UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Avaliação de modelos de IA no contexto de Credit Scoring

Bruno Soares de Melo Barreto

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Avaliação de modelos de IA no contexto de *Credit Scoring*

Bruno Soares de Melo Barreto

USP - São Carlos

2024

Bruno Soares de Melo Barreto

Avaliação de modelos de IA no contexto de *credit scoring*

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof. Dr. João Paulo Papa

USP - São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Barreto, Bruno
B273a Avaliação de modelos de IA no contexto de credit scoring / Bruno Barreto; orientador João Paulo Papa. -- São Carlos, 2024.
52 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. credit scoring. 2. machine learning. 3. classificação. 4. inteligência artificial. I. Papa, João Paulo , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

RESUMO

BARRETO, N. P. **Avaliação de modelos de IA no contexto de *credit scoring***. 2024. 52 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O mercado de crédito de crédito no Brasil representa um importante meio de fornecer poder de compra para clientes e, para isso, os modelos de *credit scoring* são importantes para entender a probabilidade de uma pessoa pagar ou não a futura dívida. O propósito do presente trabalho é avaliar algoritmos de inteligência artificial de classificação em quatro bases diferentes com o intuito de entender o desempenho deles em cada um dos cenários. Os modelos escolhidos foram: regressão logística, *random forest*, *LightGBM*, *XGBoost* e *Convolutional Neural Networks* (CNN), outro objetivo do estudo é entender se a CNN é razoável para o tema. Para isso, foi realizado uma coleta das bases de dados, após isso um pré-processamento dos dados, com utilização de normalização e modelos para balanceamento dos dados, em seguida acontece o desenvolvimento dos modelos e, por fim, a avaliação dos resultados através de três métricas, que são: AUC (*Area under the ROC Curve*), *F1-Score e* acurácia, utilizando a multiplicação das colocações para ordenar o desempenho dos algoritmos. Com isso, ao reunir as posições em cada base, foi possível fazer uma análise geral. O estudo mostra que não houve domínio de um algoritmo em todas as análises, porém o *Random Forest* demonstrou melhor desempenho em três das quatro bases, e na restante foi a regressão logística. Apesar disso, houve uma aplicação de teste de *Wilcoxon signed-rank* e foi possível concluir que não há diferença estatisticamente conclusiva entre o desempenho dos modelos. Além disso, foi possível observar que o uso da CNN é razoável para o contexto de *credit scoring*, performando de forma similar aos modelos tradicionais, sendo um estudo em potencial se aprofundar sobre a composição de modo a otimizar os resultados.

Palavras-chave: *credit scoring*; *machine leaning*; classificação; inteligência artificial.

ABSTRACT

BARRETO, N. P. **Evaluation of AI models in the context of credit scoring**. 2024. 52 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The credit market in Brazil represents an important way of providing purchasing power to customers and, for this, credit scoring models are important to understand the probability of a person paying or not the future debt. The purpose of this work is to evaluate artificial intelligence classification algorithms in four different bases to understand their performance in each of the scenarios. The models were: logistic regression, random forest, LightGBM, XGBoost, and Convolutional Neural Networks (CNN). Another study objective is to understand whether CNN is reasonable for the topic. For this, a collection of databases was carried out, after which a pre-processing of the data was carried out, using normalization and models for data balancing, then the development of the models and, finally, the evaluation of the results through three metrics, which are: AUC (Area under the ROC Curve), F1-Score and accuracy, using the multiplication of the positions to order the performance of the algorithms. With this, by gathering the positions in each base, it was possible to make a general analysis. The study shows that no algorithm dominated all analyses, but Random Forest performed better in three of the four databases, while logistic regression performed better in the remaining databases. Despite this, the Wilcoxon signed-rank test was applied and it was possible to conclude that there is no statistically conclusive difference between the performance of the models. In addition, it was possible to observe that using CNN is reasonable for the context of credit scoring, performing similarly to traditional models. A potential study could be done to delve deeper into the composition to optimize the results.

Keywords: *credit scoring*; *machine leaning*; classification; artificial intelligence.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo ilustrativo de árvore de decisão..... 10
Figura 2 – Fluxograma do Desenvolvimento do Trabalho..... 17

LISTA DE TABELAS

Tabela 1 – Resumo das bases de dados do estudo.....20
Tabela 2 – Resumo das bases de dados do estudo após os tratamentos.....20
Tabela 3 – Resultados das métricas de avaliação para a base de Taiwan.....23
Tabela 4 – Resultados com o ordenamento para a base de Taiwan.....24
Tabela 5 – Resultados das métricas de avaliação para a base de Taiwan.....25
Tabela 6 – Resultados com o ordenamento para a base da Alemanha.....25
Tabela 7 – Resultados das métricas de avaliação para a base da Austrália.....26
Tabela 8 – Resultados com o ordenamento para a base da Austrália.....26
Tabela 9 – Resultados das métricas de avaliação para a base da Aprovação de Crédito...27
Tabela 10 – Resultados com o ordenamento para a base de Aprovação de Crédito.....28
Tabela 11 – Resultados com o ordenamento geral.....29
Tabela 12 – Resultados dos p valores entre os modelos.....29
Tabela 13 – Tempos do treinamento dos modelos em segundos.....30
Tabela 14 – Consumo de memória treinamento dos modelos em mebibyte.....31
Tabela 15 – Resultados dos p valores com base no tempo de treinamento.....32
Tabela 16 – Resultados dos p valores com base no consumo de memória.....32

SUMÁRIO

1 **INTRODUÇÃO**..... 1
1.1 **Objetivo Geral**..... 3
1.2 **Objetivos Específicos**..... 3
1.3 **Hipótese**..... 4
2 **TRABALHOS RELACIONADOS**..... 5
3 **REFERENCIAL TEÓRICO**..... 8
3.1 **Técnicas de Inteligência Artificial e Métodos Estatísticos**..... 8
3.1.1 **Regressão Logística**..... 8
3.1.2 ***Random Forest***..... 9
3.1.3 ***LightGBM***..... 11
3.1.4 ***XGBoost***..... 11
3.1.5 ***Convolutional Neural Networks***..... 12
3.2 ***Credit Scoring***..... 14
3.3 **Teste de *Wilcoxon Signed-rank* (WSRT)**..... 15
4 **METODOLOGIA E DESENVOLVIMENTO**..... 16
4.1 **Fluxograma do Desenvolvimento**..... 16
4.2 **Base de dados**..... 17
4.2.1 **Base de pagamentos de consumidores de Taiwan**..... 18
4.2.2 **Base de pagamentos de consumidores da Alemanha**..... 18
4.2.3 **Base de aplicação de cartão de crédito da Austrália**..... 19
4.2.4 **Base de aprovação de cartão de crédito**..... 19
4.3 **Métricas de Avaliação**..... 20
4.4 **Tecnologias Utilizadas e Parâmetros do estudo**..... 21
5 **RESULTADOS**..... 23
5.1 **Taiwan**..... 23
5.2 **Alemanha**..... 24
5.3 **Austrália**..... 25
5.4 **Aprovação de Crédito**..... 27
5.5 **Análise Geral**..... 28
6 **CONCLUSÕES**..... 33
REFERÊNCIAS..... 35

1 INTRODUÇÃO

O mercado de crédito desempenha importante papel em proporcionar maior poder de compra para o cliente, seja para uma pessoa ter possibilidade de comprar um imóvel, carro, ou qualquer produto que seja possível pagar em outro momento, como é o caso dos cartões de crédito. O mercado está em amplo crescimento e com perspectivas bem otimistas para 2024, de acordo com Federação Brasileira de Bancos (FEBRABAN) por meio da Pesquisa de Economia Bancária e Expectativas, a projeção de crescimento da carteira de crédito total para 2024 está em 8,5%, muito por conta da queda da taxa de juros e melhora da inadimplência.

Além da perspectiva de facilidade para o cliente, também há um ganho por parte das instituições que fornecem o crédito. No caso de bancos e instituições financeiras, os ganhos diretos ocorrem, principalmente, por conta dos juros, que oscilam de acordo com as taxas negociadas com os que serão beneficiados com o crédito. Porém, não só bancos e instituições financeiras que usam o produto de crédito como uma forma de rentabilizar o seu negócio; existem outras indústrias que utilizam crédito para potencializar as vendas, como é o caso do setor de beleza, que ao trabalhar com revenda, pode fornecer um valor de crédito em produtos para o revendedor com um certo prazo para pagamento, o ganho nesse esquema seria em cima dos valores dos produtos que foram fornecidos em crédito e também em eventuais multas e juros em caso de atraso de pagamento.

Para fornecer esses valores aos clientes, o mercado de crédito utiliza métodos estatísticos para calcular o *Credit Scoring* com o objetivo de entender o perfil de pagamento da pessoa, ou seja, se há possibilidade de não pagamento desse valor no futuro. Para isso, são utilizadas diversas naturezas de dados para poder agregar ao perfil, como dados demográficos, de pagamento, de renda e até de patrimônio dependendo do que a empresa tem disponível para análise. Para além desse tipo de crédito, que seria de concessão de crédito ou de originação, também há a forma de manutenção de crédito, que seria com o cliente já tendo crédito e verificando a possibilidade de continuar ou aumentar o valor disponível para uso. Para esse segundo caso também são utilizados diversos dados para verificar as possibilidades, como de um cliente que já está em atraso conseguir efetuar o pagamento, que seria o exemplo do *collection score*.

Para criar um modelo de *credit scoring*, as principais etapas para a construção de um modelo desse tipo são (SICSÚ, 2010): (i) Planejamento e Definições, (ii) Identificação das

variáveis potenciais, (iii) Planejamento e seleção da amostra, (iv) Análise e tratamento dos dados, (v) Cálculo da fórmula de escoragem, (vi) Análise e validação da fórmula e (vii) Ajuste final do modelo.

Os modelos de *credit scoring* apresentam muitos desafios no seu desenvolvimento, isso porque há problemas com relação à acurácia desses modelos, o algoritmo preciso identificar muito bem quem vai conseguir pagar corretamente a dívida com quem não vai pagar. O erro nesses modelos pode significar uma quantia relevante para o negócio, fornecendo crédito para pessoas que não vão pagar e a empresa não consegue o retorno esperado. Da mesma forma, não fornecer crédito para clientes que poderiam pagar quer dizer que estão deixando de ter melhores resultados por conta de ineficiência do modelo.

Com todo esse desafio pontuado anteriormente, o uso de inteligência artificial tem sido essencial para que as empresas que estão concedendo crédito possam realizar o estudo de maneira mais eficiente possível, de forma mais automatizada e de melhor qualidade. A aplicação da inteligência artificial se dá exatamente no tratamento dos dados citados anteriormente e calculando as probabilidades do contexto de negócio, como o cálculo do *behavior scoring*, *fraud scoring*, dentre outros modelos de *Credit Scoring* presentes nas empresas. Para esses problemas, entre os algoritmos existentes, os mais comuns para esse contexto são (DASTILE; CELIK; POTSANE, 2020): Regressão Logística, *Naive Bayes*, *K-Nearest Neighbor* (KNN), Árvore de Decisão, *Support Vector Machines* (SVM), *Random Forest* e *Extreme Gradient Boost*.

Outros autores publicaram sobre o mesmo tema, como Abid, Masmoudi e Zouari-Ghorbel (2016), que aplicaram um modelo de regressão logística numa base de dados de um banco comercial tunisiano, tendo como resultado um erro de 0,586% na classificação de inadimplentes. Outro exemplo de trabalho relacionado é o de Peng *et al.* (2023), que aplica e compara quatro modelos de machine learning, que são: XGBoost, LightGBM, regressão logística e arvore de decisão, para uma base de dados que discrimina entre inadimplente ou não. Nesse estudo, os algoritmos XGBoost e LightGBM demonstraram melhor desempenho em relação aos outros dois. Além deles, Sadok, Sakka e Maknouzi (2022) fizeram um estudo sobre o uso da inteligência artificial na análise de crédito, citando como o uso desta técnica melhora a inclusão e acesso ao crédito para pessoas que tradicionalmente não receberiam e, também, alerta sobre o uso ético nesse contexto, dado que as variáveis de entrada para modelos desta temática são sensíveis.

Além do tema de inadimplência em crédito, outros autores estudaram sobre a aplicação na temática do financeiro. Nazareth e Reddy (2023) publicaram sobre o uso de *Machine Learning* no contexto de financeiro, avaliando 126 artigos e 44 jornais estudando dados, técnicas e contextos da aplicação de inteligência artificial. Nesse trabalho foi possível verificar outras áreas de aplicação, além do crédito, como: previsão dos preços das ações, gestão de carteira, criptomoeda, mercado cambial, previsão de crises financeiras, predição de falências de empresas e, por fim, cálculo de risco de insolvência. Outro tema estudado é sobre fraudes, onde o uso de inteligência artificial é importante para as empresas, especialmente do ramo financeiro, Rajendra *et al.* (2022) usa *Machine Learning* para analisar as detecções de fraude de cartão de crédito, utilizando métodos como: árvore de descisão, *K-Nearest Neighbor* (KNN), rede neural e regressão logística para classificar se uma transação era fraude ou não. A última tendo uma acurácia melhor, mas o KNN apresentou um melhor aprendizado com maior volume de dados.

1.1 Objetivo Geral

O objetivo geral desse estudo é avaliar modelos de inteligência artificial que determine a probabilidade de cada pessoa em ser um mau pagadora ou não, e aplicando em contextos diferentes dentro do tema de crédito. Praticando em quatro bases diferentes, com formas diferentes, para verificar quais algoritmos de classificação desempenham melhor em cada uma dessas bases de dados. Entre os algoritmos existentes, os escolhidos para esse estudo são: Regressão Logística, *Random Forest*, *LightGBM*, *XGBoost* e *Convolutional Neural Networks* (CNNs).

1.2 Objetivos Específicos

Os objetivos específicos que compõem esses estudos são os seguintes:

- Verificar se o uso de Redes Neurais Convolucionais, do inglês *Convolutional Neural Networks* (CNNs), é razoável para esse contexto de crédito, usando tratamentos iguais aos que são feitos com imagem.
- Comparar os desempenhos dos algoritmos escolhidos em cada uma das quatro bases de dados usadas nesse estudo.

1.3 Hipótese

A hipótese desse estudo é que CNNs conseguem resultados satisfatórios para calcular as probabilidades de atraso de pagamento nas diferentes bases escolhidas para esse estudo.

2 TRABALHOS RELACIONADOS

Os exemplos de aplicações de *machine learning* na literatura são numerosos e diversos, tendo uma variação grande de técnicas que resolvem problemas de uma variedade grande de área diferentes dentro do setor financeiro. É possível verificar aplicações desde predição de inadimplência até modelos para cobrança.

Dentre os trabalhos relacionados, é possível verificar os casos de revisões sistemáticas como Fernandes *et al.* (2016), que compara algoritmos de classificação, desde métodos estatísticos tradicionais até abordagens de *machine learning* mais avançadas, analisando sua eficácia, complexidade e aplicabilidade no contexto de *credit scoring*. O artigo conclui que não há um algoritmo em específico que seja o melhor, mas os métodos de *ensemble* frequentemente demonstram um resultado superior. Outro estudo que aborda análise de modelos para *credit* scoring foi Srinath e Gurujara (2022) que busca explicar como funciona os principais modelos para identificação de inadimplentes de cartão de crédito. O método para explicar os modelos foi o *model-agnostic*, através da ferramenta *DALEX* que é um modelo de *explainable artificial intelligence* (XAI), com o objetivo de aumentar a transparência e compreensão dos modelos. Os algoritmos analisados no estudo foram: *Support vector machine* (SVM), *Random Forest*, *XGBoost* e redes neurais. O estudo concluiu que o modelo com *XGBoost* desempenhou melhor do que os outros modelos, tendo pouca diferença para o SVM e *Random Forest*, mas o melhor modelo foi destacado também por usar uma lógica de mais fácil compreensão para humanos. Por fim, outro artigo que usou *explainable artificial intelligence* para melhorar modelos foi de Elhosseini *et al.* (2023), que apresenta novo método para prever inadimplência de cartão de crédito utilizando uma combinação de técnicas entre *deep learning* e XAI, apresentando um resultado de 83,5% de acurácia e com as variáveis atrasos de pagamento e valores das contas pendentes.

Outros artigos buscaram aplicar modelos de *ensemble* para resolver o problema de *credit scoring*, como Zhang *et al.* (2018), que utiliza cinco dos mais conhecidos algoritmos (que são regressão logística, SVM, redes neurais, *gradiente boosting decision tree* e *random forest*) para criar um método de classificar inadimplentes, além disso, utiliza também uma clusterização utilizando a abordagem *fuzzy*, utilizando três diferentes bases. O modelo comparado com os

tradicionais teve um desempenhou superior em acurácia em todas as bases testadas. Yoon *et al.* (2016) também utiliza o método *fuzzy* para aplicação em modelos de inteligência artificial, o estudo desenvolve um modelo de *fuzzy logistic regression* para prever inadimplentes de empréstimos para banco. Nele é possível verificar um ganho em relação à regressão logística clássica. Outro exemplo de aplicação de *ensemble* é o estudo de Abbod e Ala'raj (2016), que desenvolveram um modelo híbrido de *ensemble* para modelo de *credit scoring*, aplicando os métodos *Gabriel Neighbourhood Graph* (GNG) e *Multivariate Adaptive Regression Splines* (MARS) na fase hibrida do modelo e na fase do *ensemble* foram estudados 5 algoritmos: redes neurais, SVM, *random forest*, árvore de decisão e *Naive Bayes*. A combinação dos modelos foi realizada com *consensus approach* (ConsA) as técnicas foram aplicadas em 7 bases de dados diferentes. Wang *et al.* (2016) apresenta outra perspectiva para solução do problema de *credit scoring*, é utilizado o método de *supervised clustering*, aplicando em partições diferentes da amostra de dados, combinando os resultados de múltiplos modelos de classificação. O objetivo é ao segmentar o conjunto de dados com base em características similares e aplicar modelos específicos para cada segmento, a precisão possa ser melhorada. Como resultado foi possível verificar que o agrupamento supervisionado pode levar a uma melhoria significativa na precisão da pontuação de crédito comparada a métodos tradicionais. Por fim, Castellano e Abellán (2016) publicaram um estudo comparativo dos classificadores base para os métodos de *ensemble* para o tema de *credit scoring*, os algoritmos utilizados no estudo foram os seguintes: *bagging*, *boosting*, *random space, DECORATE* e *rotation forest*. Todos esses métodos utilizaram como base os algoritmos: regressão logística, MLP, SVM, C4.5 *decision tree* e *credal decision tree* (CDT) Utilizando seis base de dados para comparar desempenho foi possível concluir que o CDT se desempenha melhor do que as outras bases na maior parte do experimento.

Uma abordagem utilizada que se repete com bastante frequência na literatura é *XGBoost*, Liu *et al.* (2017) também utilizou esta técnica para este contexto, o artigo propõe desenvolver um modelo de *ensemble* sequenciado baseado numa variação do *XGBoost*, utilizando a abordagem de hiper parâmetros Bayesiano na etapa final de construção. O estudo concluiu que essa abordagem melhorou o desempenho comparado com abordagens mais conhecidas, como: *random Search*, *grid Search* e *manual Search.*

Ao analisar os trabalhos relacionados a predição de inadimplentes de cartão de crédito ou empréstimos bancários é possível observar que em grande parte das soluções, o balanceamento dos dados utilizados no modelo é um grande ponto de alerta a ser verificado em

qualquer aplicação nessa área Neste sentido, alguns autores buscam soluções para este problema de desenvolvimento, (DENG *et al.*, 2021) as abordagens tradicionais são facilmente influenciados pelo viés da seleção da amostra pois usam somente os cadastros aceitos, mas a população também é composta pelos cadastros rejeitados. Para resolver isso, *Reject Inference* é a técnica utilizada para inferir rótulos de bom ou mau para candidatos rejeitados. Yang *et al.* (2017) estuda propor um novo método envolvendo *Machine Learning* para resolver os problemas de inferência de rejeitados, é desenvolvido um modelo semi-supervisionado de SVM para melhorar o desempenho dos modelos de *score* comparado com modelos tradicionais, como regressão logística. Khushi *et al.* (2020) também cita que o problema de balanceamento não é resolvido pelos métodos estatísticos tradicionais, o artigo desenvolve um método para tratar desses casos. Primeiro ele normaliza os dados utilizando o método normalização Min-Max, depois utiliza métodos de *resampling* dos dados, utilizando caminhos diferentes se for uma sobreamostragem e subamostragem, como *Random Undersampling* e *K-Means SMOTE*, respectivamente, desta forma tratando os casos de dados desbalanceados. Por fim, Zhang e Cheng (2021) utilizam o método *K-Means SMOTE* para mudar a distribuição dos dados, tratando o problema de balanceamento, além disso utiliza *random forest* para calcular a importância das variáveis e substitui os valores como pesos no algoritmo de redes neurais. Comparando com cinco outros modelos de predição, o proposto desempenhou os melhores números.

Além do contexto de *credit scoring*, que foi citado pelos trabalhos mencionados anteriormente, o uso de inteligência artificial se amplia para outras áreas envolvendo o ciclo de crédito, como a área de cobrança, Kang e Kim (2016) desenvolveram cinco modelos de *Machine Learning* para prever pagamento atrasado e modelos de *scoring* de clientes para calcular a probabilidade de pagamento de quem já atrasou o pagamento. Foram utilizados modelos de *random forest,* redes neurais, árvore de decisão, SVM e modelo híbrido. Outra aplicação em cobrança foi de Moore *et al.* (2015) utilizam cadeia de *Markov* e *Hazard rate* para modelar os padrões de pagamento. Outra área que possui variedade de aplicações é a de fraudes, Elayan *et al.* (2021) utiliza rede neural profunda para detectar fraude de cartão de crédito aplicando numa empresa de transações, o resultado da classificação de fraude foi de 99,1% de *score* na área abaixo da curva ROC. Outro exemplo de aplicação nesta área é a de Khin e Khine (2020), que escolhe o método de *Online Boosting* (*OLBoost*) que usa *Extremely Fast Decision Tree* (EFDT) para detectar as fraudes de cartão de crédito.

3 REFERENCIAL TEÓRICO

A identificação precoce de inadimplentes é crucial para a gestão de riscos em instituições financeiras. Com o avanço da inteligência artificial e *big data*, modelos de *machine learning* têm emergido como ferramenta poderosa, permitindo uma análise mais precisa e profunda dos padrões de comportamento de crédito. Neste tópico, será discutido sobre os principais conceitos relacionados ao trabalho, começando com as técnicas de inteligência artificial e métodos estatísticos que serão usados, abordando pontos chaves e exemplos de uso. Além disso, é abordado o conceito de *credit scoring*, tema principal do presente trabalho e, por fim, é apresentado o conceito do teste estatístico que é utilizado no estudo.

3.1 Técnicas de Inteligência Artificial e Métodos Estatísticos

Segundo Kai-Fu Lee (2018), inteligência artificial (IA) é o campo de estudo que busca criar computadores capazes de realizar tarefas que exigem inteligência humana. Essa definição enfatiza a capacidade dos modelos de IA de executar processos que normalmente exigiriam esforço humano, como análises e tomadas de decisão. Para o estudo, serão aplicadas técnicas estatísticas, modelos de *machine learning* e *deep learning* que serão descritas a seguir.

3.1.1 Regressão Logística

Na literatura é possível observar que a regressão logística é umas técnicas mais tradicionais tanto no contexto de análise de crédito (TOKPAVI, Sessi *et al.*,2016) como problemas de classificação no geral (WEN Jinming *et al.*,2023). Segundo BRUCE, Andrew e BRUCE, Peter (2019), A regressão logística é um método conhecido por conta da sua alta velocidade de processamento, em que utiliza função de resposta logística e logito, em que é mapeado a probabilidade e, com isso, classificar entre "sucesso" (1) e "não sucesso" (0) (BRUCE; BRUCE, 2019). A função para calcular a probabilidade é dada pela função logística, que transforma uma saída linear em um resulta entre 0 e 1, a fórmula da função é (TIBSHIRANI *et al.*, 2013):

$$p(X) = \frac{e^{(\beta o + \beta 1 X)}}{1 + e^{\beta o + \beta 1 X}}$$

Onde:

- p é a probabilidade.
- β são os coeficientes de regressão.
- X são os valores das variáveis independentes.

Outro conceito relacionado ao tema de regressão logística é a de razão de chance, que se dá pela seguinte forma:

$$\frac{p(X)}{1-p(X)}$$

Adaptando com a função logística, que foi mencionada anteriormente, é encontrada a fórmula da função logit:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta o + \beta 1X$$

Com a fórmula já demonstrada, é preciso saber como estimar os coeficientes da função, uma técnica conhecida é a máxima verossimilhança. Nesta técnica, se estima os valores de βo e $\beta 1$ na função logística de modo que todos os casos de "sucesso" se aproximem a 1 e todos os casos de "não sucesso" se aproximem a 0. Esse pensamento se traduz nessa fórmula a seguir:

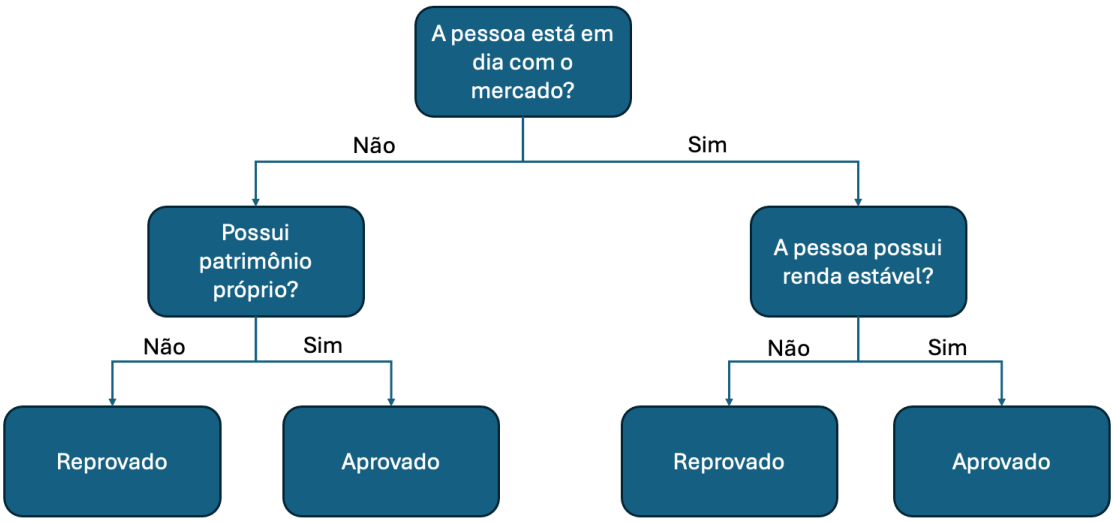
$$\ell(\beta o, \beta 1) = \prod_{i:yi=1} p(xi) \prod_{i':yi'=1} (1-p(xi'))$$

Os estimadores são escolhidos de forma que maximizem o valor da função de verossimilhança.

3.1.2 *Random Forest*

Antes de definir a técnica de *Random Forest* é preciso contextualizar sobre o que é árvore de decisão, segundo VANDERPLAS (2017), é uma técnica intuitiva de classificar ou rotular objetos, através de uma série de perguntas simples, separando de forma binária os caminhos. Um exemplo de árvore de decisão é representado na Figura 1.

Figura 1 – Exemplo ilustrativo de árvore de decisão



Fonte: Elaboração própria.

Com isso, podemos definir *random forest*, que é um modelo de *ensemble* (seria combinação uma serie de modelos, que pode ser de regressão ou classificação, com o intuito de melhorar a composição dos modelos) que combina uma série de árvore de decisão. As árvores são geradas usando uma seleção aleatória dos atributos e cada uma delas dependem da distribuição das amostras. Durante a classificação, cada árvore vota e classe mais escolhida é a retornada do modelo (PEI *et al.*, 2012).

De acordo com Friedman *et al.* (2009), o algoritmo do *random forest* se dá por essas etapas:

1. Para b=1 até B:
    a. Definir uma amostragem Z de tamanho N dos dados de treino.
    b. Aplicar um arvore de *random-forest* Tb para a amostragem dos dados, repetindo as seguintes etapas para cada nó das árvores até nó mínimo ser alcançado.
        i. Selecionar m variáveis aleatoriamente das p variáveis.
        ii. Escolher a melhor variável dentre as m.
        iii. Separa o nó em dois nós filhos.
2. O resultado é a junção das árvores
3. Para o caso de classificação, $\hat{C}b(x)$ é a classe que foi escolhida pela b-ésima árvore do *random forest*. Então,

$$\widehat{C}B(x) = maioria\ dos\ votos\ de\ \widehat{C}\ da\ primeira\ árvores\ até\ a\ b-ésima.$$

3.1.3 *LightGBM*

A técnica em questão é um algoritmo de *Gradient Boosting*, que se baseia em no conceito de *boosting*, conceito utilizado em alguns métodos de classificação e regressão, como *LightGBM*, *XGBoost* e *AdaBoost*. Nesse método, cada novo modelo é treinado para se concentrar nos erros cometidos pelos modelos anteriores. Com isso, o *Gradiente Boosting Machine* (GBM) é uma técnica de aprendizado de máquina que visa otimizar a função de perda iterativamente, combinando múltiplos modelos de aprendizado fracos para formar um modelo mais forte, usando gradientes da função de perda em relação às previsões do modelo para guiar o ajuste dos parâmetros (FRIEDMAN; 2001).

O *LightGBM* é um modelo de *boosting* baseado em árvores de decisão, ou seja, onde uma árvore ajuda a corrigir o erro das anteriores. O diferencial está em utilizar duas técnicas: *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature Bundling* (EFB). Onde a primeira é uma estratégia inteligente de amostragem de dados, mantendo todas as instancias com gradientes grandes, que são as mais críticas para o aprendizado, e realiza uma amostragem aleatória nas instâncias de gradientes pequenos, reduzindo a quantidade de dados sem comprometer a precisão. A outra técnica aborda a alta dimensionalidade de dados, agrupando recursos que são mutuamente exclusivos, ou seja, diminuindo a complexidade computacional (LIU *et al.*, 2017).

Por fim, um outro diferencial do *LightGBM* frente aos outros algoritmos de *boosting* é que a árvore verticalmente, mostrando que o algoritmo escolhe a folha que minimiza a perda para crescer. Com isso, permitindo um aprendizado mais rápido e eficiente.

3.1.4 *XGBoost*

Como o *LightGBM*, o *XGBoost* também é um algoritmo de *boosting* que utiliza árvores de decisão para ajustar os erros. Além disso, também utiliza a otimização pelo gradiente, usando o gradiente da função perda.

O diferencial do *XGBoost* frente aos outros algoritmos de mesmo tipo é que introduz um termo de regularização na função objetivo, isso tem o propósito de controlar a complexidade do modelo e evitar *overfitting.* Além disso, o *XGBoost* avalia o ganho de fazer cada possível

divisão, levando em conta a melhoria na precisão e o termo de regularização (CHEN; GUESTRIN, 2016).

A função objetivo do *XGBoost* é composta por uma função de perda L e um termo de regulação Ω, sendo formulada para uma etapa de iteração t como:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t)$$

Onde:

- y_i são os valores reais,
- $\hat{y}_i^{t-1}$ é a previsão do modelo até a etapa t-1,
- $f_t(x_i)$ é a previsão da nova árvore adicionada no passo t,
- l é a função de perda diferenciável,
- Ω é o termo de regularização, que penaliza a complexidade da árvore.

O termo de regularização Ω para uma árvore f_t é dado por:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

Onde:

- T é o número de folhas da árvore,
- ω_j são os pesos associados a cada linha,
- γ e λ são parâmetros que controlam a contribuição da complexidade da árvore para a função objetivo.

3.1.5 *Convolutional Neural Networks*

As *Convolutional Neural Networks* (CNN) são um tipo de rede neural profunda projetada para processar dados com uma estrutura de grade, como imagens. Elas se destacam pela capacidade de capturar padrões espaciais nas entradas através do uso de operações de convolução. As CNNs são especialmente eficazes em tarefas de visão computacional, como reconhecimento de objetos, devido à sua capacidade de compartilhar pesos e aprender representações hierárquicas de características visuais (GOODFELLOW; BENGIO; COURVILLE, 2016).

Sobre o funcionamento das *Convolutional Neural Networks*, as etapas são as seguintes: primeiro são as camadas de convolução, onde um conjunto de filtros (ou *kernels*) é aplicado à

entrada. Após isso, uma função de ativação não linear (como *Rectified Linear Unit* ou ReLU) é aplicada elemento a elemento ao mapa de características, introduzindo não linearidade na rede, permitindo que ela aprenda representações mais complexas dos dados. Em seguida, camadas de *pooling* são frequentemente aplicadas para reduzir a dimensionalidade dos mapas de características e tornar a representação mais invariante a pequenas translações na entrada. Ao empilhar várias camadas de convolução e *pooling*, a CNN constrói uma representação hierárquica, as primeiras camadas aprendem características simples, enquanto as camadas mais profundas combinam essas características para detectar padrões mais complexos e abstratos. Com isso, os mapas de características resultantes são achatados em um vetor e passados por uma ou mais camadas conectadas. Essas camadas agem como um classificador tradicional, mapeando características para as classes. Por fim, durante o treinamento, os pesos da CNN são ajustados através de técnicas de otimização, como gradiente descendente estocástico, para minimizar uma função de perda que mede a diferença entre as previsões da rede e os rótulos verdadeiros (GOODFELLOW; BENGIO; COURVILLE, 2016).

Aprofundando sobre a primeira etapa, a convolução é uma operação matemática que combina dois conjuntos de dados para produzir um terceiro conjunto de dados. Esses conjuntos de dados são tipicamente uma imagem de entrada e um filtro.

A convolução possui uma imagem de entrada, que é representada como uma matriz dimensional, onde cada elemento da matriz corresponde ao valor de intensidade do pixel em uma determinada posição da imagem. O filtro é outra matriz bidimensional menor, que representa um padrão específico que a rede está tentando detectar na imagem de entrada. Na operação, o filtro é deslizado pela imagem de entrada em passos definidos, calculando o produto escalar entre os valores do filtro e os valores dos pixels sobrepostos na imagem. A operação de convolução segue a seguinte fórmula (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$C(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

Onde:

- I e K são matrizes bidimensional representados de entrada e filtro, respectivamente.
- i e j são coordenadas do pixel no mapa de características C

A técnica de CNN, em primeiro momento, se aplica mais em casos de visão computacional, como identificação de imagens, porém é possível encontrar na literatura trabalhos relacionamento CNN com *credit scoring*. Song *et al.* (2023) utiliza essa técnica para

propor uma abordagem utilizando um modelo de *soft reordering one-dimensional CNN* (SR-1D-CNN), que se adapta aos dados tabulares e transformam em compatíveis com o aprendizado de CNN. Esse algoritmo foi aplicado em diferentes bases de dados de crédito obtendo um desempenho melhor do que em abordagens tradicionais. Além disso, Hosaka (2018) também aplicou CNN no tema financeiro, aplicando algoritmo baseado em *GoogLeNet* para predição de falência corporativa.

3.2 *Credit Scoring*

O *credit scoring* é uma medida de risco de crédito, ou seja, a probabilidade do crédito se tornar uma perda para quem forneceu o crédito. O propósito do modelo de *credit scoring* é prever, na data do fornecimento ou não do crédito, a probabilidade de que o crédito se transforme em prejuízo. Essa estimativa é a função das características do solicitante de crédito (SICSÚ, 2010).

O modelo, através de técnicas estatísticas, converte informações sobre um tomador de crédito em números que são combinados para formar uma pontuação. O modelo não avalia apenas se um atributo é positivo ou negativo, mas também a sua quantificação. É considerado a principal ferramenta de mensuração de risco no varejo, pois permite que bancos evitem os clientes de maior risco e filtram apenas os classificados como bons pagadores (JOHNSON, 2022).

Além disso, os modelos de *credit scoring* também são importantes por questão de custo e consistência. Maior parte dos bancos tem um número grande de clientes que movimentam, ao todo, bilhões de transações por ano, usando os modelos os bancos podendo automatizar o máximo possível os processos para menores créditos e cartão de crédito. Antes da aplicação do *credit scoring*, um analista de crédito precisava revisar a aplicação de crédito, usando a combinação de experiencia, conhecimento da indústria e sabedoria pessoal para aplicar uma decisão com base em uma quantidade grande de informações, sendo um processo com muita variação (CROUHY; GALAI; MARK, 2006).

Dentre os modelos possíveis de *scoring*, temos o *application scoring*, *behavioral scoring*, e *collection score*: no primeiro o modelo é aplicado em solicitantes de crédito com os quais o credor não teve experiencia anterior, no segundo o modelo é aplicado para clientes do credor, a principal diferença entre o primeiro e o segundo está na variáveis utilizadas, o

behavioral utiliza informações do *application*, utiliza informações relativas a créditos anteriores, com isso, tendendo a fornecer maior poder de discriminação. Por fim, o *collection score*, é um modelo onde são classificados os clientes inadimplentes em classes de acordo com o seu *score* de cobrança, com isso, aplicando diferentes estratégias para diferentes classes, melhorando relacionamento com os clientes e reduzindo despesas com cobranças desnecessárias (SICSÚ, 2010).

Entre as vantagens de medir o risco de crédito estão: consistências nas decisões (o resultado sempre será o mesmo independente de analista, agência ou filial), decisões rápidas (os recursos computacionais permitem executar o processo quase que instantaneamente), decisões adequadas, monitorar e administrar o risco de um portfólio de crédito, verifica o grau em que se atende os requisitos de órgãos reguladores, estabelecer linguagem comum entre os decisores de crédito e definir alçadas para concessão de crédito (SICSÚ, 2010).

3.3 Teste de *Wilcoxon Signed-rank* (WSRT)

Para estudos de comparação de modelos de inteligência artificial, é importante que seja realizados testes estatísticos que comprove conclusões, com o intuito de fornecer um suporte técnico para o resultado. O teste de *Wilcoxon Signed-rank* é um teste não-paramétrico, introduzido por Frank Wilcoxon em 1945, que trabalha com diferenças de pares de amostra, considerando como hipótese nula que a diferença é normalmente distribuída no zero (WILCOXON, 1945). Bandi *et. al.* (2023) complementa com a aplicação desse método em avaliação de modelos de *machine learning*, tendo dois modelos diferentes e que, nesse caso, a hipótese nula seria que o erro absoluto (ou qualquer métrica de desempenho) seja entre os modelos, por outro lado, a hipótese alternativa seria que o erro absoluto de um modelo é menor do que o outro.

4 METODOLOGIA E DESENVOLVIMENTO

Recordando que objetivo do trabalho, que é desenvolver um modelo de inteligência artificial que determine a probabilidade de uma pessoa ser mau pagadora ou não, observando se o uso de CNNs é razoável e comparando os algoritmos em quatro bases diferentes. Para cumprir com esse propósito, foi traçado uma metodologia com base no referencial teórico e trabalhos relacionados. Neste capítulo será abordado o fluxo do desenvolvimento do estudo, explicação mais detalhada de cada uma das quatro bases de dados escolhidas, as métricas de avaliação que foram escolhidas para comparar cada algoritmo em cada base e, por fim, os parâmetros escolhidos para o trabalho, as ferramentas e linguagem de programação escolhida.

4.1 Fluxograma do Desenvolvimento

Os passos do desenvolvimento foram inspirados em referências relacionadas ao trabalho e estão desenhadas na Figura 2. O estudo foi realizado com quatro bases de dados, testando cinco algoritmos de *machine learning* em cada uma delas, com isso, temos etapas que serão repetidas de acordo com o número de componentes que se tem. A primeira etapa é a coleta dos dados, na qual cada uma das bases tem um código para puxar os dados e retornar em formato ideal para tratamento.

Após isso, tem-se a etapa de pré-processamento dos dados, onde é realizado as manipulações com o propósito de fornecer dados em formato ideal para que o algoritmo consiga desempenhar melhor. Para o estudo será usado uma etapa de normalização das variáveis contínuas, transformação das variáveis categóricas em binárias e, por fim, uma técnica de balanceamento dos dados, que como base em trabalhos relacionados (ZHANG; CHENG, 2021) vamos utilizar o *K-Means SMOTE*, dado que esse é um dos problemas mais recorrentes de *credit scoring*. Importante pontuar que o método de balanceamento utilizado no estudo foi aplicado em duas das quatro bases (Taiwan e Aprovação de crédito), isso porque as outras bases, fundamentado na performance, não foram observados a necessidade da aplicação.

A próxima etapa é de desenvolvimento dos cinco algoritmos para cada uma das bases de dados, que no caso do estudo serão: regressão logística, *random forest*, *LightGBM*, *XGBoost* e CNN, com as suas respectivas escolhas de parâmetros para a execução.

Logo após o desenvolvimento do modelo, acontece a etapa de avaliação dos resultados de cada modelo em cada base de dado, avaliando em três métricas de performance do modelo,

que são: taxa de acuracidade, *F-Score* e AUC (*area under the ROC curve*). Com as métricas calculadas, é realizado uma avalição mostrando uma tabela com os algoritmos utilizados e os resultados de cada um deles para cada uma das bases.

Por fim, a última etapa é a avaliação dos resultados comparativos, que é uma conclusão com base em todos os resultados coletados. Nesta etapa que é avaliado se os objetivos do trabalho foram cumpridos e o que se pode dizer sobre ele, é possível avaliar se existe um modelo com que desempenha melhor do que os outras em todas as bases ou não.

Figura 2 – Fluxograma do Desenvolvimento do Trabalho

Fonte: Elaboração própria.

4.2 Base de dados

Neste tópico é apresentado as quatro bases de dados que são utilizadas no estudo, informando quais são as variáveis presentes em cada uma delas e o nível de balanceamento dos dados. Todos os conjuntos de dados do trabalho são bases públicas, utilizadas para competições de ciência de dados e para estudos da área, são 3 bases da *University of California, Irving* (UCI) que contém um repositório de dados para estudo. Por fim, também tem uma base do *Kaggle* utilizada para competição. Ao final desta seção é apresentado uma tabela resumo das bases de dados utilizadas na Tabela 1 e na Tabela 2 o resumo das bases após o tratamento dos dados.

4.2.1 Base de pagamentos de consumidores de Taiwan

Esta base foi extraída do repositório da *University of California, Irving* (UCI) e é do ano de 2005 e contém ao todo 25 colunas, sendo 23 variáveis, uma coluna de identificação e uma coluna de rótulo para identificar se atrasou ou não pagamento, que será o alvo do estudo. Dentre as variáveis estão presentes: limite de crédito, gênero (1 para masculino e 2 para feminino), formação educacional (1 para pós-graduação, 2 para universidade, 3 para ensino médio e 4 para outros), estado civil (1 para casado, 2 para solteiro e 3 para outro), idade, seis colunas indicando o estado de pagamento de abril a setembro de 2005, sendo uma coluna para cada mês, informando se foi pago corretamente (indicando -1), atraso de um mês (indicando 1), atraso de dois meses (indicando 2) e assim seguindo até o valor 9, que é atraso de nove meses ou mais. Além disso, também apresenta seis colunas informando o valor da dívida nos meses de abril a setembro e, por fim, mais seis colunas informando o quanto foi pago.

Sobre a volumetria, a base apresenta trinta mil linhas, sendo divida em 6636 casos como inadimplente e 23364 casos como pagamento devido, ou seja, aproximadamente 22% da base é de não pagante.

4.2.2 Base de pagamentos de consumidores da Alemanha

Esta base também foi extraída do repositório da *University of California, Irving* (UCI) contendo informações de clientes de um banco alemão do ano de 1994, apresentando 21 colunas ao todo, com 20 variáveis de comportamentais e pessoais e uma coluna identificando se a pessoa foi uma boa pagadora ou não. Explicando as variáveis envolvidas, a primeira é sobre o estado da conta corrente e duração do relacionamento com a conta. Além disso, também possui informação de crédito, informando se todos os créditos foram pagos devidamente com o dono da base ou com outra instituição financeira, outra coluna é a de propósito do crédito (compra de carro, casa, educação, entre outros). Também tem uma coluna informando o montante de crédito que possui, quantidade de poupança, tempo de emprego na atual empresa, taxa de prestação em percentual do rendimento disponível, gênero e estado civil combinados (exemplo: homem e divorciado), se existe outro devedor, tempo com a atual moradia, propriedade, idade, outros parcelamentos, moradia, número de crédito com o banco dono da base, estado do emprego, número de dependentes, se possui telefone e, por fim, se é um imigrante.

Sobre a volumetria, a base apresenta dez mil linhas, sendo divida em 300 casos como inadimplente e 700 casos como pagamento devido, ou seja, 30% da base é de não pagante.

4.2.3 Base de aplicação de cartão de crédito da Austrália

Esta base também foi extraída do repositório da *University of California, Irving* (UCI) contendo informações de clientes de uma instituição financeira australiana, esta base de dados é modificada para proteger a confidencialidade dos dados, que são classificados como sensíveis. Apesar disso, é apresentado um bom conjunto de variáveis com campos contínuos e categóricos. Apresenta 15 colunas ao todo, com 14 variáveis de comportamentais e pessoais e uma coluna identificando se a pessoa foi uma boa pagadora ou não, sendo 8 colunas categóricas e 6 contínuas.

Sobre a volumetria, a base apresenta 690 linhas, sendo divida em 307 casos como inadimplente e 383 casos como pagamento devido, ou seja, 45% da base é de não pagante.

4.2.4 Base de aprovação de cartão de crédito

Esta base foi extraída do repositório do *Kaggle*, que possui bases de dados para fins educacionais e de competição. Para este estudo essa base foi tratada para se adequar com o propósito do trabalho, apresentando 19 colunas ao todo, com 17 variáveis de comportamentais e pessoais, uma coluna de identificação e uma coluna identificando se a pessoa foi uma boa pagadora ou não. Dentre as varáveis são mostradas as seguintes informações: gênero, se possui carro, se possui moradia própria, número de filhos, renda anual, tipo de renda, escolaridade, estado civil, tipo de moradia, dias para aniversário, dias empregado (sendo uma contagem regressiva desde o dia atual, sendo número negativo indicando que a pessoa ainda está no emprego e o número positivo se a pessoa está desempregada), se possui telefone, se possui *e-mail*, tipo de ocupação e, por fim, tamanho da família.

Sobre a volumetria, a base apresenta 33110 linhas, sendo divida em 1108 casos como inadimplente e 32002 casos como pagamento devido, ou seja, aproximadamente 3% da base é de não pagante.

Tabela 1 – Resumo das bases de dados do estudo.

Base de Dados	Número de Linhas	Número de variáveis	Contínuas	Categóricas	Mau	Bom
Taiwan	30000	23	14	9	22%	78%
Alemanha	1000	19	8	11	30%	70%
Australia	690	13	5	8	45%	55%
Aprovação Crédito	33110	17	9	8	3%	97%

Fonte: Elaborada pelo autor.

Tabela 2 – Resumo das bases de dados do estudo após os tratamentos.

Base de Dados	Número de Variáveis Final	Mau Final	Bom Final
Taiwan	33	40%	60%
Alemanha	61	30%	70%
Australia	38	45%	55%
Aprovação Crédito	54	10%	90%

Fonte: Elaborada pelo autor.

4.3 Métricas de Avaliação

Com o intuito de medir o desempenho dos algoritmos em cada base de dados e concluir o estudo com base nos objetivos, foram escolhidas métricas com base em um trabalho relacionado (ZHANG; HE; ZHANG, 2018), que são: Acuracidade, AUC (*area under the ROC curve*) e *F1-score*. Ambas as métricas utilizam conceitos de Verdadeiro Positivo (VP), que seria que foi rotulado como positivo e de fato era, Verdadeiro Negativo (VN), que seria negativo e foi rotulado dessa forma também. Por outro lado, também tem o conceito de Falso Positivo (FP), que seria um negativo que foi rotulado como positivo e, da mesma forma, o Falso Negativo (FN), que seria o que foi classificado como negativo, mas era positivo na realidade.

A primeira métrica é a acuracidade, que é um indicador que calcula o quanto que o modelo acertou de forma geral, ou seja:

$$Acuracidade = \frac{VP + VN}{VP + VN + FP + FN}$$

Apesar de ser mais popular, ela pode ser enviesada em casos de bases desbalanceados, como existem no presente estudo. Por conta disso, é necessário comparar com outras métricas. A *Area Under the ROC Curve* (AUC), em uma tradução literal, quer dizer a área da curva ROC,

que seria uma curva formada por duas dimensões, que são: taxa de verdadeiros positivos e taxa de falso positivos, e dessa forma, ao calcular a área embaixo da curva formada pela combinação das duas variáveis se dá a métrica, as taxas são calculadas dessa forma:

$$Taxa\ de\ verdadeiros\ postivos = \frac{VP}{VP + FN}$$

$$Taxa\ de\ falso\ positivo = \frac{FP}{VN + FP}$$

Por fim, a métrica *F1-Score* é composta por duas outras métricas, que são: precisão e revogação (também chamada de *recall*). A primeira é uma proporção entre o verdadeiro positivo e falso positivo, enquanto a segunda é entre o verdadeiro positivo e os falsos negativos, ou seja, uma proporção de quantos positivos foram acertados dentre todos os realmente positivos. Ao calcular os dois indicadores, o *F1-Score* relaciona os dessa forma:

$$F1 - Score = 2 * \frac{Precisão * Revogação}{Precisão + Revogação}$$

Com os indicadores de performance definidos, é preciso entender como as três métricas serão comparadas. Para isso, foi utilizado o ordenamento dos algoritmos em cada uma das métricas, obtendo isso para todas as medidas, foi aplicado a multiplicação desses números e o algoritmo com o menor resultado será o primeiro lugar e, em contrapartida, o maior resultado será o último lugar. A escolha da multiplicação se dá por destacar as melhores posições, ao invés de lidar como um mesmo peso, como seria utilizando a soma. Além disso, foi realizado um teste estatístico para avaliar se os resultados demonstrados são significativos ou não. Por fim, são coletados também a velocidade de treinamento e consumo de memória para comparar com os resultados de desempenho dos modelos.

4.4 Tecnologias Utilizadas e Parâmetros do estudo

Sobre as ferramentas utilizadas no estudo, todo o desenvolvimento foi realizado com a linguagem *Python*, para o ambiente de execução foi utilizado o *Jupyter Notebook*. Em relação às bibliotecas escolhidas, são essas: *Numpy* (para executar comandos com vetores e numéricos), *Pandas* (para tratar tabelas para os algoritmos), *Matplotlib* (intuito de construção de gráficos), *Sklearn* (utilizando para aplicação de treinos e testes dos algoritmos Regressão Logística, *Random Forest* e *XGBoost*), *lightgbm* (para treino e teste do algoritmo *LightGBM*), *Torch*, *Pytorch Lightning* (construção do algoritmo de CNN) e, por fim, *scipy* (aplicação de teste estatístico).

Sobre os parâmetros utilizados, para todos os algoritmos, exceto CNN, as bases foram divididas em 25% para teste e 75% para treino, para a CNN, foram divididas em 60% para treino, 20% para validação e 20% para teste.

Para a regressão logística, para regularização foi escolhido "l2", que seria a regressão *Ridge*, taxa de tolerância para critério de parada (tol) de 0,0001 e o inverso da força regularização (C) como 1. Para o *Random Forest* foi escolhido um número de estimadores de 100, a função de mensuração de resultado é o *gini* e não há máxima profundidade da árvore. Já para o LightGBM o número máximo de ramificações é de 31, também não foi selecionado um número máximo de profundidade e a taxa de aprendizado é de 0,1. Por fim, o *XGBoost*, foi escolhido uma taxa de aprendizado de 0,3, com uma máxima profundidade da árvore de 6 e o método de *sampling* é o uniforme.

Para a CNN, como existem muitas possibilidades de desenvolvimento, o presente estudo utilizou como inspiração o trabalho relacionado de *Song et. al,* 2023, desenvolvendo uma rede neural com uma dimensão, utilizando o formato SR 1D CNN (*Soft reordering one-dimensional convolutional neural networks*). Com isso, primeiro foi realizado um ajuste nos dados de entrada inserindo nas camadas de convolução com 128 canais, além disso foram aplicadas 4 camadas de convolução de uma dimensão, com a função de ativação *ReLu*, ao final da primeira camada de convolução é aplicado o método *average pooling*, como também acontece ao final da quarta camada. Por fim são utilizadas 3 camadas de *dropout*, sendo a primeira depois do primeiro *average pooling*, utilizando taxa de 30%, o segundo é aplicada depois da terceira camada de convolução, utilizando taxa de 30% também e, por fim, o último *dropout* se encontra após o último *average pooling*, com taxa de 20%.

5 RESULTADOS

Nesta seção serão apresentados os resultados de cada modelo em cada base avaliada neste trabalho. Primeiro é demonstrado o resultado em cada um dos quatro conjuntos de dados, mostrando o desempenho de cada modelo dentro de cada métrica de avaliação e, após isso, a ordem de performance, definindo o melhor modelo para a base de dados. Por fim, será exposto uma análise englobando o estudo como um todo para definir o melhor modelo.

5.1 Taiwan

Na base de Taiwan, como podemos ver de modo resumido na Tabela 3, analisando em AUC foi verificado um desempenho de 79,29% para a regressão logística, 83,37% para *Random Forest*, 83,54% para *XGBoost*, 84,60% para o *LightGBM* e 82,84% para a CNN, mostrando um desempenho melhor para o quarto modelo e uma performance superior dos modelos que envolvem árvore de decisão. Observando em *F1-Score*, a regressão logística obteve 58,90%, enquanto o *Random Forest* teve 65,15%, o *XGBoost* 64,88%, o *LightGBM* teve 65,12% e, por fim, a CNN 61,10%, demonstrando um desempenho superior do segundo modelo nessa métrica, mas continuando com a tendência do outro indicador, que é um domínio dos modelos que envolvem árvore de decisão. Por último, analisando a acurácia, a regressão logística apresenta uma performance de 81,48%, já o *Random Forest* 83,17%, o *XGBoost* 83,00%, o *LightGBM* com 82,25% e, por fim, a CNN com 82,36%. Nesta última análise, é possível verificar novamente um desempenho superior do modelo *Random Forest*, porém mostra uma quebra de padrão observada nas outras métricas com a CNN aparecendo entre os três melhores.

Tabela 3 – Resultados das métricas de avaliação para a base de Taiwan

Algoritmo	**AUC**	**F1-Score**	**Acurácia**
Regressão Logística	79,29%	58,90%	81,48%
Random Forest	83,37%	65,15%	83,17%
XGBoost	83,54%	64,88%	83,00%
LightGBM	84,60%	65,12%	82,25%
CNN	82,84%	61,10%	82,36%

Fonte: Elaborada pelo autor.

Depois de avaliar cada métrica e os melhores modelos em cada uma delas, é necessário entender qual é o ordenamento correto levando em consideração todos os indicadores em

conjunto. Na tabela 4 é possível observar que não existe um modelo dominante em todas as perspectivas, porém o modelo *Random Forest* desempenha melhor em duas das três medidas, sendo considerado o melhor. Para as outras colocações houve um equilibro por ter trocas de posições frequente, mas aplicando o método do estudo observa-se que o *LightGBM* é o segundo melhor modelo, seguido pelo *XGBoost*, CNN e regressão logística, respectivamente. De modo geral, os modelos que têm na sua composição árvores de decisão foram os melhores modelos dessa base de dados.

Tabela 4 – Resultados com o ordenamento para a base de Taiwan

Algoritmo	**AUC**	**F1-Score**	**Acurácia**	**Pontuação**	**Ordenamento**
Regressão Logística	5	5	5	125	5
Random Forest	3	1	1	3	1
XGBoost	2	3	2	12	3
LightGBM	1	2	4	8	2
CNN	4	4	3	48	4

Fonte: Elaborada pelo autor.

5.2 Alemanha

Na base da Alemanha, analisando em AUC a regressão logística desempenhou com 73,30%, já o *Random Forest* teve 77,70%, o *XGBoost* com 74,17%, o *LightGBM* com 73,90% e a CNN com 77,50%. No primeiro indicador é possível observar um desempenho superior do segundo modelo e em seguida a CNN, com resultados próximos. Em seguida, avaliando o *F1-Score* a regressão logística pontuou com 81,10%, o *Random Forest* com 85,60%, *XGBoost* com 83,00%, *LightGBM* com 84,40% e a CNN com 81,13%. Neste segundo indicador o *Random Forest* novamente desempenhou melhor, seguido pelo quarto modelo, nesta perspectiva fica claro o domínio dos modelos que possuem árvore de decisão em sua composição. Por último, examinando a acurácia, a regressão logística foi classificada com 72,80%, o *Random Forest* com 78,40%, *XGBoost* com 75,60%, *LightGBM* com 77,20% e CNN com 75,49%. Na acurácia o comportamento foi igual ao do penúltimo indicador em relação ao ordenamento. Os resultados são mostrados de forma resumida na Tabela 5.

Tabela 5 – Resultados das métricas de avaliação para a base da Alemanha

Algoritmo	**AUC**	**F1-Score**	**Acurácia**
Regressão Logística	73,30%	81,10%	72,80%
Random Forest	77,70%	85,60%	78,40%
XGBoost	74,17%	83,00%	75,60%
LightGBM	73,90%	84,40%	77,20%
CNN	77,50%	81,13%	75,49%

Fonte: Elaborada pelo autor.

Analisando todas as métricas em conjunto é possível observar que o algoritmo *Random Forest* apresenta o melhor resultado em todas as perspectivas, mostrando domínio nessa base de dados. Como em duas das três métricas o ordenamento foi igual, então a organização final ficou parecida, com *LightGBM* em segundo e em seguida *XGBoost*, CNN e regressão logística, respectivamente. A ordenação final com a sua pontuação está presente na Tabela 6.

Tabela 6 – Resultados com o ordenamento para a base da Alemanha

Algoritmo	**AUC**	**F1-Score**	**Acurácia**	**Pontuação**	**Ordenamento**
Regressão Logística	5	5	5	125	5
Random Forest	1	1	1	1	1
XGBoost	3	3	3	27	3
LightGBM	4	2	2	16	2
CNN	2	4	4	32	4

Fonte: Elaborada pelo autor.

5.3 Austrália

Na base da Austrália, como podemos ver de modo resumido na Tabela 7, observando em AUC foi mostrado um desempenho de 91,60% para a regressão logística, 92,60% para *Random Forest*, 92,45% para *XGBoost*, 93,46% para o *LightGBM* e 92,02% para a CNN. Apresentando um desempenho melhor para o quarto modelo e uma performance superior dos modelos que envolvem árvore de decisão. Porém analisando o *F1-Score*, o comportamento se difere, com regressão logística com 85,89%, *Random Forest* com 85,35%, *XGBoost* com 85,16%, *LightGBM* com 83,87% e CNN com 85,60%. Com isso, é possível observar um desempenho melhor da regressão logística e da CNN do que os outros modelos que são

compostos por árvores de decisão. Por fim, examinando a acurácia, a regressão logística desempenho com 87,28%, o *Random Forest e XGBoost* com 86,70%, *LightGBM* com 85,54% e a CNN com 86,90%. Nesta métrica o comportamento se apresentou de forma similar com o *F1-Score*, exceto pelo empate entre *XGBoost* e *Random Forest*, mas que performaram de forma quase igual também na outra perspectiva, tendo uma diferença de 0,19%.

Tabela 7 – Resultados das métricas de avaliação para a base da Austrália

Algoritmo	**AUC**	**F1-Score**	**Acurácia**
Regressão Logística	91,60%	85,89%	87,28%
Random Forest	92,90%	85,35%	86,70%
XGBoost	92,45%	85,16%	86,70%
LightGBM	93,46%	83,87%	85,54%
CNN	92,02%	85,60%	86,90%

Fonte: Elaborada pelo autor.

Analisando todos os indicadores em conjunto é possível concluir que o algoritmo regressão logística apresenta o melhor resultado em todas as perspectivas, não com domínio total em todas as análises, mas sendo o mais bem colocado em duas das três métricas. As avaliações se diferenciaram de forma considerável entre AUC e as outras métricas, podendo ser por conta do baixo número de linhas presentes nessa base, fazendo com que o desempenho seja parecido entre os algoritmos, a diferença entre o mais bem posicionado e o pior na AUC é de menos de 2%. Por fim, a classificação final foi com a regressão logística em primeiro, seguido por CNN, *Random Forest*, *LightGBM* e *XGBoost*. A ordenação final com a sua pontuação está presente na Tabela 8.

Tabela 8 – Resultados com o ordenamento para a base da Austrália

Algoritmo	**AUC**	**F1-Score**	**Acurácia**	**Pontuação**	**Ordenamento**
Regressão Logística	5	1	1	5	1
Random Forest	2	3	3	18	3
XGBoost	3	4	3	36	5
LightGBM	1	5	5	25	4
CNN	4	2	2	16	2

Fonte: Elaborada pelo autor.

5.4 Aprovação de Crédito

Na base de aprovação de crédito, examinando o AUC, a regressão logística foi pontuada com 84,60%, enquanto o *Random Forest* ficou com 91,20%, o *XGBoost* com 89,40%, o *LightGBM* com 90,10% e a CNN com 86,20%. Apresentando um domínio dos algoritmos compostos por árvore de decisão, em especial o *Random Forest*. Em seguida, no indicador *F1-Score*, a regressão logística teve 71,70%, já o *Random Forest* teve 80,01%, *XGBoost* 78,10%, *LightGBM* com 78,70% e CNN com 70,90%. Comportamento quase parecido com o indicador anterior, porém neste a CNN teve o pior desempenho. Por último, na acurácia a regressão logística apresentou uma pontuação de 95,40%, o *Random Forest* e *LightGBM* empatados com 96,60%, *XGBoost* com 96,50% e CNN com 95,50%. Nesta última métrica tivemos um empate entre o melhor algoritmo, entre o *Random Forest e LightGBM*, de resto seguiu o mesmo padrão da AUC. Os resultados estão resumidos na Tabela 9.

Tabela 9 – Resultados das métricas de avaliação para a base da Aprovação de Crédito.

Algoritmo	**AUC**	**F1-Score**	**Acurácia**
Regressão Logística	84,60%	71,70%	95,40%
Random Forest	91,20%	80,01%	96,60%
XGBoost	89,40%	78,10%	96,50%
LightGBM	90,10%	78,70%	96,60%
CNN	86,20%	70,90%	95,50%

Fonte: Elaborada pelo autor.

Analisando de forma geral, é possível observar que os algoritmos com árvore de decisão na composição se desempenharam melhor em todos os critérios. Além disso, também pode-se concluir que o algoritmo *Random Forest* foi o que desempenhou melhor em todos os critérios, apenas na acurácia que teve a liderança dividida. Em seguida, o *LightGBM* ficou com a segunda colocação, *XGBoost* em terceiro, CNN em quarto e regressão logística em último. Os dados com a classificação de cada algoritmo em cada critério e a pontuação final está na Tabela 10.

Tabela 10 – Resultados com o ordenamento para a base de Aprovação de Crédito.

Algoritmo	AUC	F1-Score	Acurácia	Pontuação	Ordenamento
Regressão Logística	5	4	5	100	5
Random Forest	1	1	1	1	1
XGBoost	3	3	3	27	3
LightGBM	2	2	1	4	2
CNN	4	5	4	80	4

Fonte: Elaborada pelo autor.

5.5 Análise Geral

Com os ordenamentos em cada base de dados do trabalho, é possível avaliar a performance geral dos algoritmos presentes no trabalho, com o objetivo de entender se existe um algoritmo dominante ou algum outro padrão.

Na Tabela 11 é mostrada a posição de cada algoritmo em cada base de dados. A partir disso, seguindo o mesmo critério para avaliar em cada base, as colocações são multiplicadas e colocadas em ordem ascendente, assim pode-se ver o melhor algoritmo no contexto do estudo. A partir da tabela, é possível concluir que não existe um modelo dominante em todas as bases de dados, porém existe um modelo que desempenha melhor em três das quatro bases, que é o *Random Forest*, em seguida o *LightGBM*, em terceiro encontra-se a regressão logística, que foi o pior modelo na maior parte das bases, porém em uma base ela foi a melhor. Em seguida, a CNN, mostrando que não desempenhou melhor do que os modelos tradicionais em nenhuma base presente no estudo. E no final ficou o *XGBoost*, que foi o terceiro melhor em três dos quatro estudos, mas teve um último lugar em uma base e como o critério de pontuação é multiplicação isso teve peso na classificação final.

Tabela 11 – Resultados com o ordenamento geral.

Base de Dados	Aprovação Crédito	Alemanha	Taiwan	Australia	Pontuação	Ranking
Regressão Logística	5	5	5	1	125	3
Random Forest	1	1	1	3	3	1
XGBoost	3	3	3	5	135	5
LightGBM	2	2	2	4	32	2
CNN	4	4	4	2	128	4

Fonte: Elaborada pelo autor.

Após realizar os desenvolvimentos e análises dos resultados, foi aplicado o teste de *Wilcoxon signed-rank* para avaliar a significância estatística entre os resultados, buscando avaliar se houve diferenças entre as performances indicadas. Para isso, o indicador de AUC foi escolhido para a avaliação do teste, ao explorar os resultados na Tabela 12, contendo os p valores das comparações, foi possível verificar que nenhuma correspondência teve o valor menor do que 0,05, que foi a referência escolhida para o teste. Com isso, não é possível rejeitar a hipótese nula, que aponta similaridade entre os desempenhos dos modelos, ou seja, todos os resultados possuem a mesma relevância no ponto de vista estatístico.

Tabela 12 – Resultados dos p valores entre os modelos.

	Regressão Logística	*Random Forest*	*XGBoost*	*LightGBM*	CNN
Regressão Logística	-	0,25	0,25	0,25	0,875
Random Forest	0,25	-	0,125	0,125	0,25
XGBoost	0,25	0,125	-	0,625	0,25
LightGBM	0,25	0,125	0,625	-	0,25
CNN	0,875	0,25	0,25	0,25	-

Fonte: Elaborada pelo autor.

Com base nessa última análise, relacionando com o objetivo do estudo de comparar modelos de inteligência artificial em bases diferentes, foi possível verificar que não existe um modelo dominante em todas as quatro bases de dados diferentes, porém o *random forest*

apresenta um desempenho ligeiramente melhor do que os outros, mas segundo o teste estatístico não é possível afirmar que é o melhor modelo.

Além desse objetivo, conectando sobre a possibilidade do uso da CNN no contexto de *credit scoring*, nisso foi possível observar que a CNN não desempenha de forma superior aos modelos tradicionais, porém após os resultados do teste de *Wilcoxon signed-rank* é seguro afirmar que apresenta um desempenho similar aos outros modelos. Possivelmente com bases maiores e um estudo aprofundado para encontrar uma composição mais adequada para as redes neurais, pode-se chegar em um modelo com performance melhor.

Após verificar similaridade nos resultados, uma possibilidade de distinção entre os modelos pode ser a velocidade de treinamento dos modelos e o consumo de memória de cada um deles. Coletando os tempos, é possível verificar na Tabela 13 que os tempos entre os modelos tradicionais (regressão logística, *random forest, XGBoost* e *LightGBM*) apresentam uma variação baixa de tempo, oscilando entre 0,7 até 1,4 segundos para a regressão logística, *XGBoost* e *LightGBM*, já *random forest* oscila entre 1,2 até 5,9 segundos. Por último, a CNN apresenta valores muito superiores, oscilando entre 564 até 6430 segundos.

Tabela 13 – Tempos do treinamento dos modelos em segundos.

	Regressão Logística	*Random Forest*	*XGBoost*	*LightGBM*	CNN
Australia	1,3969	1,2128	1,0106	1,053	688
Alemanha	1,3487	1,3524	1,4657	1,0489	564,29
Taiwan	0,752	5,9442	0,936	0,985	6430,26
Aprovação de Crédito	0,7528	2,3583	0,9228	1,4511	3942,02

Fonte: Elaborada pelo autor.

Além do tempo de treinamento, uma informação importante na avaliação de algoritmos é o consumo de dados que cada um deles gera ao treinar em cada uma das bases. Após coletar essas informações, presentes na Tabela 14, é possível verificar que os modelos regressão logística, *random forest*, *XGBoost* e *LightGBM* apresenta um consumo similar, oscilando entre 220 e 870 MebiByte (MiB), enquanto a CNN apresenta uma oscilação entre 750 e 1027 MiB.

Tabela 14 – Consumo de memória treinamento dos modelos em mebibyte.

	Regressão Logística	*Random Forest*	*XGBoost*	*LightGBM*	CNN
Australia	241,79	241,2	245,88	232,04	752,89
Alemanha	230,61	230,14	219,53	214,26	773,08
Taiwan	870,89	864,36	805,21	810,84	1027,36
Aprovação de Crédito	370,86	360	328,4	334,04	1102,6

Fonte: Elaborada pelo autor.

Com o fim da coleta de dados de tempo de processamento e consumo de memória dos treinamentos de cada algoritmo em cada base de dados, foi necessário repassar o teste de *Wilcoxon signed-rank* em cada uma dessas análises, avaliando a relevância estatística desses resultados. Nas tabelas 15 e 16 é possível verificar os p valores e concluir que, apesar dos valores entre CNN e o resto dos modelos serem diferentes em grandeza, principalmente quando se analise os dados de velocidade, nenhum dos p valores foi abaixo de 0,05, ou seja, não recusando a hipótese nula de que os resultados são similares. Esses resultados podem ter aparecido com esse formato por conta da quantidade baixa de amostragem, talvez com uma quantidade maior seria possível provar que os resultados da CNN não similares ao restante. Por outro lado, mesmo com essas conclusões, pode-se afirmar que a CNN apresenta um consumo maior e tempo de processamento maior do que o restante e apresenta um resultado semelhante em performance, e com isso, não sendo um modelo aconselhável para uso com o formato do estudo. Apesar disso, é plausível pensar que com um aprofundamento dos detalhes da CNN com o objetivo de otimizar a performance dele no contexto de *credit scoring*, possível que seja recomendável.

Tabela 15 – Resultados dos p valores com base no tempo de treinamento

	Regressão Logística	*Random Forest*	*XGBoost*	*LightGBM*	CNN
Regressão Logística	-	0,375	0,875	1	0,125
Random Forest	0,375	-	0,25	0,125	0,125
XGBoost	0,875	0,25	-	0,625	0,125
LightGBM	1	0,125	0,625	-	0,125
CNN	0,125	0,125	0,125	0,125	-

Fonte: Elaborada pelo autor.

Tabela 16 – Resultados dos p valores com base no consumo de memória

	Regressão Logística	*Random Forest*	*XGBoost*	*LightGBM*	CNN
Regressão Logística	-	0,125	0,25	0,125	0,125
Random Forest	0,125	-	0,25	0,125	0,125
XGBoost	0,25	0,25	-	1	0,125
LightGBM	0,125	0,125	1	-	0,125
CNN	0,125	0,125	0,125	0,125	-

Fonte: Elaborada pelo autor.

6 CONCLUSÃO

A importância do crédito na vida das pessoas se concretiza, por exemplo, através da possibilidade de compra de imóveis, carros, motos e até financiamentos de estudos, como mestrado internacional e intercâmbio. Por conta disso, é fundamental desenvolver modelos cada vez melhores para que os clientes consigam ter mais poder de compra com menos risco para as instituições financeiras.

Com esse intuito, foi desenvolvida uma avaliação dos principais algoritmos de inteligência artificial, incluindo uma avaliação da aplicação de CNN dentro do contexto de *credit scoring*. Na base de Taiwan, o modelo de *Random Forest* foi o que melhor desempenhou, com 83,37% de AUC, 65,15% de *F1-Score* e 83,17% de acurácia, já na base da Alemanha o algoritmo de *Random Forest* também obteve o melhor desempenho, com 77,70% de AUC, 85,60% de *F1-Score* e 78,40% de acurácia, na base da Austrália a regressão logística teve a melhor performance com 91,60% de AUC, 85,89% de *F1-Score* e 87,28% de acurácia. Por fim, na base de aprovação de crédito, o *Random Forest* foi classificado com o melhor desempenho com 91,20% de AUC, 80,01% de *F1-Score* e 96,60% de acurácia.

Com isso, foi possível observar que não existe um modelo dominante em todas as bases, mas o algoritmo de *Random Forest* teve o melhor desempenho na maior parte das bases, sendo o mais indicado segundo o estudo, em segundo ficou o *LightGBM*, seguido por regressão logística, depois CNN e, por último, o *XGBoost*. Porém, ao realizar o teste de *Wilcoxon signed-rank* foi possível verificar que não existe diferenças estatisticamente significativas entre os modelos. Então, a CNN demonstrou resultados satisfatórios em todas as análises, não tendo resultados descolados dos modelos tradicionais.

Como o estudo teve o objetivo de avaliar modelos, não foi possível aprofundar sobre os parâmetros de cada modelo, ou seja, apesar do *XGBoost* ser classificado como pior modelo, com a melhoria de parâmetros pode ser possível que a posição dele mude bastante, mesma situação com o *LightGBM*, que com outros parâmetros poderia ser o melhor modelo. Com isso, uma limitação do estudo foi a não otimização de parâmetros dos modelos.

Para estudos futuros, uma sugestão na otimização dos parâmetros do *LightGBM* e *XGBoost*, como foi citado anteriormente, apresenta um potencial de ganho de performance, podendo ser melhor do que o modelo vencedor. Outra proposta seria encontrar uma melhor composição da CNN utilizada no trabalho, com as inúmeras possibilidades de otimização das redes neurais, é possível que se encontre uma composição que melhore o desempenho desse

algoritmo. Além disso, outra proposta seria continuar com os modelos utilizados no presente estudo e aumentar o número de bases e aplicar novamente o teste estatístico, com o objetivo de observar se existe possibilidade de mudar a conclusão. Por fim, uma última proposta seria continuar com as bases de dados e avaliar o desempenho com um modelo de *ensemble* combinando os classificadores do presente estudo.

REFERÊNCIAS

ABELLÁN, J; CASTELLANO, J. G. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, volume 73, p. 1-10, Maio 2017.

ABID, L; MASMOUDI, A; ZOUARI-GHORBEL, A. The Consumer Loan's Payment Default Predictive Model: na Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank. Journal of the Knowledge Economy, Portland International Center for Management of Engineering and Technology (PICMET), vol. 9(3), p.948-962. 2018.

ALA'RAJ, M; ABBOD, M. F; A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Systems with Applications, volume 64, p. 36-55, Dez. 2016.

ALKHATIB, K. I; AL-AIAD, A; ALMAHMOUD, M. H; ELAYAN, O. N. Credit Card Fraud Detection Based on Deep Neural Network Approach.In: 2th International Conference on Information and Communication Systems (ICICS), 2021, Valencia, Espanha, p.153-156.

BANDI, B; MANELIL, N, P; MAIYA, M,P; TIWARI, S; ARUNVEL, T. CFD driven prediction of mean radiant temperature inside an automobile cabin using machine learning. Thermal Science and Engineering Progress, Volume 37, 2023.

BRUCE, P; BRUCE, A. Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais. Rio de Janeiro: Alta Books, 2019.

CHEN, T; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785-794, Ago. 2016.

CROUHY, M; GALAI, D; MARK, R. The Essentials of Risk Management. Nova Iorque: McGraw Hill Professional, 2006.

DASTILE, X; CELIK, T; POTSANE, M. Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing, Volume 91, artigo 106263, Jun. 2020.

DUMITRESCU, E; HUÉ, S; HURLIN, C; TOKPAVI, S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, volume 297, p. 1178-1192, Mar. 2022.

FRIEDMAN, J. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, p. 1189-1232, 2001.

GOODFELLOW, I; BENGIO, Y; COURVILLE, A. Deep Learning. Cambridge: MIT Press, 2016.

HAN, J; KAMBER, M; PEI, J. Data Mining: Concepts and Techniques. 3rd ed. Massachusetts: Elsevier, 2012.

HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Nova Iorque: Springer, 2009.

HOSAKA, T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. Experts Systems with Applications, volume 117, p. 287-299, Mar. 2019.

JAMES, G; WITTEN, D; HASTIE, T; TIBSHIRANI, R. An Introduction to Statistical Learning: with Applications in R. Nova Iorque: Springer, 2013.
JOHNSON, A.B. Credit Scoring em instituição financeira digital. Dissertação (Mestrado) - Programa de Pós-Graduação em Controladoria e Contabilidade, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2022.
KAI-FU, L. AI superpowers: China, Sillicon Valley, and the new world order. Nova Iorque: HMHCO, 2018.
KANG, Y; JIA, N; CUI R; DENG, J. A graph-based semi-supervised reject inference framework considering imbalanced data distribution for consumer credit scoring. Applied Soft Computing, volume 105, Jul. 2021.
KE, G; MENG, Q; FINLEY, T; WANG, T; CHEN W; MA W; YE Q; LIU, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural information Processing Systems 30, p.3146-3154, 2017.
KHINE, A. A; KHIN, H. W. Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree. 2020 IEEE Conference on Computer Applications(ICCA), p.1-4, 2020.
KIM, J; KANG, P. Late payment prediction models for fair allocation of customer contact lists to call center agents. Decision Support Systems, volume 85, p. 84-101, Maio 2016.
LI, Q; ZHAO, Shuai; ZHAO, Shancheng; WEN J. Logistic Regression Matching Pursuit algorithm for text classification. Knowledge-Based Systems, volume 277, Out. 2023.
LI, Z; TIAN, Y; LI, K; ZHOU F; YANG, W. Reject inference in credit scoring using Semi-supervised Support Vector Machines. Expert Systems with Applications, volume 74, p. 105-114, Maio 2017.
LOUZADA, F; ARA, A; FERNANDES, G. B. Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, volume 21, p. 117-134, 2016.
MADHURYA, M. J.; GURURAJ, H. L.; SOUNDARYA, B. C.; VIDYASHREE, K. P.; RAJENDRA, A. B. Exploratory analysis of credit card fraud detection using machine learning techniques. Global Transitions Proceedings, volume 3, p. 31-37, Jun. 2022.
MAHBOOB, T. A; KAMRAN, S;HAMEED, I. A; LUO, S; SARWAR, M. U; SHABBIR, S. LI, J; KHUSHI, M. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. IEEE Access, volume 8, p. 201173-201198, 2020.
NAZARETH, N; REDDY, Y. V. R. Financial applications of machine learning: A literature review, Expert Systems with Applications, Volume 219, Fev. 2023.
QIAN, H; MA, P; GAO, S; SONG, Y. Soft reordering one-dimensional convolutional neural network for credit scoring. Knowledge-Based Systems, volume 266, Abril 2023.
SADOK, H; SAKKA, F; MAKNOUZI, M. E. H. E. Artificial Intelligence and bank credit analysis: A review. Cogent Economic & Finance. DOI: 10.1080/23322039.2021.2023262.
SICSÚ, A.L. Credit Scoring: Desenvolvimento, Implantação e Acompanhamento. São Paulo: Blucher, 2010.
SIQUEIRA, J; CASULA, D; CASTELLI, L; ALVES, J; SARDENBERG, R. Pesquisa Febraban de Economia Bancária e Expectativas. FEBRAPAN, 2023. Disponivel em: https://cmsarquivos.febraban.org.br/Arquivos/documentos/PDF/Pesquisa%20FEBRABAN%20de%20Economia%20Banc%C3%A1ria%20e%20Expectativas%20-%20Dezembro%20de%202023_v_imprensa.pdf. Acesso em: 12/01/2024.
SOHN, S. Y; KIM, D. H; YOON, J. H. Technology credit scoring model with fuzzy logistic regression. Applied Soft Computing, volume 43, p. 150-158, Jun. 2016.

SRINATH, T; GURUJARA, H.S. Explainable machine learning in identifying credit card defaulters. Global Transitions Proceedings, volume 3, p. 119-126, Jun. 2022.
TALAAT, F. M; ALJADANI, A; BADAWY, M; ELHOSSEINI, M. Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Comput & Applic, volume 36, p. 4847-4865, Dez. 2023.
THOMAS, L. C; MATUSZYK, A; SO, M. C; MUES, C; MOORE, A. Modelling repayment patterns in the collections process for unsecured consumer debt: A case studey. European Journal of Operational Research, volume 249, p. 476-486, Mar. 2016.
VANDERPLAS, J. Python Data Science Handbook: Essential Tools for Working with data. California: O'Reilly, 2017.
WILCOXON, F. Individual comparisons by ranking methods, Biometrics 1, p. 80-83. 1945.
XIA, Y; LIU, C; LI, Y; LIU, N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, volume 78, p.225-241, Jul. 2017.
XIAO, H; ZIA, Z; WANG, Y. Ensemble classification based on supervised clustering for credit scoring. Applied Soft Computing, Volume 43, p. 73-86, Jun. 2016.
ZHANG, H; HE, H; ZHANG, W. Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. Neurocomputing, volume 316, p. 210-221, Nov. 2018.
ZHANG, R; CHEN, Y. Research on Credit Card Default Prediction Based on k-Means SMOT and BP Neural Network. Complexity, volume 2021, Mar. 2021.
ZHU, X; CHU, Q; SONG, X; HU, P; PENG, L. Explainable prediction of loan default based on machine learning models. Data Science and Management, Volume 6, p. 123-133, Set 2023.