numérico	Vendas mensal do empreendimento em setembro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	5
vendas_out_2022	Numérico	Vendas mensal do empreendimento em outubro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	3
vendas_nov_2022	Numérico	Vendas mensal do empreendimento em novembro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	3
vendas_dez_2022	Numérico	Vendas mensal do empreendimento em dezembro de 2022	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	2
vendas_jan_2023	Numérico	Vendas mensal do empreendimento em Janeiro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	2
vendas_fev_2023	Numérico	Vendas mensal do empreendimento em fevereiro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	3
vendas_mar_2023	Numérico	Vendas mensal do empreendimento em março de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	2
vendas_abr_2023	Numérico	Vendas mensal do empreendimento em abril de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	1
vendas_mai_2023	Numérico	Vendas mensal do empreendimento em maio de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	1
vendas_jun_2023	Numérico	Vendas mensal do empreendimento em junho de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	1
vendas_jul_2023	Numérico	Vendas mensal do empreendimento em julho de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	1
vendas_ago_2023	Numérico	Vendas mensal do empreendimento em agosto de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	0

vendas_set_2023	Numérico	Vendas mensal do empreendimento em setembro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	0
vendas_out_2023	Numérico	Vendas mensal do empreendimento em outubro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	0
vendas_nov_2023	Numérico	Vendas mensal do empreendimento em novembro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	1
vendas_dez_2023	Numérico	Vendas mensal do empreendimento em dezembro de 2023	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_jan_2024	Numérico	Vendas mensal do empreendimento em Janeiro de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_fev_2024	Numérico	Vendas mensal do empreendimento em fevereiro de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_mar_2024	Numérico	Vendas mensal do empreendimento em março de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_abr_2024	Numérico	Vendas mensal do empreendimento em abril de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_mai_2024	Numérico	Vendas mensal do empreendimento em maio de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_jun_2024	Numérico	Vendas mensal do empreendimento em junho de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_jul_2024	Numérico	Vendas mensal do empreendimento em julho de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_ago_2024	Numérico	Vendas mensal do empreendimento em agosto de 2024	Formato númerico. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	

vendas_set_2024	Numérico	Vendas mensal do empreendimento em setembro de 2024	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_out_2024	Numérico	Vendas mensal do empreendimento em outubro de 2024	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_nov_2024	Numérico	Vendas mensal do empreendimento em novembro de 2024	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_dez_2024	Numérico	Vendas mensal do empreendimento em dezembro de 2024	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_jan_2025	Numérico	Vendas mensal do empreendimento em Janeiro de 2025	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_fev_2025	Numérico	Vendas mensal do empreendimento em fevereiro de 2025	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	
vendas_mar_2025	Numérico	Vendas mensal do empreendimento em março de 2025	Formato número. Vazio para dados faltantes	Essencial para análise temporal de desempenho comercial mês a mês.	

Fonte: O autor (2025)

APÊNDICE C – TABELA COM A DESCRIÇÃO DOS ATRIBUTOS DO DATASET UNIDADES E DISPONIBILIDADES

Tabela C.1 – Descrição dos atributos do *dataset* Unidades e Disponibilidades

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
construtora_nome	Texto	Nome da construtora principal responsável pelo empreendimento	String de texto de no máximo 100 caracteres.	Entender a percepção de marca da construtora.	Construtora Alpha	Reputação da construtora: Marca e experiência
construtora_ano_fundacao	Numérico	Ano de fundação da construtora.	Formato AAAA ex: 2024. Vazio para dados faltantes	Entender a percepção de solidez da construtora.	01/06/1976	Reputação da construtora: Marca e experiência
construtora_total_empreendimentos	Numérico	Quantidade total de empreendimentos comercializados pela construtora.	Formato númerico. Vazio para dados faltantes	Entender se a percepção de solidez da construtora faz diferença nas vendas.	22	Reputação da construtora: Marca e experiência
construtora_empreendimentos_3_anos	Numérico	Quantidade empreendimentos comercializados nos últimos 3 anos pela construtora.	Formato numérico. Vazio para dados faltantes	Entender a situação e exposição atual da construtora junto ao mercado.	3	Reputação da construtora: Marca e experiência
empreendimento_nome	Texto	Nome comercial do empreendimento	String de texto de no máximo 100 caracteres.	Reconhecer e fazer a ligação do empreendimento em cada instância das unidades com o arquivo EMPREENDIMENTOS VENDAS.	Residencial Alameda	Característica do empreendimento: Identificação e Contexto Geral
empreendimento_pavimentos	Numérico	Número de pavimentos do empreendimento.	"0 para emprendimento horizontal. 1 ate N para empreendimentos verticais. Vazio para dados faltantes.	Entender se o número de pavimentos e altura do empreendimento influencia na venda da unidade.	16	Característica do empreendimento: Estrutura e porte
empreendimento_qtde_total_unidades	Numérico	Quantidade total de unidades lançadas no empreendimento.	Formato númerico. Vazio para dados faltantes.	Entender se a quantidade de unidades lançadas e densidade do empreendimento influenciam na venda da unidade.	96	Característica do empreendimento: Estrutura e porte
empreendimento_unidades_por_andar	Numérico	Quantidade total de unidades no andar da unidade da instância tratada.	Formato númerico. Vazio para dados faltantes.	Entender se o número de unidades por andar, isto é, a densidade por andar influencia na venda da unidade.	26/05/2022	Característica do empreendimento: Estrutura e porte

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
empreendimento_data_cadastro	Data	Data de cadastro do empreendimento na base de dados lida.	Formato DD/MM/AAAA ex: 01/10/2024. Vazio para dados faltantes.	Fazer o controle dos empreendimentos que foram cadastrados depois do lançamento para identificar as vendas que não foram cadastradas no momento certo	29/04/2022	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_data_lancamento	Data	Data de lançamento do empreendimento para comercialização pelo mercado.	Formato DD/MM/AAAA ex: 01/10/2024. Vazio para dados faltantes.	Conhecer a data do lançamento do empreendimento para calcular diversos atributos relacionados a essa data.	30/11/2023	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_data_entrega	Data	Data de entrega do empreendimento para os moradores.	Formato DD/MM/AAAA ex: 01/10/2024. Vazio para dados faltantes.	Conhecer a data de entrega do empreendimento para calcular diversos atributos relacionados a essa data.	30/06/2025	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_data_totalmente_vendido	Data	Data da venda da ú;ltima unidade em estoque do empreendimento.	Formato DD/MM/AAAA ex: 01/10/2024. Vazio para dados faltantes ou se o empreendimento ainda estiver em comercialização.	Conhecer a data de quando o empreendimento foi totalmente vendido e não estar mais disponível para comercialização.	30/11/2023	Característica do empreendimento: Linha do Tempo e Comercialização
emoreendimento_numero_meses_comercializacao	Numérico	Número de meses que se passaram desde a data do lançamento para comercialização até a data da geração desse dataset ou do empreendimento estar totalmente vendido.	Formato númerico. Vazio para dados faltantes.	Entender a duração da comercialização do empreendimento	19	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_cidade	Texto	Cidade onde está localizado o empreendimento. Recife	String de texto de no máximo 100 caracteres. Vazio para dados faltantes.	Neste estudo inicial vamos estudar o comportamento apenas da cidade do Recife.	Recife	Característica do empreendimento: Identificação e Contexto Geral

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
empreendimento_bairro	Texto	Bairro onde está localizado o empreendimento.	String de texto de no máximo 100 caracteres. Vazio para dados faltantes.	Entender se a principal característica de localização (bairro) influencia na perceção de compra do cliente	Boa Vista	Característica do empreendimento: Identificação e Contexto Geral
empreendimento_categoria	Texto	Tipo de comercializaçã o do empreendimento	Residencial ou Comercial. Vazio para valores faltantes.	Entender se a categoria do empreendimento influencia na venda	RESIDENCIAL	Característica do empreendimento: Identificação e Contexto Geral
empreendimento_misto	boolean	Se o empreendimento tem unidades comerciais e residenciais.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Se o empreendimento tem unidades comerciais e residenciais.	0	Característica do empreendimento: Identificação e Contexto Geral
empreendimento_beira_mar	boolean	Empreendimento a beira mar.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Localização	0	Característica do empreendimento: Localização e acessibilidade (Diferenciais externos)
empreendimento_beira_rio	boolean	Empreendimento a margem de um rio.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Localização	0	Característica do empreendimento: Localização e acessibilidade (Diferenciais externos)
empreendimento_esquina	boolean	Emprendimento em uma esquina de ruas.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Localização	0	Característica do empreendimento: Localização e acessibilidade (Diferenciais externos)
empreendimento_rua_calcada	boolean	Empreendimento em uma rua ou avenidade asfaltada ou calçada.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Localização	1	Característica do empreendimento: Localização e acessibilidade (Diferenciais externos)
empreendimento_muro_alto	boolean	Empreendimento com muro alto	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança
empreendimento_guarita	boolean	Empreendimento com guarita de segurança	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança
empreendimento_portaria24hs	boolean	Empreendimento com portaria 24hs. Humana ou eletrônica.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança
empreendimento_portao_eletronico	boolean	Empreendimento com portão eletrônico.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
empreendimento_sistema_seguranca	boolean	Empreendimento com sistema de segurança nas áreas comuns.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança
empreendimento_interfone	boolean	Empreendimento com interfone nos apartamentos.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	1	Característica do empreendimento: Itens de segurança
empreendimento_cerca_eletrica	boolean	Empreendimento com cerca elétrica.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	0	Característica do empreendimento: Itens de segurança
empreendimento_estacionamento_visitante	boolean	Empreendimento com estacionamento para visitantes.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Segurança	0	Característica do empreendimento: Itens de segurança
empreendimento_elevador	boolean	Empreendimento com elevador.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	1	Característica do empreendimento: Itens de infraestrutura
empreendimento_poco_artesiano	boolean	Empreendimento com poço artesiano.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	0	Característica do empreendimento: Itens de infraestrutura
empreendimento_gerador	boolean	Empreendimento com gerador para áreas comuns.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	1	Característica do empreendimento: Itens de infraestrutura
empreendimento_internet	boolean	Empreendimento com internet nas áreas comuns.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	0	Característica do empreendimento: Lazer e amenidades
empreendimento_bicicletario	boolean	Empreendimento com bicicletário.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	1	Característica do empreendimento: Lazer e amenidades
empreendimento_tv_a_cabo	boolean	Empreendimento com preparação para passagem de fios para TV a cabo	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	0	Característica do empreendimento: Lazer e amenidades
empreendimento_home_office	boolean	Empreendimento com home office na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	0	Característica do empreendimento: Lazer e amenidades
empreendimento_lavanderia	boolean	Empreendimento com lavanderia na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	1	Característica do empreendimento: Lazer e amenidades

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
empreendimento_salao_convencoes	boolean	Empreendimento com salão de convenções.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Infraestrutura	0	Característica do empreendimento: Lazer e amenidades
empreendimento_sala_ginastica	boolean	Empreendimento com sala de ginástica ou academia nas áreas comuns.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Esporte	1	Característica do empreendimento: Lazer e amenidades
empreendimento_pista_cooper	boolean	Empreendimento com pista de cooper na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Esporte	0	Característica do empreendimento: Lazer e amenidades
empreendimento_quadra_poliesportiva	boolean	Empreendimento com quadras de esportes na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Esporte	1	Característica do empreendimento: Lazer e amenidades
empreendimento_piscina	boolean	Empreendimento com piscina na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_playground	boolean	Empreendimento com playground ou brinquedoteca na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_sauna	boolean	Empreendimento com sauna na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_churrasqueira	boolean	Empreendimento com churrasqueira na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_espaco_gourmet	boolean	Empreendimento com espaço gourmet na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_salao_festa	boolean	Empreendimento com salão de festas na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_salao_jogos	boolean	Empreendimento com salão de jogos na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	1	Característica do empreendimento: Lazer e amenidades
empreendimento_rooftop	boolean	Empreendimento com área de rooftop na área comum.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica Lazer	0	Característica do empreendimento: Lazer e amenidades

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
unidade_codigo	Alfa-numérico	Número da unidade listada.	Formato alfa númerico. Ex: 101 ou 101B. Vazio para dados faltantes.	Código único que representa a unidade do imóvel no empreendimento	101	Característica da unidade: Identificação e tipo
unidade_pretensao	Texto	Pretensão de comercializaçã o da unidade. Vender	Vender, Alugar, Vender ou Alugar. Vazio para dados faltantes.	Nesse caso vamos trabalhar apenas com empreendimentos para venda. Esse atributo vai ser retirado.	Vender	Característica da unidade: Identificação e tipo
unidade_andar	Numérico	Pavimento / andar da unidade.	0 para térreo, 1 para andar 1, e assim sucessivamente. Vazio para dados faltantes.	Podemos entender a influência do andar da unidade no comportamento da compra do cliente	10	Característica da unidade: Layout e tipologia interna
unidade_tipo	Texto	Tipo da unidade. Apartamento	Apartamento, Casa, Terreno, etc. Vazio para dados faltantes.	Nesse caso vamos trabalhar apenas com empreendimentos com unidades do tipo Apartamento. Esse atributo vai ser retirado.	Apartamento	Característica da unidade: Identificação e tipo
unidade_sub_tipo	Texto	Subtipo de classificação da unidade.	Para o tipo apartamento pode ser: padrão, flat, loft. Vazio para dados faltantes.	Podemos entender a influência do subtipo da unidade no comportamento da compra do cliente	Studio	Característica da unidade: Identificação e tipo
unidade_valor_imovel	Numérico	Valor da unidade sendo comercializada. Nesse caso pode ser generalizada pelo valor do tipo da unidade.	Maior que R$ 1.000/m2 e menor que R$ 40.000/m2. Vazio para dados faltantes.	Podemos entender a influência do preço da unidade no comportamento da compra do cliente	550000.00	Característica da unidade: Dimensão e valor
unidade_area	Numérico	Área privativa da unidade.	Maior que R$ 10 m2 e menor que R$ 1.000/m2. Vazio para dados faltantes.	Podemos entender a influência da área da unidade no comportamento da compra do cliente	22.00	Característica da unidade: Dimensão e valor
unidade_valor_m2_imovel	Numérico	Valor do m2 da unidade, calculado entre a razão valor/área.	Maior que R$ 1.000/m2 e menor que R$ 40.000/m2. Vazio para dados faltantes.	Podemos entender a influência do preço da unidade no comportamento da compra do cliente	25000	Característica da unidade: Dimensão e valor

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
padrao_valor_m2_imovel	Texto	Padrão de classificação da unidade (econômica .... alto luxo) de acordo com o valor do m2 da unidade.	Valores: A até 4.000/m2. B até 6.000/m2. C até 8.000/m2. D até 10.000/m2. E até 12.000/m2. F até 15.000/m2. G maior que 15.000/m2. Vazio para dados faltantes.	Podemos entender a influência do padrão do valor da unidade no comportamento da compra do cliente	Alto Luxo	Característica da unidade: Dimensão e valor
unidade_quartos	Numérico	Tipologia ou quantidade de quartos da unidade.	Mínimo 1 quarto. Máximo de 6 quartos. Vazio para dados faltantes.	Podemos entender a influência do número de quartos da unidade no comportamento da compra do cliente	1	Característica da unidade: Layout e tipologia interna
unidade_suites	Numérico	Quantidade de suítes na unidade.	Não pode ser maior que o número de quartos da unidade. Vazio para dados faltantes.	Podemos entender a influência do número de suítes da unidade no comportamento da compra do cliente	0	Característica da unidade: Layout e tipologia interna
unidade_garagem	Numérico	Quantidade de vagas de garagem da unidade.	Mínimo 0 vagas. Máximo de 6 vagas. Vazio para dados faltantes.	Podemos entender a influência do número de garagens da unidade no comportamento da compra do cliente	1	Característica da unidade: Layout e tipologia interna
unidade_salas	Numérico	Quantidade de salas (estar, jantar, íntima) da unidade.	Mínimo 0 salas. Máximo de 6 salas. Vazio para dados faltantes.	Podemos entender a influência do número de salas da unidade no comportamento da compra do cliente	1	Característica da unidade: Layout e tipologia interna
unidade_banheiros	Numérico	Quantidade de banheiros sociais, sem contar as suítes, da unidade.	Mínimo 0 banheiros. Máximo de 6 banheiros. Vazio para dados faltantes.	Podemos entender a influência do número de banheiros da unidade no comportamento da compra do cliente	1	Característica da unidade: Layout e tipologia interna
unidade_nascente	boolean	Posição da unidade em relação ao nascer do sol.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Localização	1	Característica da unidade: Layout e tipologia interna
unidade_garagem_rotativa	boolean	Se a garagem da unidade é individual ou rotativa (compartilhada) por todos os moradores.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Layout e tipologia interna

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
unidade_area_servico	boolean	Se a unidade tem área de serviço.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_wc_servico	boolean	Se a unidade tem WC de serviço.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_dep_empregada	boolean	Se a unidade tem dependência completa de empregada.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_copa	boolean	Se a unidade tem copa.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_cozinha_normal	boolean	Se a unidade tem cozinha convencional.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_cozinha_americana	boolean	Se a unidade tem cozinha americana integrada com a sala.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	1	Característica da unidade: Cômodos e funcionalidades internas
unidade_despensa	boolean	Se a unidade tem despensa.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_escritorio	boolean	Se a unidade tem escritório interno.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_lavabo	boolean	Se a unidade tem lavabo social.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_sala_estar	boolean	Se a unidade tem sala de estar.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_sala_jantar	boolean	Se a unidade tem sala de jantar.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_sala_visita	boolean	Se a unidade tem sala para visitas.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
unidade_sala_intima	boolean	Se a unidade tem sala íntima.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_varanda	boolean	Se a unidade tem varanda.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	1	Característica da unidade: Cômodos e funcionalidades internas
unidade_varanda_gourmet	boolean	Se a unidade tem varanda gourmet.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Infraestrutura	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_deposito	boolean	Se a unidade tem depósito individual.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_armarios_projetados	boolean	Se a unidade já vem com armários projetados.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	0	Característica da unidade: Acabamento e mobiliário
unidade_closet	boolean	Se a unidade tem closet.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	0	Característica da unidade: Cômodos e funcionalidades internas
unidade_box_banheiro	boolean	Se a unidade já é entregue com boxes nos banheiros.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	0	Característica da unidade: Acabamento e mobiliário
unidade_mobiliado	boolean	Se a unidade já é entregue mobiliada.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	1	Característica da unidade: Acabamento e mobiliário
unidade_hidrometro_individual	boolean	Se a unidade tem medidores de água (hidrômetro) individual.	0 para negativo, 1 para positivo, vazio para valores faltantes.	Atributo de característica de Facilities	0	Característica da unidade: Acabamento e mobiliário
unidade_venda	boolean	Descrever se a unidade está Vendida ou Disponível para venda	0 para unidade não vendida. 1 para unidade vendida, vazio para valores faltantes.	Ao classificar os empreendimentos com maiorou menor propensão a venda. Vamos descobrir porque uma unidade foi vendida ou não.	0	Dinâmica de vendas e status da unidade

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
empreendimento_qtde_estoque	Numérico	Se unidade disponível para venda. Indicar a quantidade atual do estoque do empreendimento	Formato númerico. Vazio para dados faltantes ou se a unidade estiver vendida.	Entender a influência do estoque do empreendimento.	32	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_situação_atual	Texto	Se unidade disponível para venda. Indicar a situação atual do empreendimento	Lançamento, Unidades a Venda, Totalmente Vendido. Vazio para dados faltantes ou se a unidade estiver vendida.	Entender se o momento de venda do empreendimento influencia nas vendas	Unidades a venda	Característica do empreendimento: Linha do Tempo e Comercialização
empreendimento_estagio_obra_atual	Texto	Se unidade disponível para venda. Estágio atual do empreendimento.	Projeto, Em construçao, Pronto. Vazio para dados faltantes ou se a unidade estiver vendida.	Entender se o estágio de construção do empreendimento influencia nas vendas	Em construção	Característica do empreendimento: Linha do Tempo e Comercialização
unidade_venda_data	Data	Se a unidade foi vendida. Indicar a data da venda da unidade	Formato DD/MM/AAAA ex: 01/10/2024. Vazio para dados faltantes ou se a unidade está disponível para venda.	Data da venda da unidade para onfrontar com outros aributos		Dinâmica de vendas e status da unidade
unidade_venda_estoque_empreendimento	Numérico	Se a unidade foi vendida. Número de unidades disponível para venda em estoque no empreendimento no mês da data da venda da unidade.	1 até a quantidade de unidades do empreendimento. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber se o número de unidades em estoque, abundância ou escassez, influenciam na venda.		Dinâmica de vendas e status da unidade
unidade_venda_valor	Numérico	Se a unidade foi vendida. Indicar o valor da unidade na tabela no mês da venda, se não tiver. No mês anterior a venda.	Valor real com 2 casas decimais. Ex: 580.000,00. Vazio para dados faltantes ou se a unidade está disponível para venda.	Pode ser o valor real da negociação quando tivermos ou o valor em tabela da unidade no mês antes da venda.		Dinâmica de vendas e status da unidade
unidade_venda_semestre_datalancamento	Numérico	Se a unidade foi vendida. Indicar qual semestre em relação a data de lançamento do empreendimento no mês da data da venda da unidade.	0 para antes do lançamento, 1 para semestre 1, ... 6 para semestre 6, 7 para vendas pós semestre 6. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber em que semestre após o lançamento do empreendimento foi feita a venda da unidade.		Dinâmica de vendas e status da unidade

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
unidade_venda_mes_datalancamento	Numérico	Se a unidade foi vendida. Indicar em qual mês em relação a data de lançamento do empreendimento no mês da data da venda da unidade.	0 para antes da data de lançamento, 1 para mês 1 pós lançamento, ... 36 para mês 36 pós lançamento, 37 para vendas pós mês 36. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber em que mês após o lançamento do empreendimento foi feita a venda da unidade.		Dinâmica de vendas e status da unidade
unidade_venda_semestre_dataentrega	Numérico	Se a unidade foi vendida. Indicar qual o semestre em relação a data de entrega do empreendimento no mês da data da venda da unidade.	0 para antes da data de entrega, 1 para semestre 1, ... 6 para semestre 6, 7 para vendas pós semestre 6. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber em que semestre após o lançamento do empreendimento foi feita a venda da unidade.		Dinâmica de vendas e status da unidade
unidade_venda_mes_relacao_dataentrega	Numérico	Se a unidade foi vendida. Indicar qual o mês em relação a data de entrega do empreendimento da unidade no mês da data da venda da unidade.	0 para antes da data de entrega, 1 para mês 1 pós entrega, ... 36 para mês 36 pós entrega, 37 para vendas pós mês 36. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber se quando a unidade foi vendida em relação a data de entrega do empreendimento. E qual a poporção de vendas depois da entrega.		Dinâmica de vendas e status da unidade
unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia	Numérico	Se a unidade foi vendida. Indicar a quantidade de empreendimentos na cidade e com a mesma tipologia da unidade no mês da data de venda da unidade.	Formato númerico. 0 para nenhum empreendimetno. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber se a quantidade de empreendimentos em comercializaçã o influencia nas vendas.		Dinâmica de vendas e status da unidade
unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia	Numérico	Se a unidade foi vendida. Indicar a quantidade de unidades no bairro e com a mesma tipologia da unidade vendida no mês da data de venda.	Formato númerico. 0 para nenhuma unidade. Vazio para dados faltantes ou se a unidade está disponível para venda.	Saber se a quantidade de unidades lançadas influencia nas vendas.		Dinâmica de vendas e status da unidade

Atributo	Tipo	Descrição	Valores Válidos	Motivação	Exemplo	Caracterização
emprendimento_situacao_unidade_mes_venda	Texto	Se a unidade foi vendida. Indicar a situação do empreendimento no mês da data de venda da unidade.	L para Lançamento, C para Em Construção e P para Pronto. Vazio para dados faltantes.	Pode ser Lançamento (até 3 meses da data de lançamento) / Em construção (Até a data de entrega) / Pronto (depois da data de entrega)		Dinâmica de vendas e status da unidade
empreendimento_percentual_estoque_unidade_mes_venda	Numérico	Se a unidade foi vendida. Indicar o percentual do esroque do empreendimento, isto é, unidades disponíveis para venda, no mês da data de venda da unidade.	1 para até 30% das vendas, 2 entre 30% e 60%, 3 entre 60% e 100%. Vazio para dados faltabtes	Se a venda da unidade ocorreu em que percentual do estoque: Até 30, de 30 a 60, 60 a 100.		Característica do empreendimento: Linha do Tempo e Comercialização
unidade_vendas_mes_economia_taxa_selic	Numérico	Indicar o valor da taxa selic no mês da data de venda da unidade. Se a unidade não estiver vendida aplicar mês atual	Formato numérico com 2 casas decimais.	Entender a influência econômica externa no comportamento da venda.	14.50	Dados do mercado e contexto macroeconômico
unidade_vendas_mes_economia_valor_dolar	Numérico	Indicar o valor do câmbio do dólar no mês da data de venda da unidade.. Se a unidade não estiver vendida aplicar mês atual	Formato numérico com 2 casas decimais.	Entender a influência econômica externa no comportamento da venda.	5.60	Dados do mercado e contexto macroeconômico
unidade_vendas_mes_economia_IGPM	Numérico	Indicar o valor do IGPM no mês da data de venda da unidade.. Se a unidade não estiver vendida aplicar mês atual	Formato numérico com 2 casas decimais.	Entender a influência econômica externa no comportamento da venda.	3.45	Dados do mercado e contexto macroeconômico
unidade_vendas_mes_economia_INCC	Numérico	Indicar o valor do INCC no mês da data de venda da unidade. . Se a unidade não estiver vendida aplicar mês atual	Formato numérico com 2 casas decimais.	Entender a influência econômica externa no comportamento da venda.	2.10	Dados do mercado e contexto macroeconômico

Fonte: O autor (2025)

APÊNDICE D – TABELA COM AS MÉTRICAS DE VALORES DO DATASET EMPREENDIMENTOS E VENDAS

Tabela D.1 - Métricas de valores Mínimo, Máximo, Média, Desvio Padrão, Mediana, Q1 (25%), Q3 (75%), IQR, Limite Inferior e Limite Superior

Atributo	Mín.	Máx.	Média	Desvio Padrão	Q1	Q2 Median	Q3	IQR	Limite Inferior	Limite Superio
empreendimento_qtde_total_unidades	1	576	94.61	82.23	48.0	68.0	115.0	67.0	-52.5	215.5
empreendimento_estoque_atual	0	247	14.91	32.28	0.0	3.0	15.0	15.0	-22.5	37.5
empreendimento_qtde_unidades_vendidas	0	528	79.7	72.2	39.0	60.0	100.5	61.5	-53.25	192.75
vendas_anteriores	0	527	41.03	63.94	0.0	25.0	55.0	55.0	-82.5	137.5
vendas_jan_2022	0	35	0.58	3.01	0.0	0.0	0.0	0.0	0.0	0.0
vendas_fev_2022	0	35	0.77	3.02	0.0	0.0	0.0	0.0	0.0	0.0
vendas_mar_2022	0	23	0.75	2.76	0.0	0.0	0.0	0.0	0.0	0.0
vendas_abr_2022	0	183	1.41	11.18	0.0	0.0	0.0	0.0	0.0	0.0
vendas_mai_2022	0	105	0.81	6.16	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jun_2022	0	30	0.48	1.99	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jul_2022	0	35	0.64	2.76	0.0	0.0	0.0	0.0	0.0	0.0
vendas_ago_2022	0	41	0.84	3.67	0.0	0.0	0.0	0.0	0.0	0.0
vendas_set_2022	0	53	0.64	3.25	0.0	0.0	0.0	0.0	0.0	0.0
vendas_out_2022	0	15	0.49	1.55	0.0	0.0	0.0	0.0	0.0	0.0
vendas_nov_2022	0	69	0.65	4.15	0.0	0.0	0.0	0.0	0.0	0.0
vendas_dez_2022	0	17	0.54	1.58	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jan_2023	0	41	0.53	2.64	0.0	0.0	0.0	0.0	0.0	0.0
vendas_fev_2023	0	45	0.51	2.79	0.0	0.0	0.0	0.0	0.0	0.0
vendas_mar_2023	0	51	1.02	3.79	0.0	0.0	0.0	0.0	0.0	0.0
vendas_abr_2023	0	33	0.57	2.43	0.0	0.0	0.0	0.0	0.0	0.0
vendas_mai_2023	0	113	1.37	7.83	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jun_2023	0	10	0.42	1.23	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jul_2023	0	171	1.27	10.17	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_ago_2023	0	29	0.82	2.73	0.0	0.0	0.0	0.0	0.0	0.0
vendas_set_2023	0	67	0.9	4.99	0.0	0.0	0.0	0.0	0.0	0.0
vendas_out_2023	0	118	1.29	8.5	0.0	0.0	0.0	0.0	0.0	0.0
vendas_nov_2023	0	58	1.25	5.46	0.0	0.0	0.0	0.0	0.0	0.0
vendas_dez_2023	0	114	1.23	7.72	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jan_2024	0	19	0.55	1.65	0.0	0.0	0.0	0.0	0.0	0.0
vendas_fev_2024	0	76	1.55	7.13	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_mar_2024	0	34	0.8	3.08	0.0	0.0	0.0	0.0	0.0	0.0
vendas_abr_2024	0	63	1.1	4.94	0.0	0.0	1.0	1.0	-1.5	2.5

vendas_mai_2024	0	19	0.78	2.28	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_jun_2024	0	71	0.71	4.38	0.0	0.0	0.0	0.0	0.0	0.0
vendas_jul_2024	0	84	1.15	5.99	0.0	0.0	0.0	0.0	0.0	0.0
vendas_ago_2024	0	129	1.6	8.56	0.0	0.0	0.0	0.0	0.0	0.0
vendas_set_2024	0	17	0.9	2.41	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_out_2024	0	27	0.81	2.89	0.0	0.0	0.0	0.0	0.0	0.0
vendas_nov_2024	0	27	0.82	2.92	0.0	0.0	0.0	0.0	0.0	0.0
vendas_dez_2024	0	95	1.98	7.39	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_jan_2025	0	45	0.43	3.34	0.0	0.0	0.0	0.0	0.0	0.0
vendas_fev_2025	0	25	0.76	2.71	0.0	0.0	0.0	0.0	0.0	0.0
vendas_mar_2025	0	103	1.71	9.22	0.0	0.0	0.0	0.0	0.0	0.0
vendas_abr_2025	0	49	1.42	5.16	0.0	0.0	1.0	1.0	-1.5	2.5
vendas_mai_2025	0	29	0.67	2.5	0.0	0.0	0.0	0.0	0.0	0.0

Fonte: O autor (2025)

APÊNDICE E – TABELA COM AS MÉTRICAS DE VALORES DO DATASET UNIDADES E DISPONIBILIDADES

Tabela E.1 - Estatísticas descritivas dos atributos Numéricos do *dataset* Unidades e Disponibilidades

Atributo	Mín.	Máx.	Média	Desvio Padrão	Q1	Q2 Mediana	Q3	IQR	Limite Inferior	Limite Superior
empreendimento_pavimentos	0	42	20.6	7.56	15.0	20.0	25.0	10.0	0.0	40.0
empreendimento_qtde_total_unidades	1	576	165.6	126.83	72.0	127.0	208.0	136.0	-132.0	412.0
empreendimento_unidades_por_anda	0	34	6.0	3.37	4.0	6.0	8.0	4.0	-2.0	14.0
empreendimento_numero_meses_comercializacao	2	177	49.4	40.57	18.0	40.0	70.0	52.0	-60.0	148.0
unidade_andar	0	42	10.4	7.31	5.0	9.0	15.0	10.0	-10.0	30.0
unidade_valor_imovel	0	9088762	630615.9	569785.7	340000.0	441000.0	650000.0	310000.0	-125000.0	1115000.0
unidade_area	18	536	63.1	39.2	40.8	52.5	68.0	27.2	-0.07	108.84
unidade_valor_m2_imovel	4210	28226	9882.3	3086.6	7432.3	9348.2	11491.6	4059.4	1343.2	17580.7
unidade_quartos	1	5	2.2	0.9	1.0	2.0	3.0	2.0	-2.0	6.0
unidade_suites	1	5	1.3	0.8	1.0	1.0	1.0	0.0	1.0	1.0
unidade_garagem	1	5	1.3	0.6	1.0	1.0	1.0	0.0	1.0	1.0
empreendimento_qtde_estoque	1	247	80.9	68.5	25.0	52.0	135.0	110.0	-140.0	300.0
unidade_venda_estoque_empreendimento	0	564	52.6	72.4	10.0	28.0	68.0	58.0	-77.0	155.0
unidade_venda_valor	1	9088762	632234.6	565049.6	344459.0	440000.0	650000.0	305541.0	-113852.6	1108311.5
unidade_venda_semestre_datalancamento	0	7	3.2	2.9	0.0	2.0	7.0	7.0	-10.5	17.5
unidade_venda_mes_datalancamento	0	36	7.5	8.2	0.0	7.0	8.0	8.0	-12.0	20.0
unidade_venda_semestre_dataentrega	0	7	1.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0
unidade_venda_mes_relacao_dataentrega	0	36	2.5	5.9	0.0	0.0	0.0	0.0	0.0	0.0

unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia	1	191	89.1	40.3	66.0	73.0	116.0	50.0	-9.0	191.0
unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia	0	57	10.8	13.1	2.0	5.0	18.0	16.0	-22.0	42.0
empreendimento_percentual_estoque_unidade_mes_venda	0	99	31.0	25.1	9.0	26.0	49.0	40.0	-51.0	109.0

Fonte: O autor (2025)

APÊNDICE F – TABELA ESTATÍSTICAS DESCRITAS DOS ATRIBUTOS BOLEANOS

Este apêndice apresenta os resultados detalhados da análise estatística dos **atributos booleanos** incluídos no dataset utilizado para modelagem da classe-alvo *Velocidade de Vendas*. O objetivo desta análise foi subsidiar a **seleção de atributos relevantes**, com base em critérios empíricos e estatísticos robustos.

A tabela contém, para cada atributo booleano avaliado:

- **proporcao_classe1_true**: proporção de instâncias com classe_alvo = 1 entre os casos em que o atributo possui valor True;
- **proporcao_classe1_false**: proporção de instâncias com classe_alvo = 1 entre os casos em que o atributo possui valor False;
- **valor_p**: valor-p do teste de **Qui-quadrado de independência**, utilizado para avaliar a associação estatística entre o atributo e a variável de saída;
- **% Valor 0** e **% Valor 1**: distribuição de frequência dos valores False (0) e True (1) para cada atributo, permitindo a detecção de **atributos desbalanceados** (baixa variância).

A análise combinada desses indicadores foi utilizada como **critério de decisão** para a manutenção ou exclusão dos atributos booleanos no conjunto de dados final. Foram mantidos atributos com **significância estatística (valor-p < 0.05)** ou com **importância preditiva demonstrada nos modelos**, desde que não excessivamente desbalanceados.

Esta abordagem busca equilibrar **robustez estatística, interpretabilidade e desempenho preditivo**, contribuindo para a construção de modelos explicáveis e alinhados ao conhecimento de domínio.

Tabela F.1 - Estatísticas descritivas dos atributos do tipo booleano

atributo	proporcao_classe1_true	proporcao_classe1_fals	valor_p	% Valor 0	% Valor 1
unidade_garagem_rotativa	0,01304	0,45812	0,00000	0,93650	0,06350
empreendimento_salao_jogos	0,56238	0,34987	0,00000	0,62360	0,37640
empreendimento_quadra_poliesportiva	0,62062	0,38685	0,00000	0,81610	0,18390
empreendimento_portao_eletronico	0,47945	0,28237	0,00000	0,74830	0,25170
empreendimento_rua_calcada	0,46052	0,20446	0,00000	0,88020	0,11980

atributo	proporcao_classe1_true	proporcao_classe1_fals	valor_p	% Valor 0	% Valor 1
empreendimento_salao_festa	0,49105	0,31967	0,00000	0,64290	0,35710
unidade_area_servico	0,48058	0,30056	0,00000	0,71820	0,28180
empreendimento_playground	0,48531	0,31194	0,00000	0,68010	0,31990
empreendimento_gerador	0,37621	0,54488	0,00000	0,68200	0,31800
empreendimento_sala_ginastica	0,34820	0,50422	0,00000	0,52330	0,47670
empreendimento_espaco_gourmet	0,45745	0,23256	0,00000	0,87730	0,12270
unidade_varanda_gourmet	0,45745	0,23256	0,00000	0,87730	0,12270
unidade_varanda	0,48863	0,34482	0,00000	0,59120	0,40880
empreendimento_guarita	0,51427	0,37282	0,00000	0,59690	0,40310
unidade_hidrometro_individual	0,84848	0,41942	0,00000	0,97570	0,02430
empreendimento_estacionamento_visitante	0,78873	0,42021	0,00000	0,97390	0,02610
empreendimento_bicicletario	0,46096	0,34479	0,00000	0,73220	0,26780
empreendimento_interfone	0,44654	0,32555	0,00000	0,86200	0,13800
unidade_cozinha_americana	0,00000	0,43256	0,00000	0,99370	0,00630
unidade_closet	0,22917	0,43531	0,00000	0,97350	0,02650
unidade_deposito	0,27373	0,43664	0,00000	0,95830	0,04170
empreendimento_piscina	0,43376	0,16456	0,00000	0,98550	0,01450
empreendimento_home_office	0,73043	0,42663	0,00000	0,98940	0,01060
unidade_sala_jantar	0,44584	0,37655	0,00000	0,76920	0,23080
empreendimento_internet	0,26117	0,43449	0,00000	0,97320	0,02680
unidade_cozinha_normal	0,43420	0,31579	0,00000	0,96330	0,03670
unidade_lavabo	0,36069	0,43676	0,00000	0,90910	0,09090
unidade_copa	0,00000	0,43072	0,00011	0,99800	0,00200
empreendimento_sauna	0,50615	0,42563	0,00019	0,94760	0,05240
empreendimento_pista_cooper	0,52239	0,42751	0,00240	0,97530	0,02470
empreendimento_churrasqueira	0,40942	0,43983	0,00283	0,67170	0,32830
empreendimento_muro_alto	0,43437	0,40620	0,03152	0,83950	0,16050
empreendimento_poco_artesiano	0,49558	0,42916	0,18575	0,98960	0,01040
empreendimento_lavanderia	0,42231	0,43443	0,22352	0,62200	0,37800

atributo	proporcao_classe1_true	proporcao_classe1_fals	valor_p	% Valor 0	% Valor 1
unidade_dep_empregada	0,44727	0,42849	0,32332	0,92750	0,07250
empreendimento_sistema_seguranca	0,43152	0,41774	0,35868	0,87860	0,12140
empreendimento_portaria24hs	0,43229	0,42293	0,39886	0,73900	0,26100
unidade_wc_servico	0,41872	0,43125	0,42264	0,88790	0,11210
unidade_sala_estar	0,42970	0,43040	0,97123	0,79170	0,20830
empreendimento_beira_mar	0,00000	0,42985	1,00000	1,00000	0,00000
empreendimento_cerca_eletrica	0,00000	0,42985	1,00000	1,00000	0,00000
empreendimento_esquina	0,00000	0,42985	1,00000	1,00000	0,00000
empreendimento_misto	0,00000	0,42985	1,00000	1,00000	0,00000
empreendimento_rooftop	0,42985	0,00000	1,00000	1,00000	0,00000
empreendimento_salao_convencoes	0,00000	0,42985	1,00000	1,00000	0,00000
empreendimento_tv_a_cabo	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_armarios_projetados	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_box_banheiro	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_despensa	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_escritorio	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_mobiliado	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_nascente	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_sala_intima	0,00000	0,42985	1,00000	1,00000	0,00000
unidade_sala_visita	0,00000	0,42985	1,00000	1,00000	0,00000

Fonte: O autor

APÊNDICE G – ANÁLISE ESTATÍSTICA E GRÁFICA DOS ATRIBUTOS NUMÉRICOS DE BAIXA CARDINALIDADE

Este apêndice apresenta as análises estatísticas realizadas para os atributos numéricos de baixa cardinalidade com impacto potencial na variável-alvo de velocidade de vendas. Os testes foram conduzidos para auxiliar na decisão sobre a transformação adequada de cada variável, considerando aspectos como distribuição de classes, associação estatística e interpretabilidade dos modelos.

Atributo: unidade_garagem

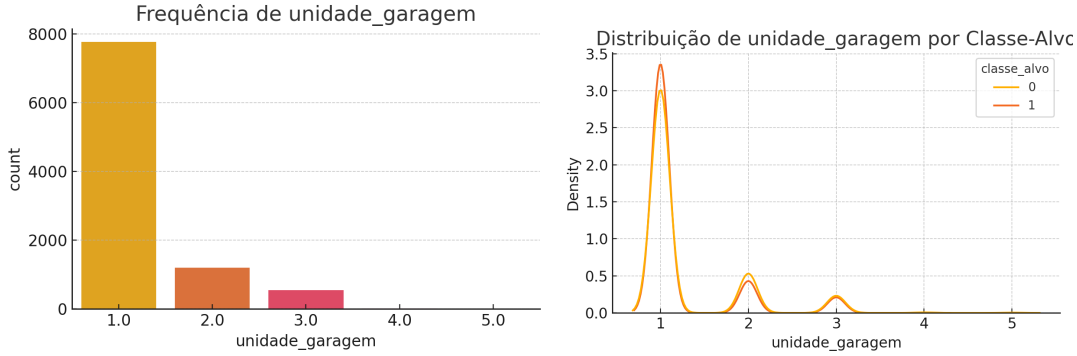
Tabela G.1 - Estatísticas descritivas do atributo garagem por classe-alvo

index	0	1
count	5434.0	4095.0
mean	1.27	1.21
std	0.59	0.53
min	1.0	1.0
25%	1.0	1.0
50%	1.0	1.0
75%	1.0	1.0
max	5.0	4.0

Fonte: O autor

Teste t de Student: t = 5.093, p = 0.0000

Figura G.1 - Frequência e distribuição por classe-alvo do atributo unidade_garagem



Fonte: O autor

Atributo: unidade_salas

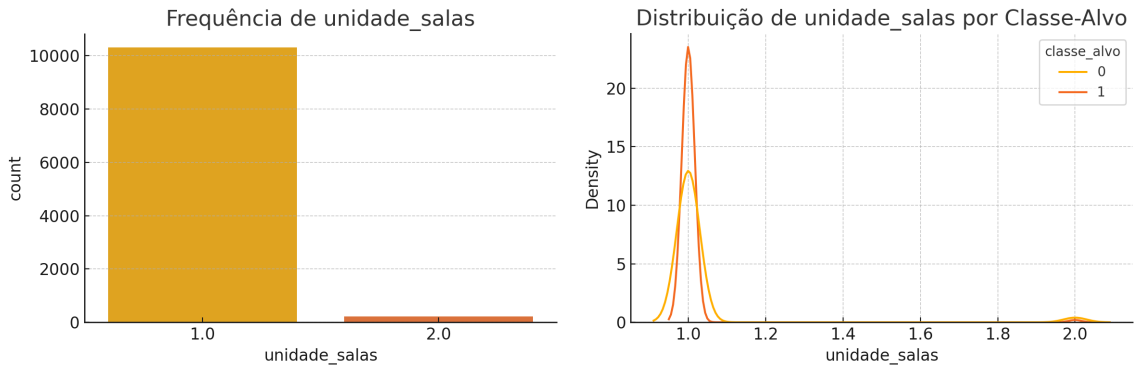
Tabela G.2 - Estatísticas descritivas do atributo unidade_salas por classe-alvo

index	0	1
count	5939.0	4579.0
mean	1.03	1.01
std	0.17	0.09
min	1.0	1.0
25%	1.0	1.0
50%	1.0	1.0
75%	1.0	1.0
max	2.0	2.0

Fonte: O autor

Teste t de Student: t = 8.328, p = 0.0000

Figura G.2 - Frequência e distribuição por classe-alvo do atributo unidade_salas



Fonte: O autor

Atributo: unidade_suites

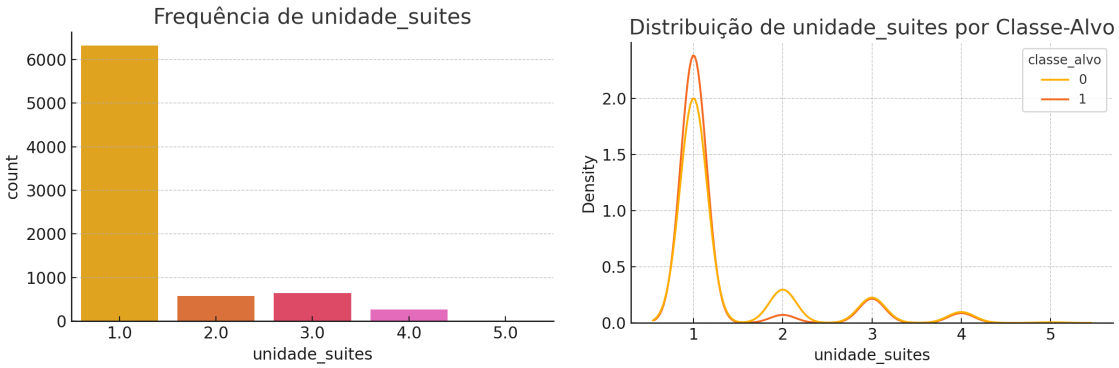
Tabela G.3 - Estatísticas descritivas do atributo unidade_suites por classe-alvo

index	0	1
count	4341.0	3479.0
mean	1.4	1.28
std	0.81	0.74
min	1.0	1.0
25%	1.0	1.0
50%	1.0	1.0
75%	1.0	1.0
max	5.0	4.0

Fonte: O autor

Teste t de Student: t = 7.174, p = 0.0000

Figura G.3 - Frequência e distribuição por classe-alvo do atributo unidade_suites



Fonte: O autor

Atributo: unidade_quartos

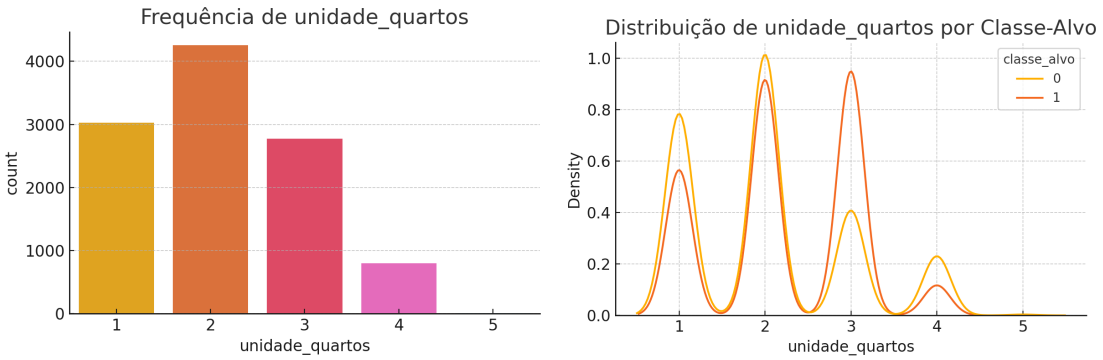
Tabela G.4 - Estatísticas descritivas do atributo unidade_quartos por classe-alvo

index	0	1
count	6193.0	4669.0
mean	2.04	2.24
std	0.94	0.85
min	1.0	1.0
25%	1.0	2.0
50%	2.0	2.0
75%	3.0	3.0
max	5.0	4.0

Fonte: O autor

Teste t de Student: t = -11.845, p = 0.0000

Figura G.4 - Frequência e distribuição por classe-alvo do atributo unidade_quartos



Fonte: O autor

Atributo: unidade_banheiros

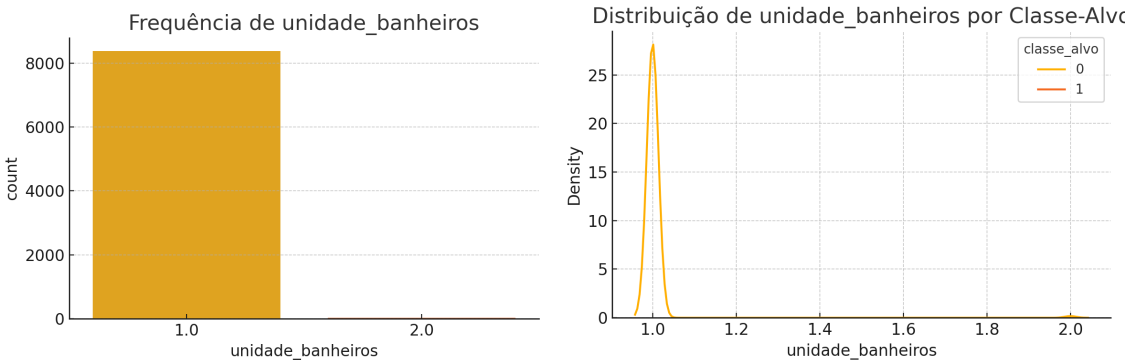
Tabela G.5 - Estatísticas descritivas do atributo unidade_banheiros por classe-alvo

index	0	1
count	4657.0	3744.0
mean	1.01	1.0
std	0.08	0.0
min	1.0	1.0
25%	1.0	1.0
50%	1.0	1.0
75%	1.0	1.0
max	2.0	1.0

Fonte: O autor

Teste t de Student: t = 5.211, p = 0.0000

Figura G.5 - Frequência e distribuição por classe-alvo do atributo unidade_banheiros



Fonte: O autor

APÊNDICE H – ANÁLISE ESTATÍSTICA E GRÁFICA DOS ATRIBUTOS NUMÉRICOS DE MÉDIA/ALTA CARDINALIDADE

Este apêndice apresenta as análises estatísticas realizadas para os atributos numéricos de média ou alta cardinalidade com impacto potencial na variável-alvo de velocidade de vendas. Os testes foram conduzidos para auxiliar na decisão sobre a transformação adequada de cada variável, considerando aspectos como distribuição de classes, associação estatística e interpretabilidade dos modelos.

Atributo: construtora_ano_fundacao

Tabela H.1 - Estatísticas descritivas do atributo do ano de fundação da construtora por classe-alvo

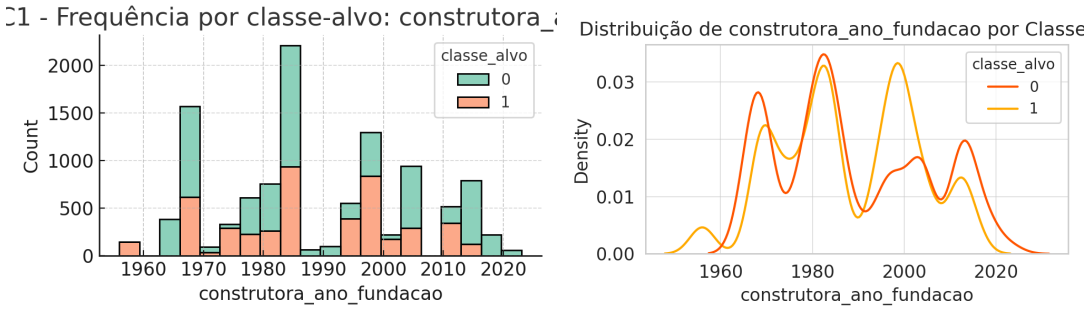
Estatística	Classe 0	Classe 1
count	6193.0	4669.0
mean	1988.83	1988.16
std	16.42	14.74
min	1966.0	1956.0
25%	1977.0	1977.0
50%	1983.0	1986.0
75%	2004.0	1999.0
max	2023.0	2015.0

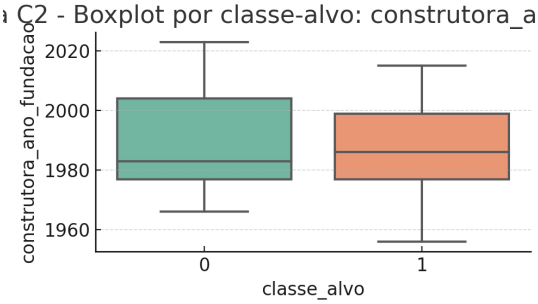
Fonte: O autor (2025)

Teste t de Student: t = 2.232, p = 0.02566

Teste Mann-Whitney U: U = 14516386.000, p = 0.71530

Figura H.1 - Frequência e distribuição por classe-alvo do atributo do ano de fundação da construtora





Fonte: O autor (2025)

Atributo: construtora_empreendimentos_entregues

Tabela H.2 - Estatísticas descritivas do atributo que indica a quantidade de empreendimentos entregues pela construtora por classe-alvo

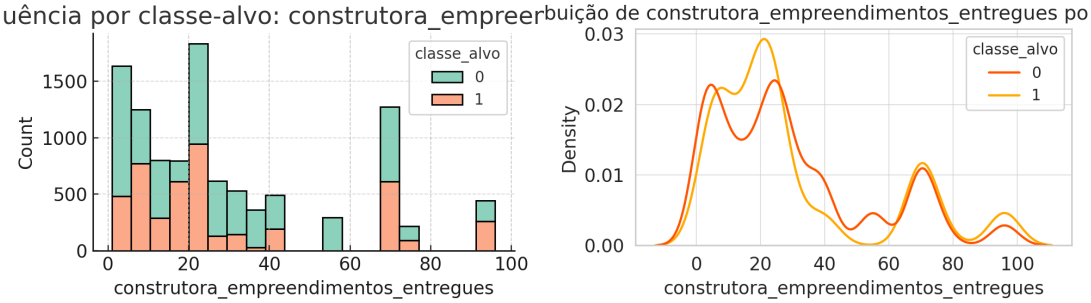
Estatística	Classe 0	Classe 1
count	5969.0	4573.0
mean	29.46	30.13
std	24.71	26.55
min	1.0	2.0
25%	8.0	10.0
50%	24.0	23.0
75%	40.0	40.0
max	96.0	96.0

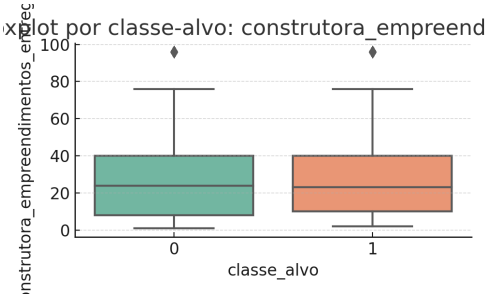
Fonte: O autor (2025)

Teste t de Student: t = -1.327, p = 0.18447

Teste Mann-Whitney U: U = 13733161.000, p = 0.58212

Figura H.2 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade de empreendimentos entregues pela construtora





Fonte: O autor (2025)

Atributo: unidade_vendas_mes_qtde_empreendimentos_bairro_tipologia

Tabela H.3 - Estatísticas descritivas do atributo que indica o estoque de unidades no bairro com a mesma tipologia no mês da unidade vendida

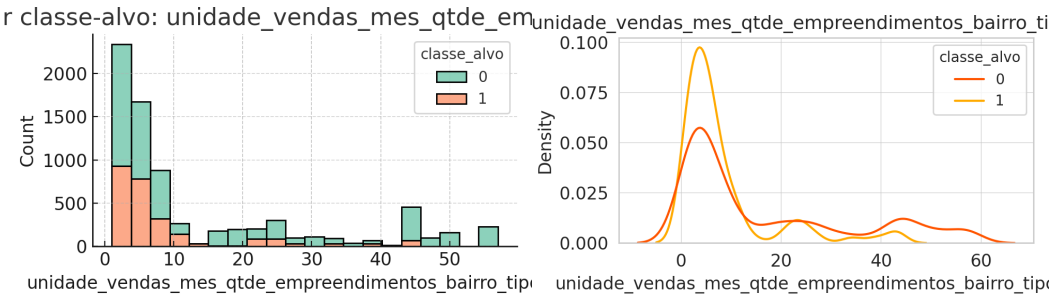
Estatística	Classe 0	Classe 1
count	4878.0	2569.0
mean	16.6	8.23
std	17.63	9.59
min	1.0	1.0
25%	3.0	3.0
50%	7.0	5.0
75%	26.0	9.0
max	57.0	43.0

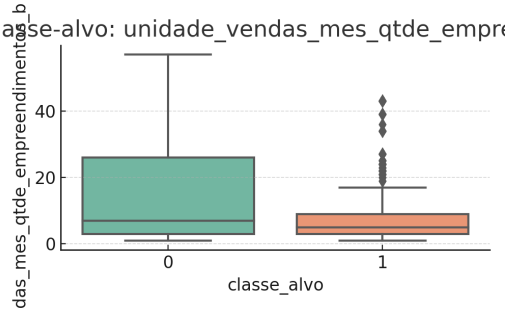
Fonte: O autor (2025)

Teste t de Student: t = 26.552, p = 0.00000

Teste Mann-Whitney U: U = 7705845.000, p = 0.00000

Figura H.3 - Frequência e distribuição por classe-alvo do atributo que indica estoque de unidades no bairro com a mesma tipologia no mês da unidade vendida





Fonte: O autor (2025)

Atributo: unidade_vendas_mes_qtde_empreendimentos_cidade_tipologia

Tabela H.4 - Estatísticas descritivas do atributo que indica o estoque de unidades na cidade com a mesma tipologia no mês da unidade vendida

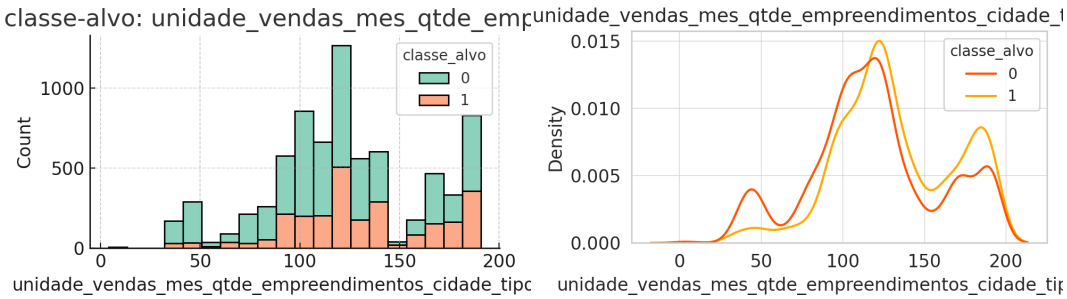
Estatística	Classe 0	Classe 1
count	4878.0	2569.0
mean	119.1	132.02
std	39.73	36.17
min	4.0	33.0
25%	97.0	109.0
50%	118.0	125.0
75%	141.0	163.0
max	191.0	191.0

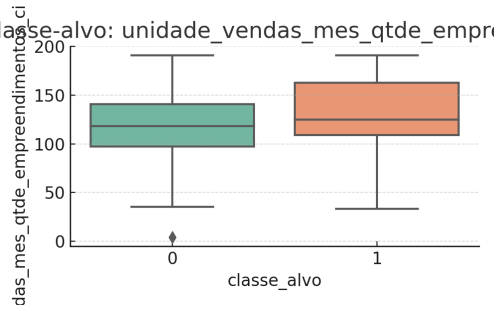
Fonte: O autor (2025)

Teste t de Student: t = -14.159, p = 0.00000

Teste Mann-Whitney U: U = 5014159.000, p = 0.00000

Figura H.4 - Frequência e distribuição por classe-alvo do atributo que indica estoque de unidades na cidade com a mesma tipologia no mês da unidade vendida





Fonte: O autor (2025)

Atributo: unidade_venda_semestre_dataentrega

Tabela H.5 - Estatísticas descritivas do atributo que indica a quantidade de vendas no semestre da venda da unidade

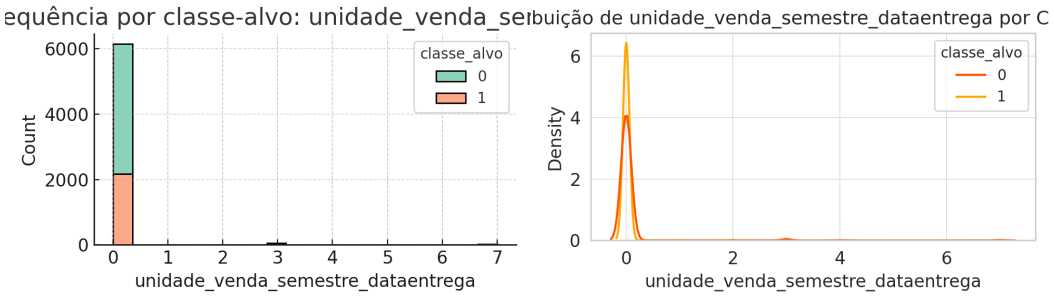
Estatística	Classe 0	Classe 1
count	4042.0	2192.0
mean	0.05	0.03
std	0.5	0.29
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	0.0
max	7.0	3.0

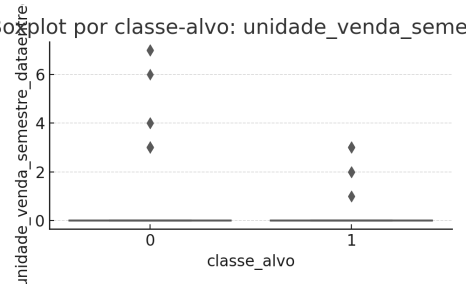
Fonte: O autor (2025)

Teste t de Student: t = 2.391, p = 0.01683

Teste Mann-Whitney U: U = 4437066.000, p = 0.59491

Figura H.5 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade de vendas no semestre da venda da unidade





Fonte: O autor (2025)

Atributo: construtora_empreendimentos_entregues_3_anos

Tabela H.6 - Estatísticas descritivas do atributo que indica a quantidade de empreendimentos no entregue ao mercado pela construtora nos últimos 3 anos

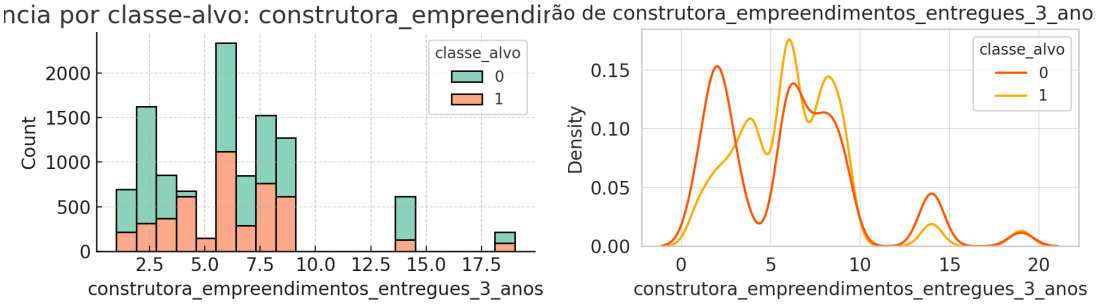
Estatística	Classe 0	Classe 1
count	6137.0	4669.0
mean	6.06	6.23
std	4.02	3.24
min	1.0	1.0
25%	2.0	4.0
50%	6.0	6.0
75%	8.0	8.0
max	19.0	19.0

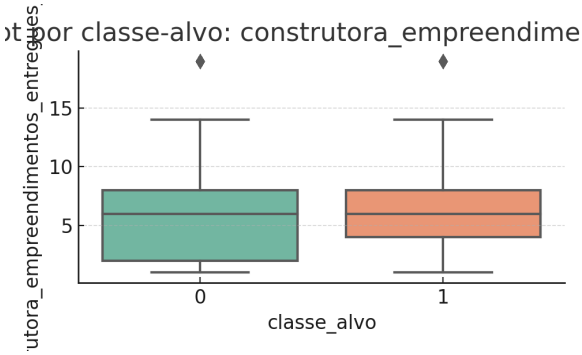
Fonte: O autor (2025)

Teste t de Student: t = -2.355, p = 0.01854

Teste Mann-Whitney U: U = 13439522.500, p = 0.00000

Figura H.6 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade de empreendimentos no entregue ao mercado pela construtora nos últimos 3 anos





Fonte: O autor (2025)

Atributo: unidade_venda_mes_relacao_dataentrega

Tabela H.7 - Estatísticas descritivas do atributo que indica em quanto meses a unidade foi vendida depois da entrega do empreendimento

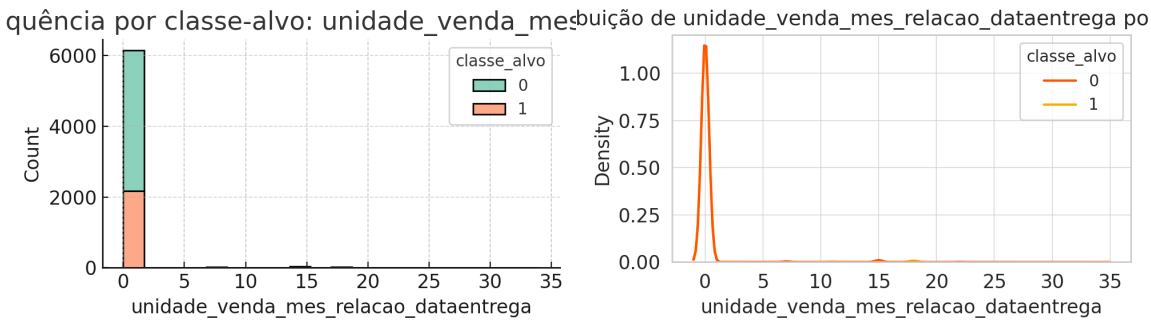
Estatística	Classe 0	Classe 1
count	4042.0	2192.0
mean	0.19	0.17
std	1.74	1.63
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	0.0
max	34.0	18.0

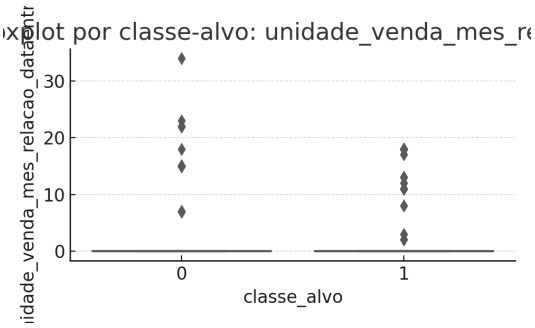
Fonte: O autor (2025)

Teste t de Student: t = 0.543, p = 0.58746

Teste Mann-Whitney U: U = 4436564.000, p = 0.62146

Figura H.7 - Frequência e distribuição por classe-alvo do atributo que indica em quanto meses a unidade foi vendida depois da entrega do empreendimento





Fonte: O autor (2025)

Atributo: empreendimento_unidades_por_andar

Tabela H.8 - Estatísticas descritivas do atributo que indica a quantidade de unidades por pavimento do empreendimento

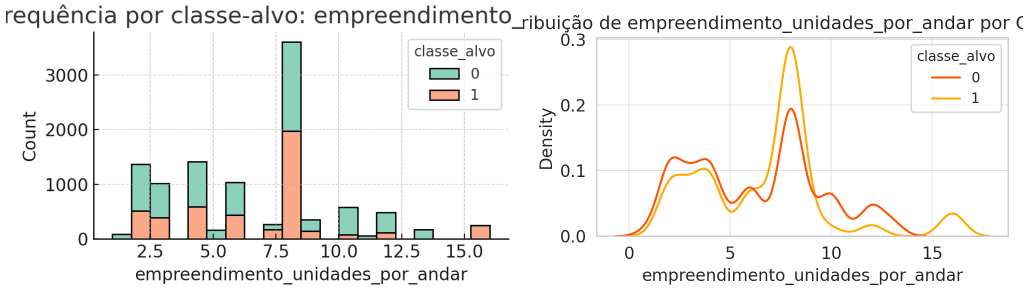
Estatística	Classe 0	Classe 1
count	6193.0	4669.0
mean	6.35	6.79
std	3.22	3.3
min	1.0	2.0
25%	3.0	4.0
50%	6.0	8.0
75%	8.0	8.0
max	13.0	16.0

Fonte: O autor (2025)

Teste t de Student: t = -6.945, p = 0.00000

Teste Mann-Whitney U: U = 13780952.000, p = 0.00002

Figura H.8 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade de unidades por pavimento do empreendimento



Fonte: O autor (2025)

Atributo: unidade_andar

Tabela H.9 - Estatísticas descritivas do atributo que indica o pavimento (andar) da unidade

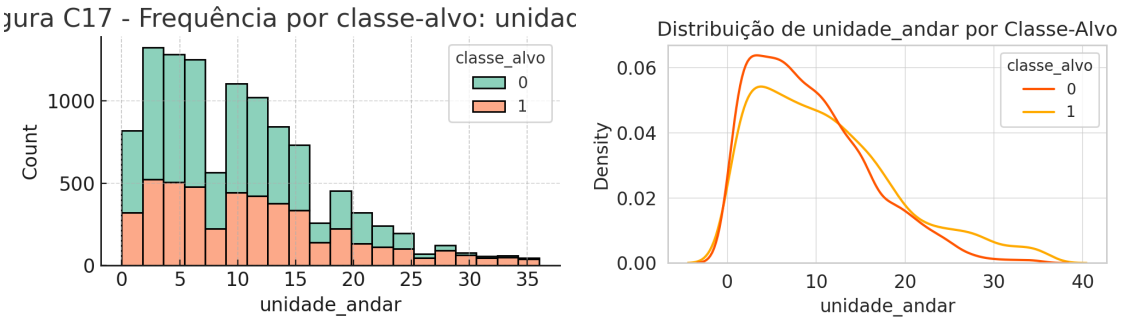
Estatística	Classe 0	Classe 1
count	6184.0	4669.0
mean	9.31	11.19
std	6.49	8.03
min	0.0	0.0
25%	4.0	5.0
50%	8.0	10.0
75%	13.0	16.0
max	35.0	36.0

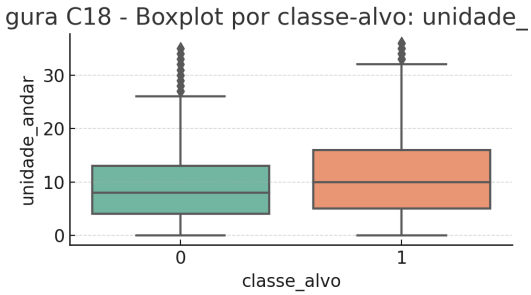
Fonte: O autor (2025)

Teste t de Student: t = -13.070, p = 0.00000

Teste Mann-Whitney U: U = 12736563.500, p = 0.00000

Figura H.9 - Frequência e distribuição por classe-alvo do atributo que indica que indica o pavimento (andar) da unidade





Fonte: O autor (2025)

Atributo: empreendimento_qtde_total_unidades

Tabela H.10 - Estatísticas descritivas do atributo que indica a quantidade total de unidades no empreendimento

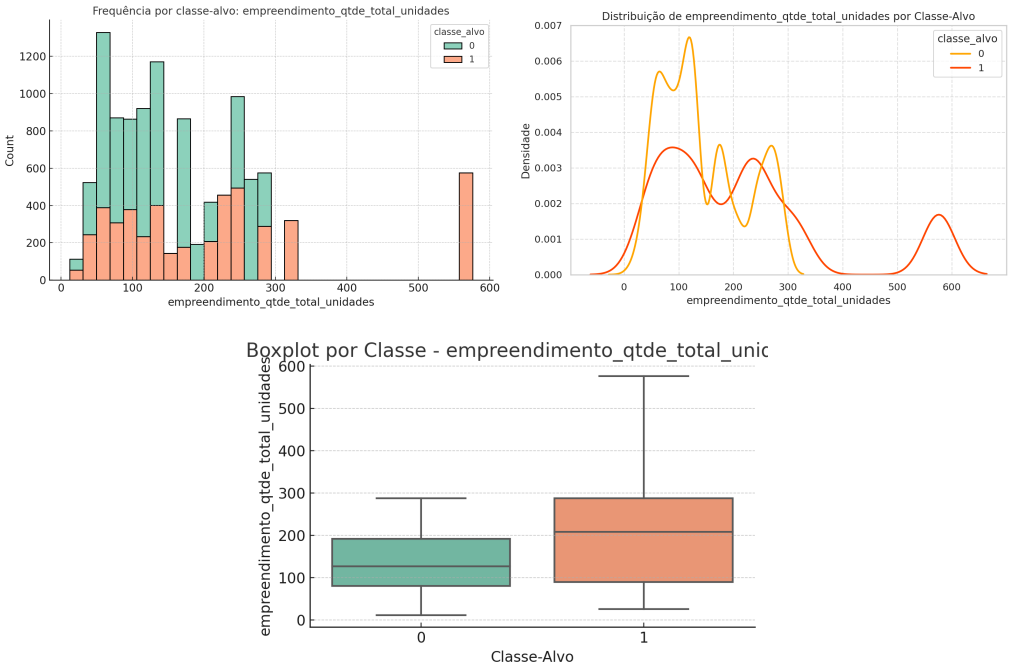
Estatística	Classe 0	Classe 1
count	6193.0	4669.0
mean	143.27	216.74
std	76.53	158.36
min	12.0	26.0
25%	81.0	90.0
50%	127.0	208.0
75%	192.0	288.0
max	288.0	576.0

Fonte: O autor (2025)

Teste t de Student: t = -29.230, p = 0.00000

Teste Mann-Whitney U: U = 10966438.500, p = 0.00000

Figura H.10 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade total de unidades no empreendimento



Fonte: O autor (2025)

Atributo: empreendimento_pavimentos

Tabela H.11 - Estatísticas descritivas do atributo que indica a quantidade de pavimentos (andares) do empreendimento

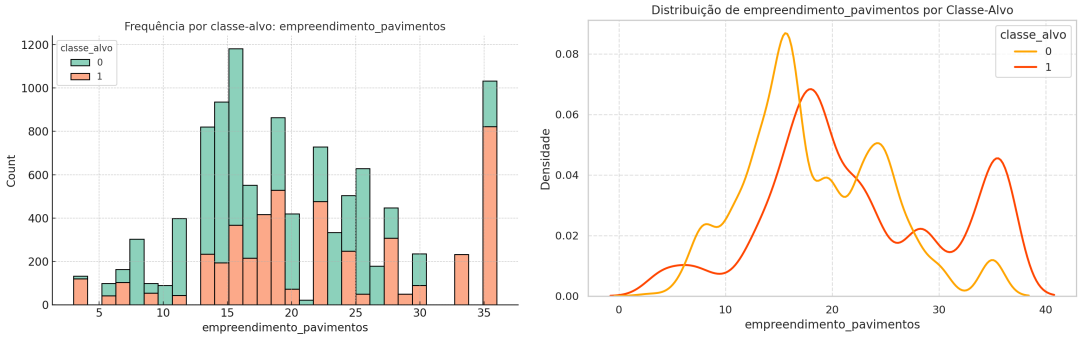
Estatística	Classe 0	Classe 1
count	6193.0	4669.0
mean	18.38	22.35
std	6.53	8.65
min	3.0	4.0
25%	14.0	17.0
50%	17.0	20.0
75%	23.0	29.0
max	35.0	36.0

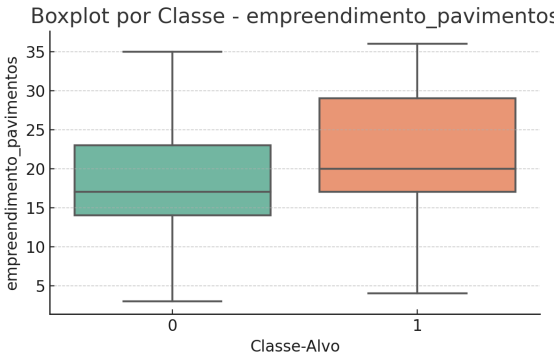
Fonte: O autor (2025)

Teste t de Student: t = -26.189, p = 0.00000

Teste Mann-Whitney U: U = 10427493.500, p = 0.00000

Figura H.11 - Frequência e distribuição por classe-alvo do atributo que indica a quantidade de pavimentos (andares) do empreendimento





Fonte: O autor (2025)

Atributo: empreendimento_numero_meses_comercializacao

Tabela H.12 - Estatísticas descritivas do atributo que indica o número total de meses para a comercialização total do empreendimento

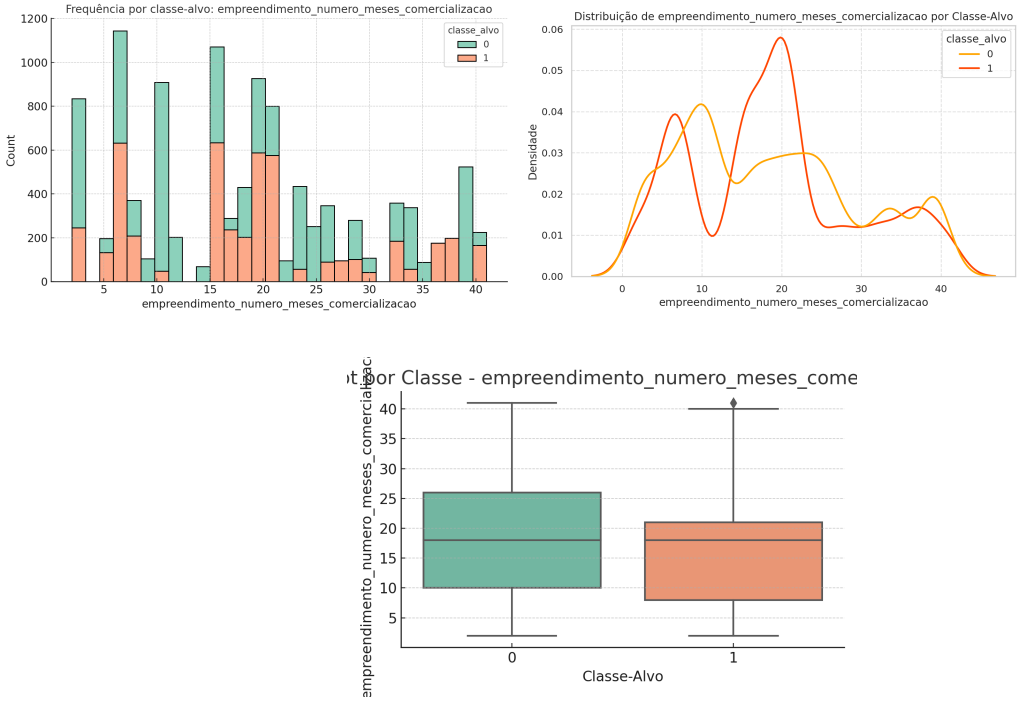
Estatística	Classe 0	Classe 1
count	6193.0	4669.0
mean	18.59	18.73
std	11.1	10.47
min	2.0	2.0
25%	10.0	8.0
50%	18.0	18.0
75%	26.0	21.0
max	41.0	41.0

Fonte: O autor (2025)

Teste t de Student: t = -0.655, p = 0.51252

Teste Mann-Whitney U: U = 14480388.500, p = 0.88768

Figura H.12 - Frequência e distribuição por classe-alvo do atributo que indica o número total de meses para a comercialização total do empreendimento



Fonte: O autor (2025)

APÊNDICE I - DECISION TREE
MÉTRICAS E GRÁFICOS DE AVALIAÇÃO DO MODELO

1. HIPÓTESE 1: VELOCIDADE DE VENDAS

1.1 MÉTRICAS DE AVALIAÇÃO

AUC-ROC: 0.9623192789304942

F1: 0.885451197053407

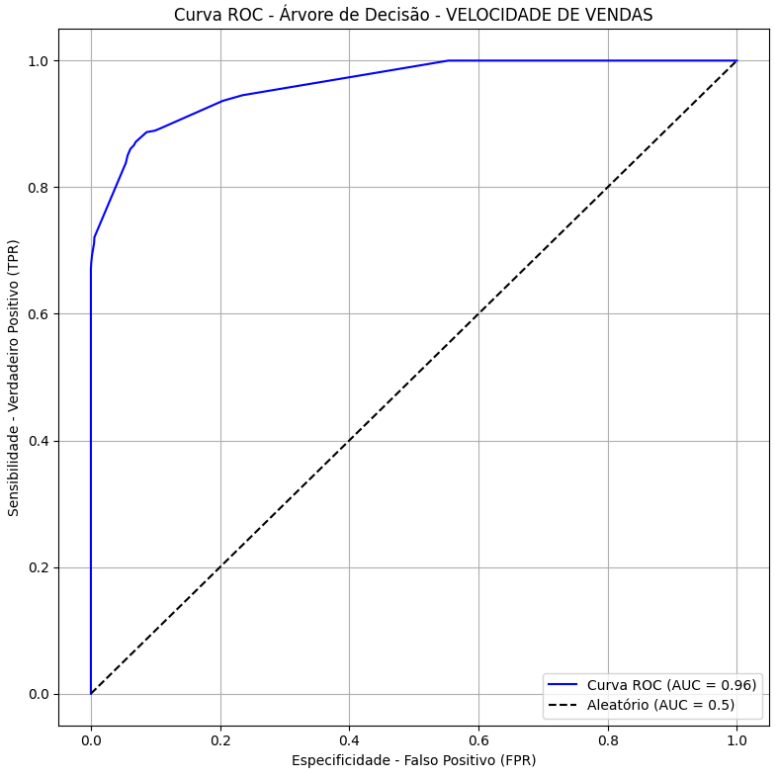
Acurácia: 0.9023547880690738

Precisão: 0.8838235294117647

Revocação (Recall): 0.8870848708487085

1.2 GRÁFICO AUC-ROC

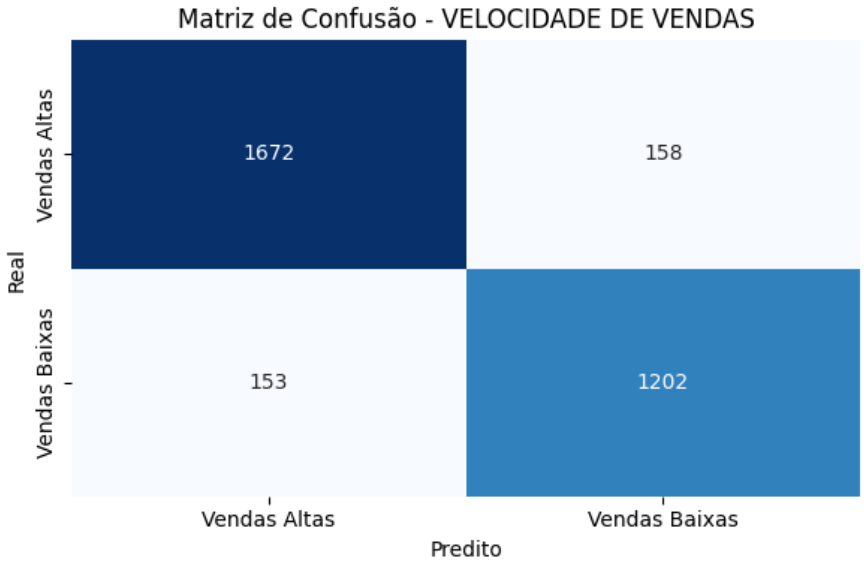
Figura I.1 - Curva ROC da árvore de decisão - Velocidade de vendas



Fonte: O autor

1.3 MATRIZ DE CONFUSÃO

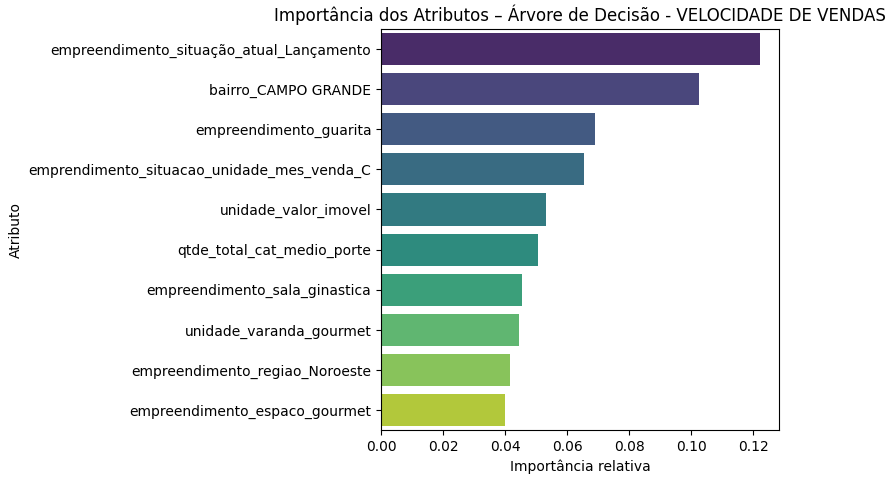
Figura I.2 - Matriz de confusão da árvore de decisão - Velocidade de vendas



Fonte: O autor

1.4 IMPORTÂNCIA DOS ATRIBUTOS

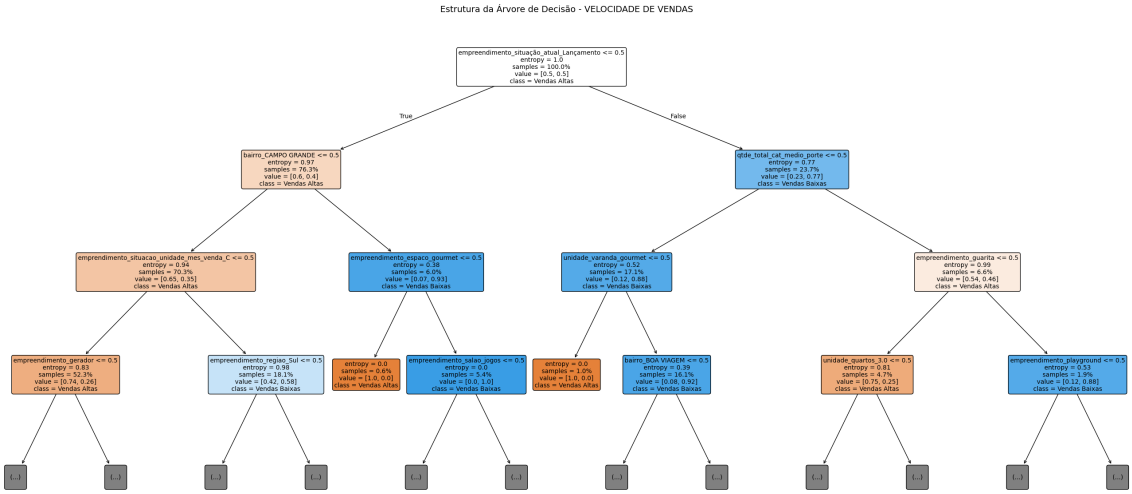
Figura I.3 - Importância dos atributos da árvore de decisão - Velocidade de vendas



Fonte: O autor

1.5 ÁRVORE DE DECISÃO

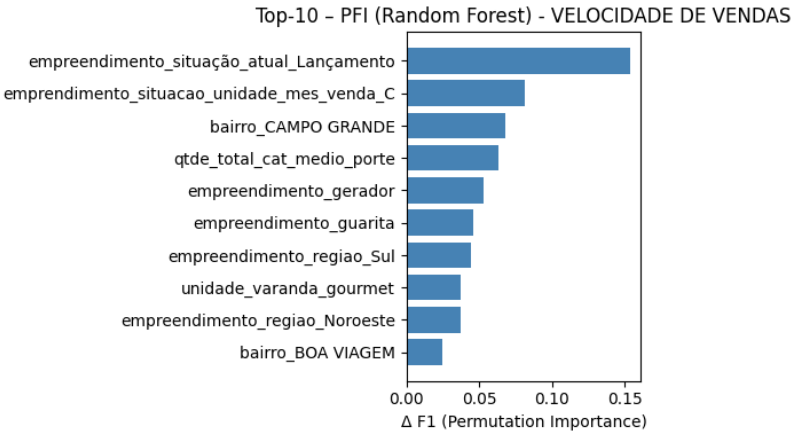
Figura I.4 - Plot da árvore de decisão - Velocidade de vendas



Fonte: O autor

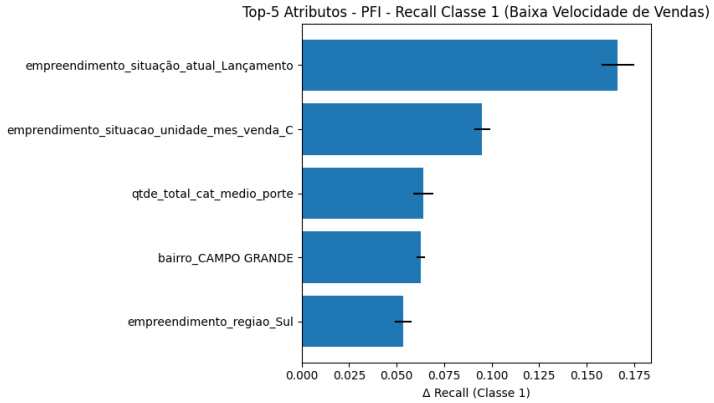
1.6 PFI

Figura I.5 - PFI Global (AUC) da árvore de decisão – Velocidade de Vendas



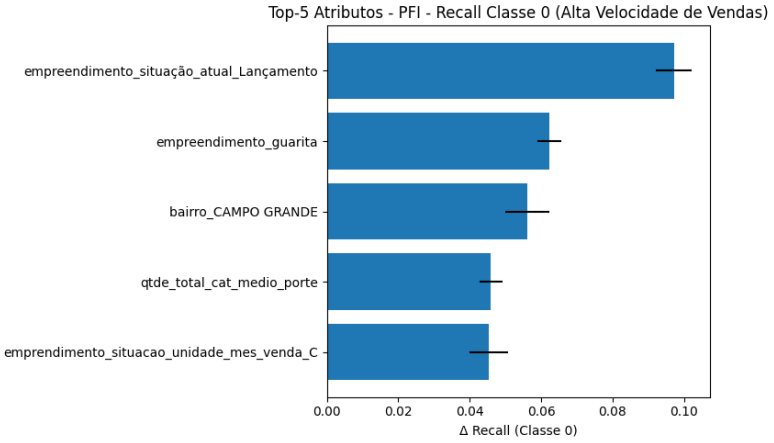
Fonte: O autor

Figura I.6 - PFI Recall Classe 1 da árvore de decisão – Velocidade de Vendas



Fonte: O autor

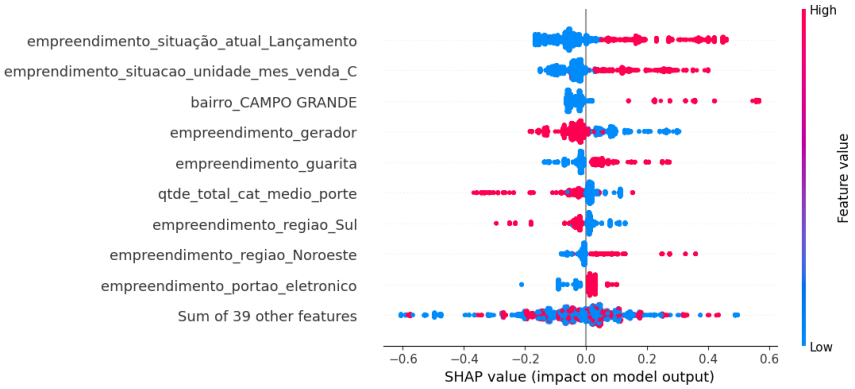
Figura I.7 - PFI Recall Classe 0 da árvore de decisão – Velocidade de Vendas



Fonte: O autor

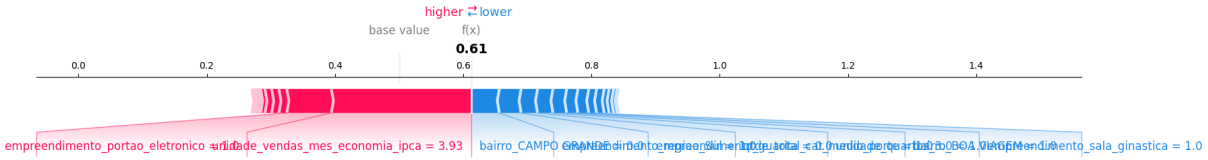
1.7 SHAP

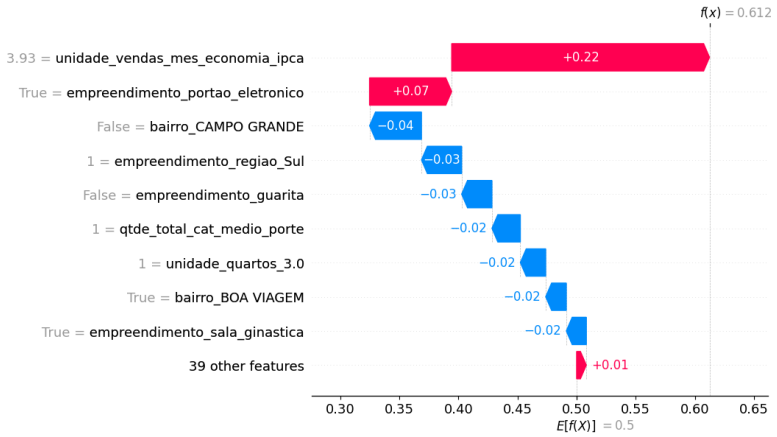
Figura I.8 - Summary Plot (SHAP values) – Velocidade de Vendas



Fonte: O autor

Figura I.9 - Force Plot (SHAP) – Instância com alta probabilidade de velocidade





Fonte: O autor

1.8 LIME

Figura I.10 - LIME: explicação local para a instância - velocidade de vendas

Fonte: O autor

2. HIPÓTESE 2: RESILIÊNCIA DE VENDAS

2.1 MÉTRICAS DE AVALIAÇÃO

AUC-ROC: 0.9894956875911852

F1: 0.9132075471698113

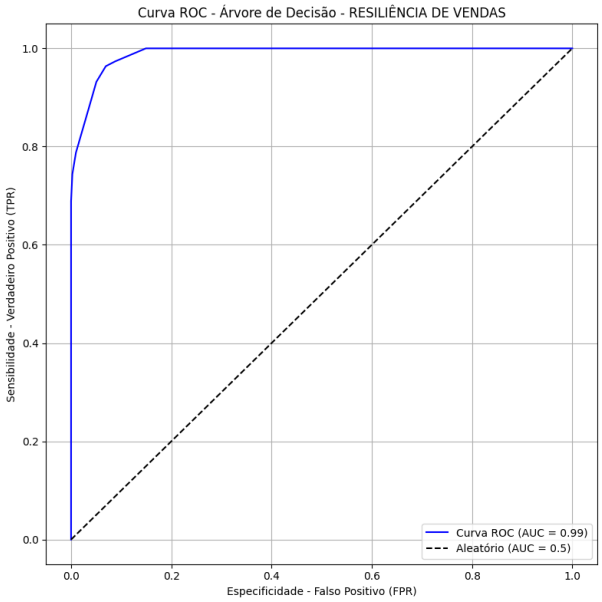
Acurácia: 0.943921978404737

Precisão: 0.8953488372093024

Revocação (Recall): 0.9317931793179318

2.2 GRÁFICO AUC-ROC

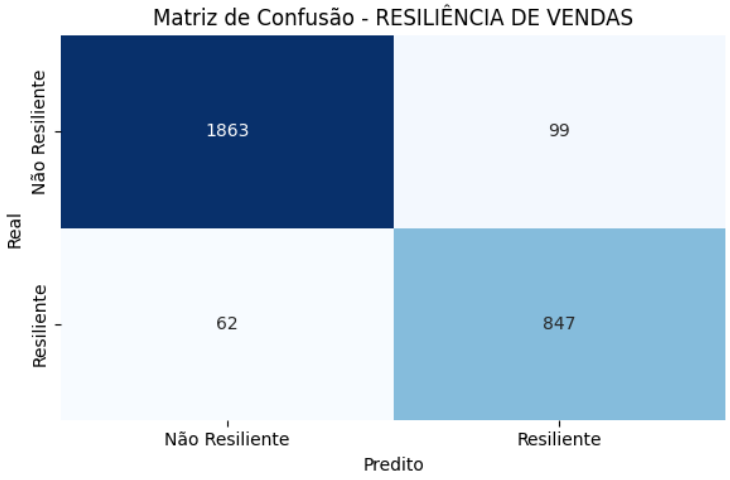
Figura I.11 - Curva ROC da árvore de decisão - Resiliência de vendas



Fonte: O autor

2.3 MATRIZ DE CONFUSÃO

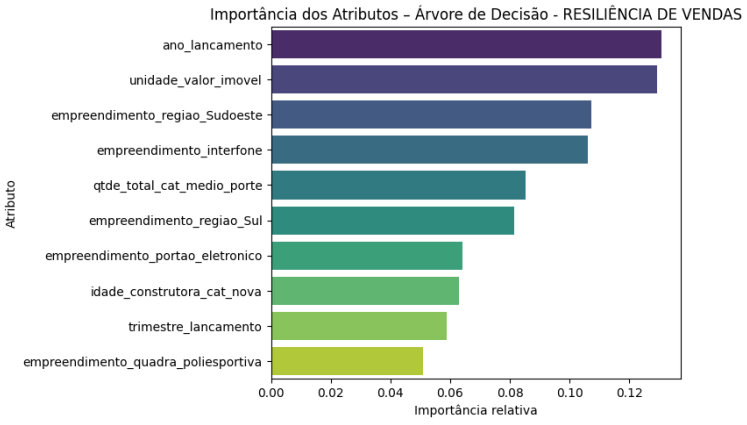
Figura I.12 - Matriz de confusão da árvore de decisão - Resiliência de vendas



Fonte: O autor

2.4 IMPORTÂNCIA DOS ATRIBUTOS

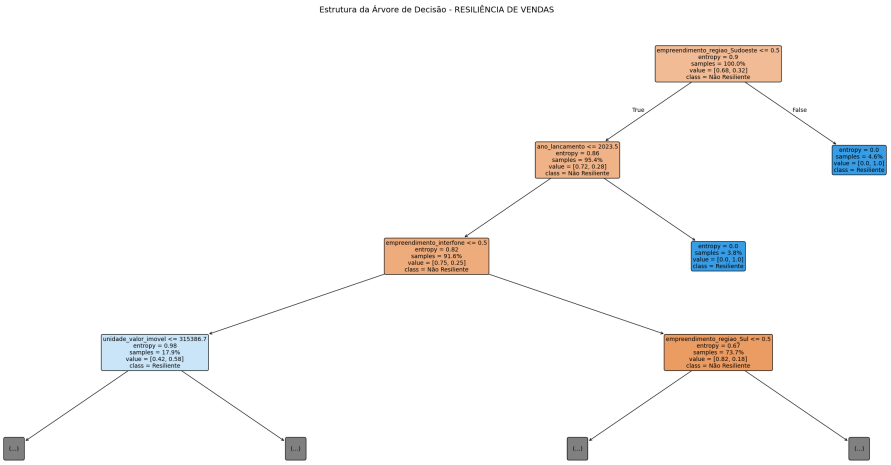
Figura I.13 - Importância dos atributos da árvore de decisão - Resiliência de vendas



Fonte: O autor

2.5 ÁRVORE DE DECISÃO

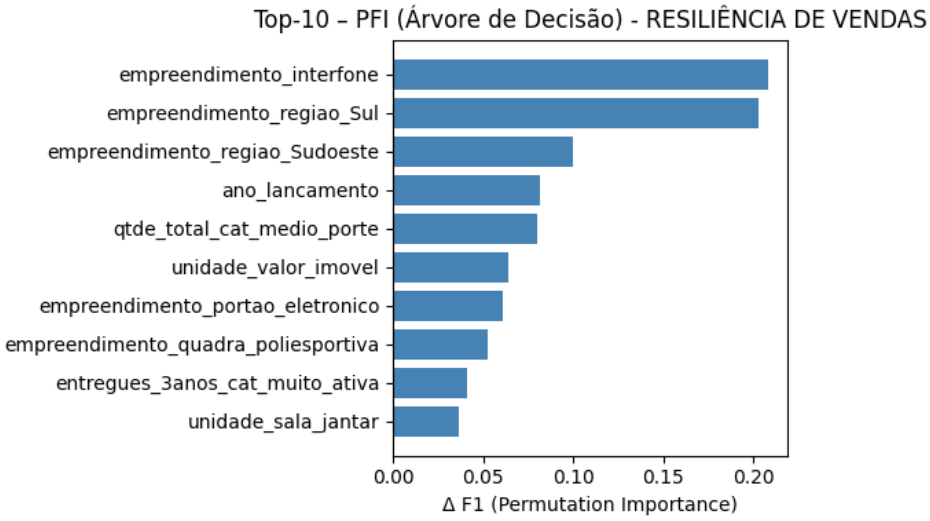
Figura I.14 - Plot da árvore de decisão - Resiliência de vendas



Fonte: O autor

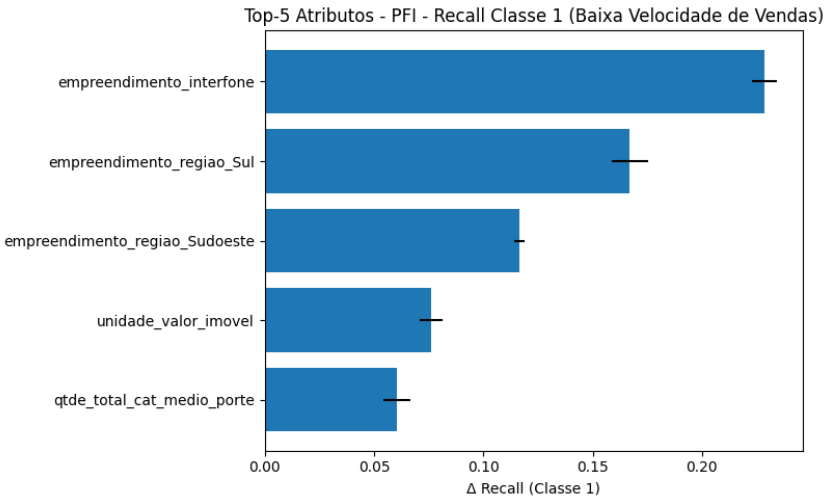
2.6 PFI

Figura I.15 - PFI Global (AUC) da árvore de decisão – Resiliência de Vendas



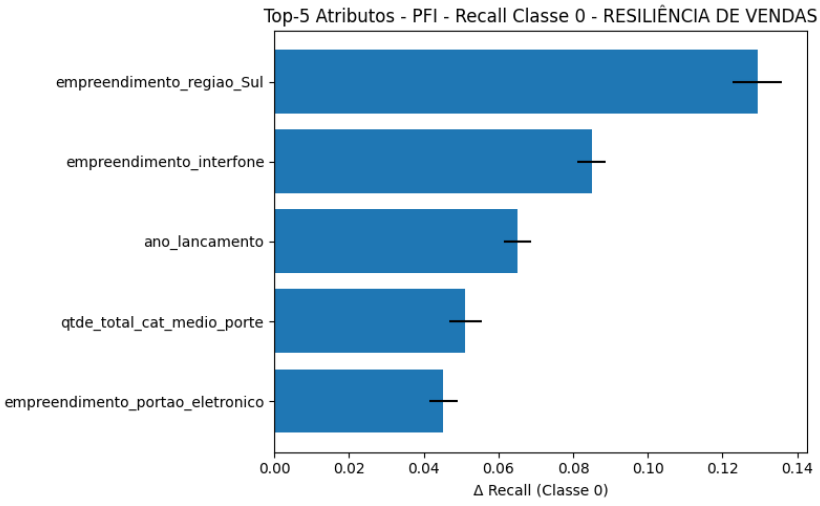
Fonte: O autor

Figura I.16 - PFI Recall Classe 1 da árvore de decisão – Resiliência de Vendas



Fonte: O autor

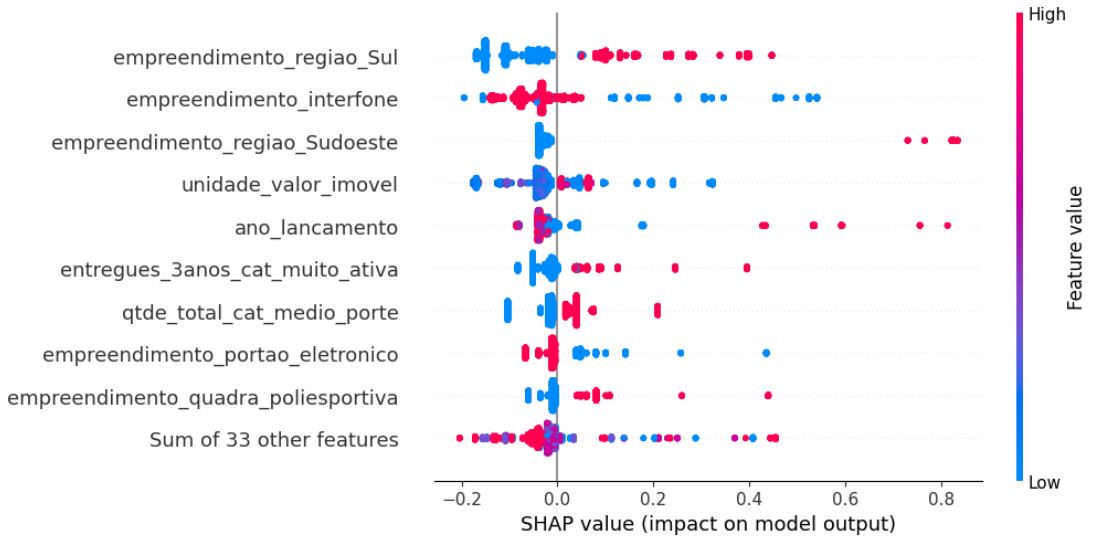
Figura I.17 - PFI Recall Classe 0 da árvore de decisão – Resiliência de Vendas



Fonte: O autor

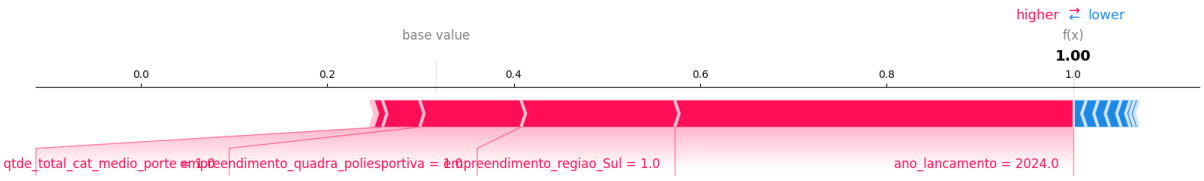
2.7 SHAP

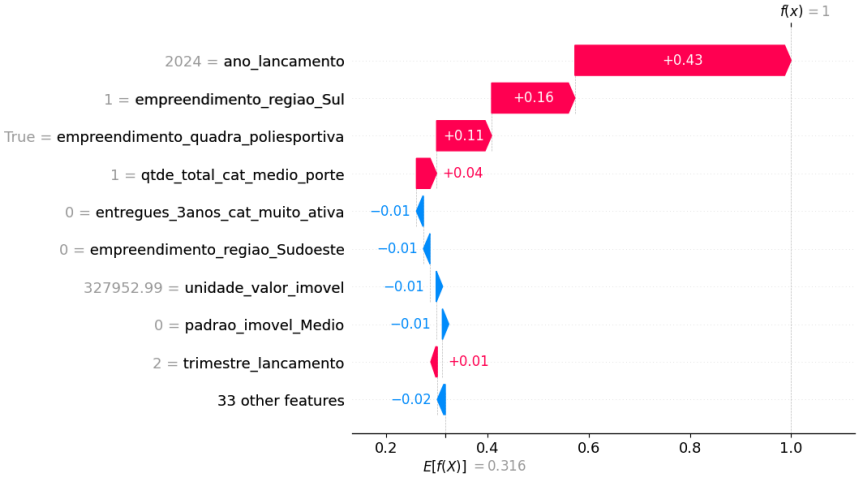
Figura I.18 - Summary Plot (SHAP values) – Resiliência de Vendas



Fonte: O autor

Figura I.19 - Force Plot (SHAP) – Instância com alta probabilidade de resiliência

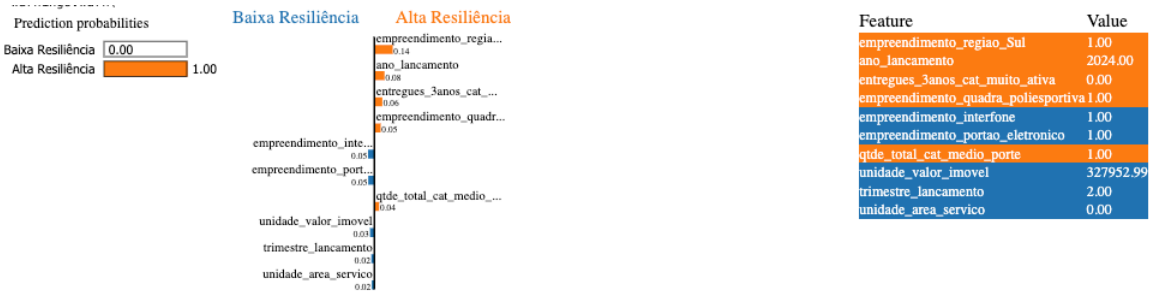




Fonte: O autor

2.8 LIME

Figura I.20 - LIME: explicação local para a instância - velocidade de vendas



Fonte: O autor

APÊNDICE J - LOGISTIC REGRESSION
MÉTRICAS E GRÁFICOS DE AVALIAÇÃO DO MODELO

1. HIPÓTESE 1: VELOCIDADE DE VENDAS

1.1 MÉTRICAS DE AVALIAÇÃO

AUC-ROC: 0.9758252172685662

F1: 0.9199565374864179

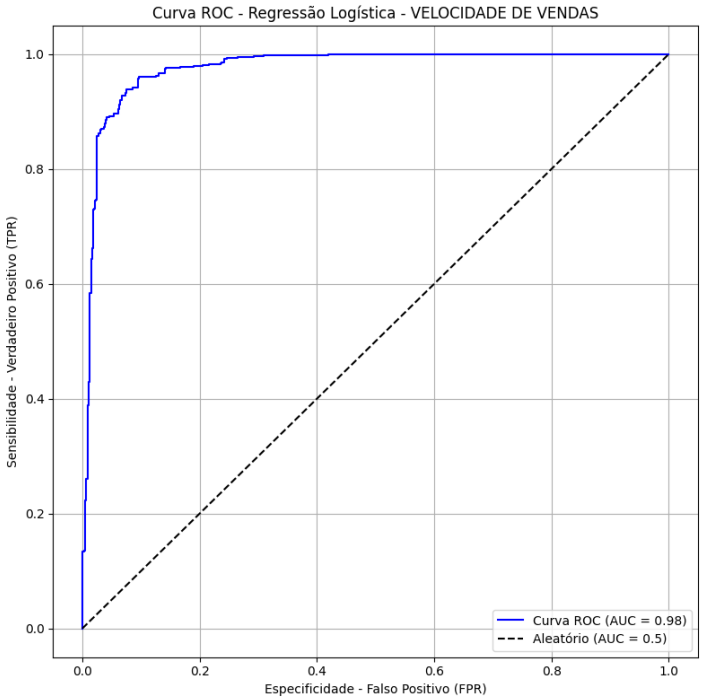
Acurácia: 0.9306122448979591

Precisão: 0.903271692745377

Revocação (Recall): 0.9372693726937269

1.2 GRÁFICO AUC-ROC

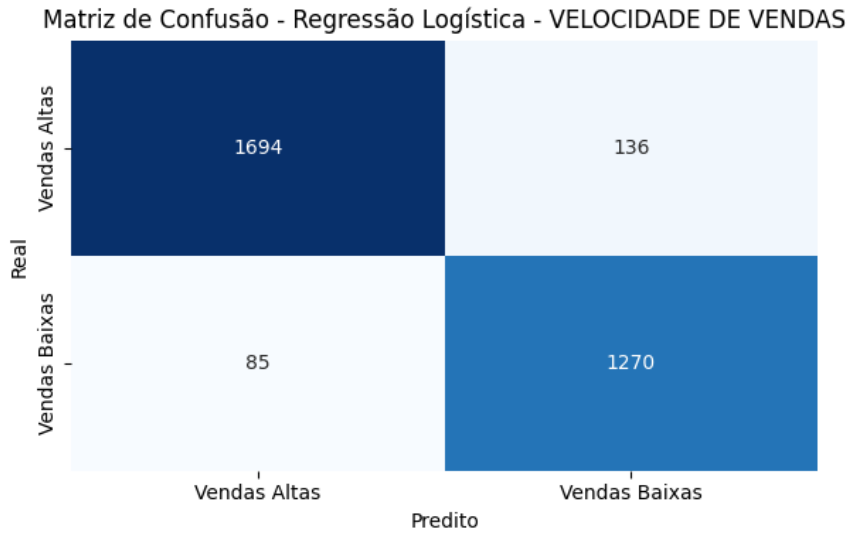
Figura J.1 - Curva ROC da regressão logística - Velocidade de vendas



Fonte: O autor

1.3 MATRIZ DE CONFUSÃO

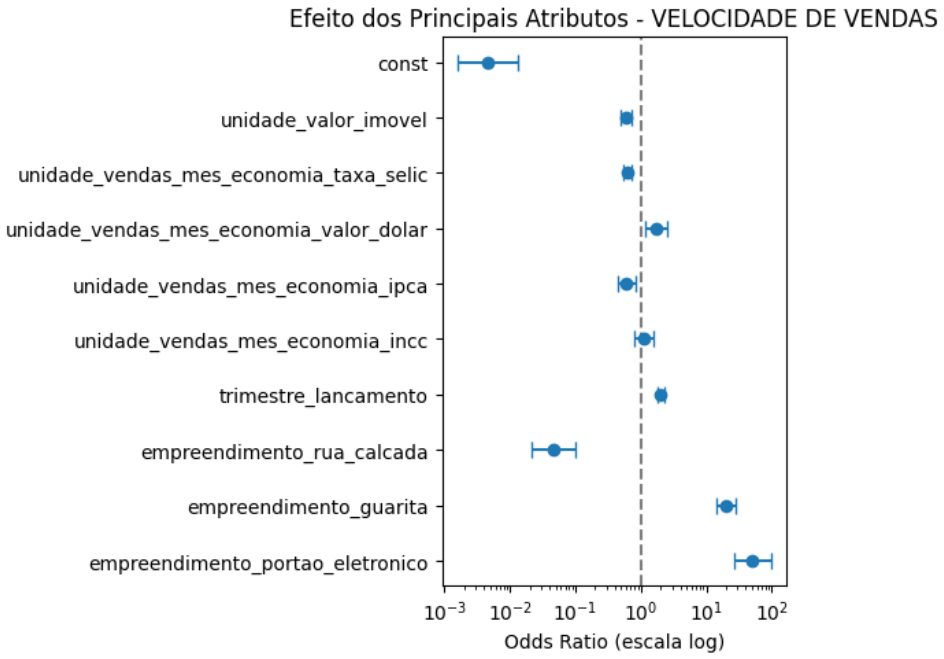
Figura J.2 - Matriz de confusão da regressão logística - Velocidade de vendas



Fonte: O autor

1.4 IMPORTÂNCIA DOS ATRIBUTOS

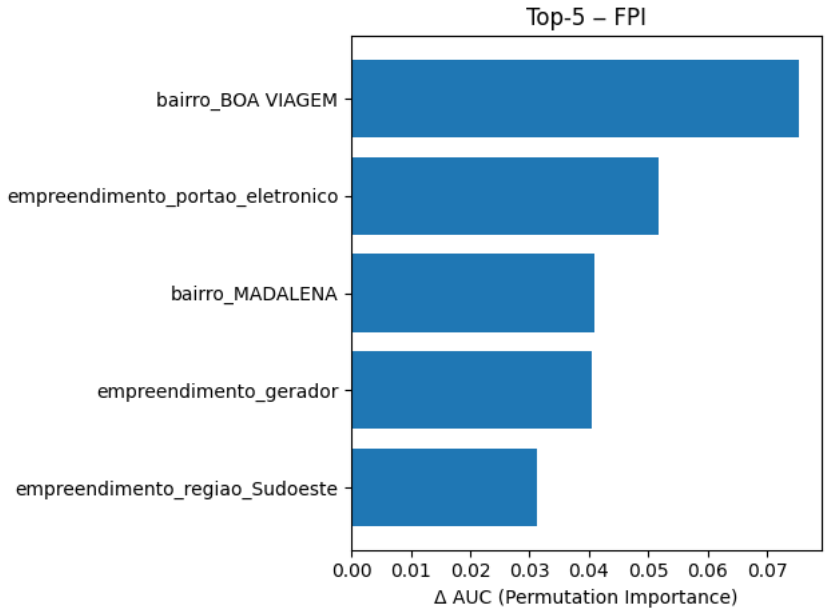
Figura J.3 - Importância dos atributos da regressão logística - Velocidade de vendas



Fonte: O autor

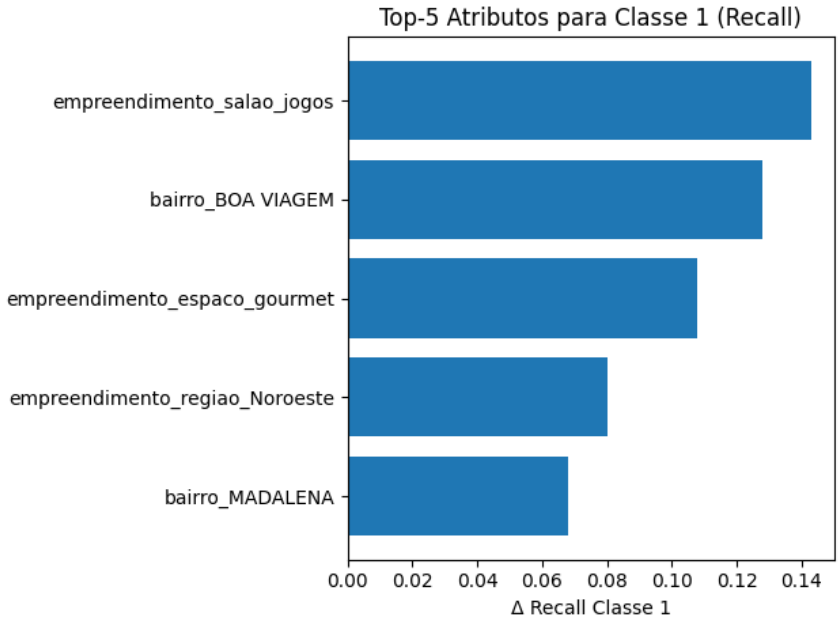
1.5 PFI

Figura J.4 - PFI Global (AUC) da regressão logística – Velocidade de Vendas



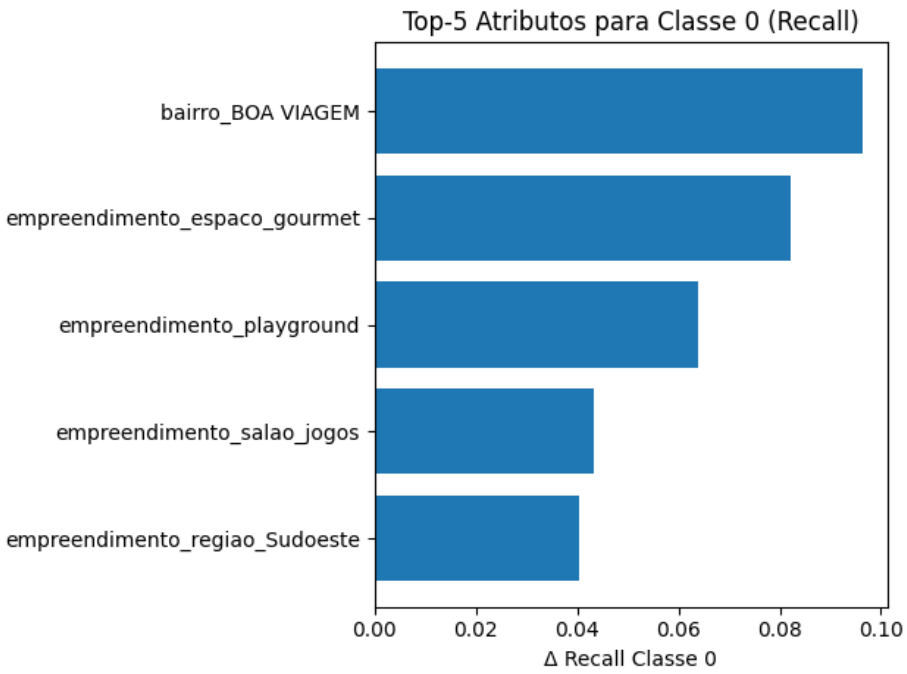
Fonte: O autor

Figura J.5 - PFI Recall Classe 1 da regressão logística – Velocidade de Vendas



Fonte: O autor

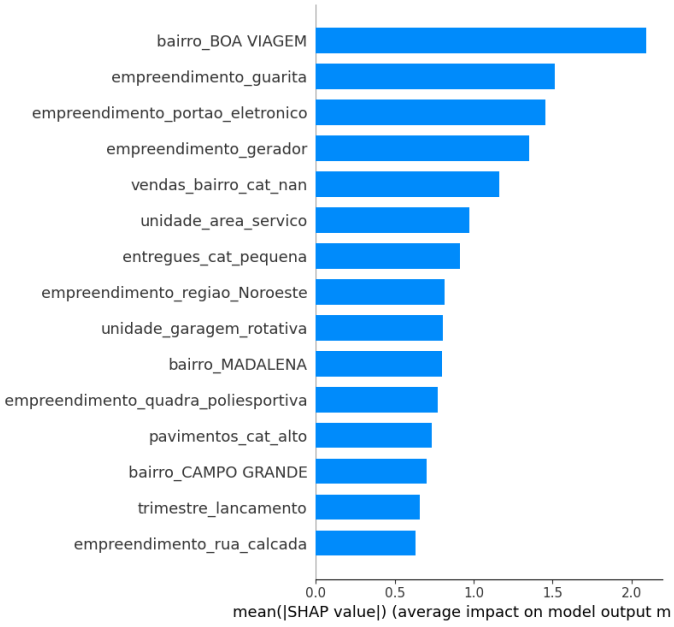
Figura J.6 - PFI Recall Classe 0 da regressão logística – Velocidade de Vendas



Fonte: O autor

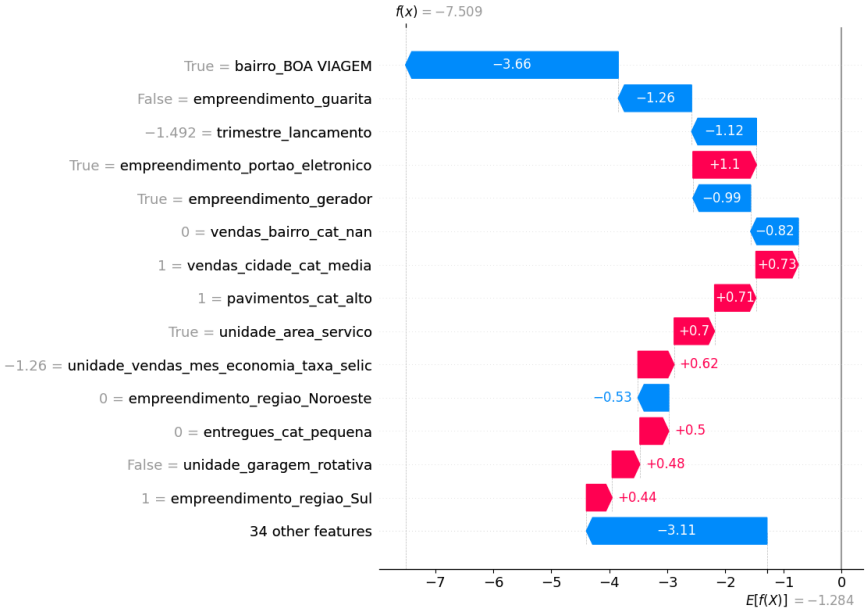
1.6 SHAP

Figura J.7 - Plot (SHAP values) Importância dos atributos – Velocidade de Vendas



Fonte: O autor

Figura J.8 - Force Plot (SHAP) – Instância com alta probabilidade de velocidade



Fonte: O autor

1.7 LIME

Figura J.9 - LIME: explicação local para a instância - Velocidade de vendas

Fonte: O autor

2. HIPÓTESE 2: RESILIÊNCIA DE VENDAS

2.1 MÉTRICAS DE AVALIAÇÃO

AUC-ROC: 0.9913936857498187

F1: 0.9460758142018153

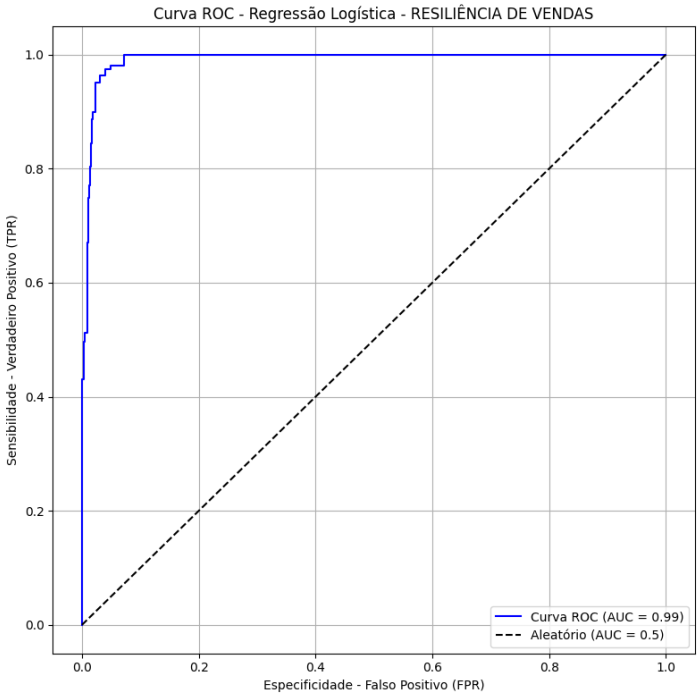
Acurácia: 0.9648206199930338

Precisão: 0.9190871369294605

Revocação (Recall): 0.9746974697469747

2.2 GRÁFICO AUC-ROC

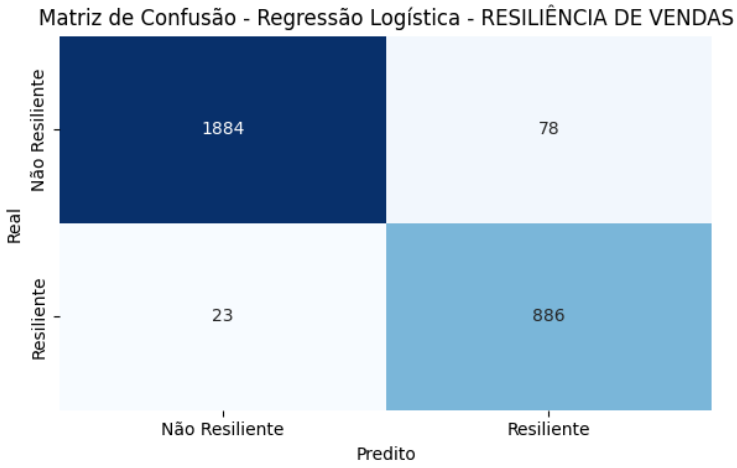
Figura J.10 - Curva ROC da regressão logística - Resiliência de vendas



Fonte: O autor

2.3 MATRIZ DE CONFUSÃO

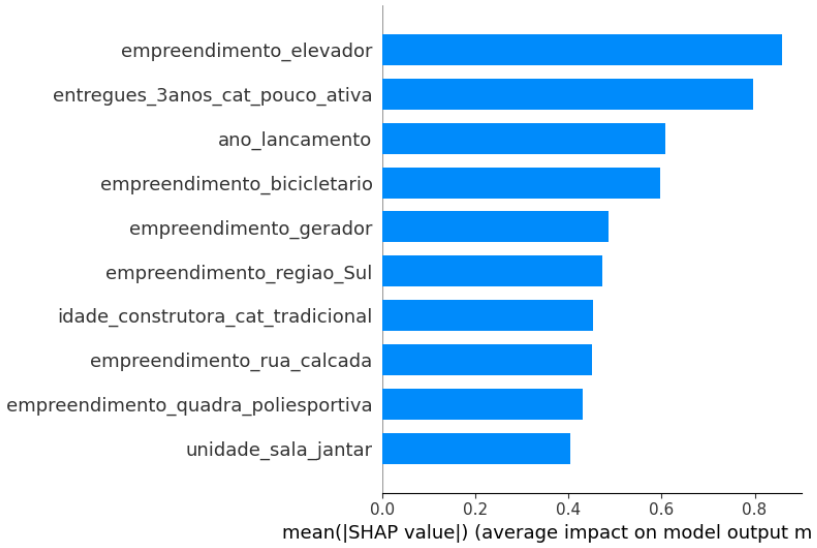
Figura J.11 - Matriz de confusão da regressão logística - Resiliência de vendas



Fonte: O autor

2.4 IMPORTÂNCIA DOS ATRIBUTOS

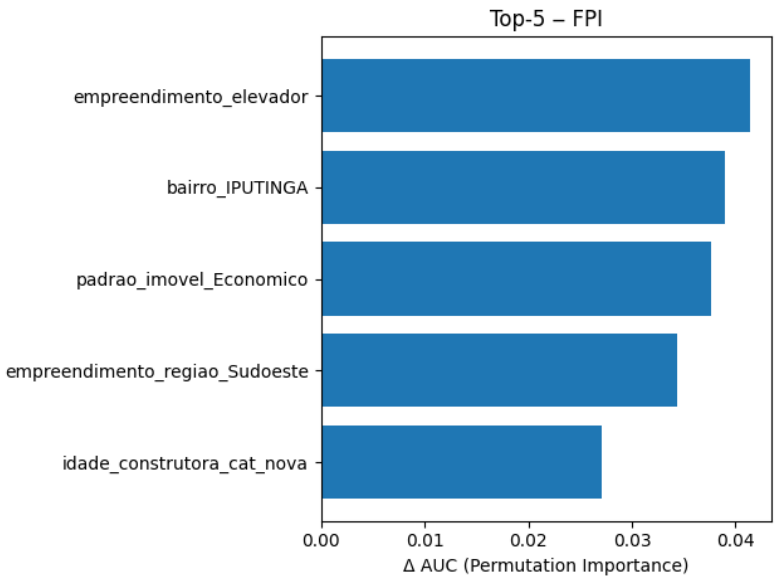
Figura J.12 - Importância dos atributos da regressão logística - Resiliência de vendas



Fonte: O autor

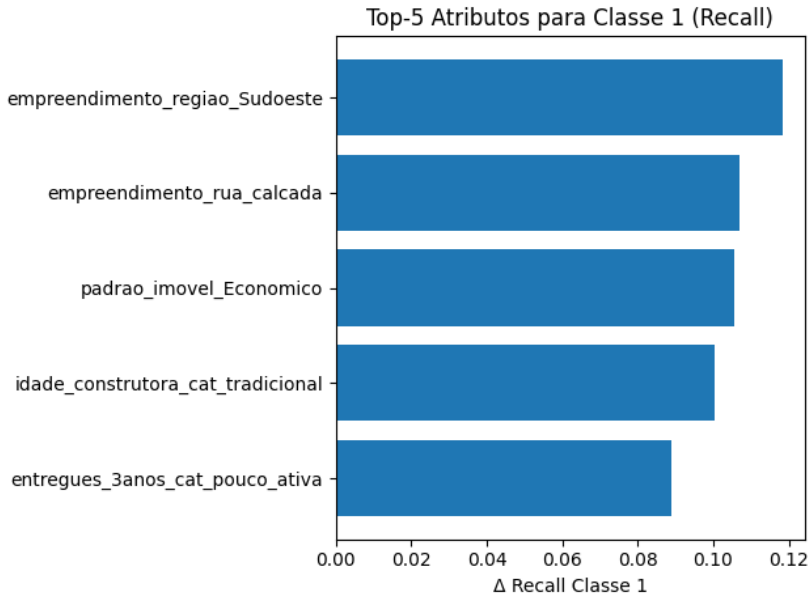
2.5 PFI

Figura J.13 - PFI Global (AUC) da regressão logística – Resiliência de Vendas



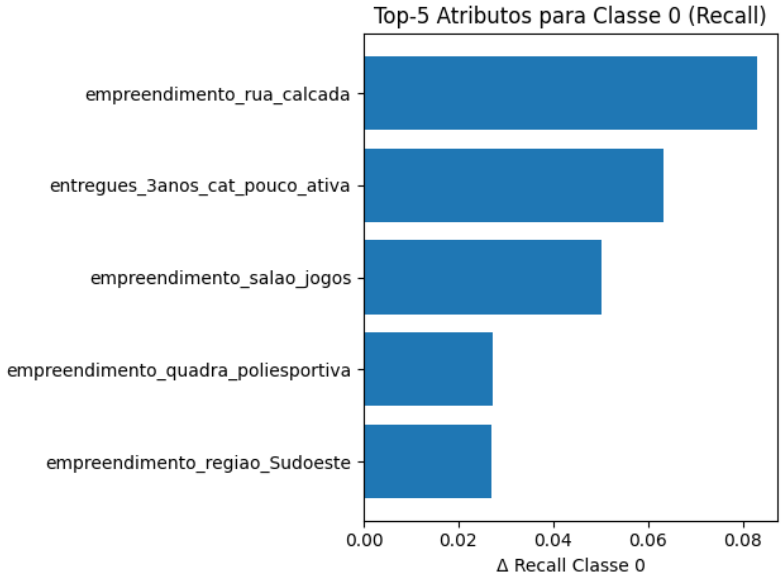
Fonte: O autor

Figura J.14 - PFI Recall Classe 1 da regressão logística – Resiliência de Vendas



Fonte: O autor

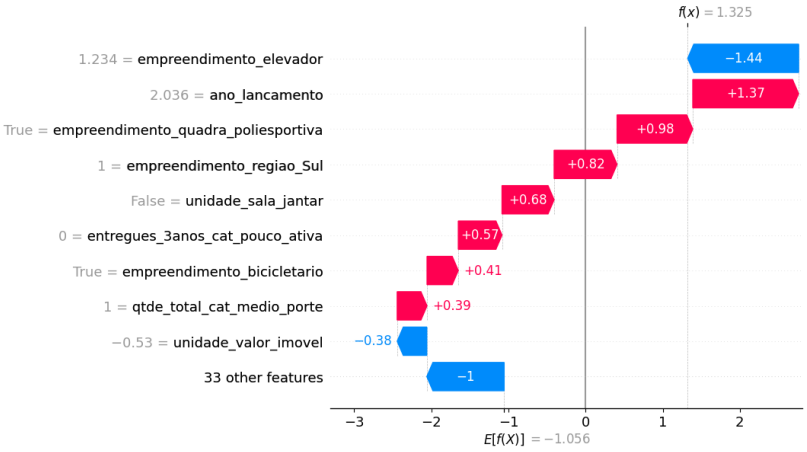
Figura J.15 - PFI Recall Classe 0 da regressão logística – Resiliência de Vendas



Fonte: O autor

2.6 SHAP

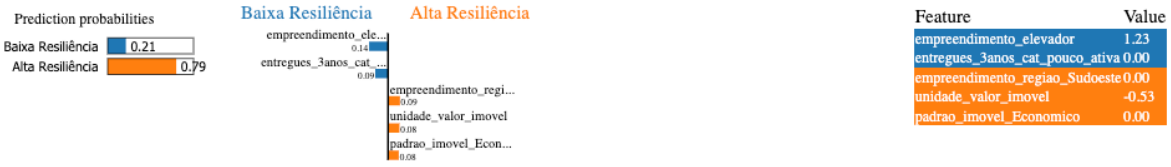
Figura J.16 - Force Plot (SHAP) – Instância com alta probabilidade de resiliência



Fonte: O autor

2.7 LIME

Figura J.17 - LIME: explicação local para a instância - Resiliência de vendas



Fonte: O autor