UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

SISTEMAS DE INFORMAÇÃO

DANILO DA ROCHA LIRA ARAÚJO

ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA APLICADOS À DETECÇÃO DE FRAUDES DE CARTÃO DE CRÉDITO

Recife

2022

DANILO DA ROCHA LIRA ARAÚJO

ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA APLICADOS À DETECÇÃO DE FRAUDES DE CARTÃO DE CRÉDITO

Trabalho apresentado ao Programa de Graduação em Sistemas de informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fernando Maciano de Paula Neto

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Araujo, Danilo.
Estudo comparativo entre algoritmos de aprendizagem de máquina aplicados à detecção de fraudes de cartão de crédito / Danilo Araujo. - Recife, 2022.
48 p. : il., tab.

Orientador(a): Fernando Maciano Neto
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2022.

1. fraudes de cartão de crédito. 2. algoritmos de aprendizagem de máquina. 3. estudo comparativo.. I. Maciano Neto, Fernando. (Orientação). II. Título.

000 CDD (22.ed.)

DANILO DA ROCHA LIRA ARAÚJO

ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA APLICADOS À DETECÇÃO DE FRAUDES DE CARTÃO DE CRÉDITO

Trabalho apresentado ao Programa de Graduação em Sistemas de informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Recife, 21 de setembro de 2022

BANCA EXAMINADORA

Prof. Dr. Fernando Maciano de Paula Neto (Orientador)
UNIVERSIDADE FEDERAL DE PERNAMBUCO

Prof. Dr. Vinicius Cardoso Garcia (2º membro da banca)
UNIVERSIDADE FEDERAL DE PERNAMBUCO

AGRADECIMENTOS

Aos professores do curso de Sistemas de Informação que fizeram parte da minha jornada de aprendizado durante todo o percurso da universidade, principalmente ao professor Fernando Neto, que foi meu primeiro professor do curso e orientador deste trabalho. Quero agradecer a todos os meus amigos que me deram apoio durante o desenvolvimento do trabalho e a minha mãe que esteve sempre presente dando todo suporte possível.

“Se você se empenhar o suficiente pode fazer qualquer história resultar.” (Saul Goodman).

RESUMO

O cartão de crédito vem se tornando um meio de pagamento cada vez mais utilizado e a preocupação com sua segurança um problema mais relevante. A fraude de cartões de crédito é definida como seu uso não autorizado por um terceiro e pode ocorrer tanto com a posse indevida do cartão físico quanto através, apenas, do uso de seus dados. Neste trabalho, serão apresentados alguns dos principais métodos de detecção de fraude que fazem uso de aprendizagem de máquina. Foi realizada uma comparação entre seus desempenhos, utilizando métricas de desempenho em problemas de classificação, como acurácia, precisão, revocação, F1-Score e área sob a curva ROC e para verificar se algum modelo foi estatisticamente superior aos demais, foi utilizado o teste de Wilcoxon. Para treinamento e testes dos algoritmos foi usada uma base de dados da Université Libre de Bruxelles contendo transações fraudulentas e legítimas.

Palavras-chave: fraudes de cartão de crédito; algoritmos de aprendizagem de máquina; estudo comparativo.

ABSTRACT

The credit card has become an increasingly used means of payment and the concern for its security is a more relevant problem. Credit card fraud is defined as its unauthorized use by a third party and can occur either with improper possession of the physical card or simply through the use of your data. In this work, some of the main fraud detection methods that make use of machine learning will be presented. A comparison was made between their performances using performance metrics in classification problems, such as accuracy, precision, recall, F1-Score and area under the ROC curve and to verify if any model was statistically superior to the others, the Wilcoxon test was used. For training and testing the algorithms, an ULB database containing fraudulent and legitimate transactions was used.

Keywords: credit card fraud; machine learning algorithms; comparative study.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

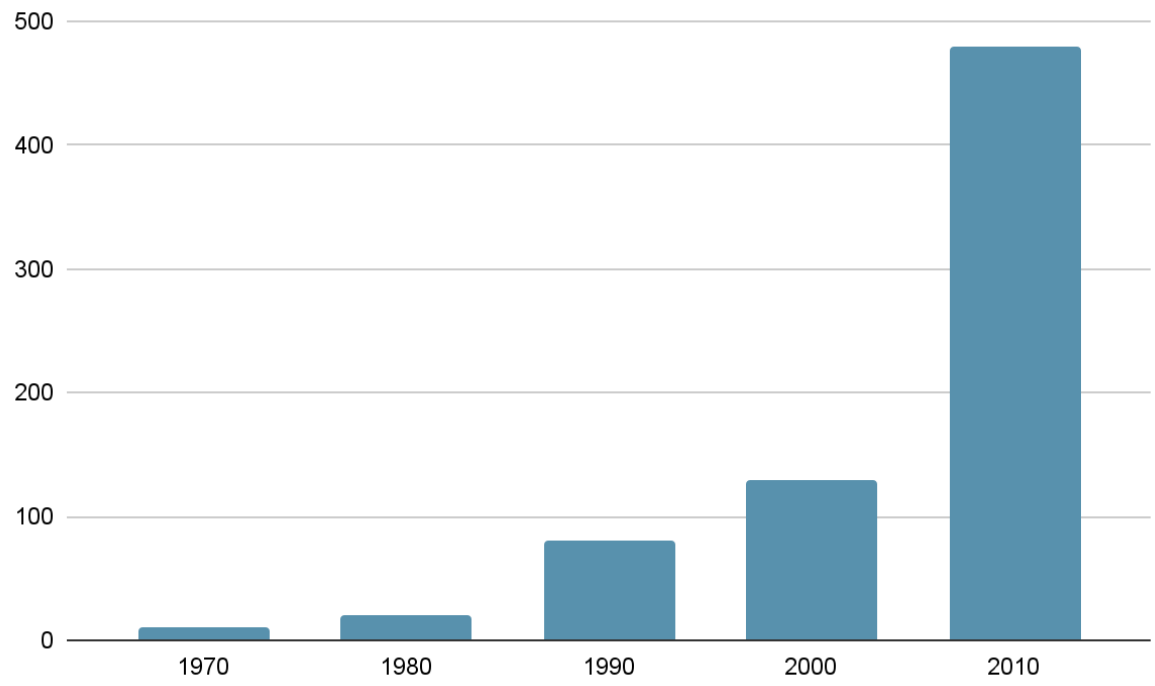
ACM	*Association for Computing Machinery*
AUC-PR	*Area Under the Curve Precision Recall*
CART	*Classification and regression trees*
CNN	*Convolutional Neural Network*
DT	*Decision Tree*
FP	Falso positivo
GB	*Gradient Boosting*
GRU	*Gated Recurrent Unit*
HMM	*Hidden Markov Model*
IEEE	Instituto de Engenheiros Eletricistas e Eletrônicos
KNN	*K-nearest neighbors*
LR	*Logistic Regression*
LSTM	*Long short-term memory*
MLP	*Multilayer Perceptron*
NN	*Neural Network*
SVM	*Support Vector Machine*
ULB	Universidade Livre de Bruxelas
VP	Verdadeiro positivo
XGB	*eXtreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da internet e o crescimento do comércio online, o número de transações feitas com cartão de crédito vem aumentando, e com isso a quantidade de fraudes também. A fraude desse meio de pagamento pode ser definida como o uso indevido do cartão físico ou de seus dados para aquisição de bens ou dinheiro, e esse tipo de fraude, apenas em 2020, gerou prejuízos de cerca de 28 bilhões de dólares a instituições financeiras, usuário de cartões e comerciantes (ROBERTSON, 2021). Com isso, o interesse sobre o tema vem crescendo ao longo dos anos e vemos isso sendo refletido pela quantidade de artigos científicos na área. Como apresentado na Figura 1, há registros de publicações que propõem soluções com algoritmos de aprendizagem de máquina desde 1970, porém a partir da década de 2010 houve um crescimento de mais de duas vezes em relação as décadas anteriores (RYMAN-TUBB, 2018).

Figura 1 – Quantidade de artigos relacionados publicados por década até o ano de 2018



Fonte: adaptada de Ryman-Tubb *et al.* (2018).

Diversos métodos de detecção de fraudes foram criados visando mitigar os prejuízos causados por esse tipo de crime. Os sistemas que são desenvolvidos com esse intuito são chamados de sistemas de detecção de fraude. Um típico sistema detecção de fraudes inclui várias camadas de controle, cada uma dessas pode ser automatizada ou supervisionada

manualmente por humanos. As partes automatizadas podem utilizar conjuntos de regras criados por especialistas ou algoritmos de aprendizagem de máquina como forma de classificação das transações (CARCILLO, 2019). A maior parte das instituições financeiras hoje utiliza a abordagem por regras, a qual possui um bom desempenho desde que não haja mudança no comportamento das transações, o que não é um cenário realista. Sendo assim, se faz necessária a utilização de uma abordagem mais adaptável e eficiente (BAHNSEN, 2013).

Sabendo desta necessidade e entendendo que a detecção de fraudes pode ser encarada como um problema de classificação binária, muitos estudos aplicando técnicas de aprendizagem de máquina foram surgindo. Vários desses utilizando abordagens diferentes, com algoritmos de aprendizagem de máquina supervisionados, não supervisionados e técnicas de comitês de classificadores. Tendo como base esses trabalhos, este artigo desenvolve um estudo comparativo entre os métodos estado da arte na detecção de fraudes de cartão de crédito utilizando algoritmos de aprendizagem de máquina.

1.1 MOTIVAÇÃO

A motivação deste trabalho é contribuir para o desenvolvimento da área de detecção de fraudes, utilizando aprendizagem de máquina, mapeando suas maiores dificuldades e as melhores propostas de solução atuais. O objetivo é entender como está o desempenho dos métodos atuais e o que justifica esse desempenho, além disso analisar a possibilidade de utilização de métodos pouco citados ou ainda não utilizados como proposta de solução do problema.

1.2 OBJETIVOS

1.2.1 Pergunta de pesquisa

Considerando que o cartão de crédito é um dos principais métodos de pagamento utilizado em todo o mundo, a detecção de possíveis fraudes pode proteger uma grande quantidade de patrimônio de pessoas físicas e entidades financeiras. O desafio pode ser resumido a identificar se uma transação é fraudulenta para aprová-la ou negá-la. Sendo assim o problema se encaixa em uma classificação binária, então as perguntas de pesquisa são: Quais métricas devem ser utilizadas para comparar o desempenho dos algoritmos analisados? Qual é

o melhor algoritmo de aprendizagem de máquina para o processo de classificação de transações?

1.2.2 Hipóteses

Tendo a pesquisa analisado diversos algoritmos de aprendizagem de máquina diferentes, utilizados para detecção de fraudes de cartão de crédito, a hipótese levantada para essa pesquisa é: O algoritmo de aprendizagem profunda *Convolutional Neural Network (CNN)* apresenta um bom desempenho na resolução do problema. As vantagens das redes neurais e do aprendizado profundo são que eles podem aproximar totalmente qualquer relacionamento não linear complexo, forte robustez e tolerância a falhas e encontrar soluções otimizadas em alta velocidade (ZHANG, 2018). Considerando que apenas um dos trabalhos citados realizou um estudo comparativo com o algoritmo de aprendizagem profunda CNN e as vantagens das redes neurais citadas anteriormente, então a hipótese foi formulada visando validar essas observações.

1.2.3 Objetivo

O objetivo deste trabalho é realizar um estudo comparativo entre os métodos de aprendizagem de máquina utilizados atualmente para solução do problema da detecção de fraudes de cartão de crédito. Os objetivos específicos que são visados para este trabalho são:

- Analisar os métodos de aprendizagem de máquina propostos para o problema;
- Avaliar as métricas utilizadas para comparação de resultados;
- Avaliar as bases de dados utilizadas para os experimentos;
- Implementar os algoritmos mais relevantes e compará-los através do uso de testes estatísticos.

1.3 ESTRUTURA DE TRABALHO

O trabalho foi dividido em sete capítulos. O primeiro capítulo é a introdução onde foi apresentado o contexto de fraudes de cartão de crédito e como se deu o crescimento do interesse por esse tema ao longo das décadas, assim como a motivação do trabalho e os objetivos definidos. Já o segundo capítulo é o referencial teórico, onde são apresentados os conceitos mais relevantes para o trabalho como fraudes de cartão de crédito e sistemas de detecção de fraudes. No capítulo três é apresentada a metodologia utilizada para o desenvolvimento do

trabalho, como os critérios de escolha dos algoritmos, métricas e conjunto de dados utilizado. Em seguida o quarto capítulo discorre sobre os trabalhos relacionados apresentando quais métodos de detecção foram explorados no estado da arte e como esses métodos foram avaliados por cada um dos autores. Já o quinto capítulo fala sobre os resultados encontrados nos testes realizados e como as métricas foram utilizadas, além disso são apresentados os gráficos para cada métrica com descrição dos melhores modelos em cada uma delas. O próximo capítulo é o de conclusão que de fato realiza uma comparação entre os modelos avaliados e traz quais foram os melhores desempenhos obtidos pelos modelos. O último capítulo é o de trabalhos futuros em que são apresentadas limitações e ameaças para os métodos de detecção utilizados neste trabalho e sugestões para novos trabalhos.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados conceitos relevantes dentro da área de detecção de fraudes de cartão de crédito, além do contexto da área.

2.1 FRAUDE DE CARTÃO DE CRÉDITO

Segundo Jurgovsky (2018), no setor de pagamentos, a fraude de cartão de crédito ocorre quando alguém rouba informações de um cartão para fazer compras sem a permissão explícita do dono e a detecção dessas transações fraudulentas é uma atividade crucial para o funcionamento dos processadores de pagamento. Já Ryman-Tubb (2018) afirma que a fraude de cartão de pagamento é o ato criminoso de engano usando um cartão físico ou os dados do cartão sem o conhecimento do titular. Em Zhang (2018), é descrito que o fraudador pode coletar informações dos usuários, como senha, idade, ocupação e outras informações para entrar em sistemas bancários. Esse tipo de comportamento fraudulento tem sido muito comum hoje, e traz grandes prejuízos para usuários, empresas e sociedade. Além disso, uma vez que esses dados de identidade são roubados, os métodos tradicionais de segurança da informação não impedirão a fraude de transações online.

2.2 SISTEMAS DE DETECÇÃO DE FRAUDE

Para determinar se uma transação de pagamento com cartão será autorizada, é utilizado um sistema de gerenciamento de fraude. Um típico sistema de detecção de fraude inclui várias camadas de controle, cada uma das quais pode ser automatizada ou supervisionada por humanos. Parte da camada automatizada abrange algoritmos de aprendizado de máquina que criam modelos preditivos com base em transações anotadas (CARCILLO, 2021). Se a transação for considerada suspeita, ela normalmente é bloqueada ou recusada e um tíquete de fraude é criado. Este bilhete de fraude contém informações suficientes para que um revisor humano analise a transação e, em seguida, tome uma decisão. Na maioria das organizações, uma equipe de revisores verifica os tíquetes de fraude e é realizada uma investigação que pode incluir o contato com o titular do cartão ou comerciante (RYMAN-TUBB, 2018)

2.3 DERIVA CONCEITUAL

O padrão das transações é mutável ao longo do tempo. Os usuários têm um padrão nas suas transações que se repete durante um período, porém esse comportamento vai sendo modificado ao longo do tempo por diversos fatores como mudanças na renda e festividades, por exemplo. Essa mudança é conhecida como deriva conceitual e os algoritmos de aprendizagem de máquina, por utilizarem informações supervisionadas, muitas vezes tem dificuldade em lidar com esse comportamento e por isso é necessário atualizar os modelos com os dados mais recentes e remover dados históricos que não representam mais o comportamento dos usuários (DAL POZZOLO, 2015). Ainda hoje a adaptação dos modelos à deriva conceitual é um desafio bastante atual.

2.4 CONTEXTO DA ÁREA

Desde o lançamento dos cartões de pagamentos na década de 1950, os vetores de fraude se estabeleceram ao longo do tempo e se tornaram bem conhecidos na indústria. Até a década de 1970, todas as transações eram processadas por meio de documentos em papel que eram lançados fisicamente. Com o desenvolvimento da tarja magnética para armazenamento dos dados dos usuários, o processo pôde ser automatizado, já que os terminais conseguiam realizar a leitura automaticamente. Foi nesse ponto que as primeiras pesquisas começaram a se concentrar na automação simples da detecção de fraudes e na criação de novos métodos usando regras. Não foi até 1994 que o primeiro trabalho significativo foi publicado neste domínio (RYMAN-TUBB, 2018).

A fraude com cartões de pagamento existe desde a introdução dos cartões no sistema de pagamento. Como o mercado de cartões cresceu rapidamente, o nível de fraude associado aos cartões de pagamento também aumentou. A cada ano, a fraude de cartão custa bilhões de dólares e os números continuam aumentando. De acordo com um relatório divulgado em janeiro de 2010 pelo Aite Group LLC, a fraude com cartão custa à indústria de pagamentos dos Estados Unidos, cerca de US$ 8,6 bilhões por ano. Embora esse valor seja apenas 0,4% dos US$ 2,1 trilhões em volume por ano, esse número continua sendo uma área preocupante para a indústria (SAKHAROVA, 2012). De acordo com o Banco Central Europeu, no continente houve um prejuízo de cerca de 1,26 bilhão de Euros, durante o ano de 2010 em transações de cartão de crédito (BAHNSEN, 2013). Já em 2016, um relatório da CyberSource apontou que o volume

de transações fraudulentas com cartão de crédito no comércio eletrônico na América Latina corresponde a 1,4% do total líquido do setor (SÁ, 2018).

Infelizmente, a sociedade em geral percebe a fraude com cartão de pagamento como um crime menor, onde seus efeitos são mitigados pelo reembolso de qualquer fraude pessoal pelo emissor. O impacto individual para a vítima da fraude é amenizado e por isso existe uma crença comum de que a fraude de pagamento afeta apenas bancos, grandes empresas e governo. No entanto, foi identificado que empresas criminosas e grupos do crime organizado usam fraudes com cartões de pagamento para financiar suas atividades, incluindo armas, drogas e terrorismo (RYMAN-TUBB, 2018). Por mais que a porcentagem de transações fraudulentas seja baixa em relação as transações legítimas, esse valor pode causar um grande impacto na sociedade e nas empresas responsáveis.

Além das perdas financeiras reais, a fraude com cartões de pagamento afeta a confiança do consumidor nos sistemas de pagamentos eletrônicos e na reputação dos emissores de cartões. Os fraudadores modernos são profissionais organizados que usam formas elaboradas para obter detalhes do titular do cartão. Os criminosos continuam a desenvolver novos métodos de ataque, usando todos os tipos de técnicas sofisticadas. Vários participantes da conferência de cartões de pagamento realizada em 2008 notaram que as redes de fraudes criminais empregam modelos que espelham modelos de negócios legítimos: bancos de dados que espelham agências de crédito; e operações de compartilhamento de tempo de banco de dados por meio de sites que fornecem acesso a informações confidenciais roubadas, por uma taxa definida. As instituições financeiras enfrentam uma busca contínua por medidas eficazes de prevenção de fraudes com cartões para minimizar os custos associados e danos à reputação da marca do banco e proteger seus clientes. (SAKHAROVA, 2012).

Os bancos e as principais instituições financeiras prestam uma ampla gama de serviços, mas a fraude é generalizada nas transações financeiras realizadas por essas instituições. Mais serviços gerarão mais dados de usuários, o que oferece uma grande possibilidade para os fraudadores roubarem as informações dos usuários para completar a fraude. Detectar transações fraudulentas com precisão e instantaneamente se tornou um problema urgente de segurança financeira para todas as instituições financeiras, incluindo bancos. O sistema tradicional de regras de especialistas é aplicado à maioria das áreas de detecção de fraude, esses sistemas são baseados nas regras de experiência existentes do setor, que podem detectar os padrões fraudulentos ocorridos e os comportamentos de fraude existentes (ZHANG, 2018). A depender da regra, se o resultado for a suspeita de uma possível fraude, a transação pode ser negada ou um alerta é emitido para investigação posterior. Esse sistema de regras funciona bem desde que

não haja novos padrões de fraude, pois fraudes repetidas são necessárias para que a equipe detecte novos padrões. (BAHNSEN, 2013).

Há outros métodos mais flexíveis para a detecção de fraudes como a detecção de anomalia e a detecção baseada em classificador. A detecção de anomalias se concentra no cálculo da distância entre os pontos de dados no espaço. Ao calcular a distância entre a transação recebida e o perfil do titular do cartão, um método de detecção de anomalia pode filtrar qualquer transação recebida que seja inconsistente com o perfil do titular do cartão. A segunda técnica utiliza alguns métodos de aprendizado supervisionado para treinar um classificador com base nas transações normais e nas fraudes dadas. O aprendizado supervisionado se concentra na extração de recursos de fraude de transações fraudulentas. No entanto, ambos têm limitações. Para a detecção de anomalias, ele não tem capacidade de retratar recursos de fraude, embora possa retratar comportamentos de transação dos titulares de cartão. Para a detecção baseada em classificador, ele não consegue distinguir diferentes comportamentos normais de diferentes titulares de cartão, embora possa detectar comportamentos de fraudadores (JIANG, 2018).

Sabendo disso o próximo capítulo apresenta a metodologia utilizada no trabalho para o desenvolvimento de um estudo comparativo entre modelos de aprendizagem de máquina baseados em classificadores. Mesmo com as limitações listadas para esse tipo de método de detecção, há modelos que desempenham melhor e conseguem bons resultados para a resolução deste problema que são apresentados nos capítulos de resultados e conclusão.

3 METODOLOGIA

Este trabalho seguiu o seguinte método científico: etapa exploratória de trabalhos relacionados, baseada em uma revisão sistemática da literatura. Em seguida, novos experimentos comparativos dos modelos encontrados e com novos modelos propostos.

A pesquisa dos trabalhos relacionados foi realizada nas plataformas IEEE, Scopus, ScienceDirect e ACM Library. Nesta pesquisa, foi utilizado o critério de proximidade com o tema de interesse, disponibilidade do artigo a partir de acesso da instituição de ensino e nota Qualis igual ou superior a B2. Para realizar a busca pelos artigos, foi utilizada a ferramenta de pesquisa avançada contida nas plataformas citadas com a *string* de busca apresentada no Quadro 1.

Quadro 1 – *String* utilizada nas ferramentas de pesquisa avançada

String de busca
"Credit card" **AND** "Fraud" **AND** "Detection" **AND** "Machine Learning"

Fonte: O autor (2022).

Para o treinamento e testagem dos modelos descritos no trabalho, foi utilizado um conjunto de dados de transações de cartão de crédito classificadas como fraudulentas ou legítimas. Este conjunto de dados foi desenvolvido pela Universidade Livre de Bruxelas (ULB) e contém transações reais de usuário de cartão de crédito europeus no ano de 2013, onde 492 delas são fraudes do total de 284,807. Na Tabela 1, tem-se uma representação do conjunto, o qual passou por um processo de transformação com análise de componentes principais, que anonimiza as transações, com isso o conjunto possui 31 colunas que vão de V1 a V28 que são as características anonimizadas da transação, além delas temos o Tempo, Valor e Classificação. A classificação é um valor binário, onde o valor 0 representa uma transação legítima e o valor 1 uma fraude.

Tabela 1 – Parte do conjunto de dados de transações de cartão de crédito da ULB

Tempo	**V1**	**...**	**V28**	**Valor**	**Classificação**
0.0	-1.359807	...	-0.021053	149.62	0
1.0	1.191857	...	0.014724	2.69	0
1.0	-1.358354	...	-0.059752	378.66	0

Fonte: O autor (2022).

Como o conjunto de dados é bastante desbalanceado, foi selecionado um método de redução deste desbalanceamento. A técnica escolhida para o balanceamento do conjunto de dados foi a subamostragem aleatória, método que foi utilizado em Jurgovsky *et al.* (2018) e Sá, Pereira e Pappa (2018). A técnica mantém todos os registros da classe minoritária e remove aleatoriamente as transações da classe majoritária, que no contexto são as transações legítimas. Realizando essa redução dos registros, obtemos um conjunto de dados menor, o que viabiliza o treinamento dos modelos.

Para o treinamento dos modelos selecionados foi utilizada a técnica de validação cruzada. A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Neste trabalho foi utilizada a técnica de validação cruzada chamada k-fold, a qual consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos de mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimar os parâmetros (KOHAVI, 1995). Foram utilizados 30 subconjuntos como parâmetro para aplicação desta técnica e a implementação da biblioteca SKLearn foi utilizada. Os resultados dessas validações foram armazenados a cada iteração para comparação utilizando os testes estatísticos de Wilcoxon.

O teste de Wilcoxon é um teste de hipótese estatística não paramétrica usado para testar a localização de uma população com base em uma amostra de dados ou para comparar as localizações de duas populações usando duas amostras combinadas (CONOVER, 1999). O teste pode ser uma boa alternativa ao teste t quando as médias populacionais não são de interesse; por exemplo, quando se deseja testar se a mediana de uma população é diferente de zero, ou se há mais de 50% de chance de que uma amostra de uma população seja maior que uma amostra de outra população.

Dado esse contexto o teste foi aplicado para todos os pares possíveis de modelos implementados. A validação cruzada foi realizada com os mesmos subconjuntos para todos os modelos e métricas visando viabilizar os testes estatísticos, já que um dos pressupostos é de que os conjuntos de dados vêm da mesma população.

4 TRABALHOS RELACIONADOS

Nesta seção, serão descritos os principais trabalhos na área que utilizam diferentes modelos de aprendizagem de máquina para resolução do problema.

Forough e Momtazi (2021) encararam o problema como uma tarefa de classificação sequencial e como solução propuseram a utilização de um comitê de classificadores de modelos *Long Short-Term Memory* e redes neurais artificiais. A maioria dos estudos até o momento utiliza o sistema de votos baseados em limites, e como ponto de melhoria os autores propuseram o uso do resultado dos classificadores básicos como mecanismo de votação. Esses classificadores são redes neurais recorrentes que tiveram os seus resultados agregados utilizando a arquitetura de rede neural direta.

Também é citado que a detecção de fraudes de cartão de crédito tem dificuldades inerentes como a falta de dados, deriva conceitual, tempo de resposta, custo da solução e pré-processamento dos dados. A deriva conceitual foi um dos problemas que os autores solucionaram a partir da combinação de uma rede sequencial profunda em conjunto com uma janela deslizante para indicar o número de transações que serão considerados no histórico, sendo assim apenas as transações mais recentes estão sendo levadas em consideração para o treinamento do modelo. Além deste, o problema do desbalanceamento dos dados também foi endereçado, já que o modelo apresentado pelos autores se mostrou bastante robusto em ambientes com alto desbalanceamento. Na Tabela 2 temos a comparação dos resultados obtidos nos testes com o modelo ℓ, que foi um modelo de comitê utilizado como base de comparação e o GRU que é outro modelo com base em redes neurais. É possível observar que o modelo utilizando LSTM superou a *performance* dos outros modelos na maioria dos critérios.

Tabela 2 – Resultados dos modelos de comitês de classificadores

Modelo	**Precisão**	**Revocação**	**F1**	**AUC-ROC**	**AUC-PR**
GRU	0.8043	0.689	0.7419	0.8413	0.5655
LSTM	0.8776	0.7144	0.7874	0.8847	0.6746
ℓ	0.9951	0.5811	0.7337	0.7905	0.5939

Fonte: adaptada de Forough e Momtazi (2020).

Já Li *et al.* (2021) focaram seus esforços em analisar o problema do desbalanceamento e sobreposição dos dados de transações de cartão de crédito. A sobreposição ocorre quando vários dados estão em posição similar entre si, de forma que não há distinção clara de comportamento entre eles. Isto ocorre no contexto de fraudes de cartão de crédito, pois há vários

usuários legítimos com diversos comportamentos diferentes, além disso os fraudadores fazem um esforço consciente para tornar suas transações o mais parecidas com transações legítimas dificultando, assim, o processo de detecção de fraudes. Segundo os autores, os métodos atuais de detecção de sobreposição utilizam comparações entre vizinhos, como a técnica KNN, porém no contexto do problema essa solução se torna inviável pelo custo computacional dessa operação ser bastante alto.

Para solucionar o problema, a técnica proposta foi dividida em duas etapas: dividir e conquistar. Na primeira parte o conjunto de dados original é dividido em dois, utilizando como critério a detecção de anomalias, onde são consideradas anomalias todas as transações legítimas. Por considerar as transações fraudulentas como normais, após a aplicação deste método será obtido um subconjunto com várias transações fraudulentas e apenas algumas transações legítimas sobrepostas, são esses dados que serão utilizados na segunda fase. Na parte de conquistar, é utilizado um classificador supervisionado, treinado a partir do subconjunto da primeira fase, para a criação de um modelo em um ambiente com bastante sobreposição. Foram testados vários modelos diferentes para otimização desta segunda etapa do processo de detecção

A conclusão foi que os métodos existentes para o problema de desbalanceamento de classes com sobreposição não levam em consideração o impacto de grandes amostras de detecção de fraude em um conjunto de dados, o que resulta em alto consumo de tempo e baixa eficiência. Além disso, as técnicas de subamostragem aplicadas nos métodos existentes podem levar à perda de informações e à degradação do desempenho geral. Na tabela 3, vemos a comparação dos resultados dos modelos selecionados para teste e a eficácia do método híbrido de redes neurais artificiais com gasto agregado foi a maior nos critérios de F1-Score e AUC_PR.

Tabela 3 – Comparação de modelos apresentados em Li *et al.* (2021)

Modelo	**F1-Score**	**AUC_PR**
RF	0.67	0.57
Tomek Links	0.60	0.50
SMOTE	0.64	0.54
OC-SVM	0.67	0.57
OSM	0.69	0.59
NB-TOMEK	0.66	0.56
Híbrido (OCSVM + RF)	0.70	0.60
Híbrido(iForest + RF)	0.71	0.61
Híbrido (AE + RF)	0.72	0.62
Híbrido (OCSVM + ANN)	0.71	0.61
Híbrido(iForest + ANN)	0.71	0.61
Híbrido (AE + ANN)	0.73	0.63

Fonte: adaptada de Li *et al.* (2021).

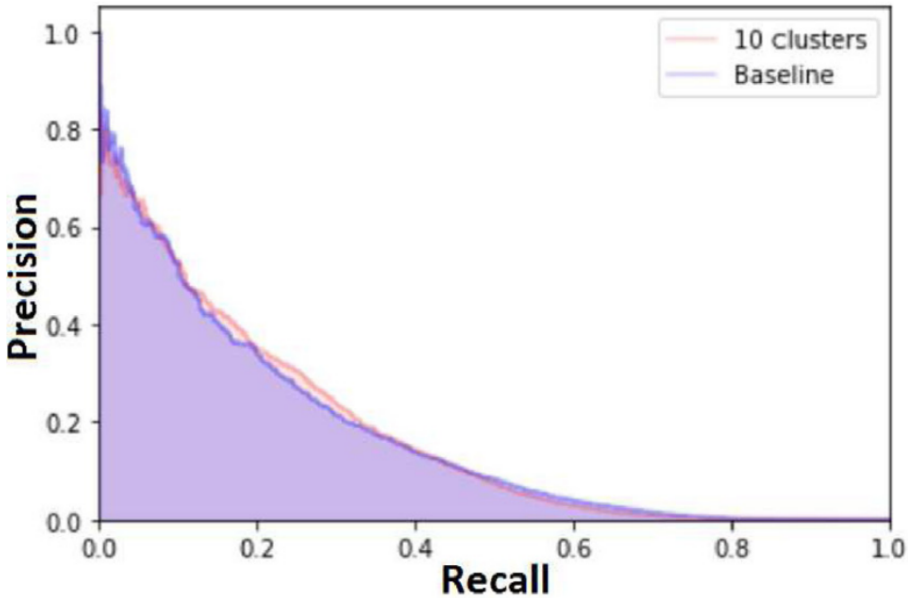
Em Carcillo *et al.* (2019) foram descritos os métodos de classificação com uso de aprendizagem de máquina supervisionada, não supervisionada e semi supervisionada. Foi observado que, em um ambiente real, as transações não são classificadas instantaneamente, sendo assim sempre há um atraso no comportamento encontrado no conjunto de dados em relação ao comportamento mais recente. Como os algoritmos de aprendizagem não supervisionada utilizam a detecção de padrões para realizar a classificação, é possível utilizar um conjunto de dados que não esteja classificado, para o treinamento de um modelo.

Foi proposto, no trabalho anteriormente citado, a utilização de técnicas combinadas de aprendizado supervisionado e não supervisionado. Há na literatura uma descrição sobre utilização dessa combinação de técnicas chamada *best-of-both-worlds*. A partir dessa técnica os autores realizaram uma adaptação para o contexto de detecção de fraudes de cartão de crédito. Analisando o estado da arte nos métodos de detecção, os autores viram que a utilização de métodos de assembleia combinando técnicas supervisionadas e não supervisionadas são bastante utilizados e possuem um resultado melhor do que as técnicas isoladas.

A proposta de solução do problema dada pelos autores é da criação de uma extensão do princípio *best-of-both-worlds* para a definição de uma pontuação de valores atípicos, levando em consideração diferentes níveis de granularidade e a integração com técnicas supervisionadas. A solução é baseada em uma pontuação de valores atípicos, que significa classificar as transações com valores que representam o quanto elas destoam das transações legítimas. Para definição de valores destoantes foram definidos três níveis de granularidade o

global, local e de conjunto. O global leva em consideração todas as transações, já o local apenas a transações feitas por um mesmo cartão e a de conjunto é uma granularidade intermediária. Os resultados não são convincentes em termos de abordagens globais e locais. A interpretação dos autores foi de que ambas as abordagens não têm o nível certo de granularidade necessário para aproveitar as informações não supervisionadas. Um resultado mais promissor foi obtido através da abordagem de conjunto, principalmente em termos de AUC-PR, embora pareça que aumentar muito os conjuntos de dados parece ser prejudicial devido ao problema de sobreajuste e variância. Na Figura 2 é possível observar a relação entre precisão e revocação para 10 conjuntos e a linha base.

Figura 2 – Curva da relação entre precisão e revocação



Fonte: Carcillo *et al.* (2019).

Jurgovsky *et al.* (2018) formularam o problema de detecção de fraudes de cartão de crédito como uma tarefa de classificação de sequencial e para solução empregaram redes de *Long Short-Term Memory*. Como os autores encararam o problema como uma classificação sequencial, foi necessária uma engenharia de características no conjunto de dados para armazenar variáveis ao longo de sequências de transações de cada portador de cartão de crédito. A rede de *Long Short-Term Memory* tinha duas camadas recorrentes, uma única camada oculta e um classificador de regressão logística empilhado no topo da última camada. O classificador de regressão logística pode ser treinado em conjunto com o modelo de transição de estado *Long Short-Term Memory* via retropropagação de erros. Para os nós do *Long Short-Term Memory,* foi aplicada a técnica de *dropout*, visando regularizar os parâmetros e o modelo foi treinado minimizando a entropia cruzada entre a distribuição de classe prevista e a verdadeira com o

algoritmo ADAM. Para os testes, a técnica proposta foi comparada com um classificador de floresta aleatória de linha de base e mostrou que o *Long Short-Term Memory* melhora a precisão da detecção em transações offline em que o titular do cartão está fisicamente presente em um comércio. Ambas as abordagens de aprendizagem sequencial e não sequencial se beneficiam fortemente das estratégias manuais de agregação de recursos. Uma análise posterior de verdadeiros positivos revelou que ambas as abordagens tendem a detectar fraudes diferentes, o que sugere uma combinação das duas. A tabela 4 apresenta os resultados obtidos a partir da comparação dos modelos RF e LSTM, das características geradas e do tipo de sequência.

Tabela 4 – Resultados dos modelos utilizando sequência longa e curta

		AUCPR	
Sequência	**Característica**	**RF**	**LSTM**
CURTA	BASE	0	0,18
	TDELTA	0,236	0,192
	AGG	0,394	0,38
LONGA	BASE	0,179	0,178
	TDELTA	0,228	0,238
	AGG	0,404	0,402

Fonte: adaptada de Jurgovsky *et al.* (2018).

A técnica de assembleia para classificação de transações foi proposta em Dal Pozzolo *et al.* (2015). O método proposto foi o treinamento de dois modelos utilizando conjunto de dados diferentes, sendo um deles um conjunto com transações classificadas manualmente por profissionais especialistas e outro formado por transações que foram classificadas por um sistema de detecção de fraudes automaticamente. As novas classificações seriam definidas pela assembleia desses dois modelos de RF com 100 árvores cada. Com o foco em superar o problema da deriva conceitual, foi adicionada a utilização da técnica de janelas deslizantes para restringir o conjunto de dados utilizados no treinamento para apenas as transações mais recentes. A conclusão dos autores foi que a utilização de métodos de classificação separados para classificações tardias e feedbacks de especialistas aumentou a resistência da solução ao problema da deriva conceitual. Os classificadores que utilizaram feedbacks dos especialistas são mais precisos do que os de classificação tardia por utilizarem dados de transações mais atualizados.

Lucas *et al.* (2019) falam sobre considerar as transações como uma sequência de eventos e não um evento único. Muitos dos artigos relacionados a detecção de fraude com aprendizagem de máquina trazem a abordagem de classificar uma transação a partir de seus atributos de forma

individual sem considerar os eventos no tempo. A abordagem proposta aqui é considerar a sequência de transações e avaliar se essa sequência contém alguma fraude, para isso foi proposta a criação de oito características no conjunto de dados, baseadas em HMM. Elas quantificam a similaridade entre o histórico de uma transação e oito distribuições aprendidas anteriormente em conjunto de sequências selecionadas de forma supervisionada para modelar diferentes perspectivas. Os autores selecionaram três perspectivas para modelar uma sequência de transações. Uma sequência pode ser feita apenas de transações históricas genuínas ou pode incluir pelo menos uma transação fraudulenta, pode vir de um titular de um cartão fixo ou de um terminal fixo, e pode consistir em valor financeiro ou de valores temporal. Os resultados do trabalho mostram um aumento na métrica de AUC-PR em 15,1% em comparação às estratégias de engenharia de recursos mais avançadas. Segundo os autores este aumento se deu devido à adição de novos recursos baseados em HMM multiperspectiva.

Sá, Pereira e Pappa (2018) propuseram um método utilizando o método de hiper-heurística evolucionário para criação de um algoritmo de classificadores de redes bayesianas customizado. A geração de um algoritmo de classificadores de redes bayesianas para um conjunto de dados de detecção de fraudes em cartões de crédito do mundo real, foi uma das principais contribuições deste trabalho. Os autores também contribuíram com o método de avaliação do desempenho do algoritmo, em termos de métricas de classificação e financeiras, aquelas utilizadas por especialistas em finanças, para avaliar os níveis de fraude. Neste trabalho são utilizadas métricas comuns para classificação de algoritmos de aprendizagem de máquina como F1, precisão e revocação e métricas financeiras para uma análise de impacto no negócio. O trabalho também consegue fazer uma análise completa do algoritmo customizado em termos de estratégias para lidar com o conjunto de dados desbalanceado, a estratégia utilizada para isso foi a subamostragem aleatória. Como resultado, o modelo desenvolvido conseguiu desempenhar melhor que as técnicas atuais utilizadas pelo PagSeguro, empresa fornecedora do conjunto de dados, em até 72,64%.

Ryman-Tubb, Krause e Garn (2018) realizaram uma pesquisa analisando historicamente os métodos de detecção de fraude de cartão de crédito e como está ocorrendo a evolução das soluções. Segundo as pesquisas, os métodos de detecção de fraude vêm crescendo a passos lentos, com ainda poucas pesquisas relevantes na área, na visão deles isso é justificado pela falta de mudança no comportamento dos fraudadores. Além disso é argumentado que, no senso comum, a fraude de cartão de crédito é um crime pequeno que afeta apenas as grandes empresas, bancos e governos, já que o prejuízo individual dos clientes é amenizado por essas

organizações. Porém, na realidade o problema é muito maior, já que vai além do prejuízo individual, pois possibilita o financiamento de instituições criminosas.

Segundo os autores, as soluções propostas não tiveram sucesso suficiente nem no corpo de trabalho pesquisado nem nas soluções implantadas, existem duas explicações para o fracasso desses métodos: A primeira é de que há pouco incentivo da indústria para melhorá-los, pois os casos de fraude são julgados como um custo do negócio e são vistos como normativos. As pesquisas do setor indicam que, apesar da validade acadêmica da pesquisa, seu impacto no setor de cartões de pagamento foi mínimo. Já a segunda explicação é que o trabalho acadêmico nesta área é difícil e marginalizado em termos de financiamento.

Concluiu-se que há uma lacuna nas pesquisas para ajudar a reduzir as fraudes de cartões de crédito na indústria e isso é refletido pelos resultados apontados no estudo. Na tabela 5 são listados os sete melhores resultados em algoritmos de detecção de fraude. Pelas datas dos trabalhos apresentados é possível perceber que não houve um grande avanço na área, pois alguns dos melhores desempenhos são de mais de 15 anos atrás. A coluna de descrição da tabela 5 especifica o tipo de modelo utilizado, o *Expert* descreve as soluções que utilizam regras criadas por especialistas em fraudes, já o *Neural* se aplica a soluções criadas com redes neurais, no caso da descrição *Eclectic* os modelos são normalmente bastante complexos e não se encaixam em nenhuma categoria anterior. Por fim, a descrição DT descreve modelos baseados em árvores de decisão. A coluna A/F descreve a quantidade de alertas para cada fraude, e foi utilizada como métrica para avaliação de desempenho dos algoritmos, quanto menor este valor, melhor o desempenho do modelo.

Tabela 5 – Resultados dos métodos analisados em Ryman-Tubb, Krause e Garn (2018)

Descrição	**Trabalho**	**A/F**	**%FP**	**%VP**	**%Erro**
Expert	Correia *et al*. (2015)	2	0.020	80.00	20.00
Neural	Ghosh and Reilly (1994)	5	0.090	40.00	12.24
Neural	Ryman-Tubb (2016)	7	0.001	75.56	24.44
Neural	Richardson (1997)	7	0.130	61.41	38.59
Eclectic	Carminati *et al*. (2014)	11	0.190	98.00	2.00
Eclectic	Salazar *et al*. (2012)	18	0.200	60.00	40.00
DT	Dal Pozzolo *et al*. (2017)	11	0.195	94.43	5.57

Fonte: adaptada de Ryman-Tubb, Krause e Garn (2018).

Em Bahnsen *et al.* (2013), os autores propõem a utilização da técnica de risco mínimo de Bayes para detecção de fraude. Neste artigo há a utilização de uma nova métrica para comparação de eficácia dos algoritmos de detecção, os autores afirmam que a utilização de

métricas clássicas de aprendizagem de máquina não é suficiente para casos de fraudes de cartão de crédito, já que o que interessa para as empresas no final é a otimização de custo. A redução dos custos com fraude pode vir de modelos que não apresentam estritamente os melhores resultados nas métricas mais convencionais, porém atingem o melhor resultado em termos financeiros. Foram selecionados os métodos RF, LR e C4.5 como base para a criação de modelos otimizados com a técnica de risco mínimo de bayes, os quais foram chamados de RF-MR, LR-MR e C4.5-MR. As avaliações confirmaram que incluir o custo real, criando um sistema sensível ao custo, usando um classificador de risco mínimo Bayes, dá origem a resultados de detecção de fraude muito melhores no sentido de maior economia. O melhor modelo otimizado pelos autores foi o RF-MR, que apresentou uma redução de custos com fraudes em 23% se comparado com os modelos clássicos

O método proposto pelos autores em Xuan *et al.* (2018) foi a utilização de duas implementações diferentes do algoritmo floresta aleatória. A escolha do método se deu pela análise da popularidade e eficácia dos modelos de árvore de decisão, por serem simples e flexíveis esses modelos se tornam uma escolha muito interessante, porém os modelos de árvore única sofrem com o sobreajuste e técnicas de assembleia podem resolver esse problema. As florestas propostas neste trabalho são diferentes pelo seu classificador base, a primeira é baseada em árvores aleatórias e a segunda baseada em CART, as duas foram treinadas a partir de um mesmo conjunto de transações, que por se tratar de um conjunto bastante desbalanceado passou por um processo de subamostragem. As conclusões foram que embora a floresta aleatória obtenha bons resultados em dados de conjuntos pequenos, ainda existem alguns problemas, como desbalanceamento de dados. O algoritmo da floresta aleatória em si deve ser melhorado, por exemplo, o mecanismo de votação assume que cada um dos classificadores básicos tem o mesmo peso, mas alguns deles podem ser mais importantes que outros. Portanto, os autores concluíram há espaço para melhoria do algoritmo aplicado ao problema. Na tabela 6 é possível analisar os resultados obtidos para cada umas das métricas avaliadas para as duas árvores citadas no trabalho.

Tabela 6 – Resultados dos métodos analisados em Xuan *et al.* (2018)

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
Random Forest I	91.96%	90.27%	67.89%	0.7811
Random Forest II	96.77%	89.46%	95.27%	0.601

Fonte: adaptada de Xuan *et al.* (2018).

Dhankhad, Mohammed e Far (2018) realizaram um estudo de avaliação e comparação de algoritmos de aprendizagem supervisionada aplicados a detecção de fraudes de cartão de crédito. Neste trabalho foram avaliados os seguintes modelos: LR, RF, DT, KNN, GBT, XGB, SVM e MLP e para isso foi utilizado o conjunto de dados de transações de cartão de crédito da Europa no ano de 2013. Com 70% dos dados voltados para realização do treinamento dos algoritmos e 30% para os testes e como técnica de balanceamento foi utilizada a subamostragem. Os autores, a partir deste cenário, chegaram à conclusão que a LR apresenta os melhores resultados na classificação, seguidos da RF e XGB. Na tabela 7 é possível analisar os resultados dos modelos que apresentaram os melhores desempenhos, segundo avaliação feita pelos autores.

Tabela 7 – Resultados dos métodos analisados em Dhankhad, Mohammed e Far (2018)

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	94.59%	95%	95%	95%
XGB	94.59%	95%	95%	95%
KNN	94.25%	91%	91%	91.83%
LR	93.91%	94%	94%	94%
GB	93.58%	94%	94%	94%
DT	90.8%	91%	91%	91%
SVM	93.23%	93%	93%	93%

Fonte: adaptada de Dhankhad, Mohammed e Far (2018).

Jiang *et al.* (2018) propõem para solução do problema de detecção de fraudes uma abordagem em quatro etapas. A intenção dos autores na criação dessas etapas é extrair o comportamento dos usuários da melhor forma possível, sendo assim a primeira etapa é o pré-processamento dos dados e envolve o agrupamento da base de usuários por seu nível de gastos nas transações. Com esse agrupamento, novas características são extraídas e o conjunto de dados é enriquecido com dados agrupados. Já a segunda etapa deste método é realizar um novo agrupamento, mas agora utilizando o critério de padrão de comportamento dos donos dos cartões de crédito. O próximo passo é classificar os padrões de comportamento analisado pelas etapas anteriores e associá-los aos usuários que o possuem. A última etapa envolve a atualização do perfil comportamental dos usuários a partir de mecanismos de *feedback*, o objetivo desta é vencer a barreira da deriva conceitual. Após todas as etapas de agregação os algoritmos selecionados para detecção foram o LR e RF. Utilizando esses algoritmos foram criados modelos utilizando o novo conjunto de dados com características comportamentais dos usuários e modelos utilizando apenas o conjunto de dados inicial. O método proposto pelos autores

apresentou, em média, melhores resultados de revocação e acurácia em relação ao treinamento dos modelos utilizando o conjunto de dados inicial. Os resultados foram ainda melhores com a utilização da técnica de feedback.

5 ESTUDO COMPARATIVO

5.1 MÉTODO PROPOSTO

Implementar e testar os modelos utilizando o conjunto de dados da Universidade Livre de Bruxelas (ULB)[1]. Para este trabalho os modelos utilizados foram selecionados a partir dos artigos de referência. Para escolher os algoritmos que seriam implementados, foram utilizados os seguintes critérios: modelos que não utilizassem a técnica de comitê de classificadores, por questões de complexidade computacional exigida e modelos que possuem implementação descrita, pois parte dos artigos descreve os resultados dos modelos, porém sem detalhes da implementação. A partir desses critérios, os modelos selecionados para implementação foram: CNN, DT, GB, KNN, LR, RF, Poly SVM, Sigmoid SVM e três arquiteturas diferentes de NN. O modelo CNN foi implementado utilizando o módulo do TensorFlow, já os outros modelos citados são provenientes do módulo SKLearn.

Como critério de comparação, foram utilizados métodos bem estabelecidos de avaliação de modelos de aprendizagem de máquina como acurácia, precisão, revocação, F1-Score e ROC AUC. No quadro 2 há um detalhamento dos critérios citados e como eles são calculados, onde VP representa a quantidade de verdadeiros positivos, o FP representa a quantidade de falsos positivos, VN a quantidade de verdadeiros negativos e FN a de falsos negativos.

Quadro 2 – Forma dos cálculos dos critérios de comparação

Critério	**Cálculo**
Acurácia	(VP +VN) / VP + FP + VN + FN
Precisão	VP / VP + FP
Revocação	VP / VP + FN
F1-Score	(2 * Precisão * Revocação) / Precisão + Revocação
ROC AUC	área do gráfico onde o eixo X é VP e o eixo Y é FP

Fonte: O autor (2022).

Para que a comparação das métricas entre cada modelo fosse realizada de forma precisa, foi utilizado o teste estatístico Wilcoxon, que é um teste não paramétrico pareado, e uma alternativa para o teste t-Student quando as amostras não seguem distribuição normal. No quadro 3, é descrito quais hipótese nula e alternativa foram utilizadas.

[1] Disponível em: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud> Acesso em: 22 set, 2022.

Quadro 3 – Hipóteses utilizadas no teste Wilcoxon

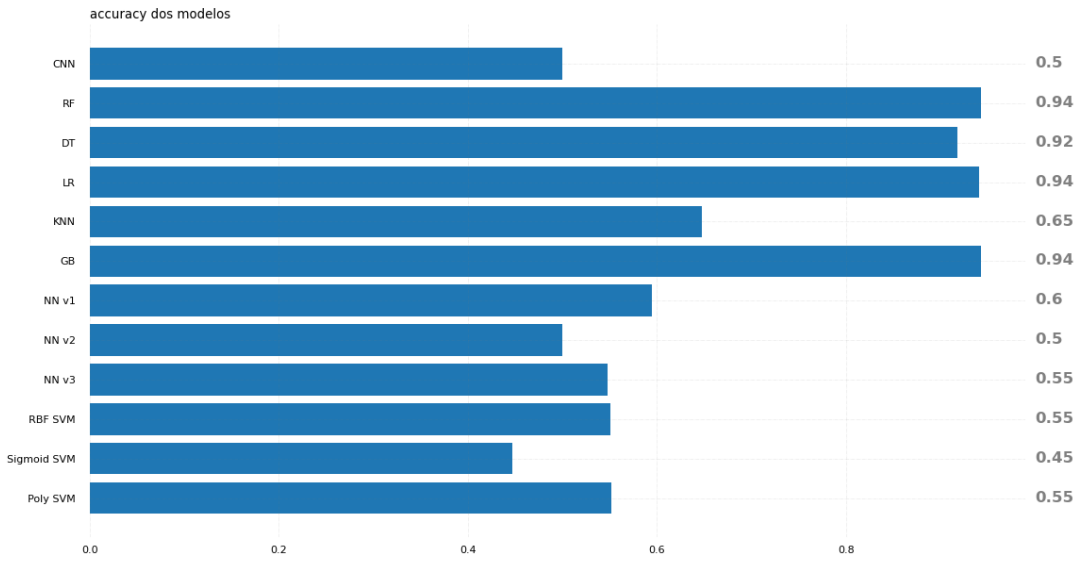
Hipótese	
H0	A distribuição da métrica avaliada é igual para o par de modelos testados
H1	A distribuição da métrica avaliada é diferente para o par de modelos testados

Fonte: O autor (2022).

5.2 RESULTADOS DA AVALIAÇÃO

Os resultados foram obtidos realizando o método proposto na sessão anterior. Para análise desses resultados, foi comparado o desempenho dos modelos para cada métrica individualmente. Os valores foram gerados a partir de testes de validação cruzada, utilizando 30 subconjuntos, com esses resultados diagramas de caixa foram criados, utilizando todos os valores obtidos. Gráficos em barra também foram gerados a partir da média dos valores obtidos em cada métrica, para cada modelo.

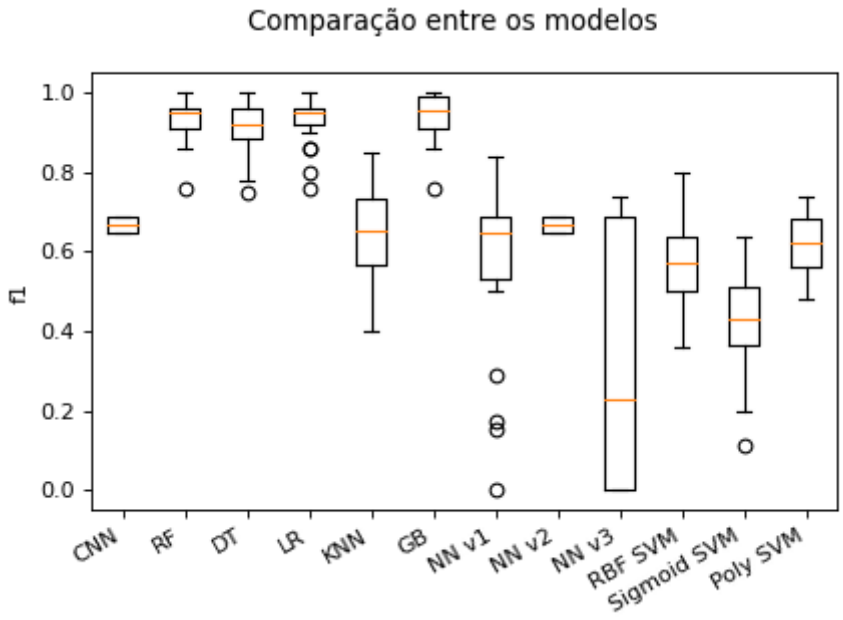
Figura 3 – Gráfico em barra da métrica acurácia para os modelos testados



Fonte: O autor (2022).

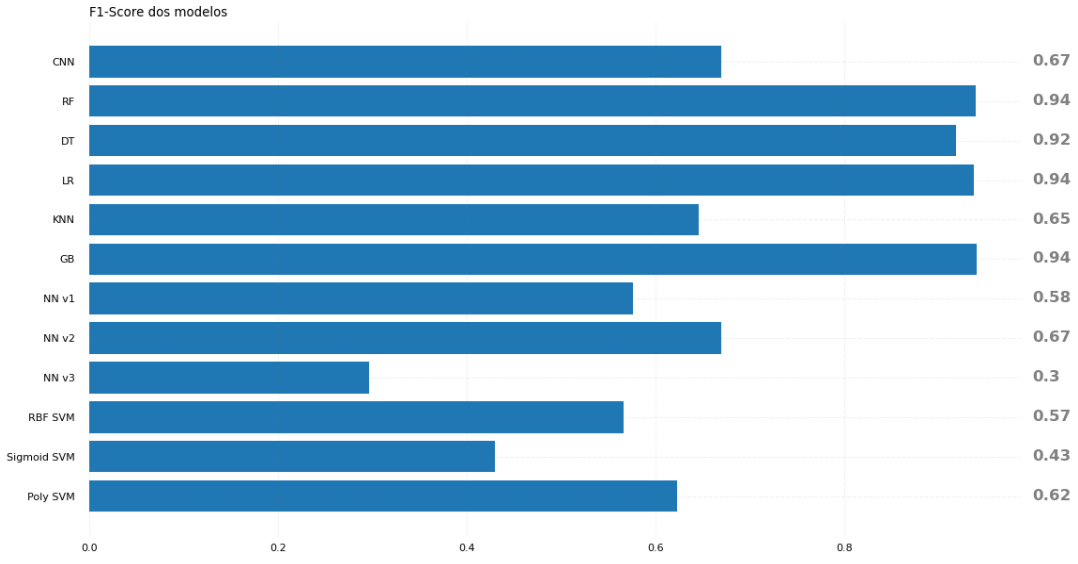
A acurácia é a métrica que avalia qual percentual das previsões foi realizado corretamente, por isso seu cálculo é feito a partir do número de previsões corretas sobre o número total de previsões. Na figura 3 é evidenciado que em relação à acurácia, há resultados bastante favoráveis para os modelos como RF, LR, GB que obtiveram o maior resultado com pontuação de 0.94, já o modelo DT obteve um resultado bastante similar de 0.92.

Figura 4 – Gráfico da métrica F1-Score para os modelos testados



Fonte: O autor (2022).

Figura 5 – Gráfico em barra da métrica F1-Score para os modelos testados

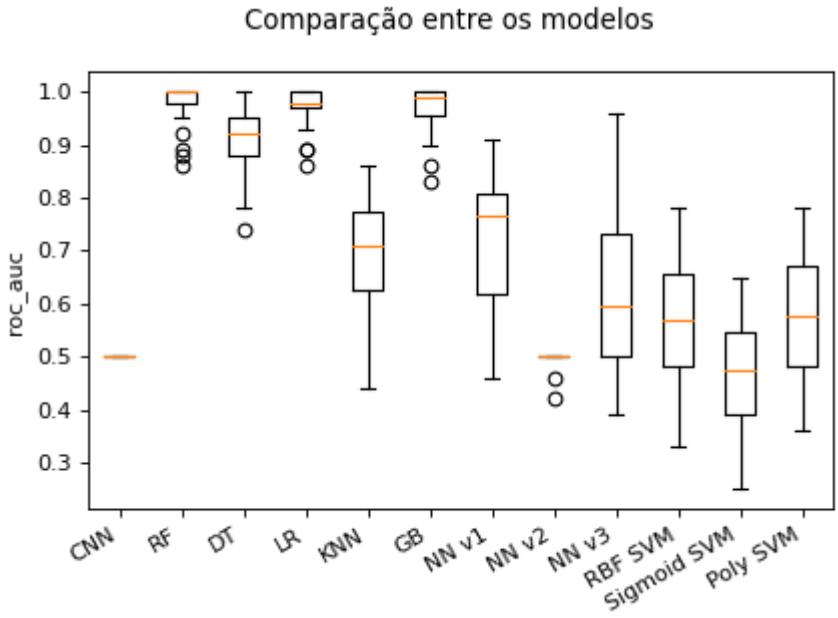


Fonte: O autor (2022).

Na figura 5 é evidenciado que em relação ao F1-Score, há resultados bastante favoráveis para os modelos como RF, LR, GB que obtiveram o maior resultado com pontuação de 0.94, já o modelo DT obteve um resultado bastante similar de 0.92. Na figura 4 é possível observar que

mesmo com bons resultados o modelo LR apresenta mais valores discrepantes em relação aos seus pares de mesma pontuação. Enquanto isso os outros modelos não apresentaram resultados muito interessantes. A pontuação F1 é a média harmônica ponderada de precisão e evocação que é uma medida abrangente e de equilíbrio (Li *et al.* 2021), sendo assim muito importante para treinamento de modelos onde o conjunto de dados é bastante desbalanceado.

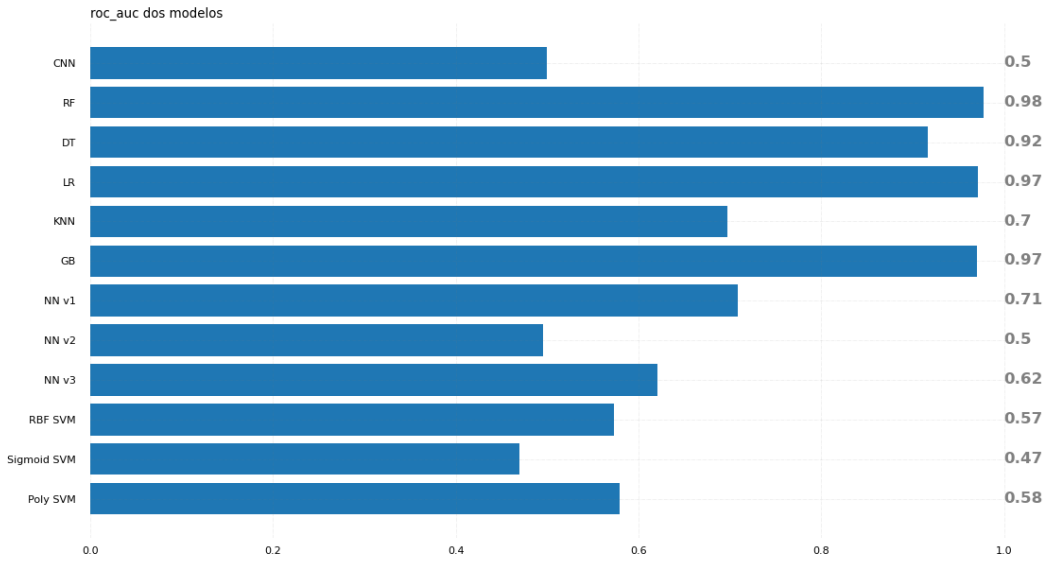
Figura 6 – Gráfico da métrica ROC-AUC para os modelos testados



.

Fonte: O autor (2022).

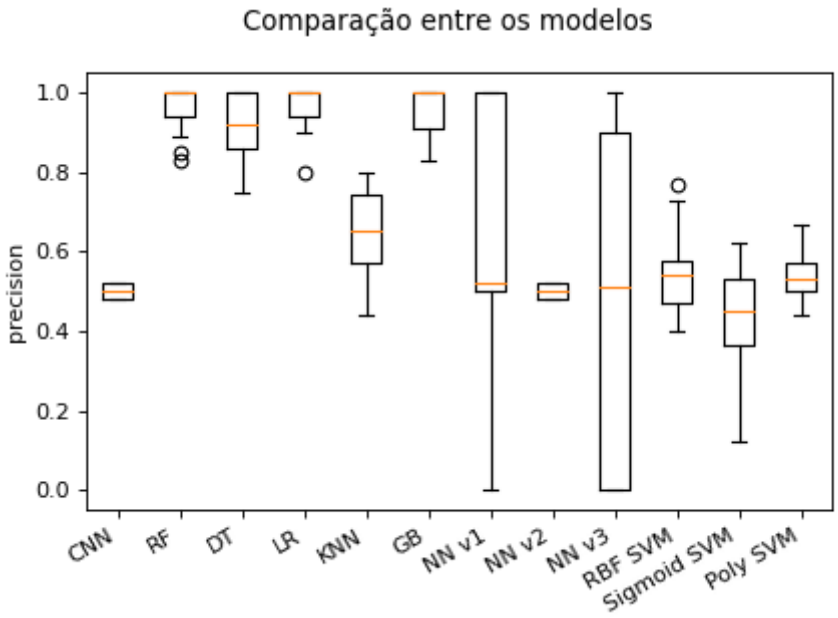
Figura 7 – Gráfico em barras da métrica ROC-AUC para os modelos testados



Fonte: O autor (2022).

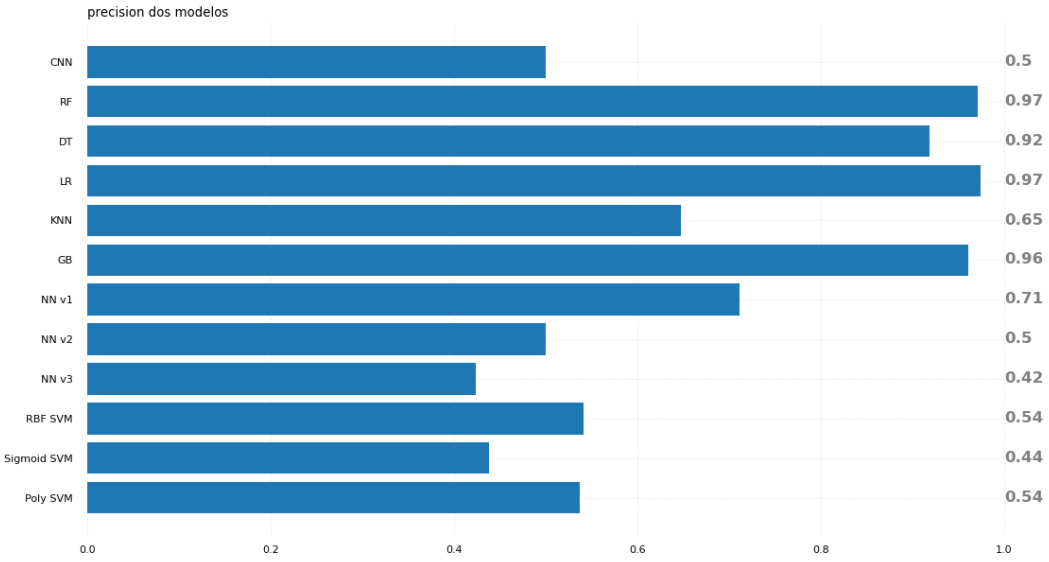
A área sob a curva ROC é uma métrica que relaciona a recordação da classe positiva (taxa de verdadeiro positivo) com a recordação da classe negativa (taxa de falso positivo). No contexto de fraude de cartão de crédito, o número de exemplos negativos excede em muito o número de exemplos positivos. Consequentemente, uma grande mudança no número de falsos positivos seria refletida pelo ROC por meio de uma mudança muito pequena na taxa de falsos positivos (JURGOVSKY, 2018). É possível ver a partir da figura 7 que os modelos que apresentam os melhores resultados são os modelos mais simples e que também possuíram bons resultados no F1-Score, como RF, LR, GB. Na figura 6 é possível observar que o RF, apesar de possuir a maior mediana, também possui um número elevado de *outliers*, já o GB possui uma mediana melhor se comparada ao LR, porém com maior dispersão.

Figura 8 – Gráfico da métrica precisão para os modelos testados



Fonte: O autor (2022).

Figura 9 – Gráfico da métrica precisão para os modelos testados

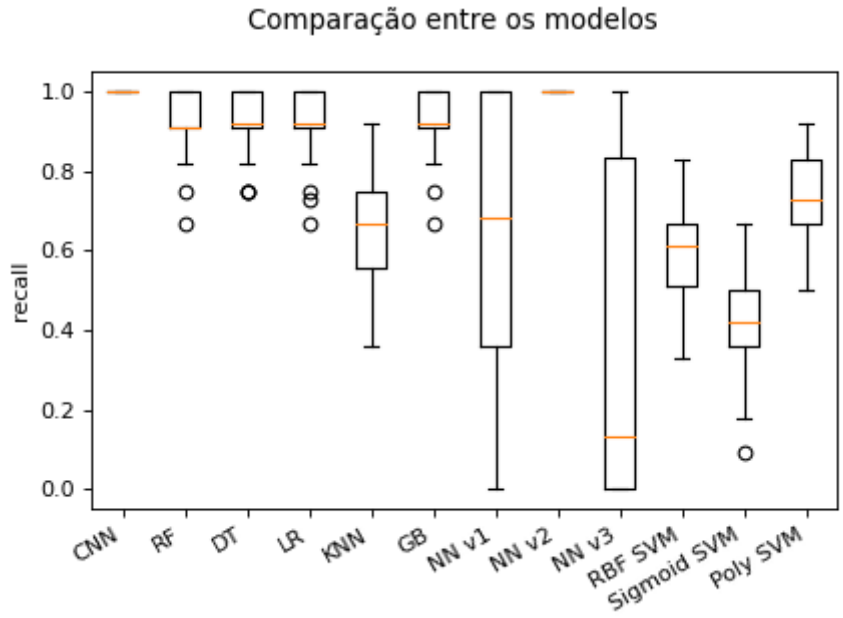


Fonte: O autor (2022).

A precisão das transações de fraude é o número de transações fraudulentas identificadas corretamente, sobre o número total de transações classificadas como fraude. A taxa de falsos positivos é o valor que representa o número de transações genuínas que foram erroneamente identificadas como fraude dentre todas as transações genuínas conhecidas (RYMAN-TUBB, 2018) e esse valor é utilizado para o cálculo da métrica de precisão. A partir da figura 9 é

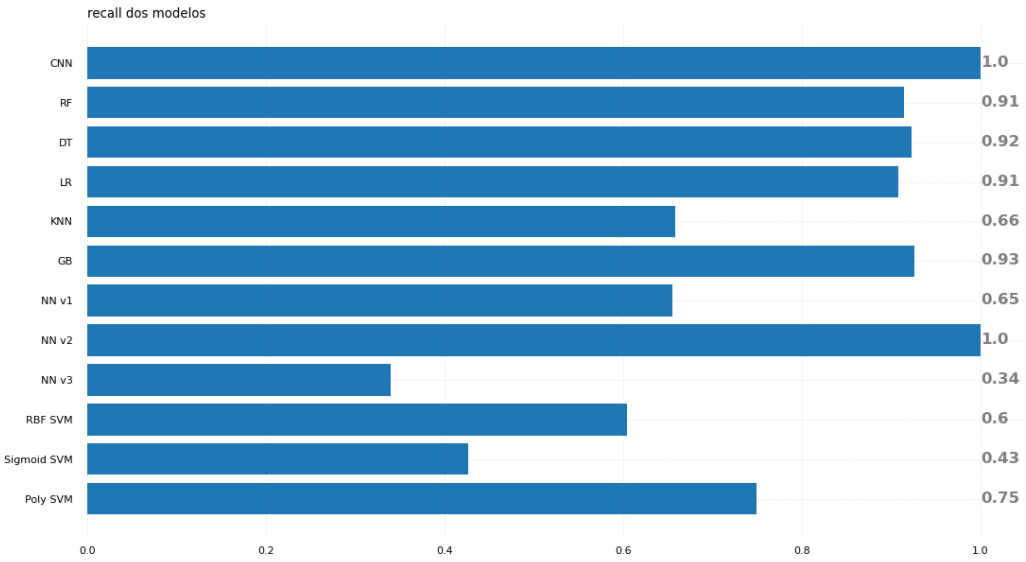
possível concluir que os modelos que apresentaram os melhores resultados são RF, LR e GB com valores bastante similares aos das métricas citadas anteriormente. Na figura 8 é possível ver que o modelo RF apresenta resultados muito próximos aos do modelo LR, porém com um maior número de *outliers*, já o modelo GB possui uma dispersão maior em relação aos outros dois modelos.

Figura 10 – Gráfico da métrica revocação para os modelos testados



Fonte: autor (2022).

Figura 11 – Gráfico em barras da métrica revocação para os modelos testados



Fonte: O autor (2022).

A taxa de revocação mede a taxa de detecção de todos os casos de fraude (XUAN, 2018). A figura 11 apresenta as porcentagens de revocação para cada um dos modelos e nele é possível ver modelos que não apresentaram resultados expressivos nas outras métricas, ganhando destaque como é o caso do CNN e NN v2. Na figura 10, é possível observar que os modelos CNN e NN v2 possuem resultados com mediana no valor máximo e nenhum *outlier*. Apresentar um bom resultado nesta métrica significa que o modelo tem uma boa taxa de identificação de fraudes, porém ela não consegue representar a taxa de falsos positivos, por isso é importante a utilização das métricas de forma conjunta.

5.3 RESULTADOS DOS TESTES ESTATÍSTICOS

Com os resultados de cada métrica avaliada, foram realizados os testes estatísticos, os quais são apresentados nas tabelas a seguir. Os registros destacados em negrito são os valores de alfa menores que 0.05, o que indica que os modelos avaliados no teste não possuem a mesma distribuição entre si. Para os conjuntos de modelos que não possuem a mesma distribuição, foram realizadas comparações das médias dos valores de cada métrica para determinar os modelos que obtiveram os melhores resultados

Tabela 8 – Resultados dos testes de Wilcoxon para métrica F1-Score

	CNN	**RF**	**DT**	**LR**	**KNN**	**GB**	**NN v1**	**NN v2**	**NN v3**	**RBF SVM**	**Sigmoid SVM**	**Poly SVM**
CNN	1.0											
RF	**0.0**	1.0										
DT	**0.0**	0.05	1.0									
LR	**0.0**	0.97	0.06	1.0								
KNN	**0.0**	**0.0**	**0.0**	**0.0**	1.0							
GB	**0.0**	0.73	**0.01**	0.93	**0.0**	1.0						
NN v1	**0.01**	**0.0**	**0.0**	**0.0**	0.84	**0.0**	1.0					
NN v2	1.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.01**	1.0				
NN v3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0			
RBF SVM	**0.0**	**0.0**	**0.0**	**0.0**	0.38	**0.0**	0.43	**0.0**	**0.0**	1.0		
Sigmoid SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.01**	**0.0**	1.0	
Poly SVM	**0.0**	**0.0**	**0.0**	**0.0**	0.13	**0.0**	0.51	**0.0**	**0.0**	**0.01**	**0.0**	1.0

Fonte: O autor (2022).
Nota: Os dados destacados em negrito representam os valores que não passaram na hipótese nula

A tabela 8 apresenta os resultados do teste estatístico feito com a métrica F1-Score. Segundo o resultado dos testes realizados, os modelos CNN e NN v2 apresentaram a mesma distribuição entre si, assim como o conjunto de modelos RF, LR, DT e GB, além deles os modelos KNN, NN v1, RBF SVM e Poly SVM também apresentaram a mesma distribuição entre si. Comparando as médias dos modelos pertencentes a cada conjunto, chega-se à conclusão de que os modelo RF, LR, DT e GB apresentaram os melhores resultados para essa métrica.

Tabela 9 – Resultados dos testes de Wilcoxon para métrica revocação

	CNN	**RF**	**DT**	**LR**	**KNN**	**GB**	**NN v1**	**NN v2**	**NN v3**	**RBF SVM**	**Sigmoid SVM**	**Poly SVM**
CNN	1.0											
RF	**0.0**	1.0										
DT	**0.0**	0.27	1.0									
LR	**0.0**	0.8	0.23	1.0								
KNN	**0.0**	**0.0**	**0.0**	**0.0**	1.0							
GB	**0.0**	0.2	0.99	0.15	**0.0**	1.0						
NN v1	**0.0**	**0.01**	**0.0**	**0.01**	0.08	**0.01**	1.0					
NN v2	1.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0				
NN v3	**0.0**	**0.0**	**0.0**	**0.0**	**0.01**	**0.0**	**0.01**	**0.0**	1.0			
RBF SVM	**0.0**	**0.0**	**0.0**	**0.0**	0.25	**0.0**	0.43	**0.0**	**0.0**	1.0		
Sigmoid SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.01**	**0.0**	0.33	**0.0**	1.0	
Poly SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.16	**0.0**	**0.0**	**0.0**	**0.0**	1.0

Fonte: O autor (2022).
Nota: Os dados destacados em negrito representam os valores que não passaram na hipótese nula

A tabela 9 apresenta os resultados do teste estatístico feito com a métrica revocação. Segundo o resultado dos testes realizados, os modelos CNN e NN v2 apresentaram a mesma distribuição entre si, assim como o conjunto de modelos RF, LR, DT e GB, além deles os modelos KNN, NN v1 e RBF SVM também apresentaram a mesma distribuição entre si, assim como os modelos NN v3 e Sigmoid SVM. Comparando as médias dos modelos pertencentes a cada conjunto, chega-se à conclusão de que os modelo CNN e NN v2 apresentaram os melhores resultados para essa métrica.

Tabela 10 – Resultados dos testes de Wilcoxon para métrica acurácia

	CNN	**RF**	**DT**	**LR**	**KNN**	**GB**	**NN v1**	**NN v2**	**NN v3**	**RBF SVM**	**Sigmoid SVM**	**Poly SVM**
CNN	1.0											
RF	**0.0**	1.0										
DT	**0.0**	**0.01**	1.0									
LR	**0.0**	0.9	**0.03**	1.0								
KNN	**0.0**	**0.0**	**0.0**	**0.0**	1.0							
GB	**0.0**	1.0	**0.01**	0.89	**0.0**	1.0						
NN v1	**0.0**	**0.0**	**0.0**	**0.0**	0.4	**0.0**	1.0					
NN v2	1.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0				
NN v3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.1	**0.0**	1.0			
RBF SVM	**0.02**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.06	**0.02**	0.95	1.0		
Sigmoid SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0	
Poly SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.08	**0.0**	0.91	0.82	**0.0**	1.0

Fonte: O autor (2022).
Nota: Os dados destacados em negrito representam os valores que não passaram na hipótese nula

A tabela 10 apresenta os resultados do teste estatístico feito com a métrica acurácia. Segundo o resultado dos testes realizados, os modelos CNN e NN v2 apresentaram a mesma distribuição entre si, assim como o conjunto de modelos RF, LR, e GB, além deles os modelos KNN e NN v1 também apresentaram a mesma distribuição entre si, assim como os modelos NN v1, NN v3, RBF SVM e Sigmoid SVM. Comparando as médias dos modelos pertencentes a cada conjunto, chega-se à conclusão de que os modelo RF, LR, e GB apresentaram os melhores resultados para essa métrica.

Tabela 11 – Resultados dos testes de Wilcoxon para métrica ROC AUC

	CNN	**RF**	**DT**	**LR**	**KNN**	**GB**	**NN v1**	**NN v2**	**NN v3**	**RBF SVM**	**Sigmoid SVM**	**Poly SVM**
CNN	1.0											
RF	**0.0**	1.0										
DT	**0.0**	**0.0**	1.0									
LR	**0.0**	0.14	**0.0**	1.0								
KNN	**0.0**	**0.0**	**0.0**	**0.0**	1.0							
GB	**0.0**	0.12	**0.0**	0.55	**0.0**	1.0						
NN v1	**0.0**	**0.0**	**0.0**	**0.0**	0.05	**0.0**	1.0					
NN v2	0.5	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0				
NN v3	**0.0**	**0.0**	**0.0**	**0.0**	0.14	**0.0**	**0.01**	**0.0**	1.0			
RBF SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.03**	1.0		
Sigmoid SVM	0.14	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.17	**0.0**	**0.0**	1.0	
Poly SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.04**	0.42	**0.0**	1.0

Fonte: O autor (2022).
Nota: Os dados destacados em negrito representam os valores que não passaram na hipótese nula

A tabela 11 apresenta os resultados do teste estatístico feito com a métrica ROC AUC. Segundo o resultado dos testes realizados, os modelos CNN, Sigmoid SVM e NN v2 apresentaram a mesma distribuição entre si, assim como o conjunto de modelos RF, LR, e GB, além deles os modelos KNN, NN v1 e NN v3 também apresentaram a mesma distribuição entre si, assim como os modelos NN v2 e Sigmoid SVM e os modelos RBF SVM e Poly SVM também. Comparando as médias dos modelos pertencentes a cada conjunto, chega-se à conclusão de que os modelo RF, LR, e GB apresentaram os melhores resultados para essa métrica.

Tabela 12 – Resultados dos testes de Wilcoxon para métrica precisão

	CNN	**RF**	**DT**	**LR**	**KNN**	**GB**	**NN v1**	**NN v2**	**NN v3**	**RBF SVM**	**Sigmoid SVM**	**Poly SVM**
CNN	1.0											
RF	**0.0**	1.0										
DT	**0.0**	**0.0**	1.0									
LR	**0.0**	0.92	**0.0**	1.0								
KNN	**0.0**	**0.0**	**0.0**	**0.0**	1.0							
GB	**0.0**	0.22	**0.0**	0.32	**0.0**	1.0						
NN v1	**0.0**	**0.0**	**0.0**	**0.0**	0.1	**0.0**	1.0					
NN v2	1.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0				
NN v3	0.51	**0.0**	**0.0**	**0.0**	**0.01**	**0.0**	**0.0**	0.51	1.0			
RBF SVM	**0.03**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.01**	**0.03**	0.07	1.0		
Sigmoid SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.95	**0.0**	1.0	
Poly SVM	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.03**	**0.0**	**0.04**	0.64	**0.0**	1.0

Fonte: O autor (2022).
Nota: Os dados destacados em negrito representam os valores que não passaram na hipótese nula

A tabela 12 apresenta os resultados do teste estatístico feito com a métrica precisão. Segundo o resultado dos testes realizados, os modelos CNN, NN v2 e NN v3 apresentaram a mesma distribuição entre si, assim como o conjunto de modelos RF, LR, e GB, além deles os modelos KNN e NN v1 também apresentaram a mesma distribuição entre si, assim como os modelos NN v2 e NN v3 e os modelos NN v3, RBF SVM e Sigmoid SVM também. Comparando as médias dos modelos pertencentes a cada conjunto, chega-se à conclusão de que os modelo RF, LR, e GB apresentaram os melhores resultados para essa métrica.

6 CONCLUSÃO

Este trabalho avalia o desempenho dos seguintes modelos de aprendizagem de máquina: CNN, DT, GB, KNN, LR, NN, Poly SVM, Sigmoid RVM, RBF SVM e RF utilizados para superar o desafio de detecção de fraudes de cartão de crédito. A identificação de fraudes possui diversas dificuldades que ainda não foram superadas de maneira satisfatória como o grande desbalanceamento das classes, deriva conceitual e dificuldade na obtenção de base de dados. Sendo assim neste trabalho foi realizada uma comparação entre os modelos que apresentavam os melhores desempenhos de detecção mesmo em um contexto com tantos desafios.

Os modelos GB, LR e RF que podem ser considerados modelos simples, apresentaram bons resultados com o desempenho superior a todos os outros modelos testados. Na maioria das métricas utilizadas, esses modelos apresentaram resultados superiores em relação aos outros. Na métrica F1-Score, os três modelos obtiveram o resultado de 94% que foi, na média, 2% maior que DT e 27% maior que o CNN e NN v2. Já na métrica ROC AUC, o RF se apresentou como melhor modelo, sendo 1% melhor em relação ao GB e LR, empatados com 97%, e 6% maior que o quarto melhor modelo DT. Para a métrica de precisão, os modelos RF e LR empataram com média de 97%, que foi superior em 1% em relação ao resultado obtido com GB, 5% maior em relação ao modelo DT, e 26% superior ao modelo NN v1. Já na revocação, temos um cenário diferente, onde o CNN e NN v2 obtiveram o resultado de 100%, e foram os modelos que melhor desempenharam nessa métrica, com valor 7% maior que o modelo GB e 8% maior em relação ao DT.

A partir dos testes estatísticos foi possível perceber que os modelos GB, LR e RF, que apresentaram os melhores desempenhos, possuem a mesma distribuição para todas as métricas avaliadas. Com isso pode-se concluir que os três modelos empataram no primeiro lugar, já a hipótese de que o modelo CNN possui bom desempenho para resolução do problema, não pode ser comprovada, pois o modelo apresentou bons resultados em apenas uma das métricas.

7 TRABALHOS FUTUROS

Com base no trabalho realizado, os trabalhos futuros podem seguir a utilização de uma base de dados com características não anonimizadas para que se possa realizar uma análise mais detalhada do impacto de cada característica no processo de detecção de fraude, possibilitando também a engenharia de recursos combinando características. Além disso, a utilização de diferentes arquiteturas de modelos de aprendizagem profunda como o CNN, que mesmo não obtendo bons resultados no trabalho é um modelo ainda pouco explorado para esse tipo de solução. Outro ponto que pode ser considerado é que para a aplicação de um modelo de aprendizagem de máquina em um ambiente real é importante que o tempo de processamento seja considerado como critério de avaliação, pois é necessário que o modelo seja rápido o suficiente para realização de uma análise transacional em tempo hábil.

Como limitações do método proposto é possível listar a deriva conceitual, já que os modelos são treinados a partir de um conjunto de dados em um momento específico, porém o comportamento dos usuários vai se modificando ao longo do tempo e isso faz com que os modelos treinados não se adaptem a essa mudança sem um novo treinamento. Além disso, como o conjunto de dados utilizado passou por um processo de análise de componentes principais, não é possível identificar quais são as características mais relevantes para o processo de detecção de fraudes e isso limita a análise dos fatores relevantes para identificação de transações fraudulentas. Um importante ponto de ameaça a ser analisado é que as transações fraudulentas são identificadas a partir de um padrão comportamental, porém esse padrão não está individualizado, sendo assim os modelos apresentados podem causar fricção na experiência de usuários com padrões comportamentais destoantes da maioria dos usuários, mesmo havendo um padrão seguido para as transações desse mesmo cliente.

REFERÊNCIAS

BAHNSEN, A. *et al.* Cost sensitive credit card fraud detection using bayes minimum risk. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS, 12., 2013, Miami, FL. **Proceedings** [...]. Piscataway, NJ: IEEE, 2013. p. 333-338. DOI: https://doi.org/10.1109/ICMLA.2013.68. Disponível em: http://ieeexplore.ieee.org/document/6784638/. Acesso em: 28 maio 2022.

CARCILLO, F. *et al.* Combining unsupervised and supervised learning in credit card fraud detection. **Information Sciences**, v. 557, p. 317-331, 2021. DOI: https://doi.org/10.1016/j.ins.2019.05.042. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0020025519304451. Acesso em: 28 maio 2022.

Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). John Wiley & Sons, Inc. ISBN 0-471-16068-7., p. 350

DAL POZZOLO, A. *et al.* Credit card fraud detection and concept-drift adaptation with delayed supervised information. *In*: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2015, Killarney, Ireland. **Proceedings** [...]. Piscataway, NJ: IEEE, 2015. p. 1-8. DOI: https://doi.org/10.1109/IJCNN.2015.7280527. Disponível em: http://ieeexplore.ieee.org/document/7280527/. Acesso em: 28 maio 2022.

DHANKHAD, S.; MOHAMMED, E.; FAR, B. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. *In*: IEEE INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION, 2018, Salt Lake City, UT. **Proceedings** [...]. Piscataway, NJ: IEEE, 2018. p. 122-125. DOI: https://doi.org/10.1109/IRI.2018.00025. Disponível em: https://ieeexplore.ieee.org/document/8424696/. Acesso em: 28 maio 2022.

FOROUGH, J.; MOMTAZI, S. Ensemble of deep sequential models for credit card fraud detection. **Applied Soft Computing**, v. 99, p. 106883, 2021. DOI: https://doi.org/10.1016. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494620308218. Acesso em: 28 maio 2022.

JIANG, C. *et al.* Credit card fraud detection: a novel approach using aggregation strategy and feedback mechanism. **IEEE Internet of Things Journal**, v. 5, n. 5, p. 3637-3647, 2018. DOI: https://doi.org/10.1109/JIOT.2018.2816007. Disponível em: https://ieeexplore.ieee.org/document/8316850. Acesso em: 28 maio 2022.

JURGOVSKY, J. *et al.* Sequence classification for credit-card fraud detection. **Expert Systems with Applications**, v. 100, p. 234-245, 2018. DOI: https://doi.org/10.1016/j.eswa.2018.01.037. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417418300435. Acesso em: 15 maio 2022.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **International joint Conference on artificial intelligence**. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

LI, Z. *et al.* A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. **Expert Systems with Applications**, v. 175, p. 114750, 2021. DOI: https://doi.org/10.1016/j.eswa.2021.114750. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417421001913. Acesso em: 28 maio 2022.

LUCAS, Y. *et al.* Multiple perspectives HMM-based feature engineering for credit card fraud detection. *In*: ACM/SIGAPP SYMPOSIUM ON APPLIED COMPUTING, 34., 2019, LIMASSOL CYPRUS. **Proceedings** [...]. Limassol Cyprus: ACM, 2019. p. 1359-1361. DOI: https://doi.org/10.1145/3297280.3297586. Disponível em: https://dl.acm.org/doi/10.1145/3297280.3297586. Acesso em: 28 maio 2022.

MACHINE LEARNING GROUP - ULB. Credit Card Fraud Detection. In: **Kaggle**. 3. [S. l.], 2017. Disponível em: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Acesso em: 22 set. 2022.

PEDREGOSA, F. *et al*. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning** Research, v. 12, p. 2825, 2011. Disponível em: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 22 set. 2022

ROBERTSON, David. Card fraud losses worldwide. In: **Nilson Report**. [S. l.], 7 dez. 2021. Disponível em: https://nilsonreport.com/upload/content_promo/NilsonReport_Issue1209.pdf. Acesso em: 22 set. 2022.

RYMAN-TUBB, N. F.; KRAUSE, P.; GARN, W. How artificial intelligence and machine learning research impacts payment card fraud detection: a survey and industry benchmark. **Engineering Applications of Artificial Intelligence**, v. 76, p. 130-157, 2018. DOI: https://doi.org/10.1016/j.engappai.2018.07.008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0952197618301520. Acesso em: 28 maio 2022.

SÁ, A. G. C.; PEREIRA, A. C. M.; PAPPA, G. L. A customized classification algorithm for credit card fraud detection. **Engineering Applications of Artificial Intelligence**, v. 72, p. 21-29, 2018, ISSN 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2018.03.011. Disponível em: https://www.sciencedirect.com/science/article/pii/S0952197618300605. Acesso em: 28 maio 2022.

SAKHAROVA, Irina. Payment card fraud: Challenges and solutions. *In*: **2012 IEEE International Conference on Intelligence and Security Informatics**. Washington, DC, USA: IEEE, 2012, p. 227–234. Disponível em: http://ieeexplore.ieee.org/document/6284315/. Acesso em: 26 set. 2022.

XUAN, S. *et al.* Random forest for credit card fraud detection. *In*: INTERNATIONAL CONFERENCE ON NETWORKING, SENSING AND CONTROL, 15., 2018, Zhuhai. **Proceedings** [...]. Piscataway, NJ: IEEE, 2018. p. 1-6. DOI: https://doi.org/10.1109/IRI.2018.00025. Disponível em: https://ieeexplore.ieee.org/document/8361343/. Acesso em: 28 maio 2022.

ZHANG, Z. *et al.* A model based on convolutional neural network for online transaction fraud detection. **Security and Communication Networks**, v. 2018, p. 1-9, 2018. Disponível em: https://dl.acm.org/doi/abs/10.1155/2018/5680264. Acesso em: 29 maio 2022.