ANA ALICE PEREGRINO PINTO

Uma Análise Comparativa de Modelos para a Predição de Fluxos Pendulares para Planejamento Urbano

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2024

ANA ALICE PEREGRINO PINTO

Uma Análise Comparativa de Modelos para a Predição de Fluxos Pendulares para Planejamento Urbano

Dissertação 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 Ciências da Computação.

Área de Concentração: Inteligência Computacional.

Orientador: Prof. Dr. Nivan Roberto Ferreira Júnior

Recife
2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Pinto, Ana Alice Peregrino.
Uma análise comparativa de modelos para a predição de fluxos pendulares para planejamento urbano / Ana Alice Peregrino Pinto. - Recife, 2024.
79f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2024.
Orientação: Nivan Roberto Ferreira Júnior.
Inclui referências.

1. Mobilidade urbana; 2. Previsão de fluxos; 3. Redes neurais para grafos. I. Ferreira Júnior, Nivan Roberto. II. Título.

UFPE-Biblioteca Central

AGRADECIMENTOS

Agradeço imensamente à minha família, cujo apoio incondicional foi fundamental para que eu chegasse até aqui. Ao meu noivo, por toda a paciência e suporte nas fases mais difíceis da escrita, e por ter tentado me ajudar a levar esse capítulo da minha vida com mais leveza. Aos meus amigos que continuaram insistindo em me ver mesmo depois de tantas respostas parecidas com "não vou, preciso trabalhar nas atividades do mestrado"que foram ouvidas nos últimos anos.

Também não poderia deixar de agradecer ao meu orientador, professor Nivan, por toda a paciência e dedicação na orientação durante a pesquisa. E por fim, ao Centro de Informática, pela oportunidade de ingressar num programa de pós-graduação de excelência. Me sinto muito privilegiada por ter vivido esta experiência, e sei que irei finalizar este ciclo com conhecimentos valiosos que irão me acompanhar por toda minha jornada profissional.

RESUMO

Com o crescimento desenfreado dos centros urbanos, o planejamento urbano se tornou imprescindível na criação e gestão de cidades a fim de melhorar a qualidade de vida de seus habitantes. Mobilidade urbana é uma das frentes desse planejamento de grande importância em grandes cidades. Problemas de mobilidade estão associados a questões como mal uso de recursos materiais e de tempo de deslocamento das pessoas, além da emissão elevada de gases de efeito estufa. Fluxos pendulares, em particular, que são os fluxos entre casa e trabalho feitos diariamente, são bastante estudados por sua importância para a qualidade de vida dos cidadãos. Neste contexto, técnicas de modelagem de fluxos de deslocamento são ferramentas importantes para auxiliar os tomadores de decisão na definição das melhores estratégias de planejamento. Tais ferramentas são importantes não só para entender os atuais padrões nas cidades, mas também para realizar planejamento de cenários, que consiste em tentar antever como possíveis ações irão impactar a mobilidade urbana. Com o crescimento da disponibilização de dados urbanos, modelos baseados em técnicas de aprendizagem de máquina têm se destacado por seu desempenho em prever fluxos de mobilidade. O objetivo do presente trabalho é realizar uma análise comparativa dos principais modelos baseados em aprendizagem de máquina propostos para previsão de fluxos pendulares, com foco em tarefas relacionadas ao planejamento de cenários. Os modelos selecionados são divididos entre abordagens lineares, modelos baseados em árvore e redes neurais. Para realizar a análise foram usadas métricas de desempenho utilizadas na literatura para o tipo de problema estudado, técnicas de interpretabilidade das saídas dos modelos, além de cenários de uso que simulam o planejamento de cenários. Os resultados indicam um desempenho superior dos modelos baseados em redes neurais para grafos, mas também mostram padrões interessantes para modelos menos complexos, o que poderia tornar esses modelos competitivos como uma ferramenta de planejamento de cenários, devido ao seu menor tempo de treinamento e simplicidade na implementação.

Palavras-chave: Mobilidade urbana. Previsão de fluxos. Redes neurais para grafos.

ABSTRACT

With the unbridled growing of urban centers, urban planning has become essential in the creation and management of cities in order to improve the quality of life of their inhabitants. Urban mobility is one of the fronts of this planning of great importance in large cities. Mobility problems are associated with issues such as misuse of material resources and people's travel time, in addition to high greenhouse gas emissions. Commuting flows, in particular, which are flows between home and work carried out daily, are extensively studied due to their importance for citizens' quality of life. In this context, displacement flow modeling techniques are important tools to assist decision makers in defining the best planning strategies. Such tools are important not only for understanding current patterns in cities, but also for carrying out scenario planning, which consists of trying to predict how possible actions will impact urban mobility. With the growth in the availability of urban data, models based on machine learning techniques have stood out for their performance in predicting mobility flows. The objective of the present work is to carry out a comparative analysis of the main machine learning-based models proposed for predicting commuting flows, focusing on tasks related to scenario planning. The selected models are divided between linear approaches, tree-based models and neural networks. To carry out the analysis, performance metrics were used in the literature for the type of problem studied, interpretability techniques for model outputs, as well as usage scenarios that simulate scenario planning. The results indicate superior performance of models based on graph neural networks, but also show interesting patterns for less complex models, which could make these models competitive as a scenario planning tool, due to their shorter training time and simplicity in implementation.

Key words: Urban mobility. Flow prediction. Graph neural networks.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As nações unidas projetaram, em 2018, que em 2050 duas de cada três pessoas estarão vivendo em cidades ou outros centros urbanos, conforme publicado em (Department of Economic and Social Affairs, United Nations, 2018). Considerando esse cenário, é inegável que o investimento em um planejamento urbano eficiente e robusto já vem se estabelecendo nas últimas décadas como uma ferramenta mantenedora do bem estar populacional. Dentre as frentes que o planejamento urbano precisa equilibrar e organizar, a mobilidade urbana chama a atenção pelas responsabilidades que carrega consigo. O relatório de 2023 do Painel Intergovernamental de Mudanças Climáticas (*IPCC*, na sigla em inglês) reporta dados de 2019 sobre a porcentagem de emissão de gases de efeito estufa entre os setores de maior impacto econômico, e o setor de transportes é citado como responsável por 15% dessas emissões (IPCC et al., 2023). O aumento de modais de transporte nas ruas que emitem esses gases, além dos congestionamentos gerados por eles, são fatores que tem potencial para aumentar esta estatística, o que faz com que seja fundamental pensar em estratégias inteligentes de desenvolvimento da mobilidade urbana. Um outro ponto de grande importância surge ao considerar que uma parcela significativa do tempo que um indivíduo passa em deslocamento (PERO; STEFANELLI, 2015) (ou seja, sem efetivamente engajar em alguma atividade, seja ela profissional, de lazer ou descanso) é também fruto de um projeto de mobilidade mais ou menos eficaz para a população, e essa parcela tende a crescer conforme a cidade se expande e soluções inteligentes de tráfego e transporte não acompanhem esse crescimento.

Uma alternativa para apoiar os tomadores de decisão na proposta de políticas eficientes de mobilidade utiliza o conceito de planejamento de cenários, que visa antecipar os possíveis desfechos futuros de potenciais decisões estratégicas tomadas no presente. No contexto dos fluxos pendulares (que são os fluxos realizados diariamente entre casa e trabalho), esse conceito pode ser aplicado utilizando ferramentas capazes de predizer a distribuição de fluxos realizando mudanças no fluxo de trânsito, por exemplo, ou construindo um novo condomínio na cidade.

Existem diversos modelos propostos na literatura que são capazes de realizar a predição desses fluxos pendulares, como mostrado nos trabalhos de (SIMINI et al., 2021; YIN et al., 2023; FRANCETIC; MUNFORD, 2021; LIU et al., 2020), e que poderiam, portanto, funcionar como candidados a ferramentas de planejamento de cenários. Com o objetivo de aprofundar o entendimento sobre os resultados gerados pelos modelos, uma proposta de estudo é analisar comparativamente os modelos a partir de performance, capacidade de generalização, dentre outros resultados que forneçam informações pertinentes que auxiliem na escolha do modelo mais apropriado para a predição dos fluxos. Este tipo de análise é reportado na literatura em problemas relacionados a área da saúde, como no trabalho

de (ANTOR et al., 2021b), engenharia, no trabalho de (NANEHKARAN et al., 2023), e, mais próximo ao conteúdo deste trabalho, aplicações voltadas a uma gestão urbana eficiente, como reportado em (AMEER et al., 2019).

Dentro deste contexto, este trabalho se propõe a trazer uma contribuição à área de planejamento de mobilidade urbana estendendo o trabalho realizado em (LIU et al., 2020) e realizando uma análise comparativa de modelos de aprendizado de máquina encontrados na literatura com o objetivo de predizer fluxos de deslocamento, com foco no deslocamento pendular, que podem ser deslocamentos entre cidades, estados ou países. Neste estudo, o foco será no deslocamentos dentro de uma mesma cidade. A motivação para pesquisar neste tema vem da premissa de que, munidos de informação sobre como as distribuições desses fluxos se comportam numa determinada área, seja possível prover ferramentas de apoio à decisão que agentes governamentais possam utilizar para tomar decisões mais estratégicas no que diz respeito a melhorias relacionadas a mobilidade numa determinada cidade.

Os modelos selecionados para este estudo são treinados em uma base de dados coletada e processada de acordo com a metodologia disponível nas seção 3, e então, avaliados utilizando métricas e ferramentas propostas na literatura para o problema abordado. Em seguida, é realizada uma análise de dados nos resultados obtidos com o objetivo de adquirir informações relevantes para atingir os objetivos da pesquisa.

1.1 OBJETIVOS

O objetivo geral deste trabalho é realizar uma análise comparativa de modelos de aprendizagem de máquina encontrados na literatura na tarefa de previsão de fluxos pendulares. Como objetivos específicos este trabalho tem por finalidade:

- Realizar uma análise do desempenho em métricas utilizadas na literatura para o tipo de problema estudado.
- Realizar uma análise das predições dos modelos sob a perspectiva da interpretabilidade.
- Realizar análise voltada ao planejamento de cenários futuros, onde são realizadas mudanças nos dados de infraestrutura da região com o objetivo de analisar como os modelos performam e quais os *insights* que os resultados conseguem fornecer para o problema.

Os capítulos restantes da dissertação encontram-se estruturados da seguinte forma:

2 REFERENCIAL TEÓRICO: apresenta os trabalhos realizados na literatura nas áreas de estudo que englobam esta pesquisa (planejamento de cenários urbanos, predição de fluxos pendulares e interpretabilidade de modelos de aprendizagem de máquina).

3 MATERIAIS E MÉTODOS: descreve a metodologia utilizada para atingir os objetivos gerais e específicos do trabalho.

4 AVALIAÇÃO COMPARATIVA DA PERFORMANCE DE MODELOS PARA PREDIÇÃO DE FLUXOS DE DESLOCAMENTO: descreve e discute os resultados das análises efetuadas para a avaliação comparativa entre os modelos escolhidos para a realização deste trabalho.

5 CASOS DE USO: apresenta um caso de uso onde o modelo que apresentou melhor performance nas métricas selecionadas é utilizado para obtenção de informações sobre o efeito que mudanças na infraestrutura podem causar nos fluxos de deslocamento.

6 CONCLUSÃO: discorre sobre as considerações finais, conclusões obtidas e potenciais trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 PREDIÇÃO DE FLUXOS DE DESLOCAMENTO PENDULAR

O censo norte-americano descreve os fluxos pendulares como os fluxos de pessoas que se locomovem entre casa e local de trabalho principal diariamente ((BUREAU, 2023)). Essa locomoção pode ser entre cidades, estados, ou até mesmo países. A predição de fluxos de deslocamento pendulares é uma área de estudo bastante ativa há algumas décadas. O *Gravity Model*, introduzido na década de 1940 por (ZIPF, 1946), foi o primeiro modelo proposto para a distribuição de fluxos entre dois lugares, utilizando uma analogia à lei da gravidade de Newton. O modelo foi bastante utilizado também no estudo do estabelecimento de relações comerciais, como no trabalho de (MARIMOUTOU; PEGUIN; PEGUIN-FEISSOLLE, 2009), e possui aplicação direta na área de predição de fluxos pendulares.

A proposta do *Gravity Model* é calcular o volume de fluxo entre duas regiões como proporcional ao produto da população da região de origem e da região de destino e inversamente proporcional à distância entre estas regiões, conforme descrito na equação 2.1. Nessa equação, P_i e P_j são as populações de origem e destino e r_{ij} a distância entre as duas regiões, e ela pode ser generalizada conforme a equação 2.2. Nela, K é uma constante, m_i e m_j podem ser interpretados como a quantidade de viagens que deixam a região i e a quantidade de viagens que são atraídas pela região j (parâmetros que na literatura geralmente são funções da população) e $f(r_{ij})$ é a chamada *deterrence function*, uma função da distância (BARBOSA et al., 2018). O *Gravity Model* é utilizado na literatura geralmente a partir de adaptações e estimativas empíricas dos seus parâmetros considerando as regiões em análise, como descrito em (LI et al., 2021), onde os autores estimaram os valores dos parâmetros do modelo para a tarefa de predição do fluxo de bicicletas *dockless*, ou como reportado em (SOHN, 2005), onde foi utilizada uma adaptação do *Gravity Model* adicionando variáveis à equação com o objetivo de conseguir melhorar a explicação da variação dos padrões de fluxo pendular, que são alguns exemplos disponíveis na literatura sobre o tema.

$$T_{ij} \propto \frac{P_i P j}{r_{ij}} \tag{2.1}$$

$$T_{ij} = K m_i m_j f(r_{ij}) \tag{2.2}$$

Propostas posteriores baseadas no *Gravity Model* também buscaram levar em conta atributos como o uso do solo, como no trabalho de (LIU; FANG; JING, 2020), para obter uma maior acurácia na modelagem. Entretanto por se tratar de um modelo linear, existem limitações na modelagem de relações complexas entre as regiões, como interações entre

dados sobre infraestrutura e a mobilidade humana. Além disso, permanece a dependência na estimativa dos parâmetros do modelo. Essa necessidade da estimativa de parâmetros únicos para cada tipo de tarefa e região é considerado um ponto negativo do *Gravity Model*. Em (SIMINI et al., 2012), os autores discorrem sobre, além dessa, outras limitações, dentre elas: 1) a necessidade de informações anteriores sobre o tráfego, o que impossibilita a predição de fluxos em regiões onde não haja uma base robusta, 2) discrepâncias encontradas entre a modelagem e os cenários reais (por exemplo, dois pares de cidades similares em número de habitantes e distância não necessariamente terão fluxos similares entre si, embora de acordo com a equação 2.2, devessem ter) e 3) falta de um limiar superior para o fluxo resultante da equação que seja coerente com as populações de origem e destino, uma vez que, sem este limitador, existem situações onde a equação pode prever fluxos acima da quantidade de habitantes da região de origem.

Com o objetivo de mitigar os problemas inerentes ao *Gravity Model*, foi proposto também em (SIMINI et al., 2012) o *Radiation Model*, um modelo livre de atributos parametrizáveis e cuja fórmula é mostrada na equação 2.3. Nela, m_i e n_j são as populações das regiões de origem e destino que estão a uma distância r_{ij} uma da outra, s_{ij} é a população total num círculo de raio r_{ij} com centro na região i (excluindo a população das regiões de origem e destino), e T_i é o número total de pessoas que iniciam seu deslocamento pela região i.

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{2.3}$$

No trabalho de (MASUCCI et al., 2013), os autores realizaram uma análise comparativa entre o *Gravity Model* e o *Radiation Model* na predição de fluxos entre países (Inglaterra e País de Gales) e dentro de uma cidade (Londres). Um dos objetivos da pesquisa seria verificar a proposta de "modelo universal"que o *Radiation Model* traz. Os autores argumentam a respeito da equação proposta para o *Radiation Model*, que seria apropriada para um sistema de população infinita, e propõem um ajuste. Dentre os resultados reportados, é mostrado que os dois modelos enfrentam problemas na realização de uma boa estimativa dos fluxos dentro da cidade, e que para esta granularidade, ainda é necessária a existência de um modelo que apresente boa performance na tarefa de predição.

Para avaliar técnicas que permitam a incorporação de atributos além dos relacionados à distância e população, modelos de aprendizagem de máquina também foram propostos para modelar a distribuição das viagens. Modelos baseados em árvore são largamente encontrados na literatura relacionada. Até onde foi esta pesquisa, os modelos baseados em árvore mais encontrados na predição de fluxos pendulares são o *Random Forest*, *Gradient Boosting Regressor* e *XGBoost*. Os modelos baseados em árvore tem em comum a utilização de uma árvore de decisão como base do algoritmo, um classificador fraco, e usam a aprendizagem do tipo *ensemble*, que une múltiplos classificadores fracos e cria um classificador forte. No *Random Forest*, o método de *ensemble* utilizado é o *bagging*, que

em linhas gerais treina várias árvores de decisão de forma paralela com diferentes partes do dado de entrada, e define a predição final do modelo a partir do voto majoritário considerando os resultados de cada árvore. Já o *Gradient Boosting* e o *XGBoost* utilizam a técnica de *boosting*, que trabalha de forma sequencial: os modelos vão sendo treinados de forma que o próximo minimize o erro do modelo anterior. Embora redes profundas atualmente tenham alcançado o estado da arte em diversas tarefas complexas, existem estudos recentes apontando a superioridade de performance de modelos baseados em árvore em problemas que envolvem dados tabulares e possuem uma quantidade média de dados disponível (aproximadamente 10000 registros), como reportado em (GRINSZTAJN; OYALLON; VAROQUAUX, 2022).

Em (POUREBRAHIM et al., 2019), o *Random Forest* e uma arquitetura de rede neural foram escolhidos para avaliar suas performances em relação ao *Gravity Model* utilizando dados de redes sociais, tendo o *Random Forest* performado melhor dentre os três. No trabalho de (SPADON et al., 2019), os autores propuseram, a partir de dados urbanos coletados do IBGE, uma avaliação comparativa de 44 modelos de regressão na predição dos fluxos de deslocamento pendular entre 5565 municípios brasileiros. Dentre os modelos selecionados, houveram abordagens lineares, baseados em vizinhança e baseados em árvore, e a melhor performance foi reportada com o *XGBoost*, seguido de perto pelo *Gradient Boosting*. Em (MORTON; PIBURN; NAGLE, 2018), foi realizada uma análise comparativa desta vez entre o *XGBoost*, o *Gravity Model* e o *Radiation Model* e os resultados obtidos mostram a melhor performance deste modelo baseado em árvore também quando reduzindo a granularidade das regiões de saída e chegada dos fluxos, sendo, neste caso, cada região representada por um bloco censitário, a menor unidade geográfica presente no censo norte-americano.

Abordagens lineares também foram propostas para resolução do problema, tanto de forma individual, como em (SPADON et al., 2019), como também em modelos cuja arquitetura se dá em duas etapas, sendo a primeira a geração de um vetor de representações a partir dos atributos selecionados, e a segunda uma regressão realizada com as representações geradas, como no trabalho de (KIM; YOON, 2022). Os artigos citados tem em comum a utilização de modelos lineares que aplicam alguma técnica de regularização, como a regressão Ridge e a regressão Lasso. Ambas aplicam um termo extra na equação de perda do modelo linear, e tem como vantagem lidar com dados de treino onde existe multicolinearidade, entretanto, o fazem de formas diferentes: na primeira (também conhecida como regularização L2), proposta em (HOERL; KENNARD, 1970), o termo adicionado penaliza a partir do quadrado dos coeficientes da equação linear (equação 2.4). Já na segunda (também conhecida como regularização L1), proposta em (TIBSHIRANI, 1996), o fator de regularização penaliza o valor absoluto dos coeficientes (equação 2.5). Nas equações 2.4 e 2.5 o β representa os coeficientes da equação linear, e o λ é o parâmetro de regularização, e controla o quanto de regularização é aplicada. A regressão Lasso em particular, além de tratar da multicolinearidade, também performa uma seleção de atributos, tendendo a

zerar os coeficientes de alguns atributos que o modelo julgue de menor importância. Esta é uma grande diferença em relação a Regressão Ridge, que reduz o valor dos coeficientes, porém sem nunca chegar a zero.

$$\lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.4}$$

$$\lambda \sum_{j=1}^{p} |\beta_j| \tag{2.5}$$

Mais recentemente passaram a ser propostas abordagens envolvendo redes neurais. O trabalho de (SIMINI et al., 2021) foi proposto como uma extensão do *Gravity Model* para geração de fluxos. Os autores argumentam que a escolha da arquitetura do modelo é uma extensão natural do *Gravity Model*, porém adicionando não-linearidade a partir das camadas ocultas da rede.

Outra técnica que tem reportado resultados promissores modela o problema utilizando grafos. Os autores de (WU et al., 2022) definem grafos como estruturas compostas por um conjunto de nós e um conjunto de arestas, onde os nós representam as entidades e as arestas representam as relações entre entidades, formando sua estrutura, que pode ser enriquecida com informações de atributos, sejam eles dos nós, das arestas ou do grafo como um todo. Para utilização de redes neurais cuja entrada é uma estrutura do tipo grafo, são utilizadas redes neurais para grafos, que modelam a estrutura de nós e arestas típica deste tipo de dado, como nos trabalhos de (LIU et al., 2020; KIM; YOON, 2022; SHI et al., 2024). Além disso, tem sido largamente explorada a área de aprendizagem por representação com grafos para resolver este tipo de problema.

As propostas que modelam o problema de predição de fluxos pendulares a partir de grafos muitas vezes utilizam redes neurais baseadas em grafos para aprender representações dos vetores de atributos utilizados para a predição. A esta técnica é dado o nome de aprendizagem por representação com grafos, e o seu objetivo é aprender vetores de representações latentes (os *embeddings*, no original em inglês) que podem ser geradas para quaisquer elementos do grafo e tem como objetivo preservar tanto sua estrutura quanto seus atributos, capturando a estrutura subjacente e/ou relacionamentos ocultos entre os dados. Esta abordagem gera representações que são utilizadas em diversas tarefas analíticas, como detecção de comunidades, como no trabalho de (LI et al., 2018), previsão de tráfego, como no trabalho de (WANG; LI, 2017) e isomorfismo de grafos, como no trabalho de (XU et al., 2018). A pesquisa nessa área vem ganhando mais e mais atenção nos últimos anos a partir da percepção de que uma grande parte dos dados no mundo real podem ser representados por grafos. Existem alguns aspectos desafiadores na obtenção de boas representações, como encontrar a dimensão adequada para o vetor final de representações gerado, e escolher a propriedade adequada do grafo a incorporar no vetor final (atributos

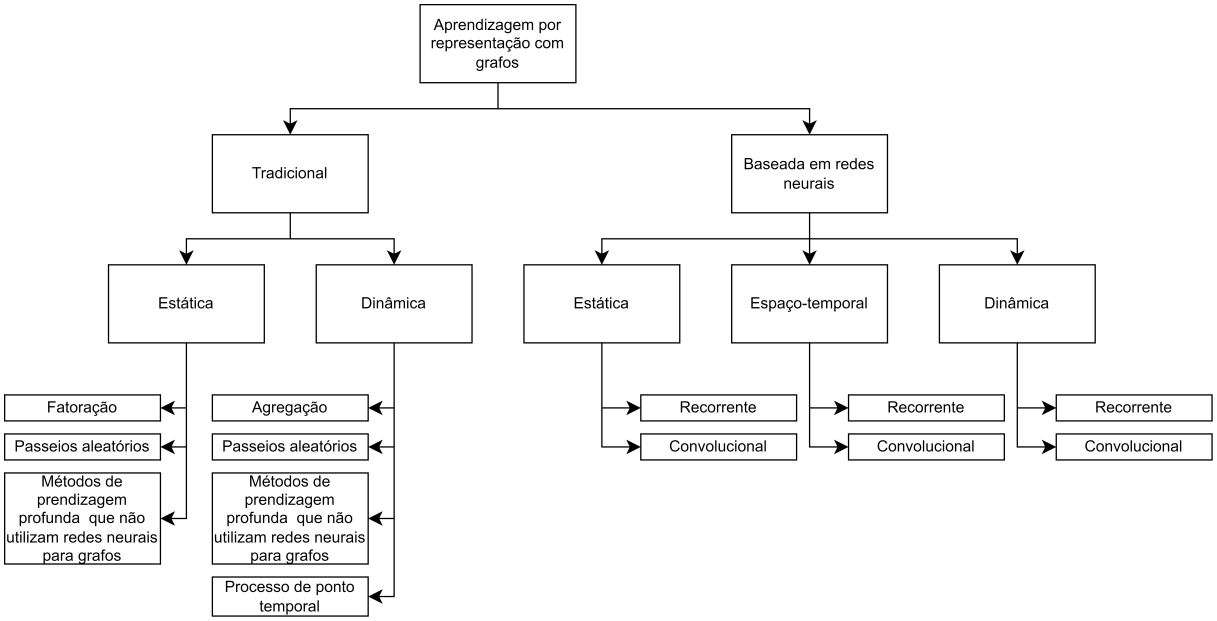

Figura 1 – Categorização dos métodos de aprendizagem por representação com grafos, proposto por (KHOSHRAFTAR; AN, 2022), traduzido pela autora.

dos nós, estruturas das arestas, entre outras), que deve ser guiada pelo tipo de aplicação que essa representação terá, conforme descrito no trabalho de (CHEN et al., 2020).

No trabalho de (KHOSHRAFTAR; AN, 2022), os autores propõem uma categorização das técnicas de aprendizagem por representação, mostrada na figura 1. De acordo com esta categorização, além das abordagens baseadas em redes neurais citadas acima, também existem os métodos tradicionais. Dentre eles, tanto para os grafos estáticos (que não mudam com o tempo e mantém a mesma quantidade de nós e arestas) quanto para os dinâmicos (grafos que mudam sua estrutura com o tempo), uma técnica que se destaca pelo sucesso na geração de representações para grafos utiliza o conceito de passeios aleatórios (*random walks*, no original em inglês). Existem diferentes implementações desta técnica, mas sua ideia geral quando aplicada a este problema consiste em gerar caminhos aleatórios pelo grafo com o objetivo de capturar sua estrutura, incluindo noções de vizinhança. Alguns modelos de grande importância na literatura se baseiam nesta técnica, como o *Node2vec*, proposto em (GROVER; LESKOVEC, 2016), o *DeepWalk*, proposto em (PEROZZI; AL-RFOU; SKIENA, 2014), e o *LINE*, proposto em (TANG et al., 2015). Estes modelos, embora consigam obter bons resultados, geram uma representação baseada na estrutura e vizinhança do grafo, não levando em consideração demais atributos que possam oferecer informações complementares para o algoritmo, que podem resultar em representações mais acuradas.

Já nas abordagens baseadas em redes neurais, novamente observando a categorização da figura 1, além das redes que atuam sobre grafos estáticos e dinâmicos, surge uma nova subcategoria, a espaçotemporal, que atua capturando as propriedades espaciais e temporais dos grafos. Dentro destas subcategorias, por sua vez, existem as técnicas baseadas

em recorrência, que remetem às redes neurais recorrentes para grafos (*RecGNNs*, na sigla em inglês), que resultram em algoritmos computacionalmente custosos, sendo este um desafio a ser vencido para este tipo de técnica. Em paralelo, devido ao sucesso das redes neurais convolucionais em tarefas de visão computacional, foram surgindo propostas para adequar o conceito de convolução para grafos, o que culminou no surgimento deste tipo de rede neural.

Na literatura, abordagens que utilizam redes convolucionais para a predição de fluxos são encontradas como nos trabalhos reportados em (NEJADSHAMSI et al., ; WANG et al., 2023; FU et al., 2023). Mais recentemente entretanto, uma técnica que ganhou rápida aceitação na literatura devido aos bons resultados reportados foram as redes de atenção para grafos (*GATs*, na sigla em inglês), uma arquitetura que utiliza o conceito dos mecanismos de atenção, que estão por trás da arquitetura *Transformer*, uma pesquisa de grande importância que impulsionou fortemente a área de processamento de linguagem natural a partir do trabalho de (VASWANI et al., 2017). No trabalho de (VELIčKOVIć et al., 2017), que propôs a arquitetura das *GATs*, os autores citam algumas vantagens: 1) se trata de uma arquitetura eficiente, uma vez que pode ser paralelizada, 2) pode ser aplicada a nós de grafos com diferentes graus (número de arestas que incidem sobre o nó), e 3) a arquitetura proposta é aplicável a problemas de aprendizagem indutiva, que são problemas onde a rede treinada precisa ser capaz de generalizar sua predição para dados não vistos. Aplicações bem-sucedidas incluem previsão de tráfego, como em (YU; LEE; SOHN, 2020), sistemas de recomendação, como em (YING et al., 2018) e descoberta de drogas, como em (KEARNES et al., 2016), além da predição de fluxos pendulares.

Como exemplos de trabalhos que utilizam *GATs* na tarefa de predição de fluxos pendulares, é possível citar o trabalho de (LIU et al., 2020), que propôs o *GMEL* (sigla para *Geo-Contextual Multitask Embedding Learner*, no original em inglês), que utiliza informações sobre infraestrutura e uso do solo para geração de representações a partir de duas *GATs* diferentes, uma para gerar representações para os setores censitários de origem e outra para fazer o mesmo para os setores censitários de destino. A este processo é dado o nome de *GMEL* pelos autores, porém a arquitetura proposta ainda possui uma segunda etapa, onde utiliza as representações geradas na predição dos fluxos pendulares utilizando o *Gradient Boosting Regressor*. Este tipo de arquitetura é encontrada em outros trabalhos que abordam o problema a partir do uso das *GATs*, como no trabalho de (KIM; YOON, 2022), onde os autores propuseram a utilização desta arquitetura em grafos heterogêneos, e levam em consideração além da dependência espacial, a temporal, o que também é feito em (SHI et al., 2024). Este tipo de abordagem (envolvendo redes neurais para grafos e geração de representações) parece ser uma linha forte de pesquisa na predição de fluxos pendulares, reportando bom desempenho, porém possuem algumas especificidades para obtenção destes resultados, que são principalmente a necessidade de uma quantidade significativa de dados de entrada para o modelo, e poder computacional para o treino su-

perior ao necessário para o treino de modelos mais simples, como os citados anteriormente nesta seção.

2.2 PLANEJAMENTO DE CENÁRIOS URBANOS

O planejamento de cenários é um termo multidisciplinar que busca antecipar resultados futuros de uma decisão tomada no presente com objetivo estratégico. A capacidade de avaliar desfechos em diferentes cenários e considerar diferentes resultados é importante em vários domínios. Na área de estudo de gestão (ou planejamento) urbana, existem na literatura diferentes estratégias que têm em comum a proposta de pensar criticamente os caminhos que o futuro pode tomar e identificar ideias-chave que podem não ter vindo à tona a partir de estratégias convencionais de tomada de decisão, como descrito no trabalho de (CHAKRABORTY; MCMILLAN, 2015). Essa etapa de tomada de decisão pode ajudar a planejar o uso do solo, detectar melhorias em rodovias, definir o local ótimo para implantação de empreendimentos de grande porte (considerando o panorama de justiça ambiental, isto é, o tratamento justo de todas as pessoas de forma que estejam todas igualmente protegidas de efeitos adversos causados por estas implantações), assim como reduzir o tempo de resposta da cidade em situações de emergência, seja ela sanitária ou causada por desastre ambiental.

Numa pesquisa realizada pela *ESI ThoughtLab* e disponível em (THOUGHTLAB, 2021), que entrevistou gestores urbanos de 167 cidades espalhadas por 82 países, 65% dos tomadores de decisão entrevistados elencaram como a maior lição aprendida com a pandemia de COVID-19 é o quão crucial programas relacionados a cidades inteligentes são para o futuro das cidades. Estas são atividades que moldam o futuro urbano e dos habitantes desses grande centros, e, portanto, devem ser decisões tomadas sob fundamentação sólida.

Atividades que fazem parte da gestão urbana precisam envolver transparência, atuação de especialistas e participação da comunidade, propondo abordagens para promover cada vez mais o envolvimento das partes interessadas, e é fundamental que as decisões tomadas levem em consideração estes pontos. Alguns trabalhos que ilustram esta visão de tomada conjunta de decisões em gestão urbana são descritas em (DORAISWAMY et al., 2018; FERREIRA et al., 2015; MIRANDA et al., 2019; ORTNER et al., 2017). Para além destes elementos basais, entretanto, também é possível trazer para auxilio à gestão os chamados sistemas de suporte à decisão, que são programas utilizados para apoiar o processo de tomada de decisão em uma organização ou negócio, podendo agir assim como uma ferramenta dê suporte ao planejamento de cenários.

É possível encontrar uma ampla gama de ferramentas de apoio gestão urbana na literatura. No âmbito da saúde pública, o trabalho de (SCHONER et al., 2018) trouxe a aplicação de planejamento de cenários urbanos propondo um sistema que modela as relações do ambiente com vários resultados de saúde para uma variedade de grupos de idade e renda. Já para a área imobiliária, no trabalho de (PETTIT et al., 2020), os autores propuseram

um *toolkit* para ajudar a determinar o provável aumento do valor da terra associado ao fornecimento de novas infraestruturas urbanas, e validam o trabalho realizado a partir de um caso de uso. Ferramentas voltadas a mobilidade urbana incluem, por exemplo, o trabalho de (MARINI et al., 2019), onde os autores utilizaram modelos de simulação baseados em agentes para investigar o crescimento de regiões residenciais, assim como a criação de atividades comerciais nessas áreas, e o trabalho de (YANG, 2020), onde o autor utiliza estatística e uma base de dados de telefonia móvel para incorporar a um *framework* uma ferramenta que fosse capaz de entender os padrões de fluxos pendulares. Outras abordagens para ferramentas de planejamento de cenários propõem abordagens tradicionais, que usam análise de regressão, como no trabalho de (NONG; DU, 2011), modelos de previsão de viagens ou modelos econométricos, como no trabalho de (BARTHOLOMEW, 2007).

Com o crescimento do poder computacional e da disponibilidade de dados urbanos, o planejamento de cenários também pôde se beneficiar de abordagens que utilizam aprendizagem de máquina, gerando trabalhos de natureza interdisciplinar que impulsionam todas as áreas envolvidas. Na literatura foram propostas redes neurais como modelos para planejamento de cenários para prever risco de inundação, no trabalho de (KIM; NEWMAN, 2020), ou como calcular melhor o uso do solo considerando a contexto espacial ao redor, como proposto em (WANG et al., 2020). Para a predição de fluxos pendulares, modelos mais recentes reportam boa performance na tarefa, como as arquiteturas propostas em (LIU et al., 2020; SHI et al., 2024; KIM; YOON, 2022), para citar algumas, o que seria um indicativo de modelos promissores para utilização em sistemas de suporte à decisão, como uma ferramenta de planejamento de cenários futuros.

2.3 INTERPRETABILIDADE EM MODELOS DE APRENDIZAGEM DE MÁQUINA

Modelos de aprendizagem de máquina vem conquistando feitos mais e mais notáveis nos últimos anos. Com os avanços obtidos no ano de 2023 no campo da inteligência artificial generativa, por exemplo, houve na literatura um movimento de divulgação de modelos aplicados a diferentes áreas de conhecimento apresentando resultados impressionantes de performance, e com potencial para proporcionar oportunidades de negócio em áreas como marketing, finanças e *supply chain*, como reportado no relatório de 2023 da *McKinsey & Company* a respeito do potêncial econômico da IA generativa, disponível em (COMPANY, 2023).

Esse movimento levanta, mais do que nunca, questionamentos sobre a responsabilidade dos mantenedores destes modelos sobre os resultados que os mesmos reportam, especialmente em ambientes que precisam de uma regulação mais rígida, como saúde e segurança pública, para citar alguns. Os autores em (CARVALHO; PEREIRA; CARDOSO, 2019) reportam em sua pesquisa que o interesse na explicação de modelos inteligentes já é registrado na literatura desde meados de 1970, mas que nos anos subsequentes veio perdendo força em comparação com a corrida por modelos com mais poder preditivo, uma tendência

discutida por autores que estudam o *trade-off* entre acurácia e interpretabilidade em modelos de aprendizagem de máquina (MORI; UCHIHIRA, 2019). Entretanto, este não é mais o cenário encontrado atualmente, e existem esforços em diversas áreas do conhecimento para entender a lógica por trás de determinada saída de um modelo, como mostrado nos trabalhos de (MARTIN et al., 2023; CHEN et al., 2023; VIJAYAKUMAR, 2023).

Esta análise do *trade-off* entre a performance de modelos de aprendizagem de máquina e a habilidade do mesmo em produzir predições interpretáveis é bem documentada na literatura, como mostrado nos trabalhos de (BRATKO, 1997; FREITAS, 2019). Entretanto, mesmo para modelos de complexidade menor que os generativos, os mesmos questionamentos seguem relevantes, uma vez que, para que um modelo possa ser utilizado em um contexto em que ele apoie uma tomada de decisão, seja esta decisão em qualquer área de conhecimento, é imprescindível que seja possível, por meio de alguma técnica, ferramenta, ou até a partir da própria natureza do modelo (considerando aqueles que já são intrínsecamente interpretáveis), obter explicações sobre quais fatores foram relevantes para que o modelo tomasse determinada decisão.

A interpretabilidade, no contexto de modelos de aprendizagem de máquina, tem diferentes conceitos encontrados na literatura. (MASÍS, 2021) conceitua interpretabilidade como a medida com a qual humanos conseguem entender a causa e efeito, entrada e saída, de um modelo de aprendizagem de máquina, e dizer que um modelo tem um alto nível de interpretabilidade significa dizer que é possível descrever, de uma forma interpretável para humanos, a razão para o resultado da sua inferência.

A definição fornecida associa o conceito de interpretabilidade à capacidade de entendimento humana do que leva a determinada predição do modelo. Dessa forma, só seria relevante para esta área de estudo explicar os detalhes do funcionamento de um modelo na medida em que isso auxiliasse o indivíduo que utiliza esse modelo a entender como ele realiza suas predições. Existem algoritmos de predição que, intrinsecamente, possuem um alto grau de interpretabilidade (os chamados modelos *white-box*), e são considerados modelos de arquitetura mais simples, o que justificaria a maior simplicidade em explicar seus resultados. As regressões lineares, por exemplo, realizam sua inferência a partir da soma ponderada dos atributos de entrada. Por isso, seus coeficientes (os pesos) da equação ajustada aos dados de treinamento já representam quais atributos foram mais relevantes na obtenção do modelo preditor. As árvores de decisão, também, são baseadas num fluxograma de decisões tomadas a partir de um limiar calculado para cada bifurcação da árvore, que pode ser interpretado sem apresentar um elevado grau de complexidade. Entretanto, para modelos de arquiteturas mais rebuscadas (os modelos chamados *black-box*, e aqui é possível citar como exemplos as redes neurais profundas e os modelos baseados em *ensemble*, como mostrado no trabalho de (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020)), essa interpretação de resultados, mesmo quando possível, se torna cada vez mais custosa e não intuitiva, o que motivou o crescimento da comunidade que

estuda a interpretabilidade e seus tópicos correlatos, propondo uma série de técnicas com o objetivo de entender as informações que motivaram um algoritmo de aprendizagem de máquina a predizer determinada saída.

Em (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020) é descrita uma estruturação dos tipos de técnicas utilizados para a avaliação da interpretabilidade, com base em parâmetros como o tipo de dado que é recebido como entrada nessas técnicas, ou se as técnicas tratam de visões globais ou locais de interpretabilidade. Na figura 2, os autores criaram o que chamaram mapa mental de taxonomia de técnicas de interpretabilidade de aprendizagem de máquina. É possível observar quatro subdivisões. A principal a ser analisada é a que divide as técnicas de interpretabilidade pelo seu propósito: criação de modelos intrínsecamente interpretáveis, onde o objetivo é propor novas arquiteturas cujas saídas sejam facilmente interpretáveis por humanos, como proposto em (USTUN; RUDIN, 2016); explicação de modelos complexos já existentes, onde os modelos existentes são submetidos às técnicas propostas pós treino para obter informações que auxiliem na interpretabilidade dos seus resultados (os chamados métodos *post-hoc*, como os propostos nos trabalhos de (LUNDBERG; LEE, 2017; RIBEIRO; SINGH; GUESTRIN, 2016); aumento da justiça (*fairness*) do modelo, onde o foco está nos impactos éticos e sociais das inferências geradas, e são propostas abordagens para mitigar esses impactos, como mostrado nos trabalhos de (BELLAMY et al., 2018; LANDERS; BEHREND, 2023); testar a sensibilidade das predições, onde o objetivo é avaliar a capacidade de modelos de aprendizagem de máquina de manter a estabilidade nas suas predições a partir de mudanças súbitas e intencionais nas entradas do modelo, como no trabalho de (ANKENBRAND et al., 2021).

As demais divisões propostas na figura 2 (quanto ao tipo de dados, se a o método de interpretabilidade é local ou global e se ele é específico por modelo ou independente do modelo) se entrelaçam nos propósitos da interpretabilidade citados acima na medida em que vão complementando a classificação da técnica proposta. Além da divisão em análise, os métodos de interpretabilidade também podem ser subdivididos pela natureza da abordagem utilizada. O trabalho proposto por (RIBEIRO; SINGH; GUESTRIN, 2016), apresentado pelos autores como LIME (*Local Interpretable Model-Agnostic Explanations*, no original em inglês), é uma técnica representante das abordagens baseadas na perturbação (GOMEZ; MOUCHÈRE, 2023), que são métodos que efetuam perturbações nos dados de entrada com o objetivo de avaliar o efeito na saída do modelo. Ela é uma técnica *post-hoc*, e propõe um método que gera interpretações para as saídas dos modelos para qualquer classificador (sendo dessa forma também um representante das técnicas de interpretabilidade que independem do tipo de modelo). O *LIME* gera, para cada instância e predição associada, uma amostra simulada dos dados ao redor da instância em análise. Novas predições são realizadas a partir da amostra gerada, e ponderadas pela sua proximidade à instância em análise, e esse resultado é utilizado como entrada para o treino de um modelo interpretável. A proposta do *LIME* é de que, sendo este modelo final interpretado, o mo-

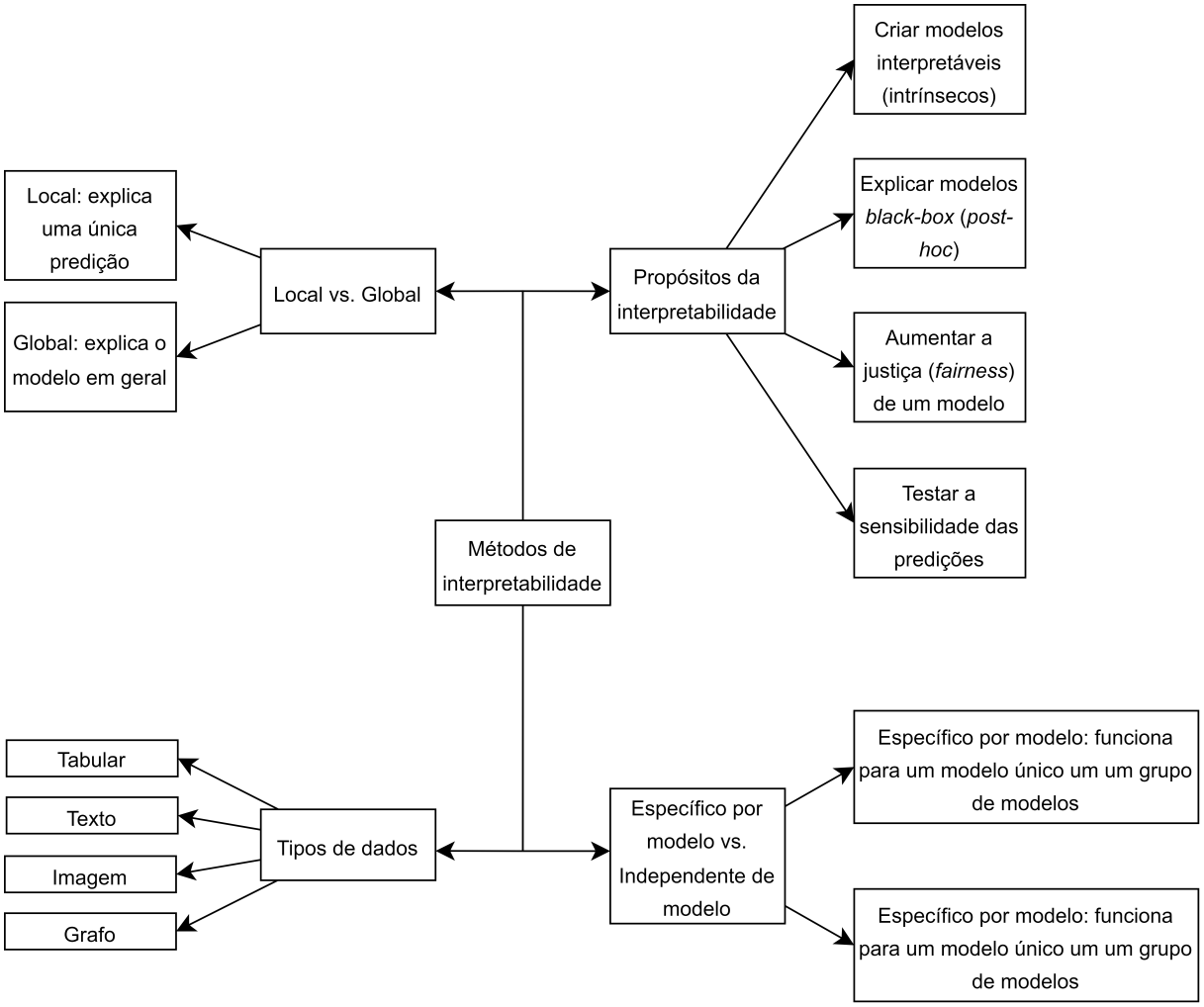

Figura 2 – Mapa mental de taxonomia de técnicas de aprendizagem de máquina, proposto por (LINARDATOS; PAPASTEFANOPOULOS; KOTSIANTIS, 2020), traduzido pela autora.

delo original também o será. O *LIME* foi largamente utilizado na literatura para tarefas de interpretabilidade, como nos trabalhos de (WU et al., 2023; CICCO et al., 2019), entretanto, foram também apontadas na literatura algumas limitações, como instabilidade nas gerações das interpretações, o que resultou na proposta do DLIME, uma versão determinística do LIME, proposta em (ZAFAR; KHAN, 2019), que substitui a geração aleatória de perturbação nos dados por um mecanismo baseado em agrupamento hierárquico para definição do grupo mais relevante de atributos para a instância que está sendo explicada.

Além dos métodos baseados em perturbação, existem os baseados em retropropagação, que são modelos que utilizam o gradiente gerado por uma saída específica em relação à entrada usando retropropagação como uma medida da contribuição de cada atributo. O método mais simples desse tipo de abordagem consiste em visualizar os gradientes nos chamados "mapas de gradiente", ou "mapas de sensibilidade"(GOMEZ; MOUCHÈRE, 2023). O *DeepLift*, proposto em (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017), é um representante desta categoria. A ideia proposta no artigo é utilizar a retropropagação das redes neurais

para comparar a ativação de cada neurônio com uma ativação de referência e, assim, conseguir gerar uma pontuação de contribuição do atributo na predição.

Visando propor uma técnica que unificasse métodos utilizados na tarefa de interpretabilidade, foi proposto o *SHAP* no trabalho de (LUNDBERG; LEE, 2017). O *SHAP* é uma ferramenta *post-hoc* e também baseada na perturbação dos dados de entrada, e teve sua origem na teoria de jogos, a partir dos valores Shapley, que representam a distribuição marginal média de um atributo considerando todas as possíveis permutações desse atributo. A intuição por trás desses valores se baseia na distribuição justa de ganhos entre jogadores, e no caso da interpretabilidade de modelos, os jogadores são os atributos dos modelos e os ganhos são a importância de cada atributo nas predições.

O artigo que apresenta o *SHAP* define uma classe de métodos que calculam a importância dos atributos de um modelo, os chamados métodos aditivos de atribuição de atributos. Os métodos que fazem parte dessa classe tem em comum a mesma aproximação interpretável do modelo original (sendo essa aproximação interpretável conhecida na literatura como *explanation model*), e seria, para esta classe, representados por uma função linear de variáveis binárias, como mostrado na equação 2.6, onde z'_i são as variáveis, ϕ_i é o efeito de cada atributo e M representa a quantidade de atributos. Segundo os autores, os métodos que fazem parte dessa classe tem como solução os resultados vindos da teoria dos jogos (relacionados aos valores Shapley), e, portanto, o *SHAP* é uma unificação dos métodos que se enquadram na classe definida.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2.6}$$

O *SHAP* propõe métodos otimizados para a explicação de um determinado tipo de modelo e/ou determinado tipo de *framework* em que o modelo foi implementado. Ele se destaca por esse fator, assim como pela sua robustez: por se tratar de uma abordagem exaustiva, este método consegue garantir propriedades que dão consistência ao resultado obtido. Entretanto, este último fator faz também com que o *SHAP* seja uma abordagem mais computacionalmente custosa, e, portanto, a escolha de qual método utilizar necessita de uma avaliação do cenário de análise. Além disso, algumas técnicas de interpretabilidade tem na sua concepção a suposição de que existe independência entre os atributos, o que em muitas bases de dados não se prova realidade. A multicolinearidade é um conceito estatístico onde existe uma alta correlação entre duas ou mais variáveis. Este é um problema já bem descrito na literatura, e no contexto da interpretabilidade, pode causar distorções nos resultados de importância dos atributos obtidos, uma vez que os atributos não são independentes e, portanto, perturbações e/ou permutações realizadas num atributo que esteja altamente correlacionado com outro(s) pode causar efeitos inesperados. Os efeitos da multicolinearidade na interpretabilidade de modelos e potenciais soluções são discutidos em (BASU; MAJI, 2022).

Embora as técnicas citadas nesta seção consigam abranger muitos cenários de modelos e *frameworks* diferentes, o avanço na especialização e complexidade dos modelos inevitavelmente produz avanços nas áreas de conhecimento onde a aplicação dessas técnicas nas arquiteturas dominantes não é possível. Nas redes neurais cuja entrada são grafos existem esforços sendo realizados para adaptar métodos de interpretabilidade para os tipos de dados de entrada e de modelo mais utilizados. Algumas abordagens que buscam tratar o problema na literatura são o GNNExplainer, proposto em (YING et al., 2019), o *PGExplainer*, proposto em (LUO et al., 2020), e o *SubgraphX*, proposto em (YUAN et al., 2021), todas técnicas baseada em perturbação, e também o *GraphLIME*, proposto em (HUANG et al., 2020), e *PGM-Explainer*, proposto em (VU; THAI, 2020), que são técnicas baseadas nos *surrogate models*, modelos simples e interpretáveis que são utilizados para aproximar as predições de modelos complexos nas vizinhaças da instância que se deseja interpretar, como descrito em (YUAN et al., 2022). O *GraphLIME*, em específico, foi concebido como uma extensão do *LIME* que aprende um modelo interpretável e não linear num subgrafo do nó sendo explicado. Alguns autores também buscaram realizar adaptações do SHAP, como o *GraphSVX*, proposto em (DUVAL; MALLIAROS, 2021), o *Shapley Flow*, proposto em (WANG; WIENS; LUNDBERG, 2021), e o *EdgeSHAPer*, proposto em (MASTROPIETRO et al., 2022). Essas adaptações possuem como limitações alguns aspectos: muitas vezes estão limitadas a um *framework* específico, e são também geralmente limitadas pelo tipo de tarefa que conseguem interpretar (classificação de nó, classificação de aresta, entre outras).

3 MATERIAIS E MÉTODOS

Neste capítulo, a metodologia aplicada na coleta e processamento dos dados é descrita, assim como o processo de treinamento utilizado para os modelos que compõem a análise comparativa. Também é realizada uma descrição sobre natureza dos dados coletados, com o objetivo de fornecer informações que auxiliem no entendimento dos resultados reportados no capítulo 4.

3.1 COLETA E PROCESSAMENTO DE DADOS

As bases de dados utilizadas para realização dos experimentos em sua maioria seguiram o processo de coleta de (LIU et al., 2020), que será descrito nas subseções a seguir. A cidade escolhida para a coleta dos dados que foram utilizados nos experimentos reportados no capítulo 4 é Nova Iorque, que é vastamente utilizada na literatura devido a quantidade significativa de bases de dados abertas e de boa qualidade disponibilizadas por diferentes órgãos, como os disponibilizados em (YORK, 2022; PLANNING, 2024a), o que gera um ambiente propício para experimentação em diversas áreas de conhecimento. Para problemas relacionados a mobilidade, algumas bases de dados de grande importância para pesquisa são a base de táxi da Comissão de Táxi e Limousine da cidade de Nova Iorque ((COMMISSION, 2024)), a base de bicicletas fornecida pelo Citi em parceria com a empresa de transporte *Lyft*, disponível em (BIKE, 2024) e a base de estatística de fluxos de origem e destino disponibilizada pelo Censo dos Estados Unidos, disponível em (BUREAU, 2024), esta última utilizada nos experimentos cuja metodologia será descrita a seguir.

Os dados coletados para realização deste trabalho contemplam, além de dados de fluxo entre duas regiões, informações sobre infraestrutura e uso do solo dessas regiões e informações sobre a distância entre as mesmas. Nas subseções a seguir, serão detalhadas cada fonte de dados citada.

3.1.1 Fluxos de deslocamento

Os fluxos de deslocamento para a cidade de Nova Iorque foram extraídos a partir do *LODES* (*LEHD Origin-Destination Employment Statistics*, no original em inglês), uma base de dados mantida pelo censo dos Estados Unidos e que disponibiliza informações que auxiliam no entendimento da dinâmica da força de trabalho no país (disponível em (BUREAU, 2024)). No momento de coleta desta base, o *LODES7* era a versão mais atual disponível. Ele utiliza como unidade geográfica o *census block*, ou bloco censitário, que é a menor área estatística definida para fins de censo, delimitada por recursos visíveis ou não.

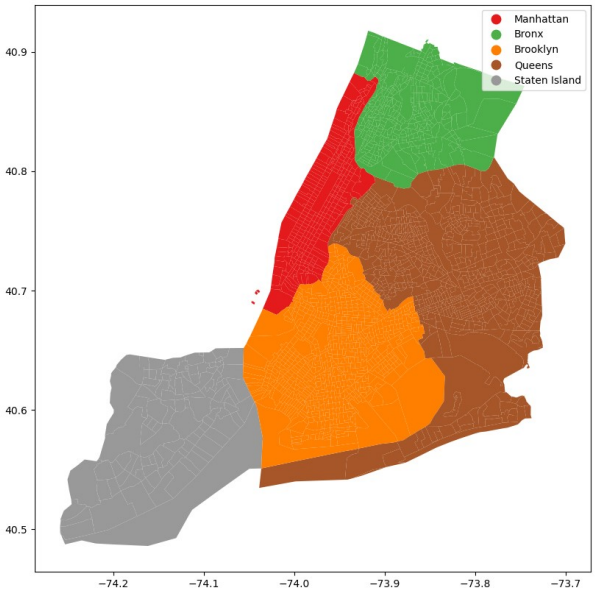

Figura 3 – *Boroughs* na cidade de Nova York. Fonte: a autora (2024).

O *LODES7* utiliza como referência o censo norte-americano de 2010, e possui informações a respeito de fluxos de deslocamento pendulares e informação a respeito da distribuição de empregos por região para todos os estados dos Estados Unidos. Para os experimentos descritos no capítulo 4, foi utilizada a base de fluxos de deslocamento pendular. Para a maior parte dos experimentos reportados, o ano utilizado foi o de 2015 (seções 4.1 e 4.2, exceto subseção 4.2.3). Já para os experimentos que analisam mudanças temporais (subseção 4.2.3, foram coletados dados referentes aos anos de 2010 a 2018.

O processamento da base consistiu na filtragem pelas informações de fluxo que fazem parte da cidade de Nova Iorque, e agrupamento dos fluxos em *census tracts*, os setores censitários, que são as unidades geográficas utilizadas para os experimentos descritos. É uma alternativa interessante utilizar setores censitários como a unidade geográfica dos experimentos, porque além de ser uma divisão já bem consolidada da cidade (o que facilita a obtenção de bases de dados na mesma granularidade), os setores censitários também são definidos em colaboração com grupos locais, o que resulta em delimitações mais significativas entre regiões, como observado na apresentação disponível em (BUREAU, 2015).

A cidade de Nova Iorque possui características particulares em relação a suas subdivisões geográficas, e portanto, os setores censitários são representados para estes experimentos pelo *BoroCT*, que consiste no código do setor censitário com a adição do *borough code*, código referente a uma das cinco divisões administrativas que compõem a cidade (*Manhattan*, *Bronx*, *Brooklyn*, *Queens* e *Staten Island*). A distribuição dos *boroughs* de Nova Iorque pode ser visualizada na figura 3.

Foram gerados dois tipos de dados de fluxo: o primeiro, que é a base utilizada nos

Fluxo entre setores de origem e destino	Representação na base (%)
1	53.7
2	16.9
3	8.3
4	4.9
5	3.2
>5	13.0

Tabela 1 – Distribuição dos fluxos para a base de dados desconsiderando os pares de fluxos de origem e destino iguais a zero.

Fluxo entre setores de origem e destino	Representação na base (%)
0	62.4
1	20.2
2	6.4
3	3.1
4	1.8
5	1.2
>5	4.9

Tabela 2 – Distribuição dos fluxos para a base de dados considerando os pares de fluxos de origem e destino iguais a zero.

experimentos reportados no trabalho de (LIU et al., 2020), leva em consideração apenas dados de fluxo onde o par de setores censitários de origem e destino possua um fluxo maior que zero entre eles, enquanto que o segundo adiciona a esta base também pares de setores censitários cujo fluxo entre si é igual a zero. Os experimentos são, em sua maioria, realizados considerando a primeira base de dados, a fim de dar continuidade aos experimentos reportados em (LIU et al., 2020). Entretanto, também são reportados e discutidos os resultados de performance dos modelos obtidos para o segundo tipo de dado. A distribuição dos fluxos para os dois tipos de dados podem ser observada nas tabelas 1 e 2. Para ambos, a divisão em treino, teste e validação seguiu a proporção estabelecida de 6:2:2.

3.1.2 Atributos urbanos

As características geográficas de cada setor censitário foram coletadas na base de dados *PLUTO* (disponível em (PLANNING, 2024a)), mantida pela cidade de Nova Iorque, que possui extensas informações sobre uso do solo a nível de lote fiscal para toda a cidade. O governo da cidade de Nova Iorque define como um lote fiscal uma parcela de terreno identificado com uma divisão administrativa, um bloco censitário e o número de um lote exclusivo para fins de imposto sobre a propriedade, conforme descrito em (PLANNING,

Ano	Quantidade de registros
2010	860541
2011	860320
2012	859329
2013	859372
2014	858914
2015	859464
2016	858370
2017	859223
2018	858982

Tabela 3 – Quantidade de registros na base de dados *PLUTO* para todos os anos coletados.

2024b), o que torna viável o agrupamento dessas informações em unidades geográficas como setores ou blocos censitários.

A base de dados *PLUTO* traz informações anuais sobre características do lote fiscal, características das construções, e dados a respeito dos distritos da cidade de Nova Iorque. Até o momento de escrita deste trabalho, haviam dados disponíveis para o ano de 2002 até o ano de 2023, e o dado em questão vai sendo atualizado pelos mantenedores da base com o passar do ano. Foram extraídos dados também referentes aos anos de 2010 a 2018. Para os anos coletados, é possível verificar a quantidade de registros disponíveis na base na tabela 3.

Os atributos representativos de cada setor censitário foram definidos para os experimentos reportados seguindo (LIU et al., 2020). Estes atributos são divididos em categorias, que são descritas a seguir, assim como a quantidade de atributos gerados para cada categoria. Entre parênteses, encontra-se a nomenclatura utilizada para cada categoria nas análises realizadas no capítulo 4:

- ***Classe da construção (buildingclass):*** a documentação da base de dados *PLUTO* define a classe de construção como um código que descreve o uso majoritário das estruturas dentro de um lote fiscal. São 25 classes reportadas na documentação. Os atributos desta categoria são calculados contando a quantidade de lotes fiscais por classe de construção para um setor censitário, a contagem de cada classe representando um atributo (25 atributos);
- ***Uso do solo (landuse):*** o uso do solo se refere ao tipo de atividade que ocorre no solo e nas estruturas que o ocupam. Na documentação da base de dados, esta categoria de atributos é relacionado a da classe da construção, tendo sido realizado um mapeamento para a classe que melhor se adeque ao tipo de uso do solo. São reportados 11 tipos de uso do solo, e o cálculo é realizado da mesma forma que o da categoria anterior (11 atributos);

- ***Razão da área de solo (landarearatio):*** calcula a razão entre a soma de áreas de interesse (comerciais, residenciais, de escritórios, etc.), dentro de determinado setor censitário e a área total do setor censitário, gerando atributos numéricos (10 atributos);

- ***Número de construções (numobj):*** calcula a razão entre a quantidade de prédios, quantidade de pavimentos, quantidade de unidades residenciais e quantidade de unidades totais dentro de determinado setor censitário e a área total do setor censitário (4 atributos);

- ***Ano da construção (yearbuilt):*** realiza a contagem de estruturas construídas dentro do setor censitário por década, de 1910 a 2010 (11 atributos);

- ***Distritos históricos, marcos históricos (histdist):*** determina se existe ou não um distrito histórico ou um marco histórico dentro do setor censitário (2 atributos);

- ***Estatísticas (far_builtfar):*** média e desvio padrão da razão de área construída, isto é, a razão entre a área total construída e a área do lote, calculada considerando todos os lotes fiscais dentro de um setor censitário (2 atributos);

Após coleta das bases, o processamento dos dados consiste na aplicação da normalização LQ (*Location Quotient*), um conceito que veio como uma ferramenta de investigação de impactos econômicos regionais, como descrito em (ISSERMAN, 1977), representando a razão entre a quantidade de indústrias numa região (um setor censitário de Nova Iorque, por exemplo) e a quantidade total de indústrias em uma região de referência (a cidade de Nova Iorque, por exemplo). Devido ao seu conceito amplo, o LQ foi aplicado em diferentes áreas de conhecimento nos anos posteriores, como no trabalho de (XU; CHENG; XU, 2018), onde o conceito é aplicado como uma "razão de especialização", um elemento de comparação entre a distribuição de um determinado fator e a média geral desse fator. Para os experimentos realizados neste estudo, o LQ representa uma medida do quão saliente é o valor de um atributo em contraste com todo o conjunto de valores desse atributo. Essa etapa foi executada replicando o que foi descrito pelos autores em (LIU et al., 2020).

3.1.3 Duração da viagem

A matriz que guarda a duração das viagens entre as unidades geográficas foi obtida utilizando a *API* do *OSRM* (*Open Source Routing Machine*), um projeto de código aberto criado a partir do trabalho de (LUXEN; VETTER, 2011) como uma ferramenta que busca encontrar o caminho mais próximo entre dois pontos considerando redes rodoviárias. O *OSRM* permite encontrar a distância considerando diferentes modais de transporte, como carro, bicicleta ou a pé. Para os experimentos realizados, foi utilizado o modal carro.

3.2 ESCOLHA DOS MODELOS

A escolha dos modelos que compõem esta análise comparativa teve como base para escolha:

- Principais modelos que são citados na literatura disponível sobre o problema abordado,
- Modelos que poderiam trazer uma perspectiva diferente sobre o problema.

Os modelos escolhidos envolvem abordagens lineares, baseadas em árvore, redes neurais e modelos de aprendizagem por representação. As regressões de Lasso e Ridge foram escolhidas como as abordagens lineares utilizadas nos experimentos realizados nesta pesquisa, já tendo sido citadas na literatura em problemas de pesquisa relacionados a predição de fluxos pendulares. Mais que isso, entretanto, sua importância reside no fato de serem regressões lineares com um fator regularizador, e que atuam, de formas distintas, na redução do problema da multicolinearidade, como mostrado no trabalho de (ALTELBANY, 2021). Sendo a base de dados coletada do *PLUTO* uma base que possui fortes indícios de multicolinearidade (é uma base com atributos que possuem algum grau de dependência entre si), o objetivo da escolha destes modelos é também avaliar se soluções lineares (que são mais simples) que aplicam regularização seriam competitivas em relação a modelos mais complexos.

Com relação às abordagens baseadas em árvore, foram selecionados o *Random Forest* e *Gradient Boosting Regressor*, ambos propostos na literatura como arquiteturas para a predição de fluxos pendulares, como descrito nos trabalhos de (POUREBRAHIM et al., 2019) e (SPADON et al., 2019). Além disso, modelos baseados em árvore se destacam pela sua robustez numa série de tarefas, possuindo resultados superiores até mesmo à redes neurais para tarefas e quantidade de dados disponíveis específicas, como mostrado em (GRINSZTAJN; OYALLON; VAROQUAUX, 2022).

Na escolha dos modelos baseados em redes neurais, foram selecionadas abordagens cuja entrada é um dado tabular e cuja entrada é um dado do tipo grafo. Para a primeira abordagem, o modelo selecionado foi o *Deep Gravity*, e teve como objetivo trazer da literatura propostas baseadas em redes neurais que procurassem resolver o mesmo problema, ou um problema similar. O *Deep Gravity* apresentou resultados interessantes em uma das métricas que será utilizada para reportar os resultados desta análise comparativa (o *CPC*), trazendo em sua metodologia um dificultador, que é o de prever os fluxos de uma região utilizando um modelo treinado com dados de outra. Dessa forma, foi realizado um treino do modelo com os dados coletados, para que fosse possível avaliar sua performance em bases de dados diferentes da utilizada nos experimentos realizados em (SIMINI et al., 2021). Já para a segunda, foi selecionado o *GMEL*, proposto em (LIU et al., 2020), uma arquitetura de rede neural que utiliza aprendizado por representação para gerar um vetor

de representações a partir dos dados de infraestrutura presentes em cada nó do grafo, que representa uma região (para os experimentos descritos neste capítulo, a região seria o setor censitário). O GMEL foi proposto para a tarefa de predição fluxos pendulares e atingiu resultados bastante satisfatórios, e, portanto, foi selecionada para que também seja analisada sua performance considerando outras abordagens.

Por fim, foi selecionado o *Node2vec*, proposto em (GROVER; LESKOVEC, 2016), como uma arquitetura de aprendizagem por representação que não se baseia em redes neurais e gera suas representações a partir das características estruturais do grafo, utilizando o conceito de *random walks*. O *Node2vec* é um modelo extensamente citado e utilizado em diversas áreas do conhecimento como detecção de comunidades ((HU et al., 2020)), predição de doenças ((PENG; GUAN; SHANG, 2019)) e detecção de fraude ((ZHOU et al., 2021)), além de ter sido utilizado em (LIU et al., 2020) como modelo *baseline* para predição de fluxos pendulares, e portanto foi selecionado para estender a análise da sua performance nesta tarefa. Arquiteturas que implementam esta forma de aprendizagem parece ser um dos caminhos seguido por pesquisadores nos trabalhos mais recentes na modelagem do problema de predição de fluxos pendulares, e, portanto, foram selecionadas para este trabalho.

3.2.1 Dados de entrada

Para cada modelo (ou conjunto de modelos), é necessário um tipo de entrada, que é um subconjunto dos dados coletados e processados descritos acima, e estas entradas são descritas a seguir.

- Regressão Ridge, Regressão Lasso, *Random Forest*, *Gradient Boosting* e *Deep Gravity*

  Para esse conjunto de modelos, que possuem entrada tabular, os dados de entrada consistem nos dados de fluxo de deslocamento e atributos urbanos. É realizado um processamento que mapeia os vetores de atributos urbanos para cada setor censitário de origem e destino na base de fluxos de deslocamento. A base de entrada final consistirá nos atributos urbanos referentes aos setores censitários de origem, os atributos urbanos referentes aos setores censitários de destino, a distância entre os dois setores censitários e o fluxo entre eles, conforme mostrado na figura 4, item a.

- *GMEL*

  Para o *GMEL*, os atributos urbanos e matriz de duração da viagem são utilizados na modelagem do grafo, que é o dado de entrada da primeira etapa do modelo, que aprende o vetor de representação. Na construção desta estrutura de dados, a duração da viagem entre dois setores censitários é modelada como um atributo das arestas entre nós, e os nós são cada setor censitário dentro da cidade de Nova Iorque.

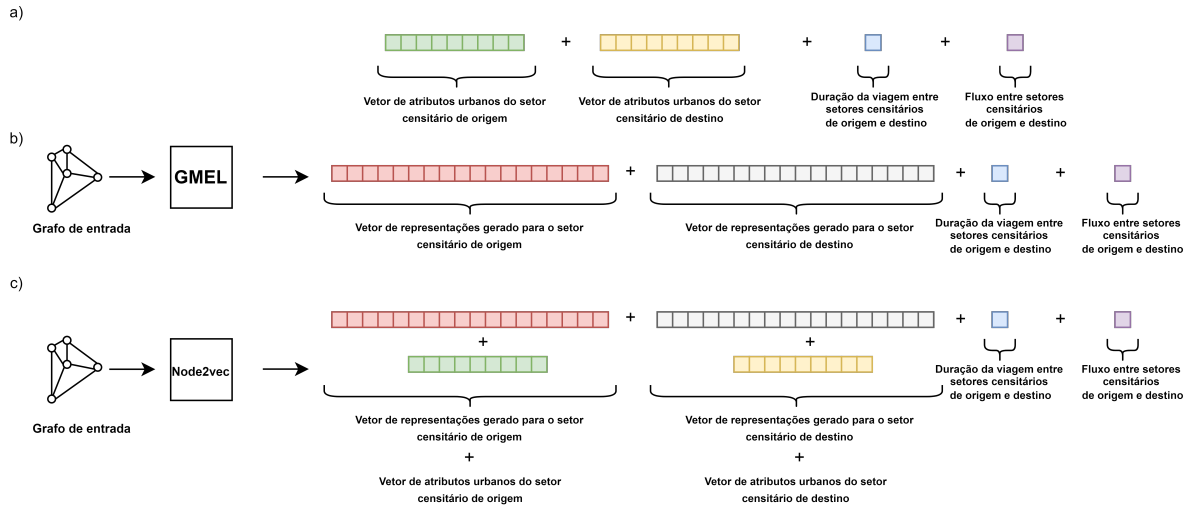

Figura 4 – Formato de entrada dos dados para os modelos: a) Regressão Ridge, Regressão Lasso, *Random Forest*, *Gradient Boosting* e *Deep Gravity*. b) *GMEL*, e c) *Node2vec*. Fonte: a autora (2024).

O grafo que é recebido como entrada do modelo é representado utilizando uma matriz de adjacência, que é uma forma de representação desta estrutura. Se o grafo não for ponderado, isto é, se suas arestas não tiverem nenhum peso associado, a matriz de adjacência será composta por 1 e 0, onde 1 representa que dois nós são adjacentes entre si. Caso o grafo seja ponderado, isto é, se as arestas possuírem um peso associado, a matriz de adjacência será construída substituindo os valores 1 pelo valor do peso entre os dois nós adjacentes. O grafo que é modelado para este experimento é ponderado pela distância entre dois setores censitários, que são os nós do grafo, e portanto a matriz de adjacência que representa este grafo será ponderada pela distância entre os nós, que é obtida a partir dos dados descritos na seção 3.1.3.

O resultado desta primeira etapa da arquitetura são os vetores de representação que serão utilizados na tarefa de predição de fluxos pendulares, a segunda etapa da arquitetura. O processo de obtenção dos dados de entrada desta segunda etapa é ilustrado na figura 4, item b.

- *Node2vec*

  O *Node2vec*, embora seja um modelo cuja entrada é um grafo, diferente do GMEL, não utiliza os atributos dos nós na geração da sua representação, gerando-a a partir das informações estruturais do grafo, como noções de vizinhança. Para os experimentos propostos, entretanto, a arquitetura em que o *Node2vec* está inserido também funciona em duas etapas, primeiro gerando o vetor de representação e, em seguida, utilizando este vetor na predição dos fluxos com o *Gradient Boosting Regressor*. Além dos vetores de representação gerados, os atributos de infraestrutura são tam-

bém utilizados para compor a base de entrada para a segunda etapa da arquitetura, como uma forma de trazer a informação presente nos nós para auxiliar na tarefa de predição. A geração dos dados de entrada para este modelo é ilustrada na figura 4, item c.

3.3 TREINAMENTO DOS MODELOS

Para reportar os resultados voltados a performance dos modelos treinados apenas com fluxos entre origem e destino diferentes de zero (resultados na subseção 4.1.1), foram realizadas 5 execuções de validação cruzada considerando 4 *folds*. O critério de escolha para a quantidade de *folds* foi manter a quantidade de dados de treino e teste, evitando, desta forma, que os resultados sofressem alterações também causadas pela diferença na quantidade dos dados de treino entre os experimentos realizados neste trabalho e em (LIU et al., 2020). A quantidade de execuções foi baseada na literatura, em trabalhos como os de (ANTOR et al., 2021a; BANSAL; GOYAL; CHOUDHARY, 2022). Os demais experimentos foram realizados considerando uma execução de treinamento dos modelos, devido ao tempo necessário para realização de validação cruzada.

Para os modelos que geram representações (*GMEL* e *Node2vec*), a validação cruzada foi realizada já na segunda etapa da arquitetura, pós geração das representações, também pelo tempo e recursos necessários para a execução dos modelos completos. Dessa forma, a etapa de geração das representações foi executada apenas uma vez, e a etapa onde o modelo de regressão é treinado tendo como entrada as representações geradas é treinado realizando também 5 execuções de validação cruzada considerando 4 *folds*.

O *Deep Gravity* utiliza o conceito de épocas (o ciclo onde todos os dados de treino passam pela rede neural) para o treino do modelo. Desta forma, para este modelo, foram utilizadas 20 épocas para o treinamento, replicando o que foi feito pelos autores em (SIMINI et al., 2021). Outro ponto importante é que no artigo os autores dividem a região utilizada nos experimentos em *grids* de 25km por 25km, dentro dos quais estariam os setores censitários. Os *grids* são divididos em treino e teste na proporção 5:5, de forma que o modelo seja testado em *grids* (e consequentemente setores censitários) que não tenham sido vistos previamente pelo modelo. Para os experimentos realizados neste trabalho, mantida esta configuração, foram obtidos 7 *grids* para cobrir a cidade de Nova Iorque. Foram realizados experimentos para verificar se a redução dessa área (e consequentemente aumento na quantidade de *grids*) auxiliaria na melhora da performance do *Deep Gravity*, porém os resultados não mostraram nenhuma melhora significativa, e portanto foi mantida a mesma configuração utilizada no artigo original.

3.4 AVALIAÇÃO DOS MODELOS

Todos os modelos foram avaliados utilizando métricas selecionadas para o tipo de problema em estudo. São elas:

- *RMSE* (Raiz do Erro Quadrático Médio)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}} \tag{3.1}$$

- *MAE* (Erro Médio Absoluto) 3.2

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\widehat{y_i} - y_i| \tag{3.2}$$

- *CPC* (Parte em Comum dos Passageiros) 3.3

$$CPC = \frac{2 \sum_{i=1}^{n} min(\widehat{y_i} - y_i)}{\sum_{i=1}^{n} \widehat{y_i} + \sum_{i=1}^{n} y_i} \tag{3.3}$$

Onde $\widehat{y_i}$ representa o fluxo pendular predito, y_i representa o fluxo pendular real e n é a quantidade total de elementos no vetor.

As duas primeiras métricas são largamente utilizadas em problemas de regressão na literatura. Elas trazem informações complementares, uma vez que o *RMSE* 3.1 tem como característica penalizar mais valores fora da distribuição (os *outliers*), já que eleva ao quadrado os resíduos calculados entre os valores preditos e reais, sendo uma métrica enviesada. Já o *MAE* 3.2, uma métrica não enviesada, consegue dar um panorama mais geral da distribuição do erro entre os valores preditos e reais (sendo menos sensível que o RMSE aos valores fora da distribuição).

Já o *CPC* (equação 3.3) é uma métrica mais especificamente utilizada em problemas de predição de fluxo pendular. Ele é baseado no índice de *Sørensen* (SORENSEN, 1948), que buscou encontrar uma equação que medisse a similaridade entre dois conjuntos. Diferente das métricas baseadas em erro citadas acima, o *CPC* é uma métrica baseada na concordância entre os fluxos preditos e os reais. Dessa forma, enquanto que para o *RMSE* e o *MAE*, quanto menor o valor obtido, melhor, para o *CPC* a regra se inverte, e quanto maior o resultado, melhor. O *CPC* varia de 0 a 1, onde 1 representa uma concordância perfeita entre o vetor de resultados preditos e o de resultados reais, e representa a falta total de concordância entre eles.

4 AVALIAÇÃO COMPARATIVA DA PERFORMANCE DE MODELOS PARA PREDIÇÃO DE FLUXOS DE DESLOCAMENTO

Nesste capítulo serão discutidos os resultados obtidos com as análises realizadas utilizando os modelos treinados seguindo a metodologia descrita no capítulo 3. Para fins de análise, os modelos foram avaliados a partir de três pilares: o primeiro deles é a performance obtida nas métricas descritas na seção 3.4, largamente aplicadas em problemas de regressão e/ou de predição de fluxos de deslocamento origem-destino; o segundo é voltado ao entendimento dos resultados obtidos pelos modelos sob a ótica da interpretabilidade; e o terceiro, por fim, é a sua robustez a mudanças temporais na base de dados, simulando mudanças temporais de infraestrutura, que é uma característica desejável ao selecionar um modelo para tarefas relacionadas ao planejamento de cenários futuros.

4.1 ANÁLISE DE PERFORMANCE

Os resultados reportados nessa seção foram obtidos conforme descrito na seção 3.3. Nas subseções a seguir, foram adotadas duas estratégias quanto aos dados de fluxo origem-destino que são recebidos como entrada pelo modelo: considerando apenas pares origem-destino cujo fluxo entre si é maior que zero e considerando os fluxos entre todos os pares origem-destino na base de dados.

4.1.1 Pares origem-destino com fluxo diferente de zero

Os resultados mostrados na tabela 4 foram obtidos seguindo a metodologia da seção 3.3. Foi observado a partir dos dados disponibilizados no trabalho de (LIU et al., 2020) que os experimentos realizados não levaram em consideração como entrada os fluxos de origem-destino iguais a zero, sendo esses registros inexistentes nos dados de treino, validação e teste. Nesta subseção, esta estratégia é utilizada para os dados que alimentaram os modelos.

É importante ressaltar que, embora os dados de treino sejam os mesmos que os reportados em (LIU et al., 2020), a metodologia utilizada se difere ao treinar os modelos utilizando validação cruzada, o que pode gerar mudanças nos resultados.

Para os modelos que também se encontram no comparativo apresentado por (LIU et al., 2020) (*Random Forest*, *Gradient Boosting Regressor*, *GMEL* e *Node2vec*), podemos observar resultados muito similares aos reportados para as três métricas, o que é o esperado, e segue destacando o GMEL como o modelo que apresenta o melhor desempenho dentre os analisados. Quanto aos modelos que foram adicionados à análise: a regressão linear de Lasso e a de Ridge reportam resultados muito similares nos resultados, especialmente no que diz respeito ao *CPC*, mas também com relação às métricas de regressão, o que

Modelo	RMSE	MAE	CPC
Deep Gravity	6.266	2.641	0.123
Lasso	6.369	2.520	0.623
Ridge	6.010	2.531	0.621
Random Forest	6.029	2.405	0.640
Node2vec	5.438	1.981	0.706
Gradient Boosting	5.202	1.953	0.707
GMEL	4.783	1.718	0.745

Tabela 4 – Performance dos modelos considerando apenas fluxos diferentes de zero

é um indicativo de que as estratégias de regularização que diferenciam esses dois modelos da regressão linear não resultam em diferenças significativas nos resultados, quando comparadas uma a outra.

O pior resultado dentre os modelos analisados é reportado pelo *Deep Gravity*. A arquitetura utilizada pelo modelo parece apresentar dificuldade em obter resultados que estejam em concordância com o valor real (sendo esse grau de concordância medido pelo *CPC*), porém podemos observar que os valores de *RMSE* e *MAE*, se observados de forma isolada se mantém competitivos com os demais. Uma hipótese para explicar esse comportamento é que o modelo não foi capaz de aprender corretamente os padrões dos dados, que possuem um desbalanceamento considerável (por exemplo, pares de origem e destino que possuem apenas um fluxo entre si representam aproximadamente 54% da base).

Realizando uma investigação sobre a base de dados de fluxo em que o *Deep Gravity* foi treinado em (SIMINI et al., 2021), esta hipótese do desbalanceamento ganha força ao observar um padrão mais balanceado em comparação com a distribuição da base de fluxos utilizada para os experimentos reportados neste trabalho (tabela 1), mostrado na figura 5. Além disso, outra hipótese para os resultados inferiores de performance é o tamanho da região em que o modelo está sendo treinado. Em (SIMINI et al., 2021), os experimentos foram realizados em áreas da grandeza de estados e países, o que aumenta a quantidade de setores censitários disponíveis para treino, e consequentemente a quantidade de *grids* (conceito explicado na seção 3.3) para dividir entre as bases de treino e teste. Ao utilizar o modelo numa área menor, a de uma cidade nos experimentos deste trabalho, é possível que não tenham sido fornecidos dados o suficiente para que o modelo fosse capaz de generalizar.

4.1.2 Pares origem-destino adicionando os pares cujo fluxo entre si é zero

O objetivo deste experimento foi avaliar qual o impacto de considerar nos dados de entrada do modelo também os pares de origem e destino que não possuem um fluxo associado entre si (fluxo igual a zero). O efeito dessa mudança no balanceamento dos dados é significativo, uma vez que os registros com essa característica representam aproximadamente 62% da

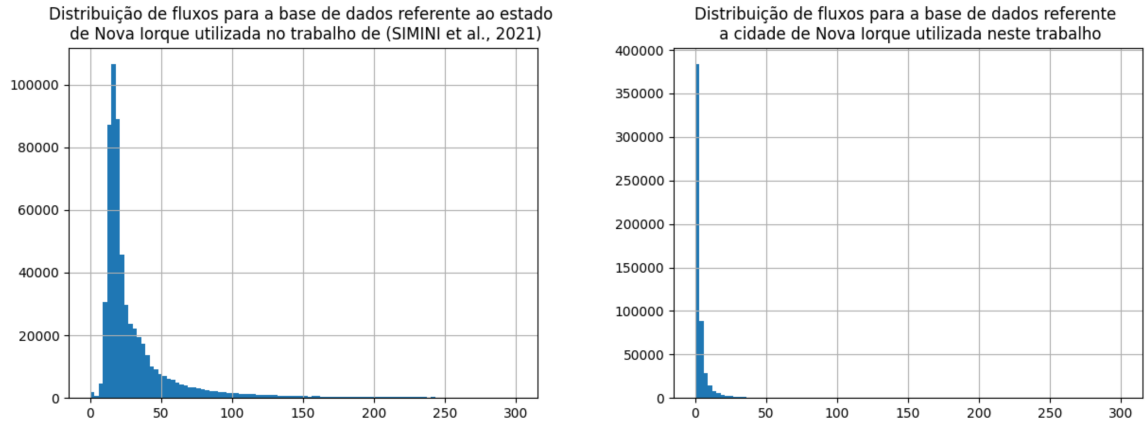

Figura 5 – Distribuição dos fluxos para a base de dados utilizada em (SIMINI et al., 2021) e a base de dados utilizada neste trabalho. Fonte: a autora (2024).

Modelo	**RMSE**	**MAE**	**CPC**
Deep Gravity	9.769	2.130	0.069
Lasso	4.238	1.580	0.414
Ridge	3.891	1.507	0.442
Random Forest	3.759	1.408	0.479
Node2vec	3.286	1.250	0.537
Gradient Boosting	3.250	1.233	0.544
GMEL	2.711	1.070	0.604

Tabela 5 – Performance dos modelos considerando também fluxos iguais a zero

base.

Se compararmos os resultados mostrados na tabela 5 com os apresentados na tabela 4, o primeiro aspecto que vem à tona é a diminuição dos resultados para as métricas propostas. Para o *RMSE* e *MAE*, é possível notar uma diminuição do erro, o que indicaria uma redução na magnitude dos resíduos obtidos entre os valores reais e preditos. Porém, o menor valor de CPC para todos os modelos mostra a dificuldade dos modelos em lidar com um maior grau de desbalanceamento de dados.

Trazendo a análise para o modelo com pior desempenho, é possível observar que o *Deep Gravity* segue apresentando dificuldade em aprender a distribuição dos dados de entrada, o que é demonstrado não apenas pelo resultado de *CPC* ainda menor que o reportado na subseção anterior, mas também pelo alto *RMSE* seguido por um *MAE* bem menos expressivo, o que passa a ideia de que o modelo teve dificuldade em prever valores mais extremos na distribuição. Isso também pode estar relacionado ao aumento no grau de complexidade dos dados de fluxo, uma vez que a arquitetura da rede neural proposta neste modelo é simples.

A partir dos resultados, é possível concluir também que o *GMEL* mantém a sua posição como melhor desempenho, apesar do impacto sofrido com a distribuição de dados con-

sideravelmente mais desbalanceada que a descrita na subseção anterior, como podemos observar pela redução do *CPC* em comparação com os resultados reportados anteriormente.

4.2 ANÁLISE DE INTERPRETABILIDADE

O segundo pilar em análise foi a interpretabilidade de cada modelo. Como os modelos em análise neste estudo possuem entradas e arquiteturas bastante diferentes, os métodos empregados para calcular a importância dos atributos de cada modelo também precisaram se adequar a essas diferenças estruturais. Modelos que recebem dados que não possuem uma estrutura espacial regular, como grafos, são desafiadores devido a sua complexidade estrutural. As informações topológicas contidas no grafo precisam ser modeladas adequadamente para que os métodos que são a elas aplicadas façam sentido, e isso é verdade também quando se tratam de interpretabilidade. Também é importante notar que, quanto mais complexa a arquitetura utilizada, mais complexa também a tarefa de propor um *framework* que resolva o problema utilizando uma única estratégia, necessitando por vezes de abordagens diferentes para obtenção de informações diferentes de um mesmo modelo. Os autores de (DUVAL; MALLIAROS, 2021) também argumentam que, devido à natureza inexplorada do campo de estudo de interpretabilidade para estes modelos, continuam em construção os conceitos do que faz uma técnica ser considerada boa, e quais abordagens chegam aos melhores resultados.

Dessa forma, esse tópico subdivide os resultados obtidos a partir do tipo da entrada que eles recebem: tabular ou grafo.

4.2.1 Modelos com entrada do tipo tabular

Para analisar os modelos que possuem entrada tabular, foi utilizado o método *SHAP*, sigla para *Shapley Additive Explanations*, cujo funcionamento é detalhado na subseção 2.3. O método calcula os valores *Shapley*, um conceito que vem da teoria dos jogos como uma medida de importância de cada atributo envolvido na predição largamente utilizada para investigações relacionadas a interpretabilidade, como mostrado nos trabalhos de (SIMINI et al., 2021; WOJTUCH; JANKOWSKI; SMUSZ, 2021; WANG; PENG; LIANG, 2022). O *SHAP* suporta análise de importância de atributos global e local, e os resultados encontrados utilizando estas duas formas de análise são mostrados a seguir.

4.2.1.1 Importância global

A importância global dá um panorama geral do impacto de cada atributo na predição da base de teste como um todo. Intuitivamente, uma característica que seria esperada nos resultados obtidos é que atributos residenciais relacionados aos setores censitários de origem (aqui entendidos também como os setores censitários residenciais) fossem conside-

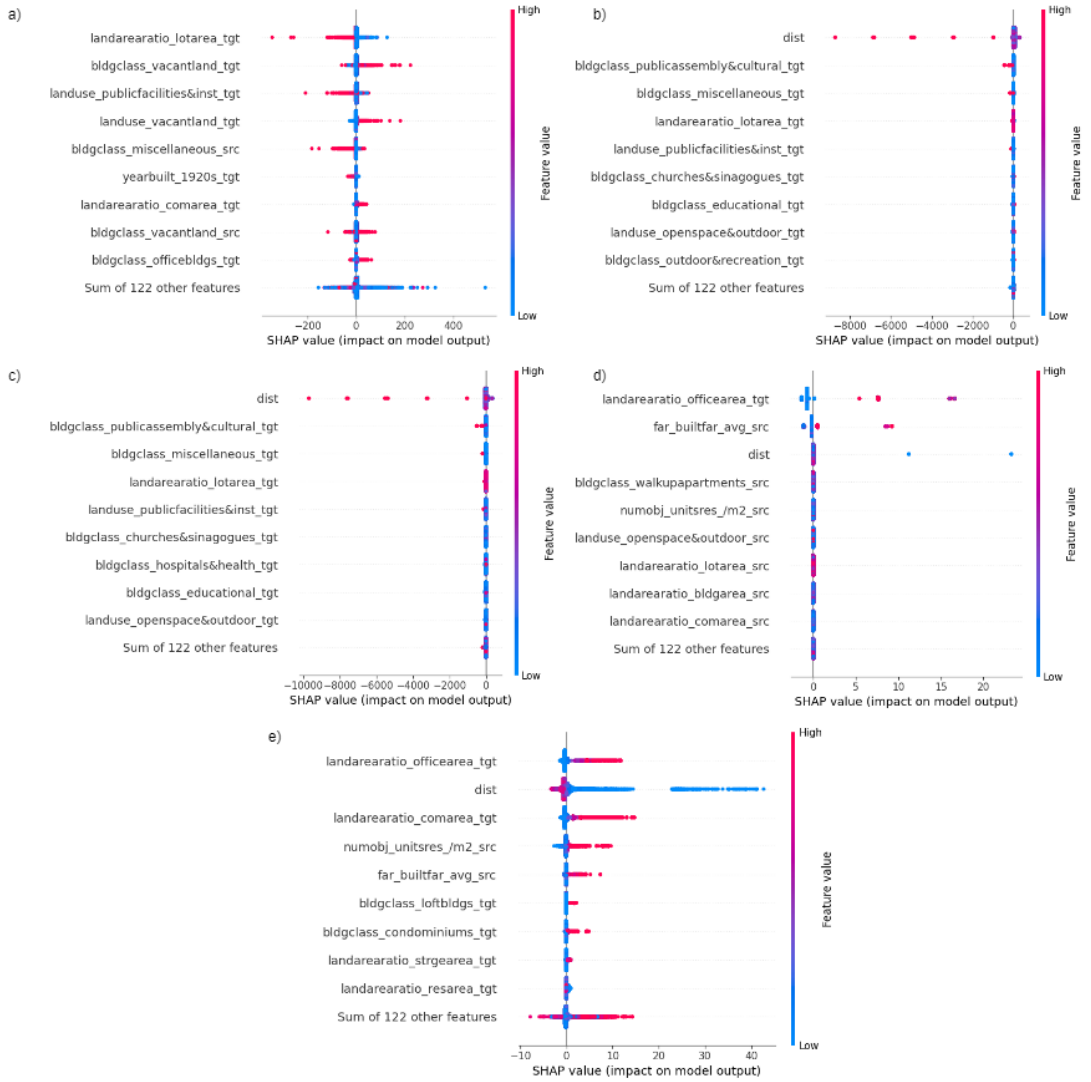

Figura 6 – Importância global dos modelos em análise: a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

rados importantes para a predição dos fluxos, e a mesma ideia se estenderia aos atributos relacionados a trabalho que fazem parte dos setores censitários de destino.

A figura 6 apresenta os resultados de importância global para todos os modelos em análise que possuem entrada tabular. A visualização escolhida para demonstrar os resultados é o *bee plot*, e deve ser interpretada a partir das seguintes considerações:

- A escala de cores representa a magnitude dos atributos. Como esse é um problema de regressão, na visualização essa informação dá um indicativo sobre o efeito na saída do modelo (aumento ou diminuição do valor final) a partir de diferentes valores de magnitude dos atributos.

- Na figura, estão dispostos os dez atributos mais importantes para a predição dos

seus respectivos modelos, em ordem decrescente de importância.

- Atributos com o final *src* são atributos que se referem ao setor censitário de origem, enquanto que atributos com o final *tgt* são atributos que se referem ao setor censitário de destino. O atributo *dist* representa a distância entre os dois setores censitários.

Considerando essas informações, é possível realizar uma análise dos resultados, iniciando pelo modelo que apresentou os piores resultados de performance. O *Deep Gravity*, que se encontra no item a) da figura, quando realizada uma comparação entre os seus atributos selecionados como mais importantes e os selecionados pelos outros modelos, é possível observar alguns pontos: em primeiro lugar, ele foi o único dentre os modelos que não elencou a distância como um atributo relevante. Na literatura é possível identificar desde o início do estudo da predição de fluxos de origem-destino (no trabalho de (ZIPF, 1946)) a proposta da distância como um atributo de grande importância na modelagem dos fluxos de deslocamento, portanto seria esperado que esse atributo de fato estivesse entre os mais relevantes.

Outro ponto de análise é o padrão aparentemente aleatório de atributos que o modelo elegeu como mais importantes, o que foge da intuição estabelecida no início desta seção. Os dois atributos relacionados a origem são o *bldgclass-miscellaneous* e o *bldgclass-vacantland*, o primeiro se referindo a construções diversas (tanto construções que podem ser associadas a um perfil residencial, como uma piscina, por exemplo, como também a prédios institucionais), e o segundo a terrenos em geral.

Já com relação aos atributos relacionados ao destino, é possível identificar o *bldgclass-officebldgs* e o *landarearatio-commarea* como atributos reportados como os mais importantes, mas não em posição de destaque.

Analisando as regressões lineares Ridge e Lasso (itens b e c da figura), é possível observar um padrão bastante similar entre os dois modelos: ambos elegeram a distância como atributo mais relevante, e ambos tem dentre seus dez atributos mais importantes (com exceção da distância) apenas atributos relacionados aos setores censitários de destino no fluxo. Para a distância, a interpretação obtida a partir dos resultados mostrados é que valores altos do atributo (ou seja, distâncias maiores entre as regiões) causam redução na saída do modelo (fluxos menores entre essas regiões), o que conversa com o primeiro modelo proposto na literatura para modelar estes fluxos, o *Gravity Model*, que supõe uma relação inversamente proporcional entre fluxo e distância. Analisando os resultados, parece que para as duas estratégias de regularização a regressão foi tratada praticamente como univariada, uma vez que a importância dos atributos relacionados a infraestrutura parece ser irrisória em relação ao atributo distância, que aparece no topo do *ranking*. Importante notar que, a partir desses resultados, mesmo formulando o problema de forma mais simples (supondo uma solução linear para a predição de fluxos pendulares, nesse caso), é possível alcançar resultados interessantes.

Para os modelos baseados em árvore, por outro lado, é possível observar padrões que se aproximam mais da intuição proposta no início da seção. Em primeiro lugar, a distância é listada como um dos atributos de maior importância. Mas além disso, também é possível notar uma divisão entre os atributos mais importantes para os setores de origem e destino, onde os atributos de origem se referem a características importantes para os setores censitários de origem (*bldgclass-walkupapartments-src*, *numobj-unitsres-/m2-src*, *landarearatio-bldgarea-src*, para citar alguns), e o mesmo ocorre para os atributos de destino (como *landarearatio-officearea-tgt*, *landarearatio-comarea-tgt*, para citar alguns). Interpretando de forma mais aprofundada os resultados apresentados para o *Random Forest*, é possível chegar a algumas conclusões: para a distância, valores menores dos atributos causam um aumento na saída do modelo, o que está em concordância com os resultados obtidos para os modelos lineares. O *landarearatio-officearea-tgt*, por sua vez, mostra valores altos do atributo associados a saídas do modelo com valores mais altos, indicando que a alta densidade de escritórios no setor censitário de destino é um fator representativo para o aumento dos fluxos para aquele setor. Já o atributo *far-builtfar-avg-src*, que é uma média da área construída por lote para o setor censitário, indica que valores mais altos do atributo (valores que representam setores censitários com alta densidade de área construída) impactam no aumento da saída do modelo, o que pode ser interpretado como um aumento na infraestrutura disponível numa região como um fator de influência para o aumento dos fluxos que saem de dessa região.

Para o *Gradient Boosting*, quase todos os atributos rankeados no top 10 possuem alguma expressividade no que diz respeito aos valores *SHAP*. Os atributos discutidos para *o Random Forest* também se encontram como relevantes para o GBRT, possuindo mais instâncias expressivas que no modelo anterior (considerando a distribuição das instâncias observada no eixo que mede os valores *SHAP*), porém com a mesma interpretação. Além deles, o *landarearatio-comarea-tgt* segue a mesma linha do *landarearatio-officearea-tgt*, indicando que alta densidade de áreas comerciais também estaria relacionada ao aumento do valor de saída do modelo. Para os atributos *bldgclass-loftbldgs-tgt* e o *bldgclass-condominiums-tgt*, os resultados indicam valores elevados destes atributos gerando saídas mais elevadas do modelo, indicando que setores censitários que possuam infraestruturas de condomínios e lofts (estruturas que possuem perfil residencial ou comercial) acabam por receber um volume maior de fluxos de destino.

Os modelos lineares possuem dificuldade em lidar com relações complexas entre os dados. Modelos baseados em árvore, por outro lado, tendem a lidar bem com relações não lineares, e esta característica poderia explicar o surgimento de outros atributos com elevada importância (em comparação com os modelos lineares) associados a valores superiores nas três métricas utilizadas para avaliar a performance dos modelos. As primeiras teorias de mobilidade que buscaram modelar os fluxos de deslocamento possuem em comum a definição da distância como um atributo decisivo para a predição dos fluxos,

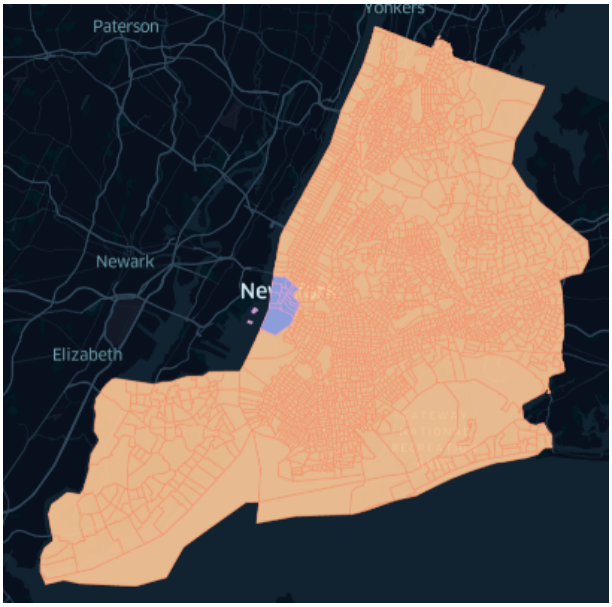

Figura 7 – Localização do distrito financeiro de *Manhattan*. Fonte: a autora (2024).

entretanto, a utilização de abordagens mais recentes mostra que é possível aumentar a capacidade preditiva dos padrões de fluxo entre regiões utilizando outras fontes de dados, conforme mostrado nos experimentos discutidos neste capítulo.

Nos tópicos a seguir, serão discutidos alguns experimentos realizados a partir dos dados de importância global obtidos, com o objetivo de aprofundar o entendimento dos resultados.

- Caso de uso: distrito financeiro de *Manhattan*

  Buscando trazer essa avaliação de importância global para o mundo real, foi realizada uma análise considerando os setores censitários que fazem parte do distrito financeiro de Manhattan. São ao todo quinze setores, e sua localização é mostrada na figura 7.

  Para este estudo foram empregados as mesmas regras de visualização descritas acima para análise da importância global dos atributos, entretanto, realizando uma etapa a mais de filtragem, onde foram selecionados três casos de análise:

  – **Caso 1: Filtra a base de dados por setores censitários pertencentes ao distrito financeiro na origem dos pares origem-destino de fluxos (fluxos gerados por pessoas que moram no distrito financeiro e trabalham fora dele)**:

    Analisando os resultados, é possível observar alguns pontos: as regressões de Lasso e Ridge se mantém bastante similares nos atributos selecionados como os mais importantes, e também mantém a escolha dos atributos relacionados ao setor censitário de destino como os mais relevantes (além da distância). Dentre os atributos relevantes podemos observar um valor (absoluto) do *SHAP*

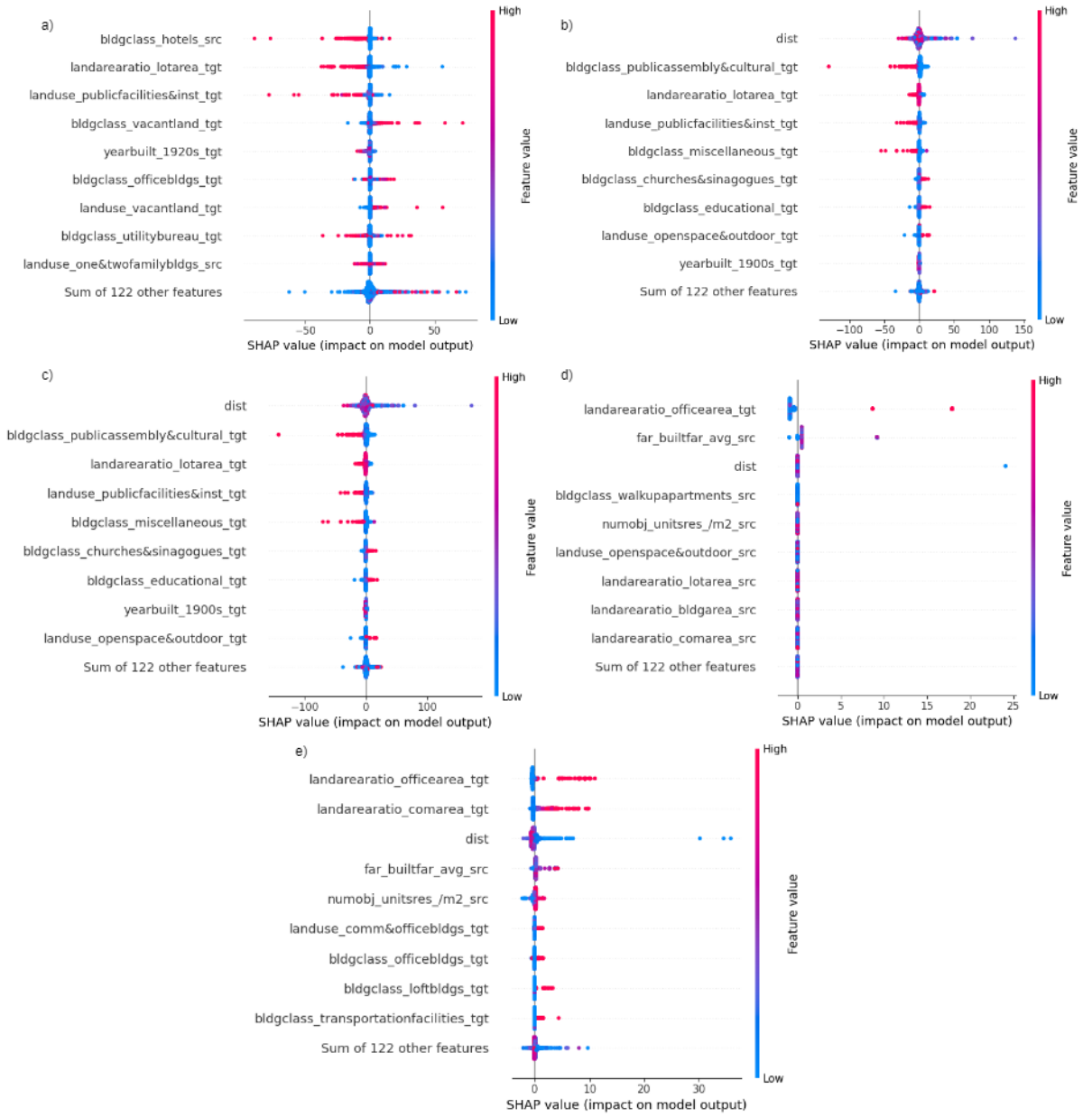

Figura 8 – Importância global dos modelos considerando apenas os setores censitários que fazem parte do distrito financeiro na origem dos fluxos: a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge, d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

mais elevado, além da distância, para os atributos *landuse-publicfacilities&inst*, *bldgclass-publicassembly&cultural* e *bldgclass-miscellaneous*, fazendo os dois últimos atributos parte do primeiro. Essa categoria de uso do solo (atriuto que inicia com *landuse*) possui uma característica de empreendimentos construídos para atender a população: seja como uma fonte de lazer (museus, piscinas, bibliotecas), religião (templos religiosos, conventos) ou centros de apoio (asilos, orfanatos), indicando certa relevância destas infraestruturas para a predição dos fluxos neste grupo.

Para os modelos baseados em árvore, quanto a performance do *Random Forest*, é possível observar uma expressividade baixa dos valores *SHAP*. Entretanto, como atributo mais relevante, é mostrado um atributo relacionado a área de escritórios no setor censitário de destino, o que é poderia indicar fluxos pendulares de quem mora no distrito financeiro para outros centros comerciais. Também é importante notar que, com exceção desse atributo e da distância, todos os outros presentes estão relacionados ao setor censitário de origem, e mais, relacionados a informações residenciais: estatísticas da quantidade de unidades residenciais, informações sobre espaços abertos (parques, por exemplo), apartamentos, área comercial, entre outros, o que poderia ser um indicativo da estrutura residencial existente no distrito financeiro. O *Gradient Boosting*, por sua vez, demonstra uma maior expressividade dos valores *SHAP* nos resultados, e segue um perfil similar ao descrito para o *Random Forest*, porém apresentando uma quantidade maior de atributos relacionados ao setor censitário de destino.

– **Caso 2: Filtra a base de dados por setores censitários pertencentes ao distrito financeiro no destino dos pares origem-destino (fluxos gerados por pessoas que moram fora do distrito financeiro e trabalham nele)**:

Na figura 9, é possível observar que as regressões Lasso e Ridge (itens b e c) se mantém com padrões similares, e também mantém a escolha dos atributos relacionados ao setor censitário de destino como os mais relevantes. A distância se mostra o atributo com mais expressividade dos valores *SHAP*, porém uma expressividade majoritariamente negativa. A leitura que pode ser feita a partir desse resultado, com relação ao atributo distância, é que valores altos de distância influenciam na diminuição da saída do modelo (o fluxo predito), e esse padrão se repete em vários momentos das análises de interpretabilidade realizada, o que condiz com o que foi proposto inicialmente no *Gravity Model* (ZIPF, 1946).

No *Random Forest* (item d na figura), o resultado mantém os mesmos atributos do caso de uso anterior. Os dois atributos mais expressivos na visualização

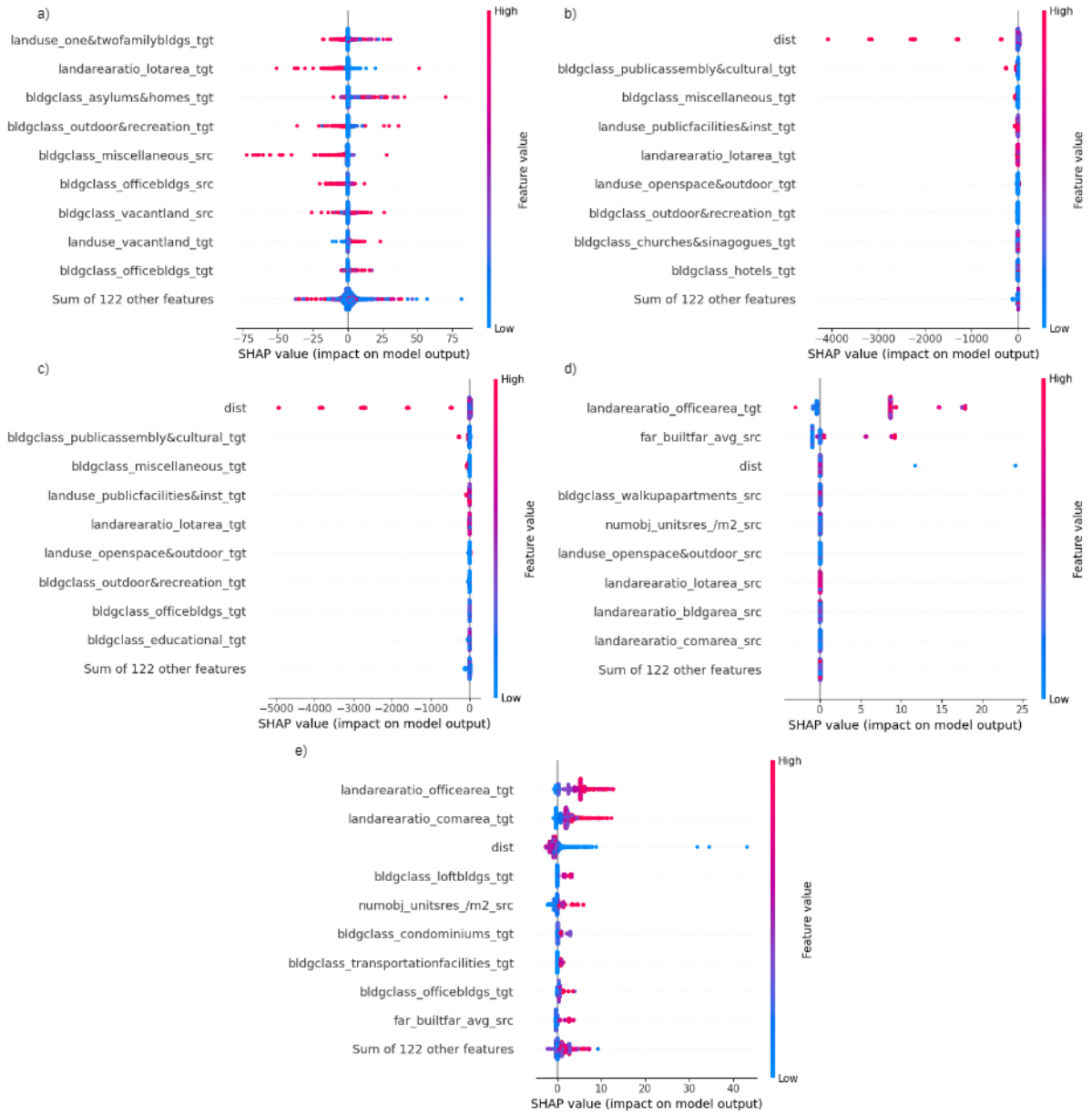

Figura 9 – Importância global dos modelos considerando apenas os setores censitários que fazem parte do distrito financeiro no destino dos fluxos: a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

seguem padrões similares: valores mais altos do atributo na base de dados representam valores maiores de fluxo predito. Já para o *Gradient Boosting* (item e), é possível observar uma quantidade maior de atributos influenciando significativamente a predição, em comparação com o *Random Forest*, porém o padrão segue similar ao apresentado na análise anterior.

– **Caso 3: Filtra a base de dados por setores censitários pertencentes ao distrito financeiro tanto na origem quanto no destino dos pares origem-destino (fluxos gerados por pessoas que trabalham e moram no distrito financeiro)**:

Nesse experimento são observados os resultados para os pares de fluxos origem-destino que tem o setor censitário de origem e de destino dentro do distrito. É interessante pontuar que as bases de dados são referentes ao ano de 2015, e que houve uma mudança nos padrões de deslocamento causada por uma mudança na vizinhança do distrito na pandemia de COVID-19, como discutido em (HAAG, 2023), com uma migração de pessoas de áreas de atuação não relacionadas ao mercado financeiro para o distrito. Entretanto, em 2015, seria esperado observar um perfil mais empresarial.

Na figura 10 é possível notar que, especialmente no que diz respeito aos resultados reportados pelos modelos baseados em árvore (que são os modelos de entrada tabular que reportaram os melhores resultados, e, portanto, são uma boa escolha de resultado para o qual é interessante olhar mais atentamente), os atributos mais importantes são os mesmos reportados nos outros experimentos, o que mostra pouca diferença entre os padrões de predição nos três casos de uso.

- Similaridade geográfica entre os vetores de importância de atributos

Nesse experimento, o objetivo é analisar padrões de similaridade entre os vetores de importância gerados pelo *SHAP* para cada setor censitário. Obtendo este vetor para cada elemento da base de dados de teste, foram realizados dois agrupamentos: o primeiro, calculando o vetor de importância médio para cada setor censitário de origem; o segundo, calculando o vetor de importância médio para cada setor censitário no destino. Esses agrupamentos resultaram em duas análises de similaridade geográfica, uma focada na origem do par origem-destino do fluxo e outra no destino do par origem-destino do fluxo.

Como estratégia para obter uma visão geral da proximidade entre os vetores resultantes, foi utilizado o *t-SNE*, proposto em (MAATEN; HINTON, 2008), uma ferramenta de visualização de dados de alta dimensionalidade que busca manter na representação a estrutura semântica entre os pontos que estão sendo representados. Para esse experimento, a medida de similaridade utilizada foi o cosseno. Para a re-

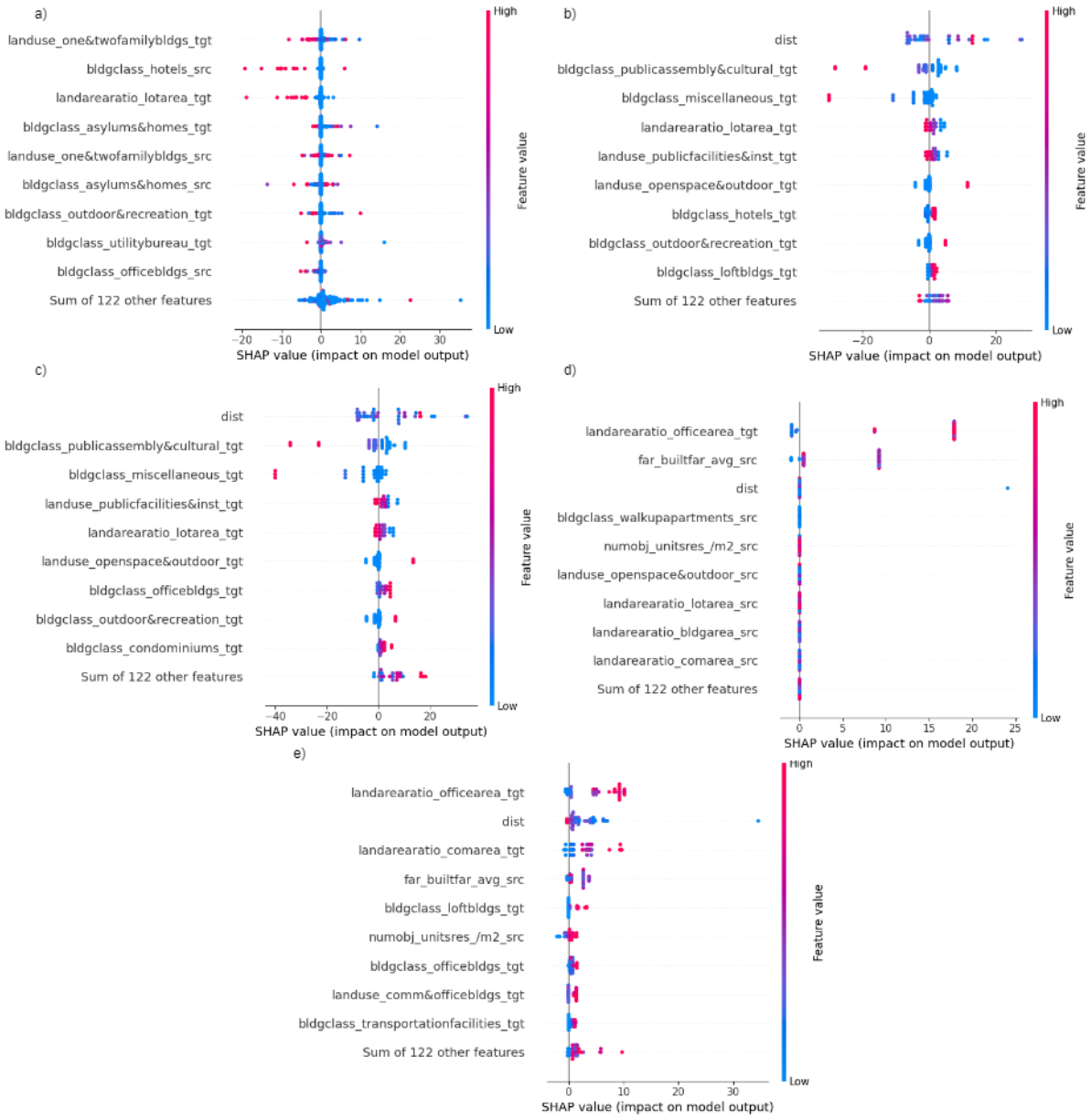

Figura 10 – Importância global dos modelos considerando apenas os setores censitários que fazem parte do distrito financeiro na origem e destino dos fluxos: a) *Deep Gravity*. b) Regressão Lasso, c) Regressão Ridge, d) *Random Forest*, e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

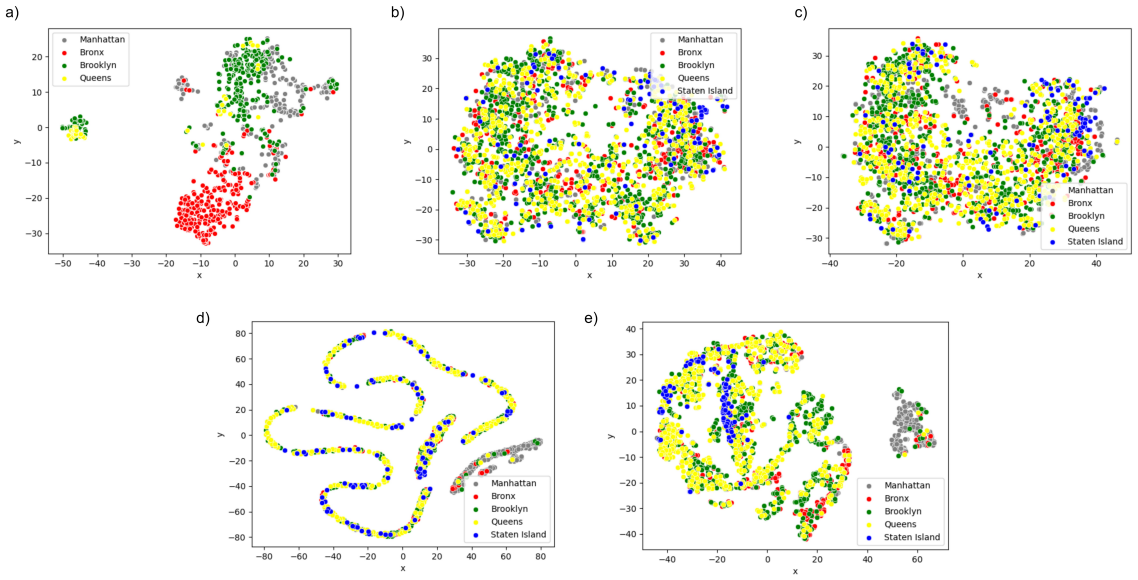

Figura 11 – Representação dos *boroughs* de Nova Iorque utilizando *t-SNE* para os setores censitários de origem. a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge, d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

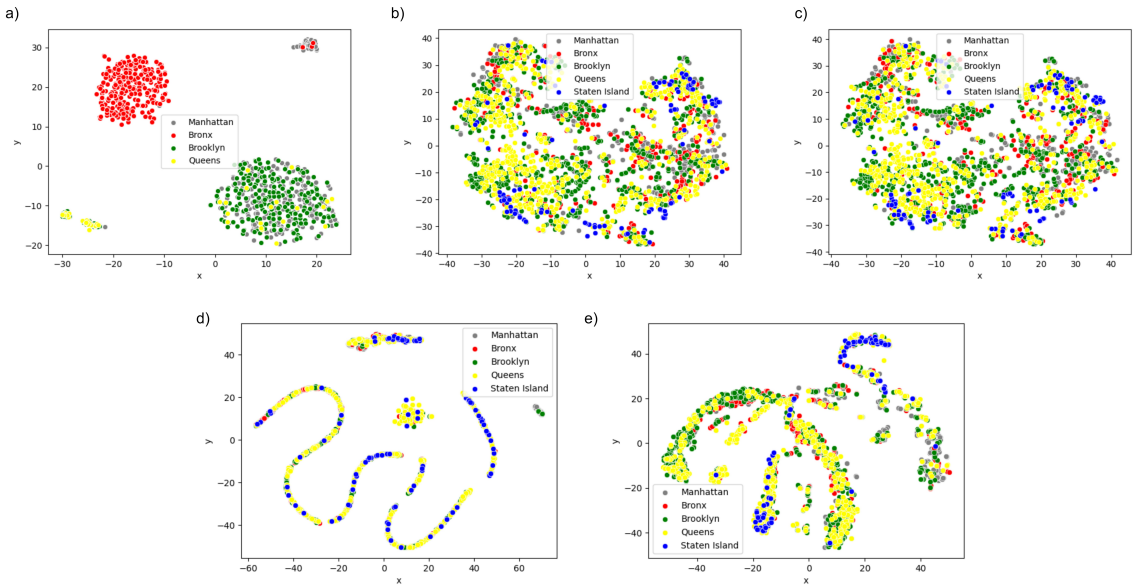

Figura 12 – Representação dos *boroughs* de Nova Iorque utilizando *t-SNE* para os setores censitários de destino. a) *Deep Gravity*, b) Regressão Lasso, c) Regressão Ridge, d) *Random Forest*, e) textitGradient Boosting Regressor. Fonte: a autora (2024).

alização desta análise, será investigada a similaridade dos vetores com relação aos *boroughs* da cidade, que são os distritos existentes em Nova York, descritos neste trabalho na subseção 3.1.1.

Nas figuras 11, 12, é possível observar o resultado do *t-SNE* considerando os setores censitários de origem e destino, separados pelo *borough*. Em geral, os resultados

Modelo	RMSE	MAE	CPC
Deep Gravity	10.185	3.019	0.035
Lasso	6.852	2.736	0.593
Ridge	6.457	2.651	0.607
Random Forest	6.381	2.535	0.624
Gradient Boosting	5.641	2.070	0.693

Tabela 6 – Performance dos modelos após filtragem dos atributos baseada no *VIF*

para os modelos mostram que os distritos, quando utilizando uma representação 2-D, não são bem separados em grupos, o que é um indicativo de que não haveria um padrão claro de vetor de importância de atributos por *borough*. Uma exceção que é possível observar acontece para os setores censitários de origem referentes ao *Gradient Boosting Regressor* (figura 11, item e), onde é possível observar uma separação significativa do *borough* de *Manhattan*.

Com a finalidade de investigar de forma mais aprofundada os resultados obtidos, foi realizada uma segunda etapa da análise, onde, utilizando os setores censitários que fazem parte do distrito financeiro de *Manhattan*, foi calculada a similaridade de cosseno entre o vetor que representa a média da importância dos atributos para os setores censitários do distrito financeiro, e a média dos vetores de importância dos atributos dos demais setores censitários. Os resultados são mostrados nas figuras 13, 14, e complementam o que foi observado anteriormente. É possível observar nos resultados gerados agrupando os vetores de importância dos setores censitários de origem, que os modelos baseados em árvore possuem um padrão de similaridade específico para a região de *Manhattan*. O *Gradient Boosting Regressor* em específico mostra um padrão de similaridade se espalhando radialmente, o que estaria em acordo com a os resultados discutidos acima, que mostram um padrão emaranhado entre a maior parte dos *boroughs*.

- Efeito das correlações entre atributos na interpretabilidade dos modelos

  Neste experimento, o objetivo é avaliar o impacto da correlação entre atributos na importância dos atributos calculada para cada modelo. Quando o objeto de estudo é a interpretabilidade, é importante levar em consideração um problema largamente descrito na literatura, a multicolinearidade. Bases de dados que possuem esta característica geralmente não possuem estabilidade no grau de importância que designam para cada atributo. Isto ocorre porque, se o valor de um atributo depende de outro (ou outros), uma mudança nele também afetará os que estão altamente correlacionados com ele. Os métodos que aqui foram utilizados para avaliar a interpretabilidade dos modelos partem do princípio de que cada atributo é independente, e, portanto, não seguir essa premissa pode levar a resultados diferentes.

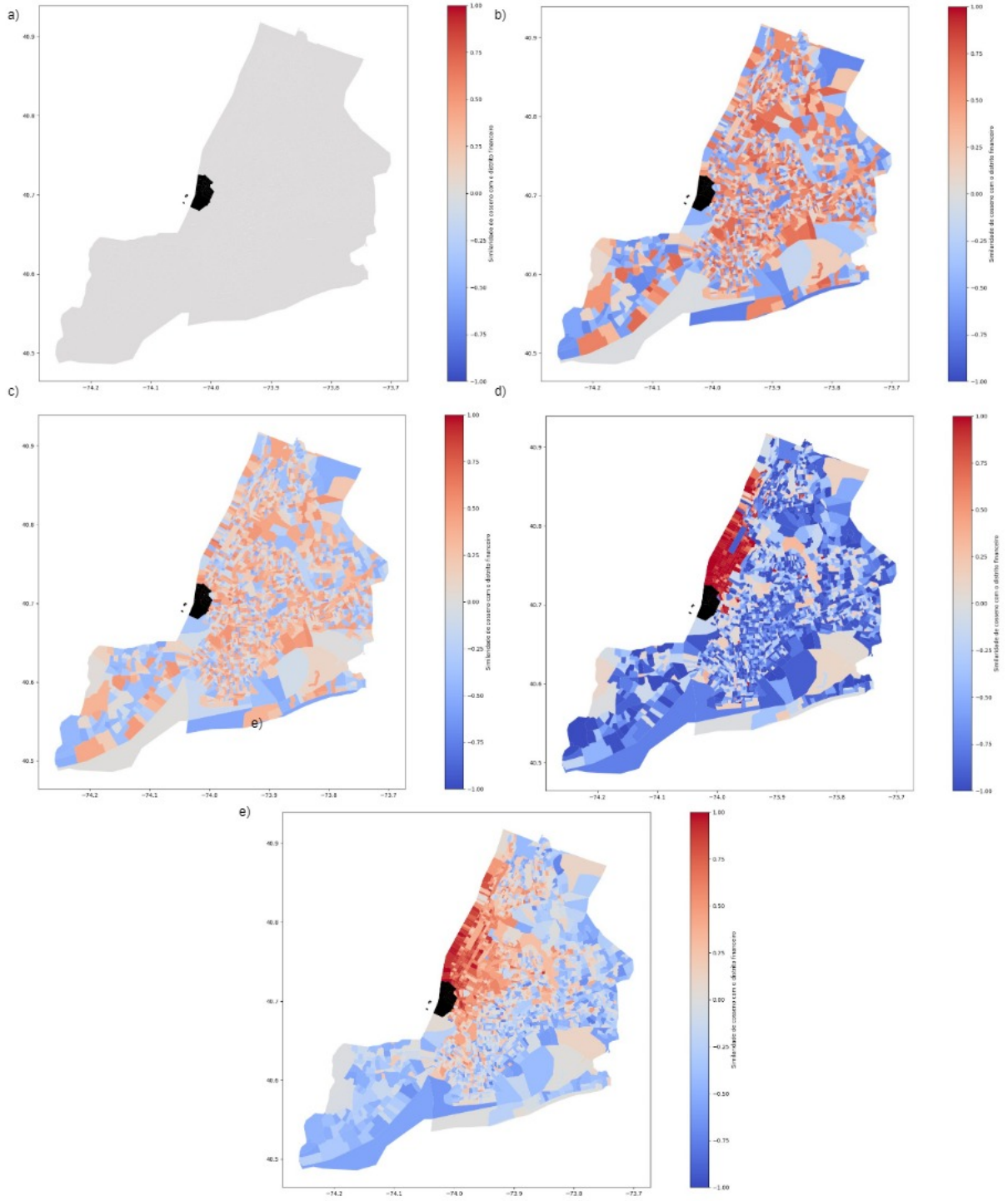

Figura 13 – Similaridade de cosseno calculada entre o distrito financeiro de *Manhattan* (em preto) e os demais setores censitários, para os setores censitários de origem nos pares de fluxos origem-destino. a) *Deep Gravity*, b) Regressão Lasso, c) Regressão Ridge, d) *Random Forest*, e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

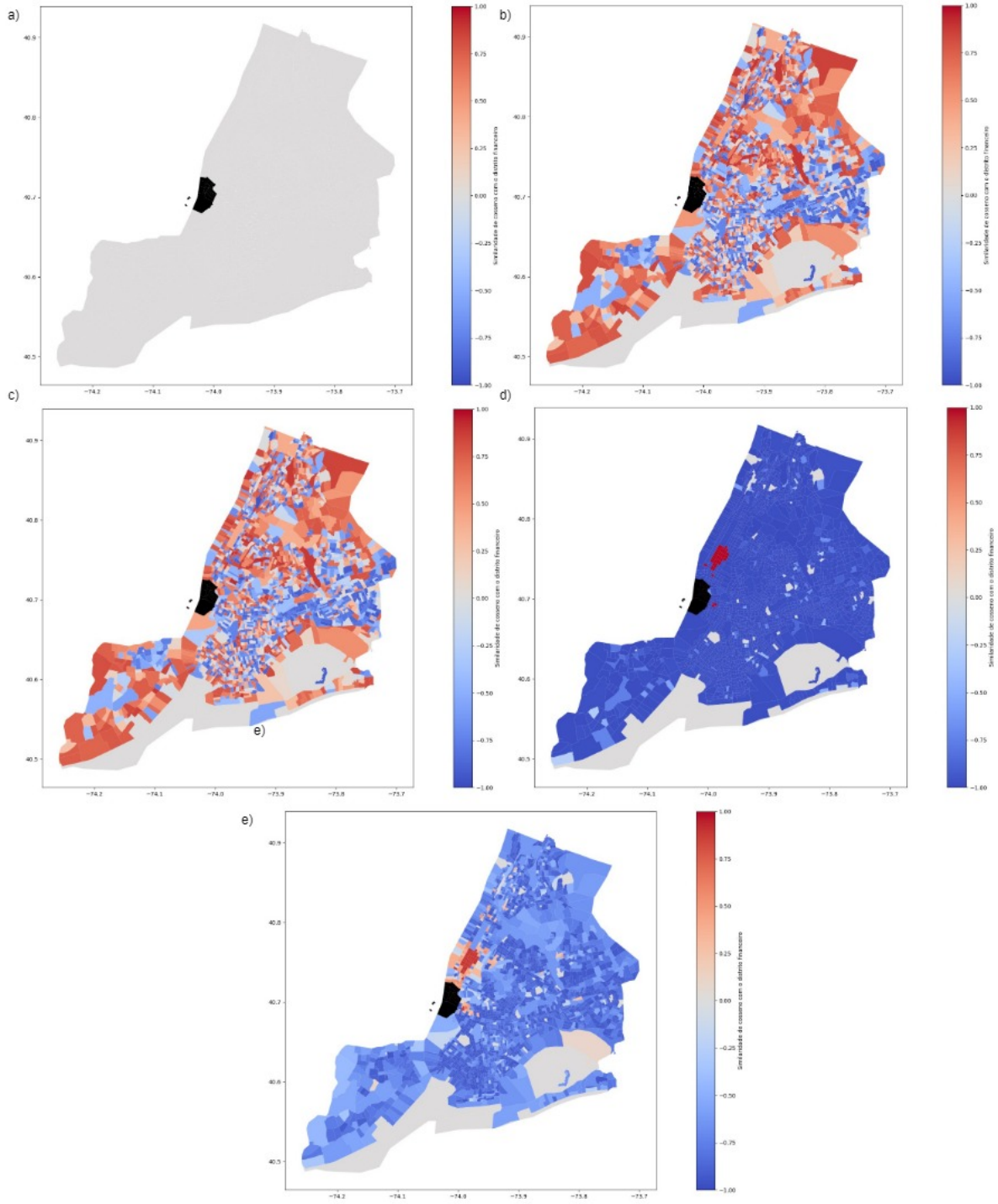

Figura 14 – Similaridade de cosseno calculada entre o distrito financeiro de *Manhattan* (em preto) e os demais setores censitários, para os setores censitários de origem no fluxo origem-destino. a) *Deep Gravity*. b) Regressão Lasso, c) Regressão Ridge, d) *Random Forest*, e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

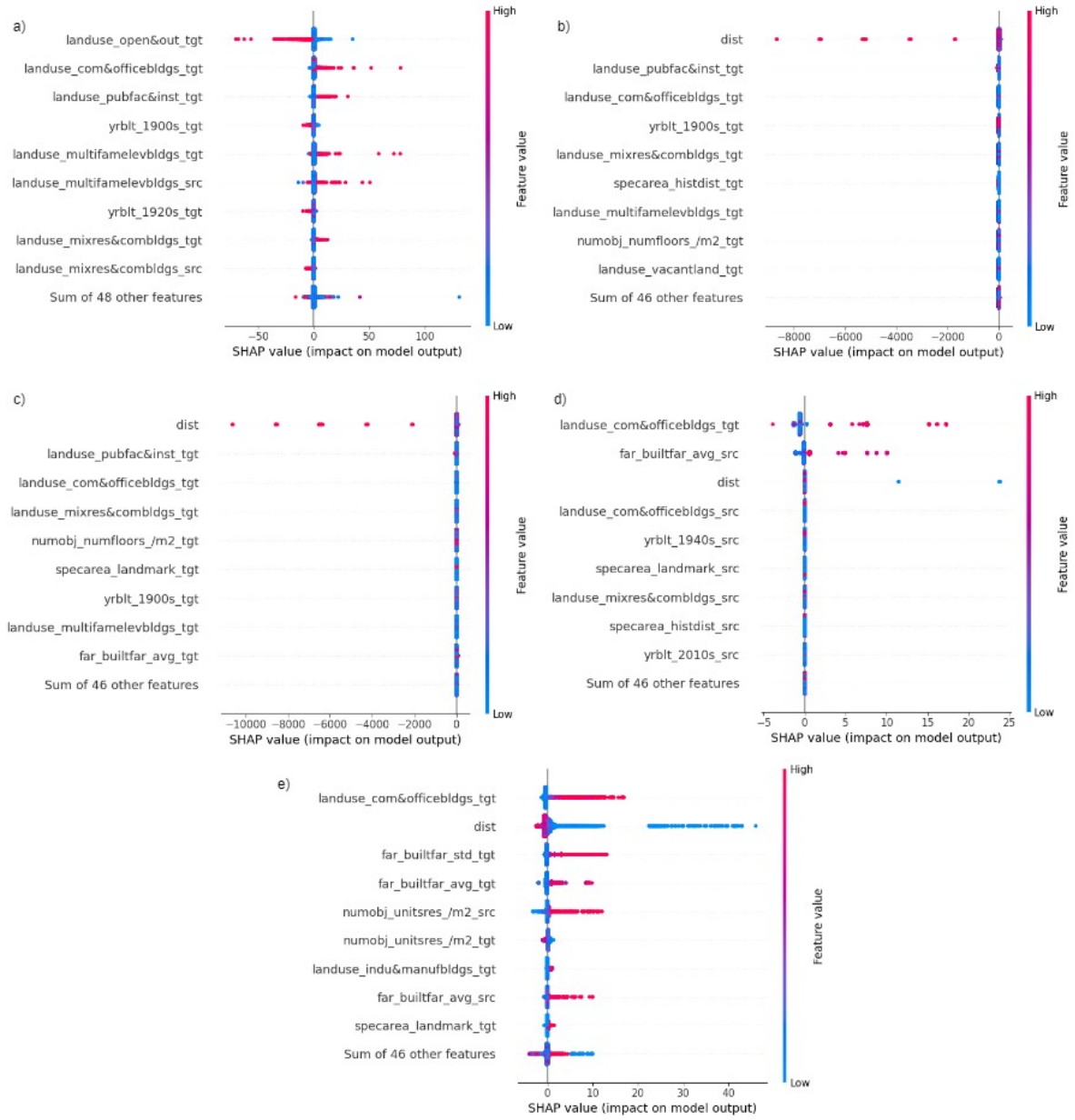

Figura 15 – Importância global dos modelos em análise, utilizando os atributos filtrados utilizando o *VIF*: a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

Para detectar os atributos que sofrem de multicolinearidade, foi utilizado o *VIF* (*variance inflation error*), uma técnica que estima o quanto que a variância de um coeficiente de regressão é inflada devido a multicolineariedade. O *VIF* permite a realização desse cálculo considerando o impacto de múltiplas variáveis simultâneamente, enquanto que o cálculo da correlação entre atributos leva em conta apenas o impacto um par de atributos por vez. A literatura utiliza geralmente 10 como um limiar para o *VIF*, como mostrado nos trabalhos de (JUARTO, 2023; CHENG et al., 2022), e portanto foi utilizado este valor como limiar. Dos 65 atributos de infraestrutura gerados após processamento da base de dados *PLUTO*, 27 ficaram abaixo deste limiar, e foram filtrados. Os atributos que foram retirados nesse filtro foram todos os atributos do tipo *bldgclass*, todos os atributos do tipo *landarearatio*, e os atributos *numobj-unitstotal-/m2*, *numobj-numbldgs-/m2*. Os atributos restantes foram fornecidos como entrada para os modelos, que foram retreinados. Na tabela 6 estão os resultados obtidos para as três métricas utilizadas neste estudo.

No resultado dos modelos treinados foi, então, aplicado o *SHAP* para investigação da importância global. Os resultados são mostrados na figura 15. Comparando estes com os mostrados na figura 6, é possível observar uma tendência a manter os mesmos padrões observados para os modelos treinados com todos os atributos disponíveis. O atributo distância, por exemplo, mantém seu posto no *ranking* de importância para todos os modelos. Também é possível observar os padrões de atributos para as regressões de Ridge e Lasso, que mantém a importância focada nos atributos voltados aos setores censitários de destino. Em geral, os modelos parecem manter a importância para informações similares quando comparando os dois resultados, por exemplo, no *Random Forest*, que tem como seu atributo mais importante o *landarearatio-officearea-tgt* na importância global considerando todos os atributos (figura 6), e que nesse experimento foi substituido pelo *landuse-comm&officebldgs-tgt* como atributo mais importante, ambos trazendo informações sobre a parte comercial no setor censitário de destino.

Existe também um padrão novo que surge para esse experimento, entretanto, que é o dos atributos do tipo *yearbuilt* como atributos relevantes para as regressões de Ridge e Lasso e o *Random Forest*. Atributos relacionados aos distritos e marcos históricos também começam a ser vistos dentre os atributos mais importantes em todos os modelos, com exceção do *Deep Gravity*.

4.2.1.2 Importância local

O *SHAP* também permite realizar análises locais de interpretabilidade. Nesse sentido, é possível avaliar um exemplo (no caso deste estudo, um par origem-destino de setores censitários com um fluxo entre eles associado), e definir quais atributos foram responsáveis

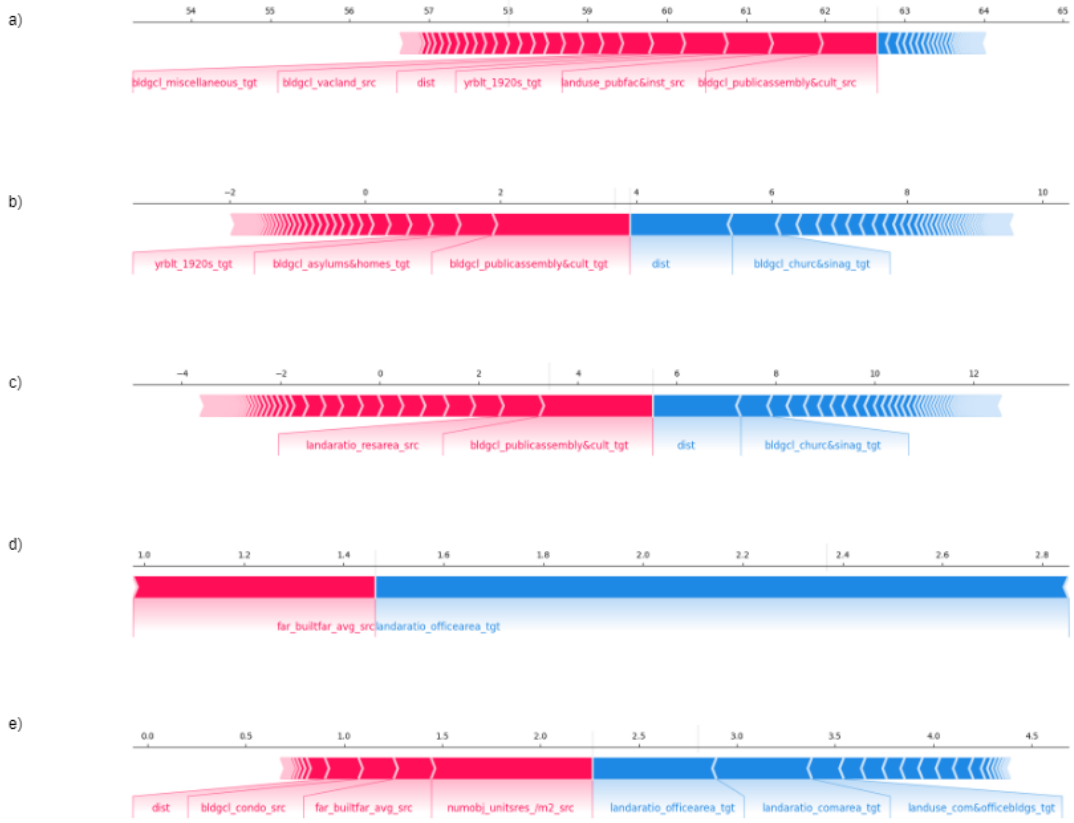

Figura 16 – Importância local dos atributos para o par origem-destino (2042100, 2028600). a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

pelo resultado predito, e o quanto cada um deles foi responsável.

A visualização utilizada para essa análise é o *force plot* e representa o processo até chegar ao valor predito sob um *layout* de adição de forças. Para a análise, foram avaliados alguns casos de uso, descritos abaixo:

- **Caso 1: *Deep Gravity* tem a melhor performance dentre os modelos**

  Esta análise é referente ao fluxo do setor censitário 2042100 para o 2028600, ambos localizados no *borough* do *Bronx*. A quantidade real de fluxos entre eles é 58, porém nos resultados mostrados na figura 16, é possível observar uma dificuldade geral dos modelos em se aproximar dele. Com exceção do *Deep Gravity*, o valor máximo das predições não chegou a 6, o que implica em valores elevados de resíduos. Entretanto, existem alguns atributos que se repetem para o *Deep Gravity*, Regressão Ridge e Regressão Lasso, como (além da distância), *yrblt-1920s-tgt* (quantidade de construções realizadas entre os anos de 1920 e 1929), *bldgcl-publicassembly&cult-tgt* (construções de interesse público) e *bldgcl-churc&sinag-tgt* (construções religiosas). Entretanto, investigando o setor censitário de destino dos fluxos, foram identificados

algumas estruturas relacionadas a medicina presentes: um hospital, uma faculdade de medicina, um instituto de pesquisa sobre câncer.

Já os modelos baseados em árvore realizam a predição desse par origem-destino a partir de atributos que trazem características residenciais e comerciais de forma mais explícita, como é possível observar a partir dos atributos relacionados a área de escritórios, unidades residenciais, área comercial, entre outros.

A partir da figura, o exemplo parece se tratar de um outlier, uma vez que o resultado que mais se aproxima do valor real de fluxo pertence a um modelo que não apresentou nas análises anteriores uma boa performance.

- **Caso 2: *Gradient Boosting Regressor* performa melhor dentre os modelos**

  Neste exemplo, o par origem-destino pertence a *Manhattan* (origem no setor censitário 1002602 e destino no setor censitário 1002100), e o fluxo real entre esse par é 23. Na figura 17 é possível observar que dentre todos os modelos, o que se aproximou mais da predição correta foi o *Gradient Boosting Regressor*. Assim como no exemplo anterior, os modelos baseados em árvore se mantém priorizando os atributos que tratam explicitamente das características residenciais e comerciais, com exceção do atributo *bldgcl-hotels-tgt*, que aparece como um atributo relevante não apenas no *Gradient Boosting Regressor*, como também na Regressão Lasso.

4.2.2 Modelos com entrada do tipo grafo

Nesta subseção foi realizada uma análise da interpretabilidade dos modelos cuja entrada é um grafo. Para este trabalho, eles são o *GMEL* e o *Node2vec*. Devido a sua arquitetura e natureza diferentes em relação aos analisados na subseção 4.2.1, a utilização de uma ferramenta como o *SHAP* se torna inviável por não suportar a arquitetura destes modelos. Tentativas foram realizadas para adaptação de técnidas que propõem o *SHAP* para grafos, entretanto, houveram dificuldades na realização desta adaptação, o que fez com que esta atividade precisasse ser postergada para trabalhos futuros.

Para que fosse possível ainda sim investigar de alguma forma os resultados obtidos pelos modelos, foi realizada uma análise voltada aos vetores de representação gerados por eles, conforme descrito a seguir.

4.2.2.1 Avaliação das representações geradas

Nesse experimento, o objetivo é avaliar se as representações geradas pelos modelos baseados em grafo são capazes de aproximar de forma eficiente grupos com características similares. Para atingir este objetivo, foram utilizados novamente os setores censitários que fazem parte do distrito financeiro de *Manhattan*, que possui um perfil empresarial.

Aqui foi realizada uma avaliação comparativa entre os atributos considerando três representações diferentes: a base de atributos de infraestrutura do *PLUTO*, que é a entrada

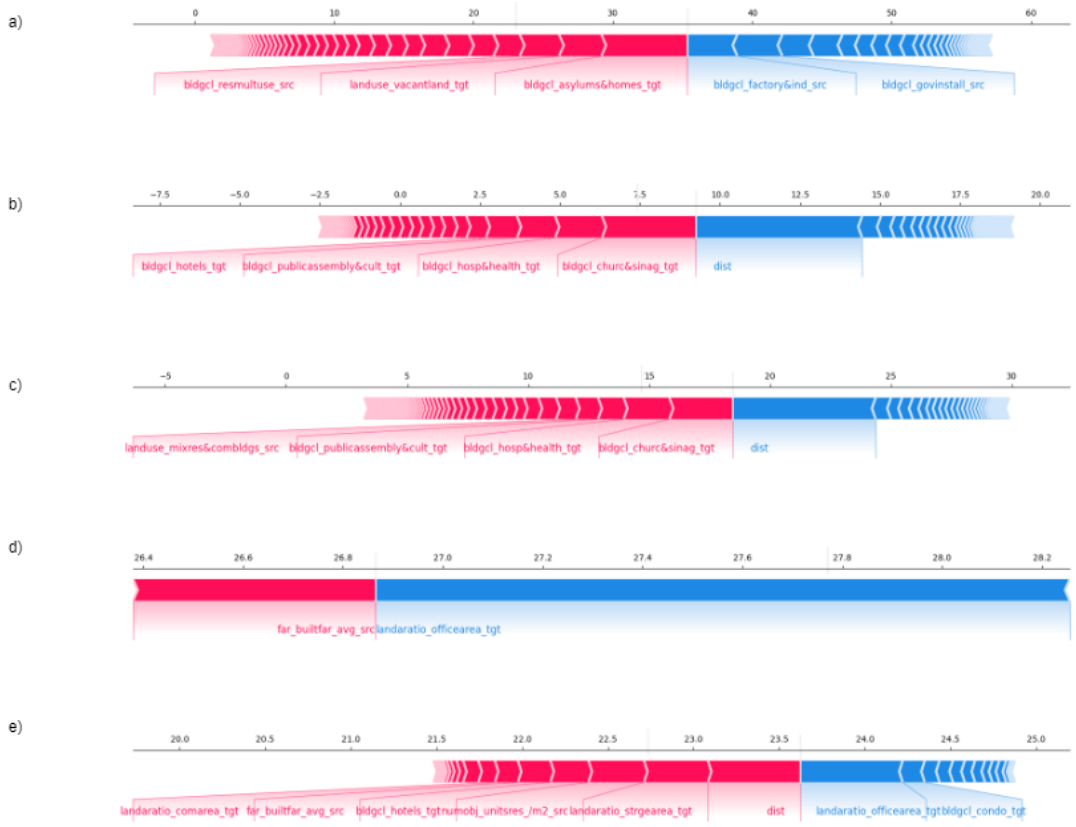

Figura 17 – Importância local dos atributos para o par origem-destino (1002602, 1002100). a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Gradient Boosting Regressor*. Fonte: a autora (2024).

para os modelos que utilizam dados tabulares, e a entrada para os modelos cuja entrada é um grafo para gerar as representações; as representações geradas pelo *GMEL* (são duas, a de origem e de destino, uma vez que o *GMEL* utiliza uma *GAT* separada para gerar cada uma), e a gerada pelo *Node2vec*. O *t-SNE* (MAATEN; HINTON, 2008) foi utilizado para reduzir a dimensionalidade das representações para o 2-D, para que fosse possível visualizar o comportamento dos vetores que representam cada setor censitário. A métrica utilizada para calcular a matrix de distância do *t-SNE* foi o cosseno.

Com o objetivo de avaliar se as representações eram capazes de separar os setores censitários de perfil mais comercial dos que possuem perfil mais residencial, foi realizado um agrupamento utilizando o algoritmo *K-Means*. A intuição seria que, aplicando o algoritmo para que os setores censitários se dividissem em 2 grupos (k=2), os setores censitários deveriam ser agrupados no grupo em que possuem maior similaridade de características, e todos os setores censitários pertencentes a um distrito deveriam pertencer ao mesmo grupo.

Os resultados são mostrados na figura 18. Os setores censitários que fazem parte do distrito financeiro estão representados pelas bordas de cor preta. No item a), é possível

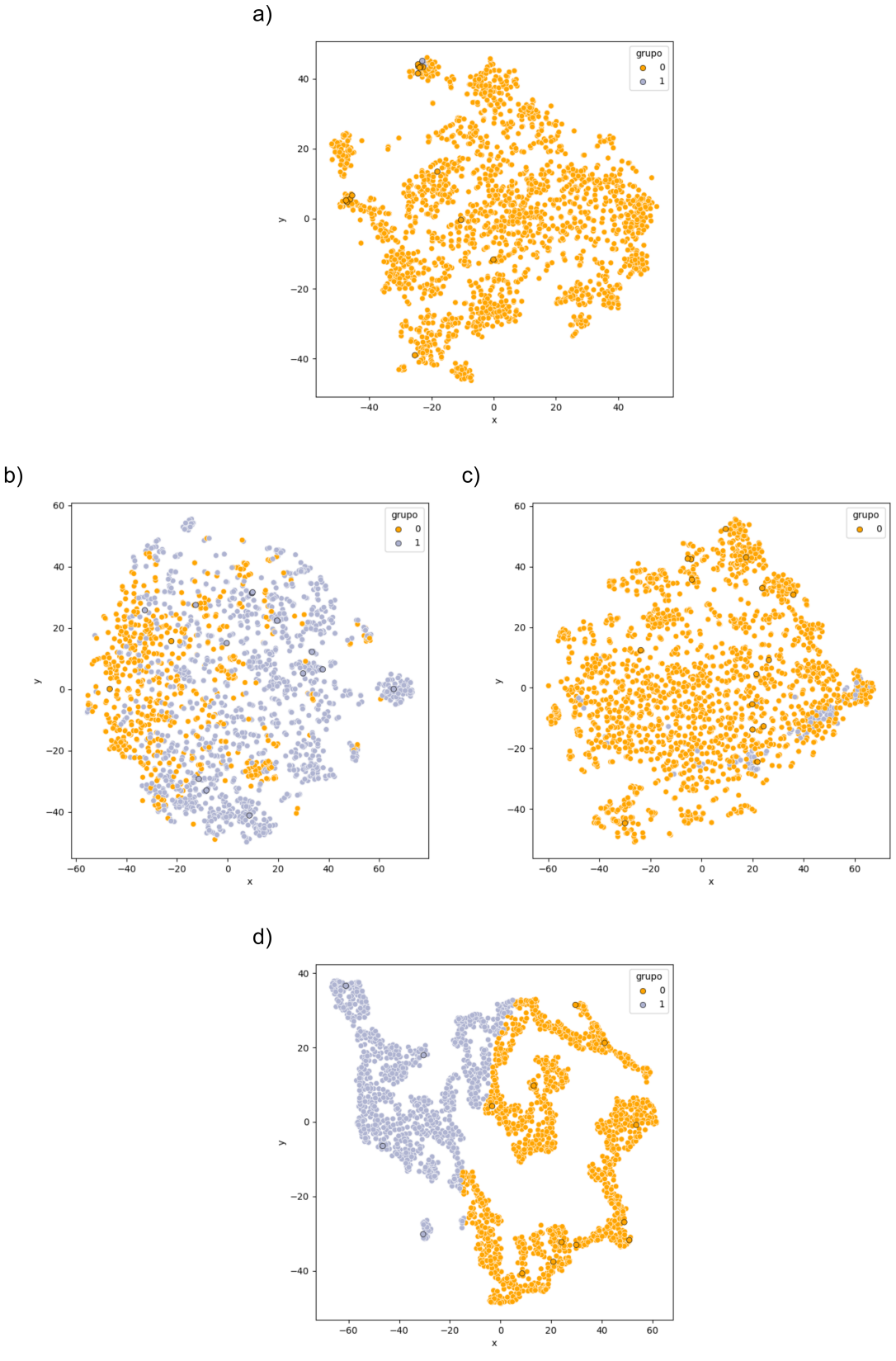

Figura 18 – Grupos gerados pelo K-Means. Em a) *PLUTO*. b) *GMEL*-origem. c) *GMEL*-destino. d) *Node2vec*. Fonte: a autora (2024).

observar a representação do *t-SNE* para os atributos do *PLUTO*. É possível observar que os setores censitários do distrito financeiro estão divididos entre os dois grupos gerados. Já nos itens b) e c), que são ambos representações geradas pelo *GMEL* voltadas a informações relevantes à origem (características residenciais) e ao destino (características empresariais e/ou comerciais), respectivamente, um padrão interessante que é possível observar nessas representações é que para a representação relacionada à origem no *GMEL*, os setores censitários estão distribuídos entre os dois grupos. Entretanto, para a representação relacionada ao destino gerada pelo *GMEL*, todos os setores censitários estão dentro de um mesmo grupo. Comparando com d), que mostra a representação gerada pelo *Node2vec* e também separa o distrito financeiro entre os dois grupos, o resultado parece indicar que os resultados gerados para o *GMEL* representariam de forma mais eficiente as características do PLUTO.

4.2.3 Análise temporal

Nessa análise, o objetivo é avaliar se o modelo apresenta robustez na predição alterando a distribuição da base de dados de teste. Aqui a alteração é feita treinando os modelos com as bases referentes a um ano e testando o modelo treinado com as bases referentes a anos posteriores. A metodologia de coleta e processamento dos dados é a mesma utilizada em todos os experimentos realizados, modificando apenas o ano de coleta da base. Foram coletados dados para os anos de 2010 a 2018.

Os resultados apresentados nas figuras 19, 20, 21 (azul representa um resultado melhor, vermelho um resultado pior) trazem alguns pontos de discussão. O primeiro deles é mais geral, onde é possível observar que, com exceção do Deep Gravity, o GMEL é responsável pelos valores extremos de performance, tanto positiva quanto negativamente, seguido de perto pelo GBRT.

Focando a análise nos resultados do GMEL em comparação com os outros modelos, é possível observar que no ano de 2010 existe uma piora significativa dos resultados nas três métricas em análise. Analisando o porquê desse padrão, foi observado que em 2010, especificamente, existe um atributo a menos, o *yearbuilt-2010*, que não está disponível na base de dados *PLUTO* para esse ano. Dessa forma, o *GMEL* gera a representação considerando um atributo a menos para os fluxos de origem e destino, treinando um modelo que apresenta uma dificuldade maior na predição dos anos subsequentes.

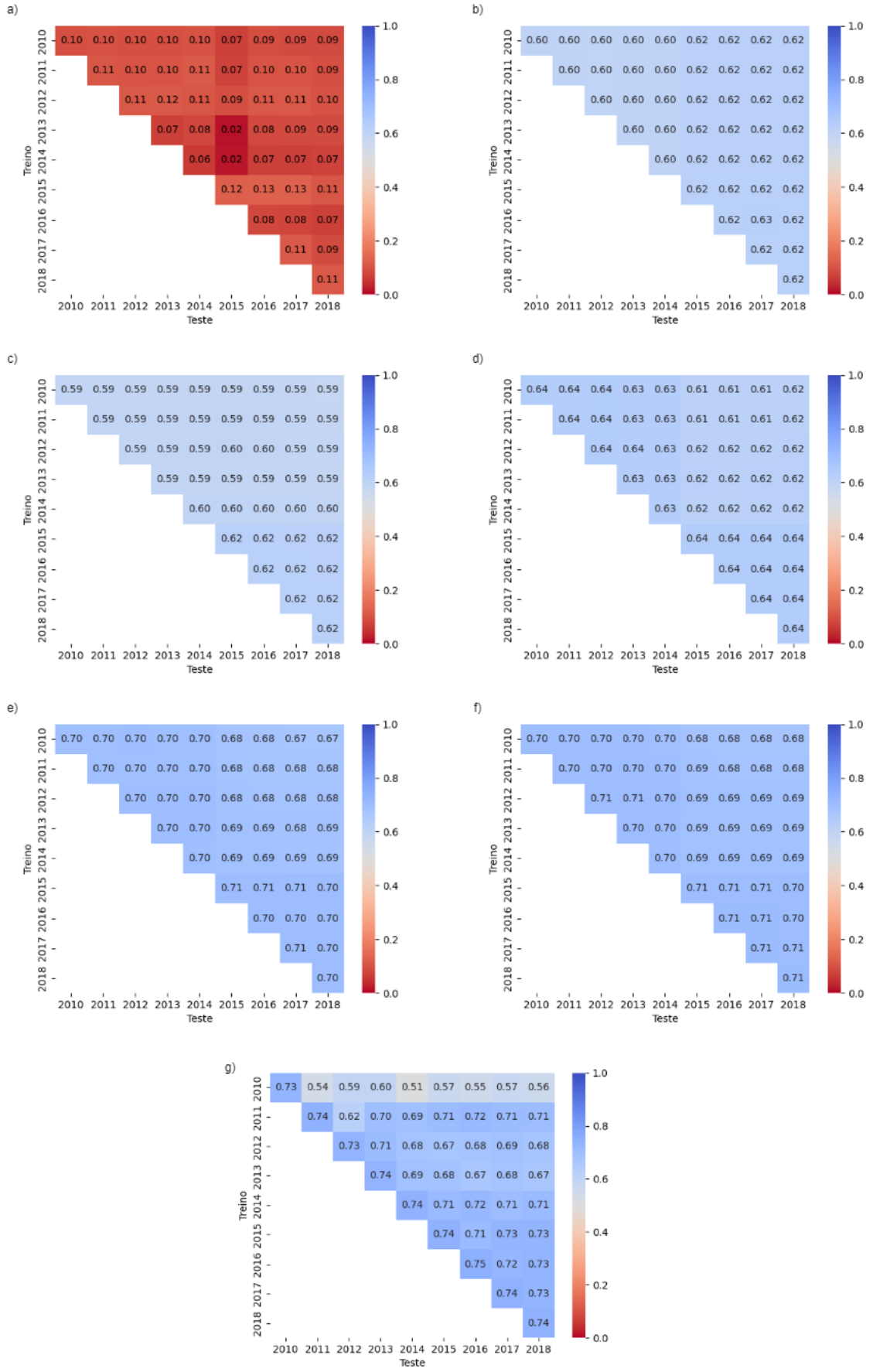

Figura 19 – Performance dos modelos na métrica *CPC* considerando mudanças temporais na base de dados. a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Node2vec*. f) *Gradient Boosting Regressor*. g) *GMEL*. Fonte: a autora (2024).

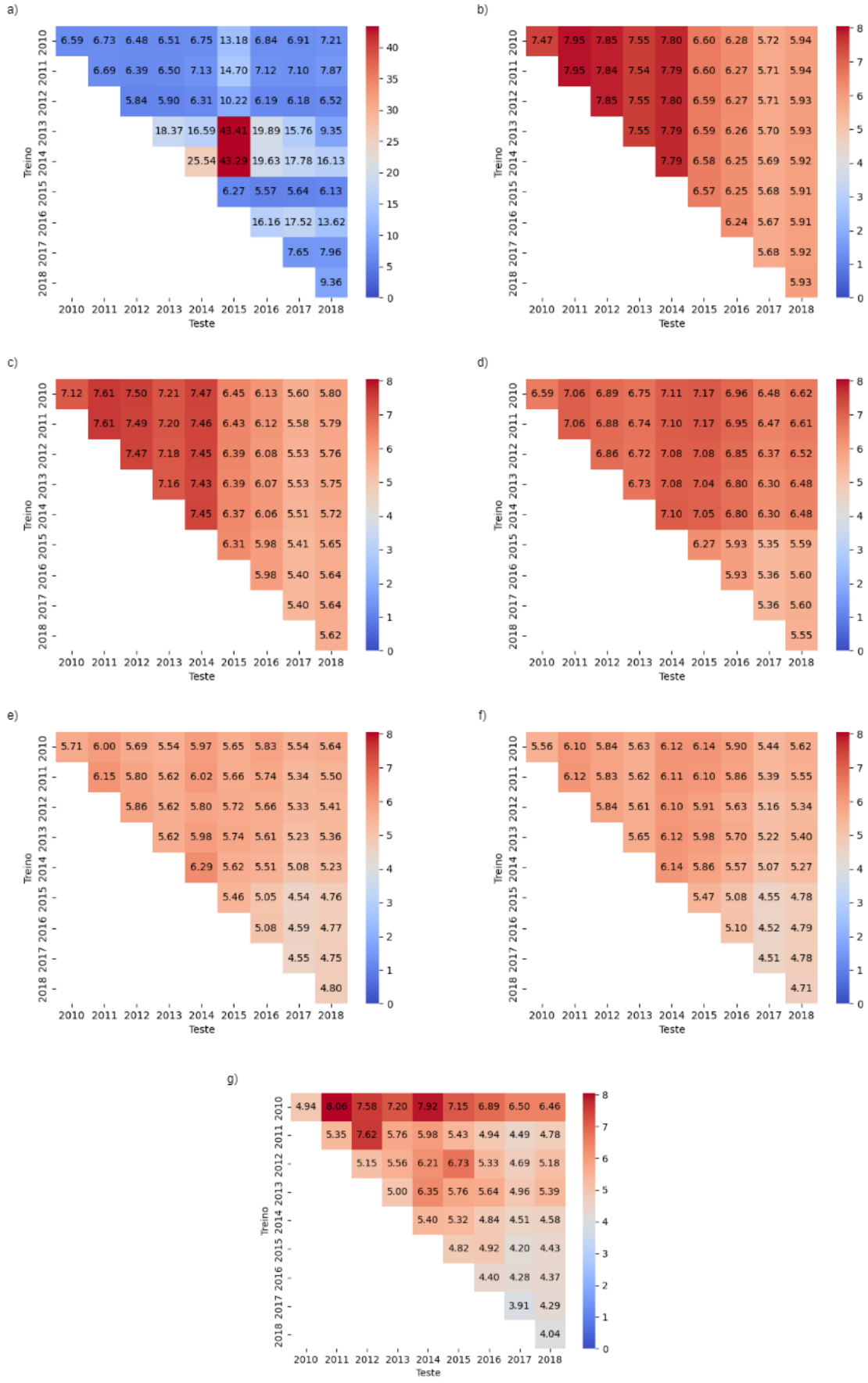

Figura 20 – Performance dos modelos na métrica *RMSE* considerando mudanças temporais na base de dados. a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Node2vec*. f) *Gradient Boosting Regressor*. g) *GMEL*. Fonte: a autora (2024).

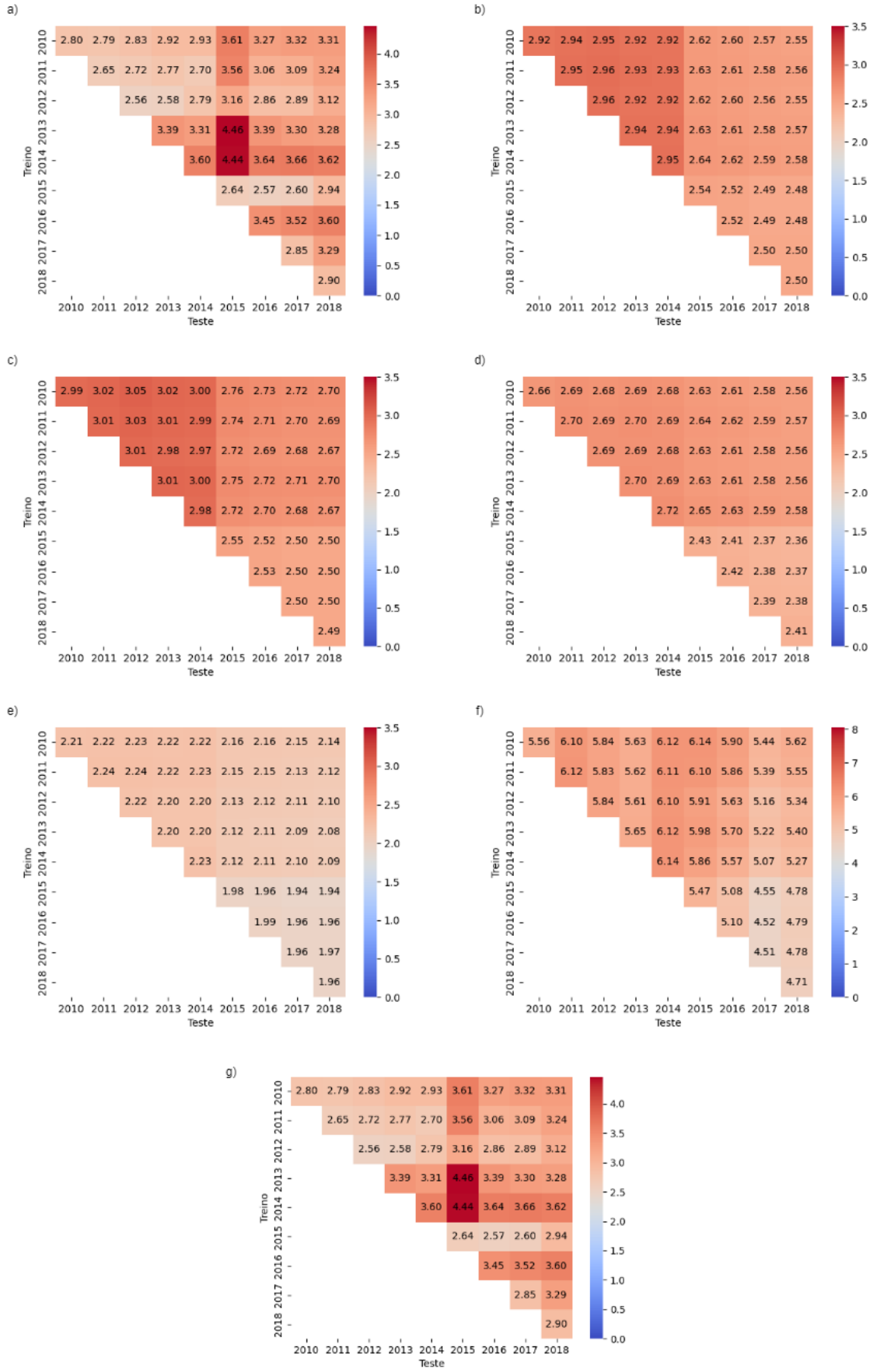

Figura 21 – Performance dos modelos na métrica *MAE* considerando mudanças temporais na base de dados. a) *Deep Gravity*. b) Regressão Lasso. c) Regressão Ridge. d) *Random Forest*. e) *Node2vec*. f) *Gradient Boosting Regressor*. g) *GMEL*. Fonte: a autora (2024).

5 CASOS DE USO

Com o intuito de avaliar o desempenho do *GMEL* (o modelo que reportou os melhores resultados de desempenho) na modelagem de cenários futuros, neste capítulo são propostos dois casos de uso onde são feitas modificações na infraestrutura da cidade do Recife, e é avaliado o impacto que essas modificações têm nos fluxos de deslocamento da cidade. A metodologia e resultados mostrados neste capítulos são reportados em (PEREGRINO et al., 2021).

5.1 MUDANÇA NA INFRAESTRUTURA DA CIDADE DO RECIFE

Recife é a capital mais antiga do Brasil, a 3ª cidade mais populosa da região Nordeste e a 9ª mais populosa do país. Além disso, também é uma das cidades que mais crescem no Brasil de acordo com (CityMayors Statistics, 2015), o que destaca a necessidade de ter as metodologias adequadas para permitir que as partes interessadas planejem melhor as intervenções urbanas realizadas. Até onde foi realizada a pesquisa bibliográfica para a escrita desta dissertação, a aplicação de modelos de aprendizagem de máquina para a predição de fluxos pendulares utilizando dados vindos da cidade do Recife é um campo pouco explorado de estudo, e nisso constitui a contribuição dos casos de uso descritos a seguir para a literatura.

5.1.1 Coleta e processamento de dados

Para treinar e validar o modelo usado para o planejamento de cenários, foram utilizados conjuntos de dados disponíveis no portal de dados abertos de Recife. Foram utilizados os setores censitários de 2020 das cidades como unidades geográficas. Para medir a distância percorrida entre os setores censitários, o *Open Source Routing Machine* (*OSRM*) foi usado, da mesma forma que descrito na subseção 3.1.3.

Para representar os fluxos de deslocamento, foi utilizado o conjunto de dados obtido a partir de uma pesquisa realizada pela prefeitura da cidade entre 2018 e 2019. Esta pesquisa capturou dados sobre movimentos típicos de deslocamento realizados pela população que reside, trabalha, estuda ou busca serviços na região metropolitana da cidade. Como indicadores urbanos, foram utilizadas informações sobre lotes individuais, bem como indicadores que mostram a presença de edifícios especiais de preservação, rotas culturais, estações de bicicletas, ciclovias e faixas exclusivas de ônibus, retirado do portal de dados abertos do Recife, disponível em (Prefeitura do Recife, 2021).

Ao fim do processo de coleta e processamento, foram obtidas informações de fluxo referentes a 1.347 setores censitários que cobrem mais da metade da cidade. Os fluxos de deslocamento foram agregados a nível de unidade geográfica, aqui representada pelos

setores censitários, resultando em 23.336 passageiros e, pós agrupamento, 15.945 pares de viagens origem-destino divididas em treino, teste e validação na proporção 6:2:2, também seguindo a metodologia apresentada no capítulo 3.

5.1.2 Metodologia

No centro da proposta desse caso de uso está a capacidade de usar uma rede neural para grafos (o *GMEL*) para prever mudanças nos fluxos de deslocamento, dadas as mudanças no uso da terra e ambiente construído. Foi seguido um conjunto de passos que permitiram o treino e validação do modelo e, em seguida, o uso desse modelo para prever fluxos modificando o cenário (os atributos das unidades geográficas afetadas).

1. **Treinamento do modelo:** foi treinado um modelo *GMEL* utilizando os dados coletados para a cidade do Recife. Com as representações obtidas pelo modelo treinado, foi utilizado o *Gradient Boosting Regressor* para predição dos fluxos, utilizando um conjunto de validação.

2. **Mudança de cenários:** depois que o modelo é treinado e testado, foram alterados alguns indicadores urbanos na cidade, seguindo cenários plausíveis de modificação urbana. Foi realizada uma atualização na rede de geo adjacência do modelo para acompanhar essas modificações e o modelo previamente treinado foi usado para gerar novas representações para a rede modificada. O *Gradient Boosting Regressor* é então utilizado nessas novas representações geradas para prever novos fluxos de deslocamento.

Foram realizadas duas alterações na infraestrutura da cidade, que serão discutidas a seguir:

- **Implantação de novas ciclovias**

  Nos últimos anos, tem havido um movimento crescente em que cidadãos se organizam para pensar em meios de se locomover que vão além dos veículos automotores. A bicicleta é um meio de transporte para pequenas e médias distâncias que traz benefícios à saúde e ao meio ambiente. Em algumas cidades do mundo, a bicicleta já é um meio de transporte amplamente utilizado por uma parte significativa da população, mas no Recife, embora a demanda por esse modal de transporte siga crescendo, a falta de infraestrutura cicloviária adequada em quantidade suficiente é um agente dificultador no aumento da quantidade de adeptos das bicicletas na cidade (Jornal do Commercio, 2021).

  Mudanças deste tipo da infraestrutura urbana devem incluir um trabalho importante de planejamento. Dentre as etapas de planejamento, seria importante estimar de alguma forma o impacto que estas mudanças causariam depois de implantadas

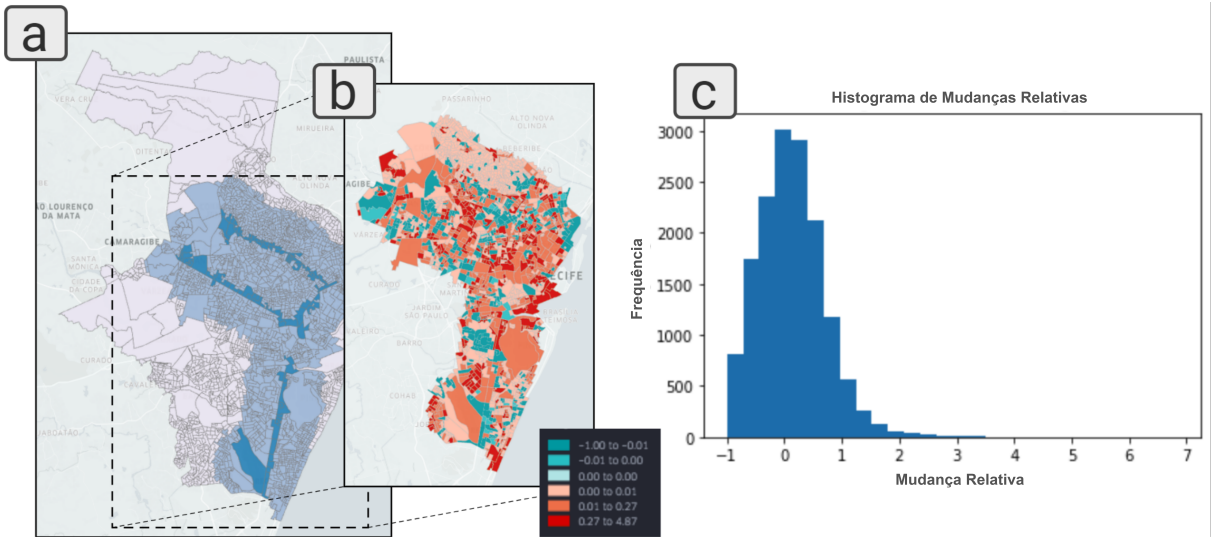

Figura 22 – Resultados do caso de uso de implantação de novas ciclovias. a) setores censitários diretamente afetados pelas mudanças de infraestrutura (azul escuro) e setores censitários afetados num raio de 2km (azul claro). b) Mapa de mudanças relativas verificadas pós mudança na infraestrutura. c) Histograma com a distribuição destas mudanças relativas. Fonte: a autora (2021).

num momento anterior a sua implantação, para os fluxos que passam pelas regiões afetadas pela mudança em estudo. Para responder esta pergunta, foram escolhidas quatro grandes avenidas que conectam importantes pontos do Recife, avenidas que têm em comum um alto índice de tráfego diário de veículos, porém atualmente não possuem uma estrutura sólida de ciclovia, colocando em risco os ciclistas que precisam se deslocar por essas avenidas para chegarem ao seu destino.

Foram distribuídos 24 km de ciclovias entre as quatro avenidas em estudo (Figura 22, item a), e foi utilizada a metodologia exposta acima para avaliar a mudança nos fluxos. Para avaliar as alterações causadas por estas modificações, foram analisados os fluxos entre unidades geográficas cujas os centros geográficos estivessem numa distância de até 2 km dos centros das unidades geográficas que receberam as novas ciclovias. Para cada fluxo, foi calculada a mudança relativa como uma forma de avaliar a quantidade de mudança de fluxo em estas áreas. Os resultados mostram um aumento médio de 13% (e desvio padrão de 0,59) dos fluxos entre essas unidades (Figura 22, itens b e c). Como o modelo não recebeu como entrada nenhum dado sobre o modo de transporte, a hipótese levantada é de que essas mudanças poderiam representar um aumento no fluxo de bicicletas nessas áreas.

- **Implantação de um condomínio**

  Construir grandes projetos em uma cidade é uma atividade que vem sempre acompanhada de uma série de impactos, seja na paisagem, economia da região e/ou no fluxo de pessoas que se deslocam de um ponto para outro. Portanto, é de grande importância para os órgãos governamentais competentes estimar esses impactos na

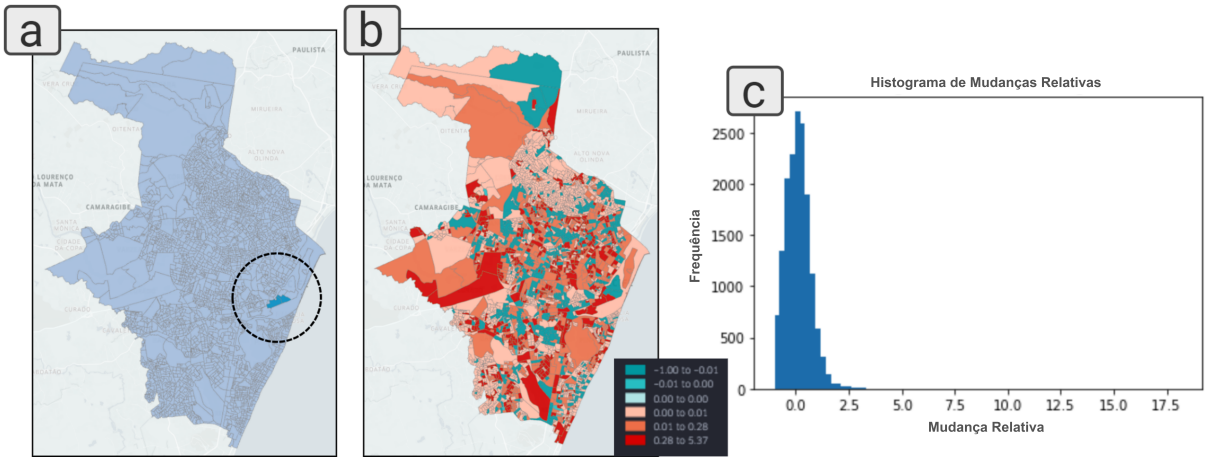

Figura 23 – Resultados do caso de uso de implantação de um condomínio. a) setores censitários diretamente afetados pelas mudanças de infraestrutura (azul escuro) e setores censitários afetados num raio de 2km (azul claro). b) Mapa de mudanças relativas verificadas pós mudança na infraestrutura. c) Histograma com a distribuição destas mudanças relativas. Fonte: a autora (2021).

infraestrutura da cidade a fim de antecipar ações para atender novas necessidades ou solucionar possíveis problemas futuros. Como forma de estimar o impacto no fluxo de pessoas devido a uma grande obra no centro do Recife, foi utilizado o caso real da construção de 13 torres, que variam de 13 a 44 pavimentos, no antigo sítio histórico do Cais José Estelita (Figura 23, item a). Os novos edifícios foram incorporados ao modelo, modificando informações relacionadas ao número de diferentes tipos de edifícios, densidade, ano de construção, relação terreno/área e relação piso/área. Semelhante ao caso de uso anterior, também foram analisadas as mudanças relativas nos fluxos pendulares que seriam causados pela modificação dos indicadores urbanos. Os resultados mostram um aumento médio de 12,5% (e desvio padrão de 0,605) dos fluxos entre essas unidades (Figura 23, itens b e c). É interessante observar que o modelo prevê mudanças relativas na cidade. No entanto, mudanças mais significativas acontecem em unidades próximas ao novo projeto. Prevê-se que estes fluxos aumentem em pelo menos 25%.

6 CONCLUSÃO

Este trabalho teve como objetivo realizar uma análise comparativa entre 7 modelos de aprendizagem de máquina na tarefa de predição de fluxos pendulares utilizando dados de infraestrutura. Para atingir este objetivo, foi realizada uma busca na literatura relacionada de forma que fossem elencados modelos promissores para uso na análise. Estes modelos foram então treinados com dados coletados da cidade de Nova Iorque, nos Estados Unidos, fornecidos por órgãos públicos locais, e foram comparados sob três pilares distintos.

O primeiro pilar buscou analisar a performance dos modelos em métricas aplicadas ao tipo de tarefa em estudo, e portanto foram obtidos os resultados para as métricas *CPC*, *RMSE* e *MAE*. O segundo pilar avaliou a seleção dos atributos mais importantes para cada modelo na predição dos fluxos, e portanto foram utilizadas técnicas de interpretabilidade propostas na literatura para entender os resultados de cada modelo e colher *insights* sobre sua performance superior ou inferior em relação aos demais modelos. Por fim, o terceiro pilar buscou investigar a performance dos modelos em relação a mudanças nos dados, que representam mudanças na infraestrutura da cidade. Para isto, foi realizada uma análise da performance dos modelos modificando o ano da base de dados utilizada para teste, de forma que fosse possível verificar a existência de degradação de performance conforme a diferença entre o ano em que o modelo foi treinado e o ano em que ele foi testado vá se distanciando (o que significaria uma quantidade cada vez maior de mudanças na infraestrutura).

Além disso, o modelo que apresentou melhor performance foi utilizado num caso de uso envolvendo mudanças na infraestrutura da cidade do Recife, no estado de Pernambuco, com o objetivo de trazer uma aplicação direta de como modelos de aprendizagem de máquina poderiam auxiliar no planejamento de cenários futuros.

Os resultados mostram uma performance superior do *GMEL*, que utiliza a aprendizagem por representação com redes neurais para grafos. O *GMEL* atingiu os melhores resultados nas três métricas utilizadas, e nas análises mostrou resultados que indicam uma geração do vetor de representações mais acurada que o *Node2vec*, a outra arquitetura proposta para esta análise. Os demais modelos em análise, entretanto, que possuem entrada tabular e se dividiram entre modelos lineares, baseados em árvores e uma rede neural *feedforward*, em sua maioria reportaram resultados competitivos em relação aos de melhor performance, e tem como vantagem seu tempo de treinamento consideravelmente menor que o tempo para treino dos modelos baseados em grafo, e simplicidade na arquitetura, o que se relaciona à facilidade com a qual o modelo pode ser interpretado, seja porque o modelo é intrínsecamente interpretável, como no caso de modelos lineares, ou porque existem ferramentas disponíveis na literatura que conseguem interpretar os resultados preditos pelos modelos sem grandes dificuldades, como nos modelos baseados em árvore e

a rede neural *feedforward*. No que diz respeito aos resultados da análise de importância de atributos utilizando a ferramenta *SHAP*, foi observado que os modelos que apresentaram melhor performance dentre os que possuem entrada tabular, que foram os baseados em árvore, em geral tem entre seus atributos mais relevantes para o setor censitário de origem atributos residenciais, enquanto que para o setor censitário de destino, atributos voltados a características comerciais ou de escritórios se destacaram. Também foi observada uma forte multicolinearidade na base, porém após reduzida e realizado novamente o treino dos modelos nesta nova base, os resultados de importância de atributos seguiram apontando padrões similares aos observados acima nos modelos de melhor performance.

Uma vez que este trabalho tem como aplicação sugerida uma ferramenta de planejamento de cenários futuros para gestão urbana, para além da performance dos modelos, os parâmetros discutidos acima (simplicidade do modelo, tempo de treinamento) também devem ser levados em consideração na escolha do modelo, especialmente se a gestão que irá utilizar esta ferramenta tiver limitações de recurso computacional disponível, ou nível técnico da equipe que irá trabalhar com a ferramenta. Desta forma, a escolha do melhor modelo deve ser definida após uma avaliação caso a caso das necessidades e recursos dos que o utilizarão. Este contexto faz com que modelos como o *Gradient Boosting Regressor*, que obteve performance muito próxima ao *GMEL* possuindo um tempo de treinamento menor, se tornem competitivos.

A partir dos pontos elencados acima, estimam-se alcançados os objetivos propostos para este trabalho.

Como trabalhos futuros, é possível elencar a adição do fator temporal nas análises a partir da utilização de modelos que recebam também uma entrada temporal, o que implica também na utilização de outras bases de dados que possuam informação temporal, e a investigação e proposição de ferramentas que sejam capazes de interpretar modelos de arquiteturas mais complexas propostos para a terefa de predição de fluxos pendulares.

REFERÊNCIAS

ALTELBANY, S. I. Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: A simulation study. *Journal of Applied Economics and Business Studies*, 2021. Disponível em: <https://api.semanticscholar.org/CorpusID:236668371>.

AMEER, S.; SHAH, M. A.; KHAN, A.; SONG, H.; MAPLE, C.; ISLAM, S. U.; ASGHAR, M. N. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, IEEE, v. 7, p. 128325–128338, 2019.

ANKENBRAND, M. J.; SHAINBERG, L.; HOCK, M.; LOHR, D.; SCHREIBER, L. M. Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac mri. *BMC Medical Imaging*, Springer, v. 21, p. 1–8, 2021.

ANTOR, M. B.; JAMIL, A.; MAMTAZ, M.; KHAN, M. M.; ALJAHDALI, S.; KAUR, M.; SINGH, P.; MASUD, M. A comparative analysis of machine learning algorithms to predict alzheimer's disease. *Journal of Healthcare Engineering*, Hindawi, v. 2021, 2021.

ANTOR, M. B.; JAMIL, A. S.; MAMTAZ, M.; KHAN, M. M.; ALJAHDALI, S.; KAUR, M.; SINGH, P.; MASUD, M. A comparative analysis of machine learning algorithms to predict alzheimer's disease. *Journal of Healthcare Engineering*, Hindawi Limited, v. 2021, 2021.

BANSAL, M.; GOYAL, A.; CHOUDHARY, A. A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, v. 3, p. 100071, 2022. ISSN 2772-6622. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2772662222000261>.

BARBOSA, H.; BARTHELEMY, M.; GHOSHAL, G.; JAMES, C. R.; LENORMAND, M.; LOUAIL, T.; MENEZES, R.; RAMASCO, J. J.; SIMINI, F.; TOMASINI, M. Human mobility: Models and applications. *Physics Reports*, Elsevier BV, v. 734, p. 1–74, mar. 2018. ISSN 0370-1573. Disponível em: <http://dx.doi.org/10.1016/j.physrep.2018.01.001>.

BARTHOLOMEW, K. Land use-transportation scenario planning: Promise and reality. *Transportation*, v. 34, p. 397–412, 06 2007.

BASU, I.; MAJI, S. Multicollinearity correction and combined feature effect in shapley values. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2022. p. 79–90.

BELLAMY, R. K.; DEY, K.; HIND, M.; HOFFMAN, S. C.; HOUDE, S.; KANNAN, K.; LOHIA, P.; MARTINO, J.; MEHTA, S.; MOJSILOVIC, A. et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

BIKE, C. *System Data*. 2024. Acessado em 21/02/2024. Disponível em: <https://citibikenyc.com/system-data>.

BRATKO, I. Machine learning: Between accuracy and interpretability. In: *Learning, networks and statistics*. [S.l.]: Springer, 1997. p. 163–177.

BUREAU, U. C. *Census Tracts*. 2015. Acessado em 04/03/2024. Disponível em: <https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf>.

BUREAU, U. C. *Commuting Flows*. 2023. Acessado em 03/03/2024. Disponível em: <https://www.census.gov/topics/employment/commuting/guidance/flows.html>.

BUREAU, U. C. *LEHD Origin-Destination Employment Statistics (LODES)*. 2024. Acessado em 21/02/2024. Disponível em: <https://lehd.ces.census.gov/data/#lodes>.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, v. 8, n. 8, 2019. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/8/8/832>.

CHAKRABORTY, A.; MCMILLAN, A. Scenario planning for urban planners: Toward a practitioner's guide. *Journal of the American Planning Association*, Routledge, v. 81, n. 1, p. 18–29, 2015. Disponível em: <https://doi.org/10.1080/01944363.2015.1038576>.

CHEN, F.; WANG, Y.-C.; WANG, B.; KUO, C.-C. J. Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, Now Publishers, v. 9, n. 1, 2020.

CHEN, Z.; XIAO, F.; GUO, F.; YAN, J. Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*, v. 9, p. 100123, 2023. ISSN 2666-7924. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666792423000021>.

CHENG, J.; SUN, J.; YAO, K.; XU, M.; CAO, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Elsevier, v. 268, p. 120652, 2022.

CICCO, V. D.; FIRMANI, D.; KOUDAS, N.; MERIALDO, P.; SRIVASTAVA, D. Interpreting deep learning models for entity resolution: an experience report using lime. In: *Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*. [S.l.: s.n.], 2019. p. 1–4.

CityMayors Statistics. *The world's fastest growing cities and urban areas from 2006 to 2020*. 2015. <http://www.citymayors.com/statistics/urban_growth1.html>.

COMMISSION, T. . L. *TLC Trip Record Data*. 2024. Acessado em 21/02/2024. Disponível em: <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.

COMPANY, M. \. *The economic potential of generative AI: The next productivity frontier*. 2023. Disponível em: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction>.

Department of Economic and Social Affairs, United Nations. *68% of the world population projected to live in urban areas by 2050, says UN*. 2018. <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html>.

DORAISWAMY, H.; ZACHARATOU, E. T.; MIRANDA, F.; LAGE, M.; AILAMAKI, A.; SILVA, C. T.; FREIRE, J. Interactive visual exploration of spatio-temporal urban data sets using urbane. In: . New York, NY, USA: Association for Computing Machinery, 2018. (SIGMOD '18), p. 1693–1696. ISBN 9781450347037. Disponível em: <https://doi.org/10.1145/3183713.3193559>.

DUVAL, A.; MALLIAROS, F. D. *GraphSVX: Shapley Value Explanations for Graph Neural Networks*. 2021.

FERREIRA, N.; LAGE, M.; DORAISWAMY, H.; VO, H.; WILSON, L.; WERNER, H.; PARK, M.; SILVA, C. Urbane: A 3d framework to support data driven decision making in urban development. In: *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. [S.l.: s.n.], 2015. p. 97–104.

FRANCETIC, I.; MUNFORD, L. Corona and coffee on your commute: a spatial analysis of covid-19 mortality and commuting flows in england in 2020. *European journal of public health*, Oxford University Press, v. 31, n. 4, p. 901–907, 2021.

FREITAS, A. A. Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In: SPRINGER. *International cross-domain conference for machine learning and knowledge extraction*. [S.l.], 2019. p. 48–66.

FU, C.; HUANG, Z.; SCHEUER, B.; LIN, J.; ZHANG, Y. Integration of dockless bike-sharing and metro: Prediction and explanation at origin-destination level. *Sustainable Cities and Society*, Elsevier, v. 99, p. 104906, 2023.

GOMEZ, T.; MOUCHÈRE, H. Computing and evaluating saliency maps for image classification: a tutorial. *Journal of Electronic Imaging*, Society of Photo-Optical Instrumentation Engineers, v. 32, n. 2, p. 020801–020801, 2023.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, v. 35, p. 507–520, 2022.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 855–864.

HAAG, M. *The N.Y.C. Neighborhood Where Families Are Filling Up Empty Offices*. 2023. Acessado em 04/03/2024. Disponível em: <https://www.nytimes.com/2023/11/17/nyregion/financial-district-office-conversions-housing.html>.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, p. 55–67, 1970. ISSN 0040-1706.

HU, F.; LIU, J.; LI, L.; LIANG, J. Community detection in complex networks using node2vec with spectral clustering. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 545, p. 123633, 2020.

HUANG, Q.; YAMADA, M.; TIAN, Y.; SINGH, D.; YIN, D.; CHANG, Y. *GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks*. 2020.

IPCC, A. et al. Intergovernmental panel on climate change. *IPCC Secretariat Geneva*, 2023. Disponível em: <https://www.ipcc.ch/>.

ISSERMAN, A. M. The location quotient approach to estimating regional economic impacts. *Journal of the American Institute of Planners*, Taylor & Francis, v. 43, n. 1, p. 33–41, 1977.

Jornal do Commercio. *Bike PE aumenta número de estações, mas quantidade de bicicletas segue a mesma, deixando muita gente sem pedalar*. 2021. <https://jc.ne10.uol.com.br/colunas/mobilidade/2021/11/13619160-bike-pe-aumenta-numero-de-estacoes-mas-quantidade-de-bicicletas-segue-a-mesma-deixar
html>.

JUARTO, B. Breast cancer classification using outlier detection and variance inflation factor. *Engineering, MAthematics and Computer Science (EMACS) Journal*, v. 5, n. 1, p. 17–23, 2023.

KEARNES, S.; MCCLOSKEY, K.; BERNDL, M.; PANDE, V.; RILEY, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, Springer Science and Business Media LLC, v. 30, n. 8, p. 595–608, aug 2016.

KHOSHRAFTAR, S.; AN, A. *A Survey on Graph Representation Learning Methods*. 2022.

KIM, N.; YOON, Y. Effective urban region representation learning using heterogeneous urban graph attention network (hugat). *arXiv preprint arXiv:2202.09021*, 2022.

KIM, Y.; NEWMAN, G. Advancing scenario planning through integrating urban growth prediction with future flood risk models. *Computers, Environment and Urban Systems*, v. 82, p. 101498, 2020. ISSN 0198-9715. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0198971520302313>.

LANDERS, R. N.; BEHREND, T. S. Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*, American Psychological Association, v. 78, n. 1, p. 36, 2023.

LI, R.; GAO, S.; LUO, A.; YAO, Q.; CHEN, B.; SHANG, F.; JIANG, R.; STANLEY, H. E. Gravity model in dockless bike-sharing systems within cities. *Physical Review E*, APS, v. 103, n. 1, p. 012312, 2021.

LI, Y.; SHA, C.; HUANG, X.; ZHANG, Y. Community detection in attributed graphs: An embedding approach. In: *AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2018.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, MDPI, v. 23, n. 1, p. 18, 2020.

LIU, Y.; FANG, F.; JING, Y. How urban land use influences commuting flows in wuhan, central china: A mobile phone signaling data perspective. *Sustainable Cities and Society*, v. 53, p. 101914, 2020. ISSN 2210-6707. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2210670718318365>.

LIU, Z.; MIRANDA, F.; XIONG, W.; YANG, J.; WANG, Q.; SILVA, C. Learning geo-contextual embeddings for commuting flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence (AAAI), v. 34, n. 01, p. 808–816, apr 2020. Disponível em: <https://doi.org/10.1609%2Faaai.v34i01.5425>.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. Disponível em: <http://arxiv.org/abs/1705.07874>.

LUO, D.; CHENG, W.; XU, D.; YU, W.; ZONG, B.; CHEN, H.; ZHANG, X. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, v. 33, p. 19620–19631, 2020.

LUXEN, D.; VETTER, C. Real-time routing with openstreetmap data. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 2011. (GIS '11), p. 513–516. ISBN 978-1-4503-1031-4. Disponível em: <http://doi.acm.org/10.1145/2093973.2094062>.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandermaaten08a.html>.

MARIMOUTOU, V.; PEGUIN, D.; PEGUIN-FEISSOLLE, A. The"distance-varying"gravity model in international economics: is the distance an obstacle to trade? *Economics Bulletin*, v. 29, n. 2, p. pp–1157, 2009.

MARINI, M.; GAWLIKOWSKA, A. P.; ROSSI, A.; CHOKANI, N.; KLUMPNER, H.; ABHARI, R. S. The impact of future cities on commuting patterns: An agent-based approach. *Environment and Planning B: Urban Analytics and City Science*, SAGE Publications Sage UK: London, England, v. 46, n. 6, p. 1079–1096, 2019.

MARTIN, S. A.; TOWNEND, F. J.; BARKHOF, F.; COLE, J. H. Interpretable machine learning for dementia: A systematic review. *Alzheimer's & Dementia*, Wiley Online Library, 2023.

MASÍS, S. *Interpretable Machine Learning with Python: Learn to Build Interpretable High-performance Models with Hands-on Real-world Examples*. Packt Publishing, Limited, 2021. ISBN 9781800203907. Disponível em: <https://books.google.com.br/books?id=eWQmzgEACAAJ>.

MASTROPIETRO, A.; PASCULLI, G.; FELDMANN, C.; RODRÍGUEZ-PÉREZ, R.; BAJORATH, J. Edgeshaper: Bond-centric shapley value-based explanation method for graph neural networks. *Iscience*, Elsevier, v. 25, n. 10, 2022.

MASUCCI, A. P.; SERRAS, J.; JOHANSSON, A.; BATTY, M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, American Physical Society (APS), v. 88, n. 2, ago. 2013. ISSN 1550-2376. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.88.022812>.

MIRANDA, F.; DORAISWAMY, H.; LAGE, M.; WILSON, L.; HSIEH, M.; SILVA, C. T. Shadow accrual maps: Efficient accumulation of city-scale shadows over time. *IEEE Transactions on Visualization and Computer Graphics*, v. 25, n. 3, p. 1559–1574, 2019.

MORI, T.; UCHIHIRA, N. Balancing the trade-off between accuracy and interpretability in software defect prediction. *Empirical Software Engineering*, Springer, v. 24, p. 779–825, 2019.

MORTON, A.; PIBURN, J.; NAGLE, N. Need a boost? a comparison of traditional commuting models with the xgboost model for predicting commuting flows (short paper). In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. *10th International Conference on Geographic Information Science (GIScience 2018)*. [S.l.], 2018.

NANEHKARAN, Y. A.; LICAI, Z.; CHENGYONG, J.; CHEN, J.; ANWAR, S.; AZARAFZA, M.; DERAKHSHANI, R. Comparative analysis for slope stability by using machine learning methods. *Applied Sciences*, v. 13, n. 3, 2023. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/13/3/1555>.

NEJADSHAMSI, S.; BENTAHAR, J.; EICKER, U.; WANG, C.; JAMSHIDI, F. A geographic-semantic context-aware urban commuting flow prediction model using graph neural network. *Available at SSRN 4701536*.

NONG, Y.; DU, Q. Urban growth pattern modeling using logistic regression. *Geo-spatial Information Science*, v. 14, p. 62–67, 03 2011.

ORTNER, T.; SORGER, J.; STEINLECHNER, H.; HESINA, G.; PIRINGER, H.; GRöLLER, E. Vis-a-ware: Integrating spatial and non-spatial visualization for visibility-aware urban planning. *IEEE Transactions on Visualization and Computer Graphics*, v. 23, n. 2, p. 1139–1151, 2017.

PENG, J.; GUAN, J.; SHANG, X. Predicting parkinson's disease genes based on node2vec and autoencoder. *Frontiers in genetics*, Frontiers Media SA, v. 10, p. 226, 2019.

PEREGRINO, A. A.; PRADHAN, S.; LIU, Z.; FERREIRA, N.; MIRANDA, F. Transportation scenario planning with graph neural networks. *arXiv preprint arXiv:2110.13202*, 2021.

PERO, V.; STEFANELLI, V. A questão da mobilidade urbana nas metrópoles brasileiras. *Revista de economia contemporânea*, SciELO Brasil, v. 19, p. 366–402, 2015.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2014. p. 701–710.

PETTIT, C.; SHI, Y.; HAN, H.; RITTENBRUCH, M.; FOTH, M.; LIESKE, S.; NOUWELANT, R. van den; MITCHELL, P.; LEAO, S.; CHRISTENSEN, B. et al. A new toolkit for land value analysis and scenario planning. *Environment and Planning B: Urban Analytics and City Science*, Sage Publications Sage UK: London, England, v. 47, n. 8, p. 1490–1507, 2020.

PLANNING, D. of C. *PLUTO and MapPLUTO*. 2024. Acessado em 21/02/2024. Disponível em: <https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page>.

PLANNING, N. D. of C. *Glossary of Zoning Terms*. 2024. Acessado em 25/02/2024. Disponível em: <https://www.nyc.gov/site/planning/zoning/glossary.page>.

POUREBRAHIM, N.; SULTANA, S.; NIAKANLAHIJI, A.; THILL, J.-C. Trip distribution modeling with twitter data. *Computers, Environment and Urban Systems*, v. 77, p. 101354, 2019. ISSN 0198-9715. Disponível em: <https://www.sciencedirect.com/science/article/pii/S019897151930119X>.

Prefeitura do Recife. *Portal de Dados Abertos do Recife*. 2021. <http://dados.recife.pe.gov.br/>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016.

SCHONER, J.; CHAPMAN, J.; BROOKES, A.; MACLEOD, K. E.; FOX, E. H.; IROZ-ELARDO, N.; FRANK, L. D. Bringing health into transportation and land use scenario planning: Creating a national public health assessment model (n-pham). *Journal of Transport Health*, v. 10, p. 401–418, 2018. ISSN 2214-1405. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214140516304406>.

SHI, Q.; ZHUO, L.; TAO, H.; YANG, J. A fusion model of temporal graph attention network and machine learning for inferring commuting flow from human activity intensity dynamics. *International Journal of Applied Earth Observation and Geoinformation*, Elsevier, v. 126, p. 103610, 2024.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: PMLR. *International conference on machine learning*. [S.l.], 2017. p. 3145–3153.

SIMINI, F.; BARLACCHI, G.; LUCA, M.; PAPPALARDO, L. A deep gravity model for mobility flows generation. *Nature Communications*, v. 12, 11 2021.

SIMINI, F.; GONZÁLEZ, M. C.; MARITAN, A.; BARABÁSI, A.-L. A universal model for mobility and migration patterns. *Nature*, Nature Publishing Group UK London, v. 484, n. 7392, p. 96–100, 2012.

SOHN, J. Are commuting patterns a good indicator of urban spatial structure? *Journal of Transport Geography*, v. 13, n. 4, p. 306–317, 2005. ISSN 0966-6923. Disponível em: <https://www.sciencedirect.com/science/article/pii/S096669230400050X>.

SORENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, v. 5, p. 1–34, 1948.

SPADON, G.; CARVALHO, A. C. d.; RODRIGUES-JR, J. F.; ALVES, L. G. Reconstructing commuters network using machine learning and urban indicators. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 11801, 2019.

TANG, J.; QU, M.; WANG, M.; ZHANG, M.; YAN, J.; MEI, Q. Line: Large-scale information network embedding. In: *Proceedings of the 24th international conference on world wide web*. [S.l.: s.n.], 2015. p. 1067–1077.

THOUGHTLAB, E. Smart city solutions for a riskier world. *How Innovation can Drive Resilience, Sustainability, and Citizen Well-Being. Version*, v. 2, 2021.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 58, n. 1, p. 267–288, 1996.

USTUN, B.; RUDIN, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, Springer, v. 102, p. 349–391, 2016.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

VELIčKOVIć, P.; CUCURULL, G.; CASANOVA, A.; ROMERO, A.; LIò, P.; BENGIO, Y. *Graph Attention Networks*. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1710.10903>.

VIJAYAKUMAR, S. *Interpretability in Activation Space Analysis of Transformers: A Focused Survey*. 2023.

VU, M.; THAI, M. T. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems*, v. 33, p. 12225–12235, 2020.

WANG, D.; FU, Y.; WANG, P.; HUANG, B.; LU, C.-T. Reimagining city configuration. In: *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. [S.l.]: ACM, 2020.

WANG, H.; LI, Z. Region representation learning via mobility flow. In: . New York, NY, USA: Association for Computing Machinery, 2017. (CIKM '17), p. 237–246. ISBN 9781450349185. Disponível em: <https://doi.org/10.1145/3132847.3133006>.

WANG, J.; WIENS, J.; LUNDBERG, S. Shapley flow: A graph-based approach to interpreting model predictions. In: PMLR. *International Conference on Artificial Intelligence and Statistics*. [S.l.], 2021. p. 721–729.

WANG, S.; PENG, H.; LIANG, S. Prediction of estuarine water quality using interpretable machine learning approach. *Journal of Hydrology*, v. 605, p. 127320, 2022. ISSN 0022-1694. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022169421013706>.

WANG, Y.; YAO, X.; LIU, Y.; LI, X. Generating population migration flow data from inter-regional relations using graph convolutional network. *International Journal of Applied Earth Observation and Geoinformation*, Elsevier, p. 103238, 2023.

WOJTUCH, A.; JANKOWSKI, R.; SMUSZ, S. How can shap values help to shape metabolic stability of chemical compounds? *Journal of Cheminformatics*, v. 13, 09 2021.

WU, L.; CUI, P.; PEI, J.; ZHAO, L.; GUO, X. Graph neural networks: foundation, frontiers and applications. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2022. p. 4840–4841.

WU, Y.; ZHANG, L.; BHATTI, U. A.; HUANG, M. Interpretable machine learning for personalized medical recommendations: A lime-based approach. *Diagnostics*, MDPI, v. 13, n. 16, p. 2681, 2023.

XU, K.; HU, W.; LESKOVEC, J.; JEGELKA, S. *How Powerful are Graph Neural Networks?* arXiv, 2018. Disponível em: <https://arxiv.org/abs/1810.00826>.

XU, N.; CHENG, Y.; XU, X. Using location quotients to determine public–natural space spatial patterns: A zurich model. *Sustainability*, MDPI, v. 10, n. 10, p. 3462, 2018.

YANG, T. Understanding commuting patterns and changes: Counterfactual analysis in a planning support framework. *Environment and Planning B: Urban Analytics and City Science*, SAGE Publications Sage UK: London, England, v. 47, n. 8, p. 1440–1455, 2020.

YIN, G.; HUANG, Z.; BAO, Y.; WANG, H.; LI, L.; MA, X.; ZHANG, Y. Convgcn-rf: A hybrid learning model for commuting flow prediction considering geographical semantics and neighborhood effects. *GeoInformatica*, Springer, v. 27, n. 2, p. 137–157, 2023.

YING, R.; HE, R.; CHEN, K.; EKSOMBATCHAI, P.; HAMILTON, W. L.; LESKOVEC, J. Graph convolutional neural networks for web-scale recommender systems. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. [S.l.]: ACM, 2018.

YING, Z.; BOURGEOIS, D.; YOU, J.; ZITNIK, M.; LESKOVEC, J. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, v. 32, 2019.

YORK, C. of N. *NYC Open Data*. 2022. Acessado em 21/02/2024. Disponível em: <https://opendata.cityofnewyork.us/>.

YU, B.; LEE, Y.; SOHN, K. Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (gcn). *Transportation Research Part C: Emerging Technologies*, v. 114, p. 189–204, 2020. ISSN 0968-090X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X19312434>.

YUAN, H.; YU, H.; GUI, S.; JI, S. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 45, n. 5, p. 5782–5799, 2022.

YUAN, H.; YU, H.; WANG, J.; LI, K.; JI, S. On explainability of graph neural networks via subgraph explorations. In: PMLR. *International conference on machine learning*. [S.l.], 2021. p. 12241–12252.

ZAFAR, M. R.; KHAN, N. M. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

ZHOU, H.; SUN, G.; FU, S.; WANG, L.; HU, J.; GAO, Y. Internet financial fraud detection based on a distributed big data approach with node2vec. *IEEE Access*, IEEE, v. 9, p. 43378–43386, 2021.

ZIPF, G. K. The p1 p2/d hypothesis: On the intercity movement of persons. *American Sociological Review*, [American Sociological Association, Sage Publications, Inc.], v. 11, n. 6, p. 677–686, 1946. ISSN 00031224. Disponível em: <http://www.jstor.org/stable/2087063>.