UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Uma análise experimental de algoritmos de aprendizado de máquina baseados em árvore de decisão para previsão de score de crédito

**Ezequiel Barazetti**

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

ICMC USP
SÃO CARLOS

**Ezequiel Barazetti**

# Uma análise experimental de algoritmos de aprendizado de máquina baseados em árvore de decisão para previsão de score de crédito

Trabalho de conclusão de curso apresentado ao 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. Flávia Bernardini

**São Carlos**

**2023**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Barazetti, Ezequiel
S856m Uma análise experimental de algoritmos de aprendizado de máquina baseados em árvore de decisão para previsão de score de crédito / Ezequiel Barazetti ; orientadora Prof. Dr. Flávia Bernardini. – São Carlos, 2023.
Monografia (MBA em Integência Artificial e BigData) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

Bernardini, Flávia, orient. II. Título.

## AGRADECIMENTOS

Agradeço primeiramente a meu senhor e salvador Jesus Cristo, que me permitiu concluir mais essa face em minha vida. Também agradeço a minha orientadora Dr. Flávia Bernardini que com muita paciência me ajudou nessa caminha.

## RESUMO

A análise experimental de algoritmos de aprendizado de máquina baseados em árvore de decisão para previsão de score de crédito, apresentam a característica de serem interpretáveis, podendo oferecer explicação sobre suas decisões, o que pode ser visto com *insights*. Uma análise exploratória utilizando diferentes algoritmos de aprendizado de máquina baseados na construção de árvores de decisão, busca determinar qual modelo apresenta melhor desempenho na previsão de scores de crédito de indivíduos. Diante disso, por meio de experimentos controlados e análises comparativas, essa pesquisa examina a capacidade de cada um dos algoritmos utilizados baseados em árvore de decisão, em lidar com complexidades, detectar padrões e generalizar informações a partir dos dados históricos de crédito. Ao considerar métricas de avaliação como acurácia, F1-score e taxa de erro total, a análise experimental contribui para uma seleção informada do algoritmo mais adequado à previsão de score de crédito, oferecendo um subsídio valioso para que as instituições financeiras otimizem suas decisões de concessão de crédito. Sendo assim, este trabalho apresenta os resultados da análise experimental realizada em um conjunto de dados público, disponibilizado na plataforma *Kaggle*, empregando os algoritmos *Decision Tree*, *Randon Forest* e *XGBoost*. Após o treinamento dos modelos, que envolveu a utilização de dados tanto balanceados quanto desbalanceados e a realização dos devidos ajustes nos hiperparâmetros, o algoritmo que apresentou melhor desempenho foi o *XgBoost*.



**Palavras-chave**: algoritmos de aprendizado de máquina, árvore de decisão, previsão de score de crédito, análise experimental.

## ABSTRACT

The experimental analysis of decision tree-based machine learning algorithms for credit score prediction presents the characteristic of being interpretable, providing explanations for their decisions, which can be seen as insights. Exploratory analysis using different machine learning algorithms based on decision tree construction seeks to determine which model performs better in predicting individuals' credit scores. Through controlled experiments and comparative analyses, this research examines the ability of each decision tree-based algorithm to handle complexities, detect patterns, and generalize information from historical credit data. Considering evaluation metrics such as accuracy, F1-score, and total error rate, the experimental analysis contributes to an informed selection of the most suitable algorithm for credit score prediction, offering valuable support for financial institutions to optimize their credit decision-making. Therefore, this work presents the results of the experimental analysis conducted on a publicly available dataset on the Kaggle platform, employing the Decision Tree, Random Forest, and XGBoost algorithms. After training the models, which involved using both balanced and unbalanced data and making necessary adjustments to hyperparameters, the algorithm that demonstrated the best performance was XGBoost.



**Keywords**: Machine learning algorithms, decision tree, credit score prediction, experimental analysis.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| GPUs | Graphics Processing Unit |
| TPUs | Tensor Processing Unit |
| SICOOB | Sistema de Cooperativas Financeiras do Brasil |
| SICRED | Sistema de Crédito Cooperativo |
| CRESOL | Cooperativa de Crédito Rural com Interação Solidária |

## SUMÁRIO

# 1 INTRODUÇÃO

O cooperativismo de crédito baseado no modelo agrícola alemão surgiu no Brasil em 1902, na cidade de Nova Petrópolis, Rio Grande do Sul. Seu objetivo era ajudar pequenas vilas e comunidades rurais, além de oferecer crédito para a compra de insumos e estruturação de propriedades agrícolas com juros mais baixos.

> O Cooperativismo de Crédito chegou ao Brasil, trazido da Europa pelo Padre Theodor Amstad, com o objetivo de reunir as poupanças das comunidades de imigrantes e colocá-las a serviço de seu próprio desenvolvimento.Foi em Linha Imperial, município de Nova Petrópolis, que o Padre precursor constituiu formalmente a primeira Cooperativa da espécie, em 28 de dezembro de 1902. (SCHARDONG, 2003, p.63).

Com o sucesso da iniciativa do Padre Theodor, várias outras cooperativas surgiram e, com elas, a necessidade de desenvolver e aprimorar técnicas para reduzir a inadimplência e melhorar o controle e gestão do risco de crédito. A análise de risco de crédito é um fator determinante para o setor financeiro, pois permite avaliar a capacidade de indivíduos e empresas em cumprir suas obrigações financeiras. Com base em uma análise criteriosa do perfil de crédito do mutuário, incluindo seu histórico financeiro dentro da instituição, ativos, passivo e outras informações, a análise de risco de crédito auxilia no processo de tomada de decisão de concessão de crédito.

Vale salientar que a importância dessa análise aumentou significativamente nos últimos anos, à medida que a economia global se tornou mais incerta e as taxas de inadimplência cresceram. Como resultado, a análise de risco de crédito tornou-se uma ferramenta valiosa para proteger os interesses das empresas que concedem crédito, a fim de mitigar o risco financeiro e garantir a sustentabilidade do setor financeiro. Nesse contexto, o uso de recursos tecnológicos se tornou indispensável para garantir uma melhor eficiência no processo de análise e concessão de crédito, o que levou as instituições financeiras a utilizar algoritmos para auxiliar nesse processo.

Dessa forma, os algoritmos de aprendizado de máquina, baseados em árvore de decisão, têm se mostrado altamente relevantes na área da previsão de score de crédito, uma vez que estes são altamente interpretáveis pelo ser humano. Esses algoritmos são capazes de analisar grandes volumes de dados históricos e identificar padrões complexos que podem ser utilizados para fazer previsões precisas sobre o comportamento futuro dos clientes em relação ao pagamento das suas obrigações financeiras.

Além disso, esses algoritmos são capazes de lidar com variáveis categóricas e numéricas, bem como com dados faltantes ou inconsistentes, o que os torna particularmente adequados para lidar com os desafios inerentes à previsão do *score* de crédito (SILVA, 2022). Esses algoritmos possuem algumas vantagens importantes nesse contexto específico. Primeiramente, eles são altamente interpretáveis, o que significa que é possível entender como as decisões são tomadas pelo modelo. Ainda, esses algoritmos são capazes de lidar com dados não lineares, o que é especialmente relevante na análise de crédito, uma vez que as relações entre as variáveis podem ser complexas e não lineares (SILVA, 2022). Segundo Coelho, Amorim e Camargos (2021), algoritmos de aprendizado de máquina, como o *Decision tree* 2.4.1, *Random Forest* 2.4.2 e *XGBoost* 2.4.3, os quais serão utilizados neste estudo, são amplamente empregados para resolver esses desafios de previsão de score de crédito.

Com base no exposto, o objetivo deste trabalho de conclusão de curso consiste em realizar uma análise experimental do uso dos algoritmos mencionados anteriormente, para previsão de score de crédito. Para isso, são considerados fatores como desempenho, vantagens e desvantagens de cada algoritmo, ao apresentar dados que possibilitam identificar se uma pessoa é ou não um bom pagador.

Para tanto, será utilizado o conjunto de dados público obtido da plataforma *Kaggle*, chamado *Credit Score Classification (Clean Data)*. Para realizar essa análise, é possível modelar o problema de duas formas, como tarefa de regressão ou classificação. Cada tipo de modelagem possui suas particularidades e se adéquam melhor a determinados cenários. Neste trabalho, foi definido modelar, como um problema de classificação.Assim, os mutuários foram classificados como: Bom, médio ou mal pagador. Tal decisão foi tomada, pois modelar o problema como classificação, permite identificar padrões e relações entre as classes à qual o mutuário pertence.

# 2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

Neste capítulo, serão abordados tópicos relacionados ao Sistema Financeiro Nacional, instituições financeiras cooperativas, crédito e risco, bem como conceitos de aprendizado de máquina para previsão de score de crédito, utilizando os algoritmos (*Decision Tree*, *Random Forest* e *XGBoost*). Também são apresentadas informações relacionadas ao processo de seleção de atributos, escolha dos melhores hiperparâmetros utilizando o método *GridSearchCV* da biblioteca *Scikit-learn*, avaliação de modelos preditivos, assim como métricas de avaliação dos modelos, como acurácia, *F1-Score* e matriz de confusão.

## 2.1 Sistema Financeiro Nacional

O Sistema Financeiro Nacional é regido pela Lei 4.595/1964, que criou o Conselho Monetário Nacional como órgão máximo do sistema, responsável por definir normas, adaptar o volume dos meios de pagamento do país e controlar o valor interno e externo da moeda, entre outras atribuições. (BRASIL, 1964) De acordo com Fortuna (1999), o Sistema Financeiro Nacional é formado por "um conjunto de instituições que se dedicam ao trabalho de propiciar condições satisfatórias para a manutenção de um fluxo de recursos entre poupadores e investidores". O Banco Central do Brasil é o órgão executivo central do sistema financeiro, responsável por fiscalizar o cumprimento das normas expedidas pelo Conselho Monetário Nacional e por atividades como emissão de papel-moeda, recebimento de compulsório dos bancos comerciais e realização de operações de redesconto e empréstimo às instituições financeiras.

Assim como os bancos comerciais, os bancos cooperativos são constituídos como sociedades anônimas fechadas, mas possuem a participação exclusiva de cooperativas de crédito singulares, centrais, federações ou confederações de cooperativas de crédito. A atuação dos bancos cooperativos é restrita aos estados onde estão situadas as sedes das pessoas jurídicas controladoras, ou seja, as centrais de cooperativas de crédito. No Brasil, existem grandes sistemas cooperativos de crédito, entre os quais se destacam: SICOOB, CRESOL e SICRED.

## 2.2 Instituição Financeira Cooperativa

O cooperativismo é um importante mecanismo de organização econômica da sociedade, que valoriza a resolução de problemas coletivos por meio da união, da ajuda mútua e da integração entre as pessoas. Um dos princípios fundamentais do cooperativismo é a busca pela correção de desigualdades e injustiças sociais, através da distribuição equitativa e harmoniosa de bens e valores pertencentes ao patrimônio da cooperativa,

conforme descrito pelo (OCERGS-SESCOOP/RS, 2020).

> As cooperativas de crédito são reconhecidas pelos seus órgãos regulamentadores como uma instituição financeira, de caráter não bancário. Nesta condição, as cooperativas atuam no mesmo mercado financeiro que os bancos comerciais, porém preservam a sua natureza cooperativa. (LOPES, 2021, p.20)

De acordo com Alves Sérgio Darcy e Soares (2006), o cooperativismo tem um papel importante na política de desenvolvimento nacional, uma vez que contribui para a expansão das pequenas e médias empresas, busca fortalecer as novas empresas e ajudar no crescimento das já existentes. A cooperação mútua entre as pessoas, comunidades e povos tem sido historicamente uma das principais forças motrizes por trás das grandes realizações. Salienta-se que a união em grupos para proteger-se de perigos, enfrentar ataques de animais e buscar alimentos são características inerentes à natureza humana.

## 2.3 Crédito e Risco

Conforme a definição de PALMUTI Claudio Silva e PICCHIAI (2012 apud SCHRICKEL, 2000) , o crédito consiste em uma disposição de alguém em ceder temporariamente parte do seu patrimônio a terceiros, com a expectativa de que essa parcela retorne integralmente após um tempo determinado.

O patrimônio pode se referir tanto a dinheiro, no caso de empréstimos financeiros, quanto a bens e mercadorias em compras parceladas. Em um sentido mais amplo, o crédito é considerado um instrumento importante para o desenvolvimento econômico, pois visa fornecer recursos financeiros para agentes econômicos, como empresas, famílias e governos de um país ou de várias nações específicas, a fim de atender às suas necessidades de consumo e investimento.

De acordo com Lopes (2021, p.30), assim como em qualquer relação comercial, a concessão de crédito envolve um pacto moral entre as partes credora e devedora, constituindo-se em um elo no qual uma depende da outra para sua existência. No entanto, é inegável que existem riscos envolvidos em tais transações, sendo o risco maior para o credor, que abre mão de um recurso na expectativa de recuperá-lo no futuro, com um ganho adicional.

O emprego de modelos matemáticos e estatísticos, que podem ser construídos manualmente ou por algoritmos de aprendizado de máquina, em dados já existentes em bancos de dados, auxiliam no processo de classificação de bons e maus pagadores. Tais modelos podem apoiar instituições financeiras no processo de tomada de decisão na

concessão de crédito (ERIC *et al.*, 2013).Neste trabalho, realizou-se estudo em modelos construídos por algoritmos de aprendizado de máquina, cujos conceitos e algoritmos utilizados são apresentados na seção a seguir.

## 2.4 Conceitos de Aprendizado de Máquina

Nesta seção, são apresentados alguns conceitos relacionados ao aprendizado de máquinas, mais especificamente relacionados ao uso de três algoritmos, Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*) e *XGBoost*.

Com a evolução dos computadores e os avanços na área de aprendizado de máquina, o que antes se resumia a programar um computador para executar tarefas humanas de forma automatizada, hoje, com o advento da inteligência artificial, as máquinas são capazes de aprender a partir de exemplos. Segundo MITCHELL (1999), o aprendizado de máquina é uma área de pesquisa em Inteligência Artificial focada no desenvolvimento de algoritmos que buscam aprimorar a execução de uma determinada tarefa por meio da análise de observações, utilizando uma métrica de desempenho como referência.

De acordo com Ris-Ala (2023), há três categorias centrais, no âmbito de aprendizado de máquina:

- aprendizado supervisionada: máquina aprende, por meio de um conjunto de dados rotulados, previamente preparados por um humano, contendo pares de entradas e suas respectivas saídas desejadas.
- aprendizado não supervisionada: máquina aprende, por meio de um conjunto de dados não rotulado, ou seja, os dados não foram previamente supervisionados.
- aprendizado por reforço: a máquina não possui um conjunto de dados pré-existente, ela precisa interagir com o ambiente, para coletar os dados.

A categoria de aprendizado utilizada neste trabalho é a de aprendizado supervisionado.

O aprendizado de máquina tem se mostrado cada vez mais importante na área financeira, especialmente no contexto da previsão de score de crédito. Isso se deve ao fato de que o processo de tomada de decisão em relação à concessão de crédito é complexo e envolve a análise de uma grande quantidade de dados. O aprendizado de máquina permite automatizar esse processo, tornando-o mais eficiente e preciso. Além disso, o uso de algoritmos de aprendizado de máquina na previsão de score de crédito possibilita a identificação de padrões e relações não triviais nos dados, o que pode levar a melhores decisões em relação à concessão ou não do crédito (ROMANI, 2017).

O aprendizado de máquina oferece soluções para uma ampla gama de problemas, incluindo regressão, agrupamento e classificação. No contexto deste trabalho, o objetivo é prever a classe à qual um mutuário pertence. Para isso, será empregado o processo de aprendizagem de máquina supervisionado, uma vez que o conjunto de dados utilizado foi previamente preparado. Isso significa que cada exemplo do conjunto de dados possui rótulos ou classes associadas, o que permite que o modelo aprenda a fazer previsões com base nesses exemplos.

### 2.4.1 Árvores de Decisão

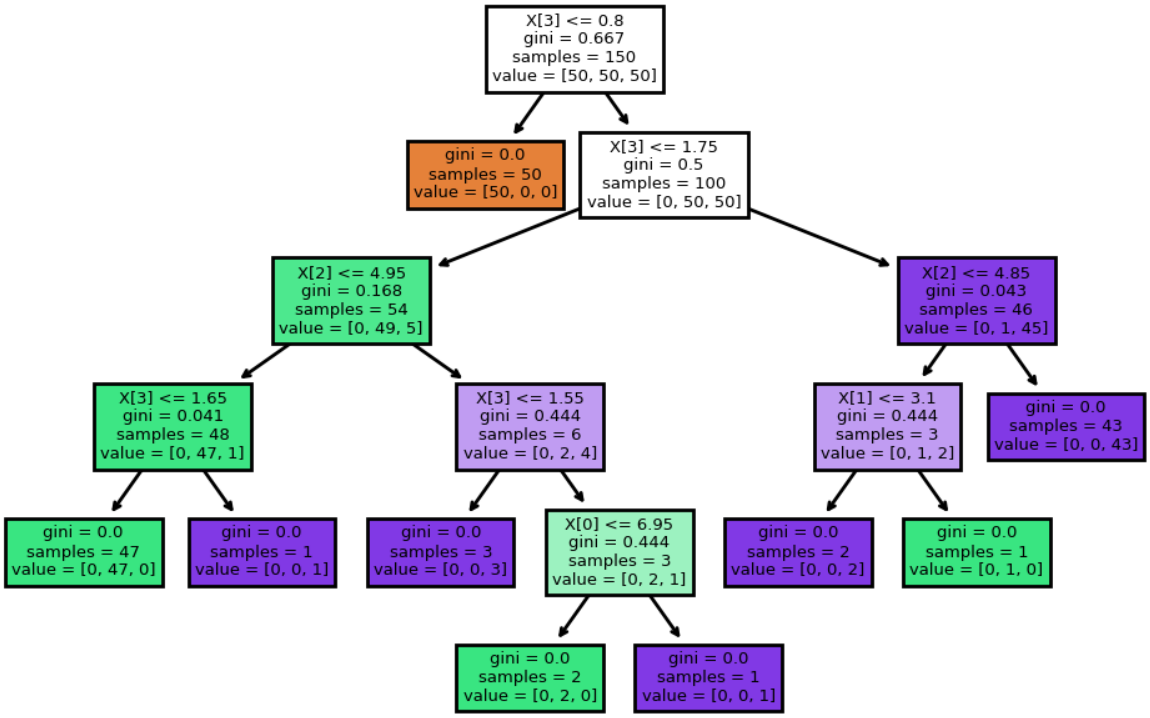
As árvores de decisão, são algoritmos de aprendizado supervisionado, os quais são utilizados para classificação e regressão. Tais algoritmos, visam criar um modelo capaz de efetuar previsões sobre o valor de uma variável alvo, aprender regras de decisão, obtidas a partir das características dos dados. (LEARN, 2023)

A seleção adequada das variáveis de entrada na construção da árvore de decisão é de suma importância para garantir a precisão das previsões do score de crédito. Variáveis irrelevantes ou redundantes podem introduzir ruído nos dados e comprometer a qualidade do modelo. Portanto, é necessário realizar uma análise cuidadosa das variáveis disponíveis e selecionar aquelas que possuem maior poder preditivo em relação ao score de crédito (GNOATTO, 2023).

Uma árvore de decisão é composta por nós que se organizam em uma estrutura de árvore enraizada. Essa estrutura inclui um nó raiz, que não tem arestas de entrada, seguido por outros nós conhecidos como nós internos ou nós de teste, que têm arestas de entrada e saída.(DAHAN; COHEN S. E ROKACH; MAIMON, 2014)

A Figura 1, apresenta a estrutura de uma árvore de decisão, utilizada para classificar espécies de flores, tendo como base um conjunto de dados amplamente reconhecido, denominado Iris.

Figura 1 – Árvore de decisão para classificação do conjunto de dados Iris




Fonte: (SCIKIT, 2021)

[1] De forma geral, os algoritmos de árvores de decisão usam a estratégia de dividir e conquistar. Dessa forma, o algoritmo divide o problema em subproblemas menores e mais simples, construindo uma árvore de decisão com base em regras lógicas que determinam como as decisões devem ser tomadas em cada nó da árvore, como pode ser visto no exemplo da Figura 1. O processo se inicia pelo nó raiz (primeiro nó da árvore), a partir dele são criadas as divisões denominadas nós, que na Figura 1 são representados por caixas. Essas caixas, contêm informações de cada teste, como número de amostras de treino do nó (*Samples*), o nível de impureza do nó (*Gini*) e quantos elementos de cada classe o nó em questão contém (*Value*).

Uma das principais vantagens dos algoritmos baseados em árvore de decisão em relação aos outros métodos de previsão do score de crédito é sua capacidade interpretativa. Ao contrário dos modelos mais complexos, como redes neurais, as árvores de decisão permitem uma interpretação direta das regras utilizadas na tomada das decisões. Isso significa que é possível entender quais variáveis são mais relevantes na determinação do score de crédito e como elas influenciam no resultado final (RODRIGUES, 2021).

[1] Uma árvore de decisão usa a estratégia dividir para conquistar para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia.(CARVALHO, 2011)

### 2.4.2 Randon Forest

O *Random Forest*, ou floresta aleatória, é um algoritmo de aprendizado de máquina supervisionado e amplamente utilizado no campo do aprendizado de máquina e da ciência de dados. Assim como o algoritmo de árvore de decisão, ele se destaca como uma abordagem eficaz para tarefas de classificação e regressão, graças à sua capacidade de lidar com uma variedade de problemas complexos:

> Em uma Floresta Aleatória, as características são selecionadas aleatoriamente em cada divisão de decisão. A correlação entre as árvores é reduzida ao selecionar aleatoriamente as características, o que melhora o poder de previsão e resulta em maior eficiência.(ALI; KHAN REHANULLAH E AHMAD; MAQSOOD, 2012)

O funcionamento do *Random Forest* consiste em três etapas principais: amostragem aleatória dos dados, construção das árvores de decisão e combinação das decisões. Na primeira etapa é realizado um processo de amostragem aleatória dos dados originais para criar diferentes subconjuntos de treinamento. Em seguida, cada subconjunto é utilizado para construir uma árvore de decisão independente. Por fim, as decisões de todas as árvores são combinadas por meio da votação majoritária, ou seja, a classe ou valor mais frequente é escolhido como resultado final(SANTOS, 2021).

O Random Forest apresenta diversas vantagens em relação às árvores de decisão tradicionais. Uma delas é sua maior robustez contra *overfitting*, pois a combinação das decisões de várias árvores reduz a tendência ao ajuste excessivo aos dados de treinamento. Além disso, o *Random Forest* é capaz de lidar tanto com dados categóricos quanto numéricos sem a necessidade de transformações prévias. Essa flexibilidade é especialmente importante na previsão de score de crédito, onde diferentes tipos de dados podem estar envolvidos. Por fim, o *Random Forest* também apresenta melhor desempenho em conjuntos de dados grandes, pois a construção paralela das árvores permite um processamento mais rápido (SOUZA D. H. M. DE E BORDIN JR, 2023).

### 2.4.3 *XGBoost*

O algoritmo *XGBoost* tem uma importância significativa na área de aprendizado de máquina para previsão de score de crédito. Sua capacidade de lidar com dados desbalanceados e sua eficiência computacional o tornam uma escolha popular para essa tarefa. O *XGBoost* é capaz de lidar com dados desbalanceados através do uso de técnicas como a ponderação das instâncias durante o treinamento e a definição de um limite mínimo para a criação de novos nós nas árvores. Além disso, sua eficiência computacional é garantida pelo

uso de estruturas de dados otimizadas, como matrizes esparsas, que reduzem o consumo de memória e aceleram o processo de treinamento (CASTRO, 2019).

A implementação do algoritmo *XGBoost* em um modelo de previsão de score de crédito requer alguns passos específicos. Primeiramente, é necessário realizar a preparação dos dados, que envolve a limpeza, transformação e seleção das variáveis relevantes. Em seguida, é preciso definir os hiperparâmetros do modelo, como o número máximo de árvores e a taxa de aprendizado. Após isso, o modelo pode ser treinado utilizando os dados disponíveis. Durante o treinamento, é importante monitorar as métricas de desempenho, como a acurácia, para ajustar os hiperparâmetros e evitar *overfitting* (SANTOS, 2021).

Ao comparar o *XGBoost* com outros algoritmos baseados em árvore de decisão, como o *Random Forest*, é possível observar diversos benefícios em seu uso. O *XGBoost* possui uma maior flexibilidade na definição dos hiperparâmetros, o que permite um ajuste mais refinado do modelo. Além disso, o *XGBoost* é conhecido por sua capacidade de lidar com dados desbalanceados de forma mais eficiente, resultando em modelos mais precisos. Outra vantagem do *XGBoost* é sua eficiência computacional, que permite treinar modelos em grandes conjuntos de dados em um tempo razoável(ROMANI, 2017).

Estudos experimentais comprovam a eficácia do algoritmo *XGBoost* na previsão de score de crédito. Esses estudos mostram resultados superiores em termos de acurácia e tempo de processamento quando comparados a outros algoritmos baseados em árvore de decisão. (JÚNIOR, 2018).

Apesar das vantagens mencionadas, o algoritmo *XGBoost* também apresenta algumas limitações. Uma delas é a necessidade de ajuste fino dos hiperparâmetros para obter os melhores resultados. Isso pode ser uma tarefa complexa e demorada, exigindo conhecimento especializado e experimentação cuidadosa. Além disso, se não for utilizado corretamente, o *XGBoost* pode levar ao *overfitting*, ou seja, um modelo que se ajusta muito bem aos dados utilizados no treinamento, mas tem dificuldade em generalizar para novos dados (SANTOS, 2021).

## 2.5 Seleção de Atributos em *Machine Learning*

A seleção de atributos desempenha um papel fundamental nos algoritmos de aprendizado de máquina baseados em árvore de decisão para previsão de score de crédito. A escolha adequada dos atributos pode melhorar significativamente a precisão e a eficiência do modelo, além de reduzir o risco de *overfitting*. Isso ocorre porque nem todos os atributos são igualmente relevantes para a tarefa de previsão do score de crédito, e incluir atributos irrelevantes pode introduzir ruído nos dados e dificultar a generalização do modelo para novos exemplos(NETO; SILVA, 2021).

Existem diferentes métodos de seleção de atributos utilizados em algoritmos de

aprendizado de máquina baseados em árvore de decisão. Entre eles, destacam-se os métodos *Wrapper*, *Filter* e *Embedded*. O método *Wrapper* utiliza um algoritmo de aprendizado para avaliar a relevância dos atributos e selecionar os mais importantes. Ele realiza uma busca exaustiva ou heurística no espaço dos subconjuntos de atributos, avaliando o desempenho do modelo com cada subconjunto através da validação cruzada ou outra técnica similar (SILVA, 2022).

O método *Filter*, por sua vez, utiliza medidas estatísticas para avaliar a relevância dos atributos e selecionar os mais importantes. Essas medidas podem ser baseadas na correlação entre os atributos e o *target*, como o coeficiente de correlação ou o teste qui-quadrado, ou podem ser baseadas na informação mútua entre os atributos e o *target*, como o ganho de informação ou a razão entre ganhos(SILVEIRA, 2022).

Já o método *Embedded* incorpora a seleção de atributos diretamente no processo de treinamento do modelo. Ele utiliza algoritmos de aprendizado que possuem mecanismos internos para selecionar os atributos mais relevantes durante o treinamento. Exemplos de algoritmos *embedded* incluem o *Random Forest*, que utiliza a importância dos atributos calculada durante a construção das árvores, e o *Gradient Boosting*, que utiliza a derivada do erro em relação aos atributos para selecioná-los (GNOATTO, 2023).

Também é possível fazer a seleção de atributos a partir da correlação dos atributos presentes no conjunto de dados, utilizando o coeficiente de correlação de *Pearson*, que mede a força da relação linear entre duas variáveis, sendo que se houver uma relação linear forte, o coeficiente de correção será mais próximo de 1 ou -1. Já se o coeficiente de correlação for zero, indica que não há relação linear(DEPREZ; ROBINSON, 2023). Segundo Saini, Lata e Sinha (2021), o coeficiente de correlação de *Pearson*, é amplamente utilizado como métrica para seleção de atributos, uma vez que essa correlação é calculada entre cada atributo e a saída alvo. Nesse sentido, os atributos que tiverem maior coeficiente de correlação de *Pearson* com o resultado alvo, são selecionados para treinamento e teste do modelo.

A seleção de atributos em algoritmos de aprendizado de máquina, baseados em árvore de decisão, apresenta vantagens e desvantagens. Entre as vantagens, destacam-se a redução da dimensionalidade dos dados, o que pode melhorar a eficiência computacional e reduzir o risco de *overfitting*, além da melhoria na interpretabilidade do modelo resultante. Por outro lado, a seleção incorreta ou inadequada dos atributos pode levar à perda de informações relevantes e à diminuição da precisão do modelo (COELHO; AMORIM; CAMARGOS, 2021).

Ao realizar a seleção de atributos em algoritmos de aprendizado de máquina, baseados em árvore de decisão para previsão de score de crédito, é importante considerar algumas questões práticas. O tempo computacional é uma delas, pois alguns métodos podem ser computacionalmente custosos, especialmente quando aplicados a conjuntos de

dados grandes. Além disso, é necessário avaliar a interpretabilidade do modelo resultante, já que nem sempre é possível compreender completamente as regras e padrões utilizados pelo algoritmo para tomar suas decisões. Portanto, é fundamental encontrar um equilíbrio entre a complexidade do modelo e sua capacidade de explicação (SANTOS, 2021).

## 2.6 Avaliação de Modelos Preditivos - validação cruzada

A avaliação de modelos preditivos desempenha um papel fundamental na previsão de score de crédito. A capacidade de avaliar a eficácia dos modelos é crucial para garantir a precisão e confiabilidade das previsões. Além disso, a avaliação adequada permite identificar possíveis problemas nos modelos e realizar ajustes necessários para melhorar seu desempenho (COELHO; AMORIM; CAMARGOS, 2021).

Dentre os diferentes métodos de avaliação de modelos preditivos, destaca-se a validação cruzada. Esse método consiste em dividir o conjunto de dados em subconjuntos de treinamento e teste, o que permite que o modelo seja treinado em uma parte dos dados e testado em outra. Salienta-se que a validação cruzada é especialmente útil quando se tem um conjunto limitado de dados disponíveis, pois permite uma melhor estimativa do desempenho do modelo (SANTOS, 2021).

A utilização de algoritmos baseados em árvore de decisão apresenta diversas vantagens na previsão de score de crédito. Esses algoritmos são capazes de lidar com variáveis categóricas e numéricas, além de serem interpretáveis e facilmente compreensíveis pelos usuários. Além disso, as árvores de decisão são robustas a *outliers* e podem lidar com grandes conjuntos de dados sem comprometer o desempenho (ROMANI, 2017)

No entanto, a avaliação de modelos preditivos para previsão de score de crédito enfrenta alguns desafios. Um dos principais desafios é a presença de desbalanceamento nos dados, ou seja, quando há uma grande disparidade entre as classes positiva (bons pagadores) e negativa (maus pagadores). Isso pode levar a uma tendência do modelo em classificar a maioria dos casos como positivos, comprometendo a precisão das previsões (NETO; SILVA, 2021).

Diversas métricas são utilizadas na avaliação dos modelos preditivos para previsão de score de crédito. A acurácia é uma medida comum que indica a proporção de casos corretamente classificados pelo modelo. Além disso, a área sob a curva ROC (*Receiver Operating Characteristic*) é amplamente utilizada para avaliar o desempenho do modelo em diferentes pontos de corte. Essa métrica fornece uma medida da capacidade do modelo em distinguir entre as classes positiva e negativa (RODRIGUES, 2021).

## 2.7 Escolha de Valores de Hiperparâmetros dos Algoritmos (*GridSearchCV*)

A escolha dos valores de parâmetros dos algoritmos de aprendizado de máquina baseados em árvore de decisão é de extrema importância para a previsão de score de crédito. Esses parâmetros determinam como o modelo será construído e como ele se ajustará aos dados disponíveis. Uma escolha adequada dos valores de parâmetros pode levar a um modelo com melhor desempenho e maior capacidade de generalização (RODRIGUES, 2021).

No entanto, a escolha dos valores de parâmetros enfrenta diversos desafios. Primeiramente, existe uma infinidade de combinações possíveis para os valores dos parâmetros, o que torna inviável testar todas elas manualmente. Além disso, diferentes valores podem afetar o desempenho do modelo de maneiras não intuitivas, o que torna difícil prever qual combinação será a mais adequada. Logo, a escolha incorreta dos valores pode levar a um modelo com baixa precisão na previsão do score de crédito e comprometer sua utilidade prática (CASTRO, 2019).

Uma técnica amplamente utilizada para encontrar os melhores valores de parâmetros é o *grid search cross-validation (GridSearchCV)*. Essa técnica consiste em definir uma grade com possíveis valores para cada parâmetro e testar todas as combinações possíveis por meio da validação cruzada. O *GridSearchCV* avalia o desempenho do modelo para cada combinação e retorna aquela que obteve os melhores resultados (CASTRO, 2019).

Durante o processo de busca por hiperparâmetros, é fundamental definir uma métrica adequada para avaliar o desempenho dos modelos. Essa métrica deve ser relevante para o problema em questão e refletir a qualidade das previsões do modelo. Métricas comumente utilizadas incluem acurácia, precisão, *recall* e *F1-Score*. A escolha da métrica correta é essencial para garantir que o modelo seja avaliado de forma adequada e que os valores de parâmetros sejam escolhidos com base em critérios relevantes (SOUZA D. H. M. DE E BORDIN JR, 2023).

A realização de validação cruzada durante o *grid search* é fundamental para evitar *overfitting* e garantir que os resultados sejam generalizáveis. A validação cruzada divide o conjunto de dados em $k$ partes iguais, chamadas *folds*, e realiza $k$ iterações do treinamento e teste do modelo. Isso permite que todas as observações sejam utilizadas tanto para treinamento quanto para teste, evitando vieses na avaliação do desempenho do modelo (JÚNIOR, 2018).

Existem diferentes estratégias de busca por hiperparâmetros, sendo as mais comuns a busca em grade (*grid search*) e a busca aleatória (*random search*). A busca em grade consiste em definir uma grade com possíveis valores para cada parâmetro e testar todas as combinações possíveis. Já a busca aleatória seleciona aleatoriamente um conjunto de valores para cada parâmetro e testa várias combinações diferentes. A vantagem da busca

em grade é que ela garante a exploração completa do espaço de parâmetros, enquanto a busca aleatória pode ser mais eficiente computacionalmente (SILVEIRA, 2022).

## 2.8 Métricas de Avaliação de Modelos Preditivos para Classificação

As métricas de avaliação desempenham um papel fundamental na área de aprendizado de máquina, especialmente quando se trata de previsão de score de crédito. Essas métricas permitem medir o desempenho dos modelos em relação aos dados de teste e fornecem informações valiosas sobre a qualidade das previsões realizadas. Dentre as métricas mais comumente utilizadas, destaca-se a acurácia, que mede a taxa de acertos do modelo em relação ao total de instâncias. A acurácia é calculada dividindo o número de previsões corretas pelo número total de instâncias e é uma medida simples e intuitiva para avaliar o desempenho geral do modelo (CASTRO, 2019).

Outra métrica importante é o *F1-Score*, que combina as métricas de precisão e *recall* para fornecer uma medida mais equilibrada do desempenho do modelo. A precisão representa a proporção de instâncias classificadas corretamente como positivas em relação ao total de instâncias classificadas como positivas, enquanto o *recall* representa a proporção de instâncias classificadas corretamente como positivas em relação ao total de instâncias verdadeiramente positivas. O *F1-Score* é calculado como a média harmônica dessas duas medidas e é especialmente útil quando há um desequilíbrio entre as classes(SANTOS, 2021).

A matriz de confusão é uma ferramenta visual que permite analisar os resultados do modelo, mostrando as classificações corretas e incorretas. Ela organiza as previsões em quatro categorias: verdadeiros positivos (instâncias corretamente classificadas como positivas), falsos positivos (instâncias incorretamente classificadas como positivas), verdadeiros negativos (instâncias corretamente classificadas como negativas) e falsos negativos (instâncias incorretamente classificadas como negativas). A matriz de confusão fornece uma visão geral do desempenho do modelo e permite identificar possíveis erros de classificação (CASTRO, 2019)..

Segundo Carvalho (2011 apud MONARD M. E BARANAUSKAS, 2003), a partir da matriz de confusão, diversas medidas de desempenho podem ser obtidas. Entre elas encontra-se:

- Taxa de erro na classe positiva

$$error_{+}(\hat{f}) = \frac{FN}{VP + FN}$$

- Taxa de erro na classe negativa

$$error_{-}(\hat{f}) = \frac{FP}{FP + VN}$$

- Taxa de erro total

$$error(\hat{f}) = \frac{FP + FN}{n}$$

- Taxa de acerto ou acurácia total

$$ac(\hat{f}) = \frac{VP + VN}{n}$$

- Precisão

$$prec(\hat{f}) = \frac{VP}{VP + FP}$$

- Sensibilidade ou revocação

$$sens(\hat{f}) = rev(\hat{f}) TVP(\hat{f}) = \frac{VP}{VP + FN}$$

- Especificidade

$$esp(\hat{f}) = \frac{VN}{VN + FP} = 1 - TFP(\hat{f})$$

A utilização de diferentes métricas de avaliação em conjunto é essencial, pois cada uma delas fornece informações valiosas sobre o desempenho do modelo em diferentes aspectos. Enquanto a acurácia mede o desempenho geral do modelo, o *F1-Score* considera tanto a precisão quanto o *recall*, proporcionando uma visão mais equilibrada. Além disso, a matriz de confusão permite uma análise mais detalhada das previsões realizadas pelo modelo (GNOATTO, 2023)

Neste trabalho, serão utilizadas as métricas de acurácia, *F1-Score* e erro total, calculadas a partir da validação cruzada utilizando os 3 algoritmos de aprendizado descritos neste capítulo. No capítulo a seguir, são descritas todas as ferramentas utilizadas para a condução da análise experimental, que implementam os conceitos ora apresentados.

## 2.9 Revisão da Literatura

Para fazer a revisão da literatura, buscou-se algumas referências bibliográficas em conferências e periódicos. Assim, são apresentadas as mais importantes.

### 2.9.1 Análise dos determinantes no grau de evidenciação do risco de crédito em centrais de cooperativas de crédito

De acordo com Brandalize, Flach e Sallaberry (2022), o segmento de cooperativas teve um aumento significativo na última década no Brasil e, assim, como em qualquer outra instituição financeira, todos os tipos de risco as ameaçam, em particular o risco de crédito, que tem especial relevância por se tratar de recursos necessários, uma vez que estas precisam receber os créditos concedidos a terceiros para cumprir com suas obrigações. Para tentar mitigar o risco de não recebimento, alguns procedimentos preliminares são adotados, como comparar as características e padrões de comportamento de um solicitando com as características de outros solicitantes de crédito e assim, verificar a proximidade das características de um indivíduo com um grupos de bons pagadores ou de inadimplentes. Dessa forma, é possível estimar uma probabilidade do não pagamento do crédito obtido.

### 2.9.2 Algoritmos e score de crédito

De acordo com Peck (2023), o uso da tecnologia está transformando os processos de tomada de decisão no âmbito financeiro, principalmente no que diz respeito à concessão de crédito, o que torna esse processo mais eficiente e garante uma melhor assertividade na classificação de eventuais inadimplentes, uma vez que, de acordo com dados divulgados pela Confederação Nacional do Comércio de Bens, Serviços e Turismo, sete em cada dez famílias brasileiras (71,4%) estão endividadas no ano de 2021. Esse mesmo problema foi detectado no ano de 2020, onde 19% dos consumidores tiveram uma queda significativa no score de crédito. Através da aplicação de algoritmos de aprendizado de máquina e inteligência artificial, são analisados diversos fatores e características do tomador, tais como: comprometimento da renda; pontualidade no pagamento das contas; idade; histórico de consumos, entre outros. Com base nessa análise se chega a uma pontuação de crédito, que pode aumentar ou diminuir as vantagens na obtenção de acesso ao crédito, em casos de compras parceladas ou possíveis financiamentos/empréstimos. No entanto, apesar do uso dos algoritmos trazerem diversas vantagens, é imprescindível que estes utilizem métodos de inteligência artificial, que possam ser explicados, permitindo ao ser humano, entender por que o algoritmo chegou à determinada conclusão. Recentemente a Comissão Europeia, apresentou uma proposta, na qual prevê que os sistemas de risco elevado (como os que avaliam a classificação do score), sejam desenvolvidos de modo a assegurar transparência de seu funcionamento.

### 2.9.3 Analisando métodos de *machine learning* e avaliação de risco de crédito

Segundo Coelho, Amorim e Camargos (2021), as instabilidades no ambiente econômico, levam tanto empresas, quanto pessoas físicas a enfrentarem dificuldades financeiras, resultando em um desestímulo das vendas via crédito e um aumento no risco de concessão de crédito no mercado financeiro. Portanto, é fundamental que o risco de inadimplência seja avaliado, não somente por critérios convencionais, mas sim por meio de outros métodos, como o uso de algoritmos de aprendizado de máquina e inteligência artificial. No entanto, para que se possa garantir uma maior assertividade nos resultados obtidos a partir de algoritmos de aprendizado de máquina, o processo de comparação de desempenho entre eles é algo fundamental.

# 3 MATERIAIS E MÉTODOS

A análise experimental de algoritmos de aprendizado de máquina tem se tornado cada vez mais importante em diversas áreas, inclusive em finanças. A previsão de score de crédito é uma das aplicações desses algoritmos, pois permite avaliar a probabilidade de um indivíduo ou empresa honrar seus compromissos financeiros. Nesse contexto, os algoritmos de aprendizado que constroem modelos baseados em árvores de decisão, se destacam como uma técnica promissora para a previsão de score de crédito, graças à sua capacidade de lidar com variáveis categóricas e numéricas e de gerar regras interpretáveis para a tomada de decisão.

## 3.1 Ferramentas Computacionais

### 3.1.1 Pandas

A biblioteca *Pandas* desempenha um papel fundamental na análise e manipulação de dados dentro do campo da ciência de dados e análise exploratória. Criada em *Python*, a *Pandas* oferece uma ampla gama de ferramentas e estruturas de dados que permitem aos cientistas de dados importar, limpar, transformar e visualizar dados de maneira eficiente e eficaz. Além disso, ela dispõe de suporte para criação de *DataFrame* e Séries. O *DataFrame* representa os dados de uma planilha, enquanto as Séries correspondem a uma única coluna do *DataFrame* (CHEN, 2018).

### 3.1.2 Numpy

A biblioteca *NumPy* foi criada com o objetivo de fornecer principalmente suporte abrangente para a componente numérica, especialmente no âmbito científico, na linguagem *Python*. Criado por *Travis Oliphant*, no ano de 2005, para ser o sucessor do pacote *Numeric* e com raízes do módulo *SciPy*, desde de sua criação foi muito bem aceita por profissionais que trabalham nas áreas de matemática, ciências e engenharia.(CHIN L. E DUTTA, 2016).

### 3.1.3 Scikit-learn

Sendo um projeto de código aberto, o que significa que sua utilização e distribuição são livres, o *scikit-learn* permanece em constante evolução e aperfeiçoamento, contando com uma comunidade de usuários ativa. A ferramenta dispõe de diversos algoritmos voltados para aprendizado de máquina, sendo utilizada tanto no meio acadêmico como industrial. (MÜLLER A.C. E GUIDO, 2016). A principal finalidade desta biblioteca reside na sua capacidade de (re)utilização de algoritmos de aprendizado aproveitando de tantos outros recursos disponíveis na linguagem *Python*.

### 3.1.4 Matplotlib

Criada por John Hunter em 2003, a biblioteca *Matplotlib*, tem como objetivo, auxiliar no processo de criação de gráficos 2D na linguagem *Python*, sendo muito utilizada por cientistas e engenheiros.(HUNTER, 2007).Ela dispõe de uma ampla gama de gráficos, como gráfico de linha, barras, pizza entre outros, que podem ser facilmente ajustados.

### 3.1.5 Seaborn

Outra biblioteca muito útil para visualizar dados na linguagem *Python* é a *Seaborn*. Segundo Guilhon *et al.* (2022): "esta biblioteca tem como base o *Matplotlib* e fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e profissionais. Por ser baseada na biblioteca *Matplotlib*, possibilita gerar diversos tipos de gráficos, entre eles o mapa de calor, o qual é utilizado neste trabalho, para apresentar a correlação entre os atributos do conjunto de dados, ora utilizado.

### 3.1.6 Imblearn

A *imbalanced-learn*, também conhecida como *imblearn*, é uma biblioteca de código livre, essencial para lidar com desequilíbrios em conjuntos de dados e em tarefas de aprendizado de máquina. De acordo com Aridas (2017) , a biblioteca oferece diferentes estratégias para resolver problemas de desbalanceamento em conjunto de dados, como:

- *Undersampling*
- *Oversampling*

De acordo com He e Ma (2013), o método *Undersampling*, visa descartar os dados de classe majoritária, até que o número de classes se iguale a classe minoritária. Já o método *Oversampling*, tem por finalidade, gerar novos dados, até se igualar ao total da classe majoritária. Neste trabalho, utilizou-se o método *Oversampling*.

### 3.1.7 Google Colab

O *Google Colab* é uma plataforma em nuvem que oferece serviços tanto pagos quanto gratuitos para a execução de códigos na linguagem *Python*. Por meio dessa ferramenta, é possível rodar algoritmos de baixa e alta complexidade que demandam ou não de um maior desempenho de hardware, uma vez que ela disponibiliza recursos que otimizam esse processo, incluindo o acesso a GPUs e TPUs. Isso resulta em uma execução muito mais rápida do código em comparação com uma máquina pessoal. Além disso, tal ferramenta possui uma interface intuitiva e de fácil utilização, tendo grande parte das bibliotecas para análise de dados e aprendizado de máquina já disponíveis, basta apenas instalar as classes das bibliotecas que se deseja utilizar.

## 3.2 Exploração e Seleção de *Dataset* para Análise de *Credit Score*

Para realização deste trabalho, como não foi possível ter acesso a um conjunto de dados real, foi necessário identificar possíveis *datasets*, que pudessem conter informações relevantes para o trabalho em questão. Isso envolveu uma busca nas principais plataformas que disponibilizam *datasets* para estudos, como *kaggle*, *Google Dataset Search* e *UCI Machine Learning Repository*, utilizando a palavra chave *Credit Score Classification*.

Depois de avaliar as várias opções, com base na abrangência das informações e número de variáveis presentes no *Dataset*, como renda salarial, número de pagamentos atrasados, empréstimos tomados, entre outras, que são conhecidas por influenciar o *credit score*, foi selecionado um conjunto de dados de crédito chamado *Credit Score Classification (Clean Data)*, disponível na plataforma *Kaggle*.

## 3.3 Análise Exploratória e Pré-processamento dos Dados

A análise exploratória de dados desempenha um papel fundamental ao utilizar algoritmos de aprendizado de máquina em um conjunto de dados. Esse processo permite identificar aspectos essenciais que irão facilitar a etapa de pré-processamento dos dados.

Para facilitar o processo de análise dos dados, a linguagem *Python* dispõe de uma vasta gama de recursos, ou seja, inúmeras bibliotecas, como *Pandas*, *NumPy*, *Matplotlib* e *Seaborn*, que tornam essa tarefa mais rápida e eficiente. Para importar o conjunto de dados selecionado e assim prosseguir com a análise, foi utilizada a biblioteca *Pandas*, descrita 3.1.1, a qual permite gerar um *DataFrame* a partir dos dados importados e, assim, obter diversas informações, como o número total de registros do conjunto de dados, o número de atributos, se há ou não valores nulos, entre outros.

A partir do *DataFrame* gerado com a importação do conjunto de dados, foi possível separar os atributos em dois grupos, sendo eles atributos do tipo numérico, compostos por números inteiros e flutuantes, e atributos do tipo categóricos, compostos por dados alfanuméricos.

Após uma análise dos atributos presentes no conjunto de dados e de suas características, foi realizada uma verificação para identificar possíveis registros duplicados. Para essa finalidade, utilizou-se uma função chamada *duplicated*, disponível na linguagem de programação *Python*. Ao aplicar essa função, confirmou-se que não existem registros duplicados.

Em seguida, foi realizada uma avaliação mais detalhada, na qual foi possível perceber que certos atributos, como *ID*, *CustomerID*, *TypeofLoan*, *SSN* e *Name*, têm um impacto limitado no processo de análise. Dessa forma, optou-se por removê-los do conjunto de dados. É importante observar que não foram encontrados valores faltantes no conjunto de dados, o que reforça a qualidade e a confiabilidade dos dados utilizados.

Além da remoção dos atributos citados anteriormente, foi efetuada a transformação dos atributos categóricos *Credit_Score*, *Credit_Mix* e *Payment_of_Min_Amount* para o tipo número, visto que os computadores trabalham melhor com números. Esta conversão de atributos categóricos para numéricos visa preparar os dados de maneira mais adequada para os algoritmos de aprendizado de máquina.

## 3.4 Escolha dos Melhores Atributos *(Features)*

As *features* são características ou atributos dos dados que estão presentes em conjunto de dados, estes podem ser do tipo quantitativo ou numéricos (Idade, Ano) ou categóricos (Nome, Endereço). O processo de seleção dos atributos é um passo fundamental, pois uma má seleção, pode afetar significativamente a precisão e a capacidade preditiva dos modelos de aprendizado de máquina.

> A extração das variáveis, *Feature Extraction* em português, é a fase onde o cientista de dados passa seu tempo escolhendo aquelas variáveis que entende serem importantes para o modelo de aprendizado de máquina. A etapa de seleção de variáveis, *Feature Selection* em português, é compreendida como sendo a remoção de variáveis que não consigam trazer informações úteis para os algoritmos de aprendizado de máquina. Geralmente, essa etapa é automatizada e remove variáveis que têm pouca variância, contêm muitos valores nulos ou não têm correlação com a variável que se deseja predizer(PELISON, 2018).

De acordo com Vasconcellos (2019), umas das formas mais simples para entender como os atributos de um conjunto de dados se relacionam, é a partir da matriz de correlação. Em tal matriz, cada linha e cada coluna representa um atributo do conjunto de dados que está sendo analisado. Assim, o valor da célula cuja linha e coluna seja correspondente a duas variáveis X e Y possui o valor da correlação entre X e Y. Quanto maior o valor absoluto da correlação maior é a dependência linear entre as duas variáveis.

Para seleção dos atributos mais relevantes, no presente trabalho, primeiramente foi obtida a correlação entre os atributos do conjunto de dados, utilizando a função *corr()*, presente na biblioteca *Pandas*. Em seguida, foi utilizada a função *heatmap*, da biblioteca *Seaborn*, para gerar o mapa de calor das correlações.

## 3.5 Validação Cruzada

Visando uma avaliação mais precisa e confiável dos modelos utilizados no presente estudo, optou-se por fazer o uso do método chamado *k-fold cross validation*, que realiza amostragem sem reposição do conjunto de dados em subconjuntos de treinamento e teste, de acordo com um número de k divisões definido pelo usuário.

O funcionamento do *k-fold cross-validation* ocorre conforme descrito a seguir:

> No método de validação *K-fold* as $N$ observações da amostra original $D$ são divididas em $k$ conjuntos distintos de observações, sendo eles $D_1$, $D_2$, ..., $D_k$, cada um de tamanho mk aproximadamente igual, tal que n = $\sum_{k=1}^{K}$ mk. A partir disso, a amostra de validação é composta pela partição Dk enquanto que a amostra de treino engloba as outras k -1 partições que não incluem a k-ésima partição, ou seja, o conjunto de treino é dado por $D_( - k) = D_1$, $D_2$, ..., $D_k - 1$, $D_k + 1$, ...., $D_K$. (VELOSO, 2022)

A escolha do valor ideal de $k$ na técnica *k-fold cross-validation* pode variar dependendo do tamanho do conjunto de dados, da quantidade de dados de treinamento disponíveis e da natureza do problema em questão. Salienta-se que não existe um valor único de $k$ que seja ideal para todos os cenários.

Segundo, Hastie, Tibshirani e Friedman (2009), o valores de K tipicamente utilizados são k=5 ou 10, onde a partir desse número de divisões, é possível supor que o modelo estime o erro esperado, uma vez que os conjuntos de treinamento em cada dobra são bastante diferentes do conjunto de treinamento original.

A etapa de execução da validação cruzada foi realizada mediante a divisão do conjunto de dados em cinco subconjuntos, aderindo o que é recomendado na literatura. Também foi usada a técnica chamada *SMOTE* para equilibrar os dados, e outra técnica chamada *GridSearchCV* para encontrar as melhores hiperparâmetros para os modelos utilizados. As duas técnicas mencionadas são abordadas com maiores detalhes nas seções 3.6 e 3.7

## 3.6 Balanceamento dos Dados

A existência de dados desbalanceados é um problema recorrente na área de classificação de dados, pois em um conjunto de dados real, o número de objetos varia em cada classe. Isso faz com que determinadas classes se sobressaiam sobre as demais (CARVALHO, 2011).

Para efetuar o balanceamento do conjunto de dados, foi utilizada a função *SMOTE*, com o método *Oversampling*, da biblioteca *imblearn*. Segundo Siahaan e Sianipar (2023), o método *imblearn . over_sampling*, é uma técnica utilizada para trabalhar conjunto de dados desbalanceados e visa criar novas amostras sintéticas para a classe minoritária, a fim de equilibrar o número de exemplos entre as classes.

## 3.7 Ajustes de Hiperparâmetros

Com o objetivo de aprimorar os resultados obtidos por meio dos algoritmos *Decison Tree*, *Random Forest* e *XGBoost*, são conduzidos, neste trabalho, testes automatizados com o intuito de explorar uma ampla gama de configurações de hiperparâmetros. Essa tarefa será executada através da aplicação da classe *GridSearchCV*, presente na biblioteca *Scikit-Learn*, que realiza uma busca sobre um dicionário de parâmetros passados como entrada, para encontrar a melhor combinação de hiperparâmetros para um modelo de aprendizado de máquina.

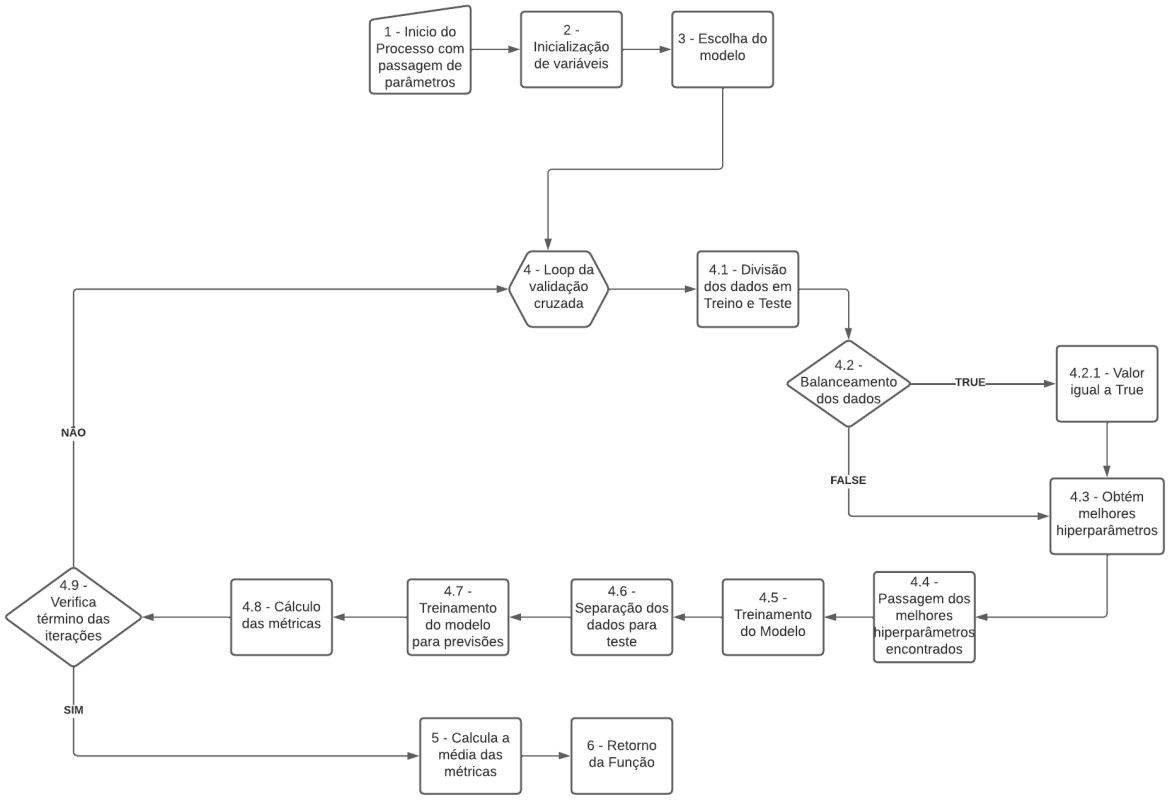
Neste processo de busca, para encontrar os melhores hiperparâmetros, é realizado a validação cruzada, que consiste em dividir os dados em partes para treinamento e validação, ajustar o modelo com diferentes combinações de hiperparâmetros e avaliar seu desempenho em dados de validação.(MOLIN S. E JEE, 2021)

Para utilizar o *GridSearchCV*, como está sendo trabalha, com 3 algoritmos de árvore de decisão, foi necessário criar dois dicionários com parâmetros distintos. O primeiro dicionário, empregado na busca pelo melhor hiperparâmetro para o algoritmo *Decision Tree*, é composto por um conjunto de valores que varia de 1 a 20, os quais foram atribuídos ao parâmetro *min_samples_leaf*, que representa o número mínimo de amostras requeridas para formar um nó folha. Já o segundo dicionário, utilizado para otimizar os hiperparâmetros dos algoritmos *Random Forest* e *XGBoost*, engloba uma série de valores na faixa de 50 a 1000. Esses valores são aplicados ao parâmetro *n_estimators*, que indica a quantidade de árvores presentes na floresta do modelo. A criação desses dois dicionários com parâmetros distintos, se deu pelo fato de que o parâmetro *n_estimators* não é aceito ao usar o *GridSearchCV* com o algoritmo *Decision Tree*. Além do dicionário criado e passado como parâmetro na função *GridSearchCV*, foram atribuídos valores aos parâmetros *verbose* e *n_jobs*, mantendo o restante dos parâmetros da função com o valor padrão da mesma.

## 3.8 Treinamento dos Modelos via Validação Cruzada

Após a análise e tratamento dos dados, iniciou-se o processo de treinamento dos modelos. Para isso, utilizou-se a técnica de validação cruzada. Ainda, a fim de executar essa tarefa de maneira eficiente e que permitisse efetuar o treinamento de todos os modelos, optou-se por criar uma função. A partir dessa, foi possível obter os valores das métricas (acurácia, *F1-score* e erro total). A Figura 2 apresentada a seguir mostra o fluxo de execução da função:

Figura 2 – Diagrama de fluxo da função de validação cruzada




Fonte: O AUTOR (2023)

Como pode ser visto no fluxograma, cada passo executado pela função recebeu um número de acordo com a sequência de execução. Todos esses passos são descritos a seguir:

1. Parâmetros da função:

   Para fazer uso da função, é necessário a passagem de 5 parâmetros :

   - Modelo:

     Valor inteiro que determina qual modelo de classificação será usado (0 para *Decision Tree*, 1 para *Random Forest* e 2 para *XGBoost*).

   - X:

     Conjunto de dados de entrada sem a presença da variável alvo.

   - Y:

     Vetor contendo as classes da variável alvo.

   - paramsGridCV:

     Dicionário de hiperparâmetros que será usado na função *GridSearchCV* para ajustar os modelos durante a validação cruzada.

   - n:

     Número de divisões (folds) a serem criadas durante a validação cruzada. O valor padrão é 5.

   - *Oversampling*:

     Valor booleano que determina se o *Oversampling* (SMOTE) será aplicado aos dados de treinamento. O valor padrão é False.

2. Inicialização de variáveis

   Cria um objeto *K-Fold* para dividir os dados em n partes.
   Cria variáveis do tipo *array*, para armazenar os resultados de (acurácia, *F1-Score* e erros totais) em cada interação.

3. Escolha do modelo

   Cria a instância do algoritmo que será utilizado de acordo com o valor passado no parâmetro "modelo".

4. *Loop* da validação cruzada

   A função entra em um *loop* que itera através das N divisões criadas pelo objeto *K-fold*.
   Em cada iteração, ocorre:

4.1. Divisão dos dados em Treino e Teste

Efetua a divisão do conjunto de dados em treinamento e teste, sendo 80% para treino e 20% para teste.

4.2. Balanceamento dos dados

Verifica o valor passado no parâmetro *Oversampling*,

4.2.1. Valor igual a *True*

Se igual a *True*, os dados de treinamento são balanceados.Caso contrario, segue para o passo 4.3.. É importante ressaltar que, nesse processo, foram utilizados apenas dados de treino, utilizando a técnica de *Oversampling SMOTE*, para adicionar dados à classe minoritária até igualizar as classes majoritárias.

4.3. Obtém melhores hiperparâmetros

Neste ponto, é feito o uso da função *GridSearchCV*, que realiza uma pesquisa em grade para encontrar os melhores hiperparâmetros para o modelo atual com base nas seguintes configurações fornecidas e atribuídas a cada parâmetro da função:

- *estimator*:

  Recebe a instância do algoritmo escolhido a partir do parâmetro "modelo".

- *param_grid*:

  Recebe o dicionário de hiperparâmetros a serem testados.

- CV:

  Recebe o número de iterações para validação cruzada.

- *verbose*:

  Recebe o *True*, para apresentar status de cada iteração.

- *n_jobs*:

  Recebe o valor -1, para utilizar todo o recurso de processamento da máquina.

4.4. Passagem dos melhores hiperparâmetros encontrados

Neste ponto, é criado uma nova instância do algoritmo, passando os melhores hiperparâmetros obtidos a partir da execução da função *GridSearchCV*.

4.5. Treinamento do Modelo

O modelo é treinado com os dados de treinamento da divisão atual.

4.6. Separação dos dados para teste

Efetuado separação dos dados para teste

4.7. Treinamento do modelo para previsões

O modelo treinado é usado para fazer previsões nos dados de teste.

4.8. Cálculo das métricas

Nesta etapa, a partir do uso de funções como *accuracy_score*, *f1_score*, foram obtidas as métricas de acurácia e *F1-Score* e armazenado os valores obtidos nos *array* apropriados, criados anteriormente. Também foi efetuado o cálculo do erro total, tendo como base os dados obtidos a partir da matriz de confusão, sendo os resultados obtidos armazenados no respectivo *array* criado para esse fim.

4.9. Verifica término das iterações

Se o número de iterações atingir o limite especificado, o processo prossegue para o passo 5. Caso contrário, ele permanece dentro do *loop* até que o processo seja concluído.

5. Calculando média e desvio padrão

Com o término das iterações, executadas com base no número de N divisões criadas pelo objeto *K-fold* e com os resultados obtidos e armazenados nos respectivos *arrays*, foi possível obter a média e desvio padrão de cada métrica.

6. Retorno da Função

Função retorna os dados armazenados nos *arrays*, bem como as médias e desvios padrões calculados.

# 4 RESULTADOS OBTIDOS

Neste capítulo, são apresentados os resultados obtidos a partir dos estudos e experimentos conduzidos com o conjunto de dados *Credit Score Classification (Clean Data)*, utilizando os algoritmos (*Decision Tree*, *Random Forest* e *XGBoost*) para a classificação do *score* de crédito. Para se obter os resultados exibidos a seguir, diversas etapas foram executadas. Isso inclui a análise e o tratamento dos dados, a escolha dos atributos com maior correlação entre si, a busca pelos melhores hiperparâmetros, o treinamento do modelo por meio do processo de validação cruzada, com e sem o balanceamento do conjunto de dados, e a avaliação do desempenho de cada modelo com base nas métricas de desempenho (acurácia, *F1-Score* e erro total).

## 4.1 Análise dos dados

O processo de análise de dados se iniciou com a verificação do número de registros presentes no conjunto de dados, bem como a identificação de registros nulos ou duplicados. Nesse processo, foram obtidos os seguintes resultados:

- Número de Registros : 23929
- Registros Nulos : 0
- Registros Duplicados : 0

Constatou-se também que os registros presentes no conjunto de dados, estão distribuídos em 28 atributos. Dentre esses, 17 são do tipo numérico e 11 do tipo categórico:

| Nome | Tipo |
|---|---|
| ID | object |
| Customer_ID | object |
| Month | object |
| Name | object |
| SSN | object |
| Occupation | object |
| Type_of_Loan | object |
| Credit_Mix | object |
| Payment_of_Min_Amount | object |
| Payment_Behaviour | object |
| Credit_Score | object |

Tabela 1 – Nomes dos 11 atributos categóricos do conjunto de dados selecionado

| Nome | Tipo |
|---|---|
| Age | int64 |
| Annual_Income | float64 |
| Monthly_Inhand_Salary | float64 |
| Num_Bank_Accounts | int64 |
| Num_Credit_Card | int64 |
| Interest_Rate | int64 |
| Num_of_Loan | int64 |
| Delay_from_due_date | int64 |
| Num_of_Delayed_Payment | int64 |
| Changed_Credit_Limit | float64 |
| Num_Credit_Inquiries | int64 |
| Outstanding_Debt | float64 |
| Credit_Utilization_Ratio | float64 |
| Credit_History_Age_Months | int64 |
| Total_EMI_per_month | float64 |
| Amount_invested_monthly | float64 |
| Monthly_Balance | float64 |

Tabela 2 – Nomes dos 17 atributos numéricos do conjunto de dados selecionado

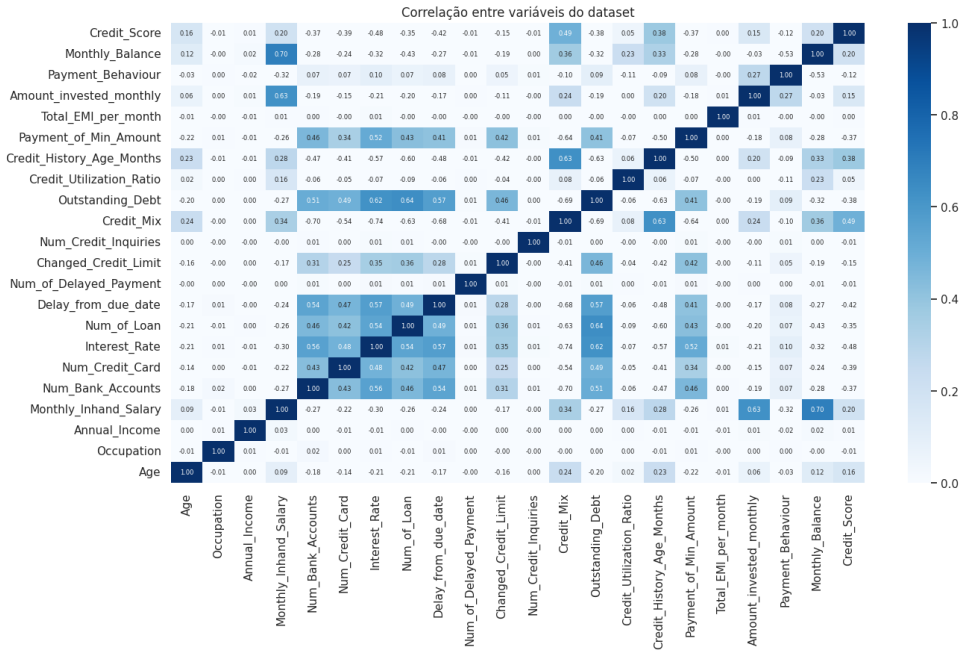
Em seguida, foi realizada uma avaliação mais detalhada, na qual foi possível perceber que certos atributos, como *ID*, *Customer_ID*, *Type_of_Loan* e *Name*, têm um impacto limitado no processo de análise, inclusive por serem atributos identificadores dos objetos. Portanto, optou-se por removê-los do conjunto de dados. É importante observar também que não foram encontrados valores faltantes no conjunto de dados, o que reforça a qualidade e a confiabilidade dos dados utilizados.

### 4.1.1 Seleção dos melhores atributos

Nesta fase, foi utiliza a função *.corr()* disponível na biblioteca *Pandas* para analisar a correlação entre os diferentes atributos do conjunto de dados. No entanto, para calcular a correlação de *Pearson* entre os atributos, foi necessário realizar a conversão dos atributos *Credit_Score*, *Credit_Mix*, *Payment_of_Min_Amount*, *Payment_Behaviour* e *Occupation*, que originalmente eram do tipo categórico, para o tipo numérico. Essa conversão foi realizada por meio da função *replace* nativa da linguagem *Python*.

Com base no resultado da função e com o objetivo de melhorar visualização dos dados, realizou-se o uso da função *heatmap*, presente da biblioteca *Seaborn*, gerando assim o mapa de calor apresentado na Figura 3 a seguir:

Figura 3 – Mapa de calor das correlações




Fonte: O AUTOR (2023)

Ao interpretar o mapa de calor, foi possível identificar os atributos com maior nível de correlação com a variável alvo, os quais são apresentados na Tabela 3

Tabela 3 – Atributos com maior nível de correlação com a variável alvo

| Atributo | Nível de correlação |
|---|---|
| Payment_of_Min_Amount | -0.37 |
| Outstanding_Debt | -0.38 |
| Credit_Mix | 0.49 |
| Delay_from_due_date | -0.42 |
| Num_of_loan | -0.35 |
| Interest_Rate | -0.48 |
| Num_Credit_Card | -0.39 |
| Credit_History_Age_Months | -0.38 |

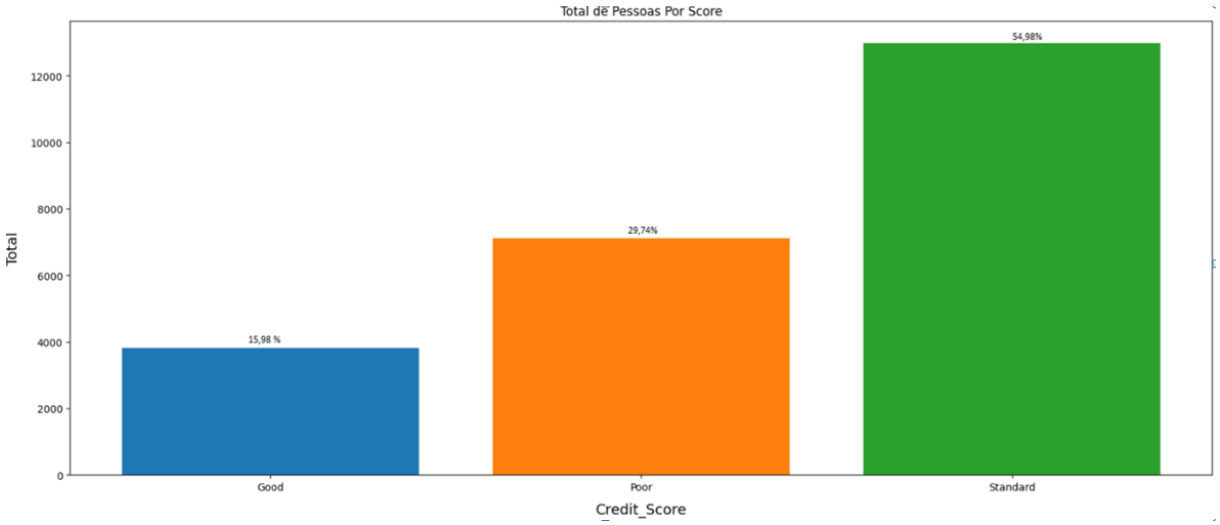
Fonte: O AUTOR (2023)

É importante salientar, que foram selecionados apenas atributos com maior nível de correlação com a variável alvo. Para isso, com base no método de correlação de *Pearson*, mencionado na seção 2.5, foram considerados apenas atributos com nível de correção na faixa (-0.35 e 0.49), pois como pode ser visto na imagem 3, os atributos que possuem uma correlação linear forte, o coeficiente de correção será mais próximo de 1 ou -1.

### 4.1.2 Balanceamento dos dados

Identificou-se também que a distribuição dos dados no conjunto não é uniforme. Cerca de 15,98% dos registros correspondem ao escore de crédito "Bom", enquanto 29,74% são classificados como "Baixo"e os restantes 54,27% são categorizados como "Regular". Essa assimetria na distribuição deve ser considerada ao interpretar os resultados das análises subsequentes.

Figura 4 – Distribuição de dados por classes




Fonte: O AUTOR (2023)

Ao analisar a Figura 1, fica evidente o desbalanceamento entre as classes. Essa disparidade pode levar o modelo gerado a produzir resultados incorretos, já que ele tende a classificar os novos dados como pertencentes à classe com maior quantidade de exemplos (classe majoritária). Essa tendência pode enganar o modelo, fazendo-o acreditar que está obtendo um bom desempenho ao classificar as novas amostras como parte da classe majoritária. O balanceamento dos dados, foi realizado durante o processo de validação cruzada, explanado na seção 3.8

## 4.2 Resultados Obtidos: Algoritmos de Aprendizado de Máquina

Nesta seção, serão apresentados os resultados obtidos a partir da execução de cada modelo, evidenciando a média dos resultados gerados em diferentes cenários. Para garantir uma análise abrangente, os experimentos foram conduzidos utilizando tanto dados balanceados quanto os não balanceados, além de considerar a influência da seleção ou não seleção de atributos preditores.

Nos testes, foram utilizados dois conjuntos de atributos como variáveis preditoras. Um com os atributos selecionados conforme apresentado na seção4.1.1 e outro sem seleção dos atributos, sendo utilizado todos os campos do conjunto de dados, com exceção da variável alvo *Credit_Score* e dos campos excluídos previamente, que foram mencionados na seção 3.3.

### 4.2.1 Resultados obtidos com e sem seleção de atributos em dados desbalanceados

Em uma primeira análise, os resultados foram obtidos sem a aplicação da técnica de *Oversampling SMOTE* para balanceamento dos dados, bem como sem realizar a seleção de atributos. Estes resultados podem ser visualizados na Tabela 4:

Tabela 4 – Resultados obtidos com e sem seleção de atributos (dados desbalanceados)

| | Acurácia | | F1-Score | | Erro Total | |
|---|---|---|---|---|---|---|
| Modelo | Com | Sem | Com | Sem | Com | Sem |
| Decision Tree | 0,6684 | 0,6707 | 0,6191 | 0,6303 | 0,3352 | 0,3293 |
| Random Forest | 0,6806 | 0,7078 | 0,6458 | 0,6775 | 0,3194 | 0,2922 |
| XGBoost | 0,6803 | 0,7079 | 0,6478 | 0,6781 | 0,3197 | 0,2921 |

Fonte: O AUTOR (2023)

### 4.2.2 Resultados obtidos com e sem seleção de atributos em dados balanceados

Para tentar melhorar o resultado dos modelos, foi aplicada a técnica de *Oversampling SMOTE* para balanceamento dos dados; no entanto, como é possível ver nos resultados apresentados na Tabela 5, não foi isso que ocorreu, o que demonstra que o balanceamento dos dados não é eficaz em todos os casos.

Tabela 5 – Resultados obtidos com e sem seleção de atributos (dados balanceados)

| | Acurácia | | F1-Score | | Erro Total | |
|---|---|---|---|---|---|---|
| Modelo | Com | Sem | Com | Sem | Com | Sem |
| Decision Tree | 0,5614 | 0,6075 | 0,5358 | 0,5828 | 0,4386 | 0,3925 |
| Random Forest | 0,6594 | 0,6927 | 0,6420 | 0,6754 | 0,3406 | 0,3073 |
| XGBoost | 0,6441 | 0,6730 | 0,6158 | 0,6450 | 0,3559 | 0,3261 |

Fonte: O AUTOR (2023)

### 4.2.3 Análise dos resultados

#### 4.2.3.1 Dados desbalanceados

Ao comparar os resultados apresentados na Tabela 4, é possível observar que os modelos *Random Forest* e *XGBoost* tiveram um desempenho aproximado na previsão do score de crédito, tanto quando os testes foram realizados com atributos selecionados, quanto não selecionados. No entanto, o algoritmo *XGBoost* apresentou uma ligeira vantagem em relação aos outros modelos avaliados. Em termos de acurácia, o *XGBoost* obteve a pontuação mais alta, tanto nos testes realizados com e sem seleção dos atributos, mas teve uma pontuação melhor, chegando a 0,7079, quando não utilizado os atributos selecionados. Ao levar em consideração o *F1-Score*, o modelo que utilizou o algoritmos *XGBoost*, também teve um desempenho melhor, quando não utilizado atributos selecionados, obtendo uma pontuação igual 0,6781. Quanto à métrica Erro Total, que reflete a proporção de previsões incorretas, o *XGBoost* apresentou o valor mais baixo, indicando que cometeu menos erros em suas previsões quando comparado aos outros modelos. Portanto, com base nas métricas apresentadas, o modelo que fez uso do algoritmo *XGBoost* se destacou em relação aos outros, obtendo um melhor desempenho nos dois cenários (com e sem seleção de atributos), quando avaliadas as métricas (acurácia, *F1-Score* e erro total) em conjunto.

#### 4.2.3.2 Dados balanceados

Neste cenário, em que ocorreu o balanceamento dos dados, é possível ver uma queda significativa no desempenho dos modelos, bem como uma inversão no resultado do modelo que teve melhor desempenho. Isso sugere que pode ter ocorrido uma sobreposição entre as classes, o que aumenta o risco de *Overfitting*, que ocorre quando um modelo se ajusta muito bem aos dados utilizados no treinamento, mas erra muito quando novos dados são introduzidos.

Ao analisar os resultados na Tabela 5, é possível identificar que o modelo que fez uso do algoritmo *Random Forest* teve um desempenho melhor em comparação aos outros, quando analisado as métricas (acurácia, *F1-Score* e erro total) em conjunto. Também foi possível identificar, que os resultados onde os testes foram realizados sem a seleção dos atributos, foram significativamente melhores, em relação aos testes que utilizaram atributos selecionados, padrão que também se repetiu nos resultados onde não ocorreu o balanceamento dos dados.

#### 4.2.3.3 Análise final

A Análise dos resultados indica que a seleção ou não dos atributos preditores, bem como o balanceamento/não balanceamento dos dados, podem influenciar significativamente nos resultados obtidos. Isso pode ser visto nos resultados apresentados nas tabelas (4 e 5, onde o *XGBoost* teve um melhor desempenho, quando realizados os testes com dados

não balanceados e sem seleção dos atributos. Já o algoritmo *Random Forest* se destacou quando os testes foram realizados com dados balanceados e sem seleção de atributos. No entanto, como dito na subseção 4.2.3.2, ao balancear o conjunto de dados, pode ocorrer sobreposição entre as classes, aumentar o risco de *Overfitting* e fazer com que os resultados não sejam realistas.

# 5 CONCLUSÕES

Este trabalho teve como objetivo realizar uma análise experimental em algoritmos de aprendizado de máquina baseados em árvore de decisão para a previsão de scores de crédito, a fim de identificar qual dos algoritmos teria um melhor desempenho no processo de classificação do score. Ao analisar o desempenho de cada algoritmo utilizado (*Decision Tree*, *Random Forest* e *XGBoost*), tendo como base para isso os resultados das métricas (acurácia, F1-Score e Erro total) em conjunto, foi possível concluir que o *XGBoost* foi o que apresentou melhor resultado nos testes realizados sem o balanceamento do conjunto de dados.

No entanto, quando realizado o balanceamento do conjunto de dados, o algoritmo que teve melhor desempenho foi o *Random Forest*. Levando em consideração que, ao balancear os dados, pode ter ocorrido uma sobreposição das classes, apresentando resultados não realistas, é possível afirmar que o *XGBoost* foi o que apresentou melhores resultados neste cenário.

É importante ressaltar que todo o processo de execução dos algoritmos foi realizado a partir da plataforma *Google Colab*, em sua versão gratuita, a qual apresenta algumas restrições de tempo, de uso e de recursos computacionais. Dessa forma, foi necessário limitar o número de iterações no processo de validação cruzada, bem como o número de hiperparâmetros a serem otimizados. Além disso, outra limitação foi o tempo disponível para realizar testes mais aprofundados, como identificar a precisão de acerto em cada classe.

Diante das conclusões e limitações desse trabalho, como trabalhos futuros pode-se mencionar uma avaliação mais aprofundada da qualidade dos modelos em cada classe. Além disso, explorar outras técnicas de seleção de atributos.

# REFERÊNCIAS

ALI, J.; KHAN REHANULLAH E AHMAD, N.; MAQSOOD, I. Random forests and decision trees. **International Journal of Computer Science Issues (IJCSI)**, International Journal of Computer Science Issues (IJCSI), v. 9, n. 5, p. 272, 2012. Acesso em : 03 ago. 2023. Disponível em: <https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees>.

ALVES SÉRGIO DARCY E SOARES, M. M. da S. Microfinanças: Democratização do crédito no brasil atuação do banco do brasil. 2006. Acesso em: 01 jun. 2023. Disponível em: <https://www.bcb.gov.br/htms/public/microcredito/NotaDC200512.pdf>.

ARIDAS, G. L. e Fernando Nogueira e C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017. Acesso em: 25 jun. 2023. Disponível em: <http://jmlr.org/papers/v18/16-365.html>.

BRANDALIZE, T.; FLACH, L.; SALLABERRY, J. D. Análise dos determinantes no grau de evidenciação do risco de crédito em centrais de cooperativas de crédito. **Revista de Gestão e Organizações Cooperativas**, v. 8, n. 15, p. 01–34, maio 2022. Acesso em: 04 nov. 2023. Disponível em: <https://periodicos.ufsm.br/rgc/article/view/e42461>.

BRASIL. **Lei No. 4.595**. 1964. <http://www.planalto.gov.br/ccivil_03/leis/l4595compilado.htm>. Acesso em: 25 ago. 2023.

CARVALHO, K. F. e Ana Paula Lorena e Joao Gama e André A.C.L.F de. **Inteligência Artificial: Uma abordagem de Aprendizado de Máquina**. [*S.l.: s.n.*]: Grupo Editorial Nacional, 2011. ISBN 9788521618805.

CASTRO, J. S. Estudo comparativo entre metodologias de aprendizado de máquina e híbridas aplicadas a risco de crédito. 2019. Acesso em: 14 set. 2023. Disponível em: <http://tede.fecap.br:8080/handle/123456789/818>.

CHEN, D. **Análise de dados com Python e Pandas**. Novatec Editora, 2018. Acesso em : 27 jul. 2023. ISBN 9788575226995. Disponível em: <https://books.google.com.br/books?id=ILFwDwAAQBAJ>.

CHIN L. E DUTTA, T. **NumPy Essentials**. Packt Publishing, 2016. Acesso em : 15 ago. 2023. ISBN 9781784392185. Disponível em: <https://books.google.com.br/books?id=RvvfDAAAQBAJ>.

COELHO, F.; AMORIM, D.; CAMARGOS, M. Analisando métodos de machine learning e avaliação do risco de crédito (analyzing machine learning methods and credit risk assessment). **Revista Gestão e Tecnologia**, v. 21, p. 89–116, 03 2021. Acesso em: 04 nov. 2023. Disponível em: <https://www.researchgate.net/publication/350079793_Analisando_metodos_de_machine_learning_e_avaliacao_do_risco_de_credito_Analyzing_machine_learning_methods_and_credit_risk_assessment/citation/download>.

DAHAN, H.; COHEN S. E ROKACH, L.; MAIMON, O. **Proactive Data Mining with Decision Trees**. Springer New York, 2014. (SpringerBriefs in Electrical and Computer Engineering). Acesso em : 25 jul. 2023. ISBN 9781493905393. Disponível em: <https://books.google.com.br/books?id=Fey3BAAAQBAJ>.

DEPREZ, M.; ROBINSON, E. **Machine Learning for Biomedical Applications: With Scikit-Learn and PyTorch**. Elsevier Science, 2023. Acesso em 18 de novembro de 2023. ISBN 9780128229057. Disponível em: <https://books.google.com.br/books?id=_4RTEAAAQBAJ>.

ERIC, B. G. *et al.* Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidadea**, v. 10, n. 20, 2013. Acesso em: 12 jun. 2023. Disponível em: <http://www.redalyc.org/articulo.oa?id=76228118008>.

FORTUNA, E. **Mercado Financeiro: Produtos e Serviços.** [*S.l.: s.n.*]: Qualitymark, 1999. v. 12. ISBN 978-8573032161.

GNOATTO, R. **Análise do desempenho de hiperparâmetros de aprendizagem de máquina aplicados na previsão da taxa de rotatividade**. 2023. Monografia (Especialização) — UNIVERSIDADE DO VALE DO TAQUARI - UNIVATES, 2023. Acesso em: 28 jun. 2023. Disponível em: <https://www.univates.br/bdu/bitstreams/764ab0f8-1b0a-4809-893e-fbe82f6ff76e/download>.

GUILHON, A. *et al.* **Jornada Python: uma jornada imersiva na aplicabilidade de uma das mais poderosas linguagens de programação do mundo**. Brasport, 2022. Acesso em : 16 ago. 2023. ISBN 9786588431481. Disponível em: <https://books.google.com.br/books?id=Nx1dEAAAQBAJ>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition**. Springer New York, 2009. (Springer Series in Statistics). Acesso em : 23 jul. 2023. ISBN 9780387848587. Disponível em: <https://books.google.com.br/books?id=tVIjmNS3Ob8C>.

HE, H.; MA, Y. **Imbalanced Learning: Foundations, Algorithms, and Applications**. Wiley, 2013. Acesso em: 10 jun. 2023. ISBN 9781118646335. Disponível em: <https://books.google.com.br/books?id=CVHx-Gp9jzUC>.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science and Engineering**, v. 9, n. 3, p. 90–95, 2007. Acesso em: 10 jun. 2023. Disponível em: <https://ieeexplore.ieee.org/document/4160265>.

JÚNIOR, L. O. **Tomada de decisões em sistemas financeiros utilizando algoritmos de aprendizado de máquina supervisionado**. 2018. Dissertação (Dissertação) — Universidade de São Paulo, 2018. Acesso em: 18 ago. 2023. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55137/tde-22032019-171747/en.php>.

LEARN scikit. **1.10. Decision Trees**. 2023. Acesso em: 05 ago. 2023. Disponível em: <https://scikit-learn.org/stable/modules/tree.html>.

LOPES, E. **Gestão e Análise de Crédito nas Instituições Financeiras Cooperativas**. Editora Confebras, 2021. Acesso em: 01 jun. 2023. ISBN 9786588748022. Disponível em: <https://books.google.com.br/books?id=fs4iEAAAQBAJ>.

MITCHELL, T. M. Machine learning and data mining. **Communications of the ACM**, v. 42, n. 11, November 1999. Acesso em: 04 jun. 2023. Disponível em: <https://www.ri.cmu.edu/pub_files/pub1/mitchell_tom_1999_1/mitchell_tom_1999_1.pdf>.

MOLIN S. E JEE, K. **Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization**. Packt Publishing, 2021. Acesso em : 25 jul. 2023. ISBN 9781800565913. Disponível em: <https://books.google.com.br/books?id=Eh4sEAAAQBAJ>.

MONARD M. E BARANAUSKAS, J. Conceitos de aprendizagem de máquina. *In*: **Rezende**. [*S.l.: s.n.*], 2003. p. 89–114.

MÜLLER A.C. E GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. O'Reilly Media, 2016. Acesso em : 18 ago. 2023. ISBN 9781449369897. Disponível em: <https://books.google.com.br/books?id=vbQlDQAAQBAJ>.

NETO, A. F. d. S.; SILVA, J. F. G. d. Predição de pagamentos atrasados através de algoritmos baseados em Árvore de decisão. **Revista de Engenharia e ...**, 2021. Acesso em: 24 jun. 2023. Disponível em: <http://revistas.poli.br/index.php/repa/article/view/1746>.

OCERGS-SESCOOP/RS, S. **Fundamentos do Cooperativismo**. 2020. Acesso em: 03 jun. 2023. Disponível em: <https://www.sescooprs.coop.br/app/uploads/2020/07/fundamentos-do-cooperativismo.pdf>.

PALMUTI CLAUDIO SILVA E PICCHIAI, D. Mensuração do risco de crédito por meio de análise estatística multivariada. **Revista Economia Ensaios**, 2012. Acesso em : 23 jul. 2023. Disponível em: <https://seer.ufu.br/index.php/revistaeconomiaensaios/article/view/14808/12192>.

PECK, P. **Algoritmos e Score de Crédito**. 2023. Acesso em: 04 nov. 2023. Disponível em: <https://febrabantech.febraban.org.br/especialista/patricia-peck-pinheiro/algoritmos-e-score-de-credito>.

PELISON, L. F. **Geração Automática de Features para Modelagem Preditiva - Predição de Empresas Brasileiras de Alto Crescimento**. 2018. Monografia (Especialização) — Universidade Federal de Santa Catarina, 2018. Acesso em: 16 set. 2023. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/200013/PFC%20Luis%20Felipe%20Pelison_2018-2.pdf?sequence=1&isAllowed=y>.

RIS-ALA, R. **Fundamentos de Aprendizagem por Reforço**. Rafael Ris-Ala, 2023. Acesso em: 01 jun. 2023. ISBN 9786500604368. Disponível em: <https://books.google.com.br/books?id=IKmtEAAAQBAJ>.

RODRIGUES, R. B. **Explicabilidade utilizando LIME: um Estudo de Caso para o Mercado Financeiro**. 2021. Monografia (Especialização) — Universidade de São Paulo, 2021. Acesso em: 30 jun. 2023. Disponível em: <https://bdta.abcd.usp.br/directbitstream/00f4b9f8-c26e-4093-b469-789e1bae6722/Robson+Buratti+Rodrigues.pdf>.

ROMANI, M. F. **Comparação de algoritmos de aprendizagem de máquina na construção de modelos preditivos para rentabilidade de clientes bancários**. 2017. TCC — Universidade de Brasília, 2017. Acesso em : 07 set. 2023. Disponível em: <https://bdm.unb.br/bitstream/10483/20501/1/2017_MateusFlachRomani_tcc.pdf>.

ROMANI, M. F. **Comparação de algoritmos de aprendizagem de máquina na construção de modelos preditivos para rentabilidade de clientes bancários**. 2017. Monografia (TCC) — UNIVERSIDADE DE BRASILIA, 2017. Acesso em: 18 ago. 2023. Disponível em: <https://bdm.unb.br/handle/10483/20501>.

SAINI, S.; LATA, K.; SINHA, G. **VLSI and Hardware Implementations using Modern Machine Learning Methods**. CRC Press, 2021. Acesso em: 19 nov. 2023. ISBN 9781000523843. Disponível em: <https://books.google.com.br/books?id=iPRPEAAAQBAJ>.

SANTOS, A. S. **Previsão de insolvência corporativa: uma análise de empresas brasileiras de capital aberto por meio de aprendizado de máquina**. 2021. Dissertação (Dissertação) — UNIVERSIDADE FEDERAL DA PARAÍBA, 2021. Acesso em: 10 set. 2023. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/22397>.

SCHARDONG, A. **Cooperativa de crédito: instrumento de organização, econômica da sociedade**. [*S.l.: s.n.*]: Rigel, 2003. v. 2. ISBN 9788573490138.

SCHRICKEL, W. k. **Análise de Crédito. Concessão e Gerência de Empréstimos**. [*S.l.: s.n.*]: Atlas, 2000. v. 5. ISBN 978-8522426461.

SIAHAAN, V.; SIANIPAR, R. **DATA SCIENCE FOR RAIN CLASSIFICATION AND PREDICTION WITH PYTHON GUI**. BALIGE PUBLISHING, 2023. Acesso em: 10 jun. 2023. Disponível em: <https://books.google.com.br/books?id=421tEAAAQBAJ>.

SILVA, D. O. **Otimização de hiper-parâmetros de algoritmos de machine learning aplicados no contexto de análise de risco de crédito**. 2022. Monografia (Especialização) — UNIVERSIDADE TECNOLOGICA FEDERAL DO PARANÁ, 2022. Acesso em: 25 jun. 2023. Disponível em: <http://riut.utfpr.edu.br/jspui/handle/1/31719>.

SILVEIRA, C. C. V. T. **Revisão e aplicação de métodos de aprendizado de máquina para a predição de Churn**. 2022. Monografia (TCC) — UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, 2022. Acesso em: 25 ago. 2023. Disponível em: <https://pantheon.ufrj.br/handle/11422/18441>.

SOUZA D. H. M. DE E BORDIN JR, C. J. Detecção de fraude de cartão de crédito por meio de algoritmos de aprendizado de máquina. **Revista Brasileira de Computação**, v. 23, n. 1, p. 1–10, 2023. Acesso em: 10 set. 2023. Disponível em: <http://seer.upf.br/index.php/rbca/article/view/13790>.

VASCONCELLOS, P. Como selecionar as melhores features para seu modelo de machine learning. **Paulo Vasconcellos**, 2019. Acesso em: 16 set. 2023. Disponível em: <https://paulovasconcellos.com.br/como-selecionar-as-melhores-features-para-seu-modelo-de-machine-learning-2e9df83d062a>.

VELOSO, L. T. **Um estudo comparativo de técnicas de validação cruzada aplicadas a modelos para dados desbalanceados**. 2022. Dissertação (Mestrado) — Universidade de São Paulo, 2022. Acesso em : 23 jul. 2023. Disponível em: <https://www.teses.usp.br/teses/disponiveis/45/45133/tde-18042022-200608/pt-br.php>.