UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
Curso de Graduação em Sistemas de Informação
Pedro Henrique Barroso

UM ESTUDO COMPARATIVO DENTRE OS PRINCIPAIS MODELOS DE APRENDIZADO DE MÁQUINA APLICADO À DETECÇÃO DE FRAUDES BANCÁRIAS

Diamantina
2025

Pedro Henrique Barroso

UM ESTUDO COMPARATIVO DENTRE OS PRINCIPAIS MODELOS DE APRENDIZADO DE MÁQUINA APLICADO À DETECÇÃO DE FRAUDES BANCÁRIAS

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Alessandro Vivas Andrade

Diamantina
2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

FOLHA DE APROVAÇÃO

Pedro Henrique Barroso

UM ESTUDO COMPARATIVO ENTRE OS PRINCIPAIS MODELOS DE APRENDIZADO DE MÁQUINA APLICADO À DETECÇÃO DE FRAUDES BANCÁRIAS

Trabalho de Conclusão de Curso apresentado ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisito parcial para conclusão do curso.

Orientador: Prof. Dr. Alessandro Vivas Andrade

Aprovado em 15 de setembro de 2025

BANCA EXAMINADORA

Prof. Dr. Alessandro Vivas Andrade
Faculdade de Ciências Exatas - DECOM - UFVJM

Profª. Dr.ª Luciana Pereira de Assis
Faculdade de Ciências Exatas - DECOM - UFVJM

Dra. Claudia Beatriz Berti
Faculdade de Ciências Exatas - DECOM - UFVJM

Documento assinado eletronicamente por **Alessandro Vivas Andrade**, **Servidor(a)**, em 06/10/2025, às 09:45, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Claudia Beatriz Berti**, **Servidor(a)**, em 06/10/2025, às 11:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Luciana Pereira de Assis**, **Servidor(a)**, em 06/10/2025, às 17:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufvjm.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1900649** e o código CRC **81A525D1**.

Referência: Processo nº 23086.000729/2023-83 SEI nº 1900649

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, pois sem Ele nada disso teria sido possível. Expresso também minha profunda gratidão aos meus pais, Stela e Barachinho, que sempre se dedicaram com esforço e zelo para que eu pudesse estudar. Em todos os momentos, estiveram ao meu lado, sem jamais medir esforços ou me deixar sozinho nessa caminhada.

Estendo meu agradecimento a toda a minha família e amigos, que sempre me apoiaram, incentivaram e tornaram esse percurso mais leve e acolhedor.

Não poderia deixar de reconhecer o apoio essencial da minha psicóloga, Kênia, que tem me acompanhado com dedicação, acolhendo cada passo e celebrando comigo cada conquista.

Agradeço, de forma especial, ao meu orientador, Prof. Dr. Alessandro Vivas, por sua orientação atenciosa e por tornar o desenvolvimento deste trabalho mais leve, agradável e até divertido.

Sou grato também a todos os professores que contribuíram para minha formação, transmitindo conhecimentos que levarei comigo por toda a vida.

Por fim, agradeço a todos os colegas de turma que compartilharam dessa jornada, em especial Victor, Ian, Mariana e Jhonathan, por tornarem esse caminho mais leve, compreensivo e repleto de bons momentos.

RESUMO

O cartão de crédito hoje é utilizado em diversas formas de transações online, devido sua praticidade e multiuso, ele se tornou o segundo meio de pagamento mais utilizado pelos brasileiros. Estas particularidades o tornam um alvo frequente de golpistas com objetivo de obter ganhos ilícitos, se tornando uma preocupação global. Assim, o presente estudo teve por objetivo realizar a comparação e avaliação de diferentes modelos de Aprendizado de Máquina para a classificação e detecção de fraude bancária, visando quantificar a diferença de desempenho entre o uso dos seguintes algoritmos aprendizado de máquina: Árvore de Decisão, Floresta Aleatória e XGBoost em uma base de dados única, possibilitando a comparação dos modelos e a compreensão das especificidades dela. Todos os modelos obtiveram desempenho excelente, se divergindo apenas na terceira casa decimal. Toda via, o modelo com melhor desempenho em relação às métricas de avaliação, considerando esta ressalva, foi o Floresta Aleatória, demonstrando sua eficiência para o objetivo proposto.

Palavras-chave: Aprendizado de Máquina. Fraude. Análise de Dados. Machine Learning. Árvore de Decisão. Floresta Aleatória. XGBoost. Classificação.

ABSTRACT

Credit cards are now used in various forms of online transactions. Due to their practicality and versatility, they have become the second most widely used payment method among Brazilians. These characteristics make them a frequent target for fraudsters seeking to obtain illicit gains, which has become a global concern. Thus, the objective of this study was to compare and evaluate different Machine Learning models for the classification and detection of bank fraud, aiming to quantify the difference in performance between the use of the following algorithms: machine learning algorithms: Decision Tree, Random Forest, XGBoost on a single database, enabling the comparison of models and understanding of its specificities. All models performed excellently, differing only in the third decimal place. However, the model with the best performance in relation to the evaluation metrics was Random Forest, demonstrating its efficiency for the proposed objective.

Keywords: Machine Learning. Fraud. Data Analysis. Machine Learning. Decision Tree. Random Forest. XGBoost. Classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Conforme descrito por Furlan (2023) e ampliado por Paixão (2023), o cartão de crédito hoje é utilizado em diversas modalidades de transações online, tais como parcelamentos de compras, assinatura de produtos e serviços, compras em aplicativos de *delivery*. Além disso, Paixão (2023) acrescenta que a sua praticidade e multiuso tornam o cartão de crédito o segundo meio de pagamento mais utilizado pelos brasileiros. Segundo pesquisa realizada pelo Serasa (2022), metade dos brasileiros possui quatro ou mais cartões de crédito. Particularidades como essas tornam o cartão de crédito um alvo frequente de golpistas, que usam deste meio para roubar dados ou realizar uso indevido de valores monetários.

De acordo com Paixão (2023), dois em cada dez titulares de cartão de crédito já foram vítimas de golpes online, conforme registrado em uma pesquisa realizada pela empresa de segurança *Kaspersky* e divulgada no blog pela Sica (2022). Seu estudo considerou apenas as compras realizadas por cartão de crédito. O Brasil em 2023, com base na matéria de Longo (2024), registrou 280 milhões de pedidos de compras pelo comércio eletrônico. Destes, ocorreram 3,7 milhões de pedidos com tentativas de fraudes. Embora este valor seja apenas 1,4% do total de pedidos, movimentou-se aproximadamente 3,5 bilhões de reais, conforme pesquisa realizada pela ClearSale (2024).

FRAUDE BANCÁRIA

Conforme exposto pelo Banco Pan (2022), **Fraude Bancária** pode ser definida como um crime cometido por terceiros que utilizam meios ilegais para furtar dinheiro ou crédito dos correntistas. Além disso, o Banco Pan (2022) também explica uma outra forma de fraude, onde indivíduos mal-intencionados roubam os dados pessoais através de links maliciosos, páginas de contato falsas, ligações telefônicas, mensagens em redes sociais, entre outros. Estes dados pessoais são utilizados pelos golpistas para furtar o dinheiro ou crédito dos correntistas, conforme descrito na definição de fraude bancária.

O autor Leonardis (2023) caracteriza a ocorrência de fraudes bancárias em duas categorias principais: fraudes por comportamento e fraudes por aplicação. As fraudes por aplicação se referem a indivíduos que solicitam um novo cartão de crédito, utilizando dados fraudulentos. Já as fraudes por comportamento estão relacionadas diretamente ao roubo ou clonagem de cartões de créditos já existentes. A fraude bancária por cartão de crédito é um problema que afeta consumidores, varejistas e bancos, causando prejuízos significativos para todos os envolvidos, afinal, o cliente muitas vezes perde a confiança na compra online, a varejista precisa estornar o valor e, por fim, o banco precisa arcar com o prejuízo. Dessa forma, os investimentos anuais realizados pelos bancos em novas tecnologias e apoio a novas técnicas com o objetivo de inibir as tentativas de fraudes se aproximam dos R$3 bilhões segundo relatório da Federação Brasileira de Bancos (2023).

ENTENDIMENTO E CONTEXTO

Segundo o Relatório Varejo 2023, realizado pelas empresas Opinium Research e Censuswide, o ano de 2022 registrou um aumento significativo das fraudes bancárias no Brasil, impactando até 36% dos varejistas (BigDataCorp, 2023). Esta ameaça emergente apresenta maior risco para o setor, atrás apenas da redução do poder de compra brasileiro. As fraudes financeiras são uma preocupação global. Segundo o Relatório Global de impacto de fraude financeira da IBM, em 2022, a fraude de cartão de crédito é o tipo mais comum de fraude financeira em todos os países pesquisados. Outro ponto a destacar da pesquisa é que 91% da população brasileira entrevistada está inclinada a optar por utilizar empresas que adotassem medidas de proteção contra fraudes BigDataCorp (2023).

Diante desse cenário, a adoção de medidas eficazes contra estas fraudes se torna essencial. A tecnologia tem se tornado grande aliada nestes casos; campos de pesquisa como o da Inteligência Artificial, através de modelos de aprendizado de máquina *(Machine Learning)* e Redes Neurais Artificiais, estudam o reconhecimento de padrões com o objetivo de classificar e/ou predizer determinados comportamentos a partir de uma base de dados (SILVA, 2020). A aplicação destes algoritmos possibilita a detecção de transações atípicas antes que ocorram, garantindo maior confiabilidade, segurança e, principalmente, economia de tempo e dinheiro para todos os envolvidos neste processo.

LEVANTAMENTO DE HIPÓTESES

A partir do conjunto de dados utilizado, espera-se identificar relações relevantes entre as colunas da base de dados que alcance o melhor resultado pelos diferentes modelos de classificação binária entre **“É fraude”** ou **“Não é fraude”**.

1.1 OBJETIVO

Este trabalho tem como objetivo de realizar um estudo comparativo dos algoritmos de aprendizado de máquina: Árvore de Decisão, Floresta Aleatória e XGBoost, aplicado à detecção de fraude bancária de cartão de crédito, um problema comum nos dias atuais, como dito anteriormente.

1.2 OBJETIVOS ESPECIFICOS

Este estudo considerou a utilização de uma mesma base de dados sobre fraudes bancárias ocorridas na Europa em 2023 e disponibilizada na plataforma *Kaggle*, uma plataforma para estudo e pesquisa em ciência de dados. Dessa forma, torna-se possível a quantificação da diferença entre o uso dos algoritmos e garante a descoberta sobre qual método apresenta uma melhor performance. É importante ressaltar que este estudo traz uma perspectiva de apoio do processo de tomada de decisão a cerca de qual método adotar ao desenvolver uma solução.

2 REFERENCIAL TEÓRICO

Neste capítulo, serão apresentados os conceitos sobre os métodos e algoritmos utilizados no trabalho.

2.1 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (AM), segundo Amorim (2008), pode ser definido como uma subárea da Inteligência Artificial. Ela é responsável pela criação e desenvolvimento de algoritmos e técnicas que permitem à máquina realizar o processo de aprendizagem — daí o nome Aprendizado de Máquina.

Ele ainda destaca que este aprendizado inclui principalmente aprimorar seu desempenho em determinada atividade. E também complementa em seu trabalho que este processo de aprendizagem possui relação com áreas, como, mineração de dados e estatística, focando em particularidades de métodos estatísticos. É válido ressaltar que este complemento está totalmente relacionado ao presente trabalho de monografia.

O Santos *et al.* (2019) traz outra definição para o aprendizado de máquina, partindo de uma perspectiva técnica e formalizando o tema de aprendizado de máquina, como:

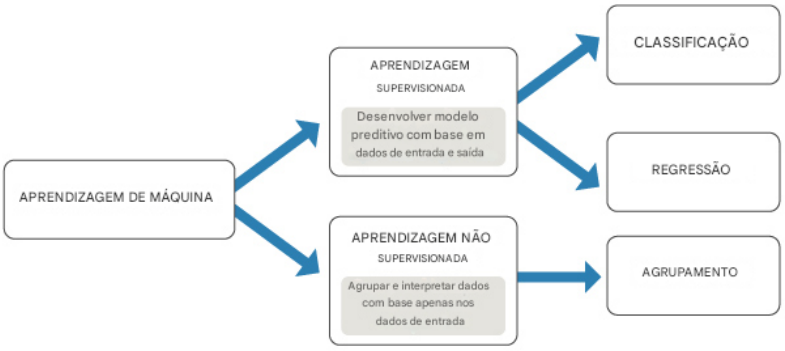
> "O Aprendizado de Máquina é um campo da ciência da computação que representa a evolução dos sistemas de reconhecimentos de padrões e permitindo aos computadores aprender a partir dos erros e fazer previsões". (SANTOS *et al.*, 2019, pag. 387–396)

Além disso, o Amorim (2008) acrescenta que para que seja possível realizar a etapa de aprendizado de máquina é necessário a existência de uma quantidade massiva de dados, para que seja possível realizar de maneira eficiente o treinamento e teste dos modelos, além dos dados, ele destaca que é importante para a máquina obter as respostas destes dados, a fim de realizar o treinamento de maneira eficiente.

O Sarker (2021) propõe que os tipos de algoritmos podem ser divididos em 4 grandes grupos principais, sendo elas: Supervisionado, Não Supervisionado, Semi-Supervisionado e por fim de Aprendizado por Reforço. Além disso, ele propõe que as técnicas e métodos de aprendizado de máquina podem ser divididos em Classificação, Regressão, Agrupamentos de Dados, Engenharia de Recurso e Aprendizado por Reforço.

Como ilustrado na Figura 1 abaixo.

Figura 1 – Tipos de Aprendizado de Máquina.



Fonte: Adaptado de MathWorks (2024).

Ademais, o Sarker (2021) diz que os dados podem seguir diversas formas distintas, sendo elas estruturados, semi-estruturados, não estruturados e metadados. Ele ainda define os ”dados estruturados”como um conjunto de dados que possui uma estrutura bem definida, seguindo uma ordem lógica e padrão, seu armazenamento segue um banco de dados relacional ou de maneira tabular, como em uma planilha excel. Ademais, ele define os ”dados não estruturados”, como um tipo de dados que não segue uma ordem ou sequência lógica, essa ausência de estrutura lógica dificulta a captação, processamento e análise destes dados.

Segundo Sarker (2021), dados de sensores, imagens, posts de blogs podem ser considerados como dados não estruturados. Os dados semi-estruturados são um tipo de dados que não são armazenados em banco de dados relacionais, porém eles seguem uma sequência ou forma lógica de organização, exemplos de dados deste tipo são HTML, XML e JSON. Por fim, Sarker (2021) define os metadados como uma forma diferente dos dados, sendo expressos por “dados sobre dados”, ou seja, são os dados característicos sobre um determinado dado. Como em uma publicação, o dado principal é a publicação e os metadados podem ser classificados como a data da publicação, autor, ano de publicação, local de publicação etc.

2.2 CLASSIFICAÇÃO OU REGRESSÃO

Conceituado sobre o que é e os tipos principais de algoritmos de aprendizado de máquina, na sequência, será abordado o tipo de aprendizado supervisionado, pois foi a técnica utilizada neste trabalho, além disso, detalha-se a seguir, o funcionamento dos algoritmos utilizados.

Sobre o aprendizado supervisionado, Sarker (2021) explica seu funcionamento base de forma sucinta como um processo de aprendizado baseado na utilização de dados rotulados para treinamento e utilização de uma coleção de exemplos de treinamento para inferir uma função. Ou seja, ele é realizado quando determinados objetivos identificados podem ser alcançados a partir de um conjunto de entradas. Suas principais tarefas são de classificação e regressão, Expoẽ-se a seguir sobre estas técnicas e posteriormente explicaremos os algoritmos utilizados.

O Sarker (2021) define a tarefa de classificação como um método de aprendizado supervisionado e geralmente é associado a problemas de modelagem preditiva, ou seja, um tipo de problema em que um rótulo da classe é previsto em um determinado cenário. Os métodos de classificação podem ser categorizados em classificação binária (que vamos focar neste trabalho), classificação multi-classe e multi-rótulos. A **classificação binária** refere-se a uma classe de problemas de classificação em que existem basicamente, dois rótulos da classe como “True” e “False” ou “Yes” e “No”. Um exemplo clássico de classificação binária é a classificação de e-mails como spam ou não-spam, a classe normal pode ser spam e a anormal ser não-spam, desta forma, o spam pode ser relacionado como True e a não-spam como False.

Ademais, Sarker (2021) define a tarefa de regressão como um método de aprendizado de máquina que permite prever uma variável de resultado contínuo (y) a partir de uma ou várias variáveis preditoras (x). A maior diferença entre classificação e regressão é que a classificação prevê os rótulos de classes distintas e a regressão facilita a previsão de uma quantidade contínua de rótulos. Alguns algoritmos podem ser utilizados tanto para problemas de classificação quanto de regressão. Além disso, os modelos de regressão mais clássicos são os algoritmos de regressão linear e regressão polinomial.

Para exemplificar de maneira visual estes conceitos de classificação e regressão, na Figura 2 abaixo. Expõe-se as duas técnicas, sendo a de classificação representada por meio da divisão de duas classes ”A”e ”B”. No que se refere a técnica de regressão, sua representação é por meio de uma reta formada pela função linear viabilizando a previsão de valores através da reta.

Figura 2 – Classificação Ou Regressão.

Fonte: Adaptado de Ali (2024).

2.3 ALGORITMOS DE APRENDIZADO DE MÁQUINA

Conforme exposto anteriormente, O aprendizado de máquina (AM) é uma grande área de estudos que explora os conceitos de aprendizado supervisionado, não supervisionado

utilizando técnicas de classificação ou regressão. Nesse contexto, o presente trabalho envolve as técnicas de classificação contempladas pela área de aprendizado supervisionado explorada pela grande área de aprendizado de máquina. Sob esta ótica, exploraremos os algoritmos de Árvore de Decisão, Floresta Aleatória e XGBoost.

2.3.1 ÁRVORE DE DECISÃO

O Autor Santos (2023) Define o algoritmo de árvore como:

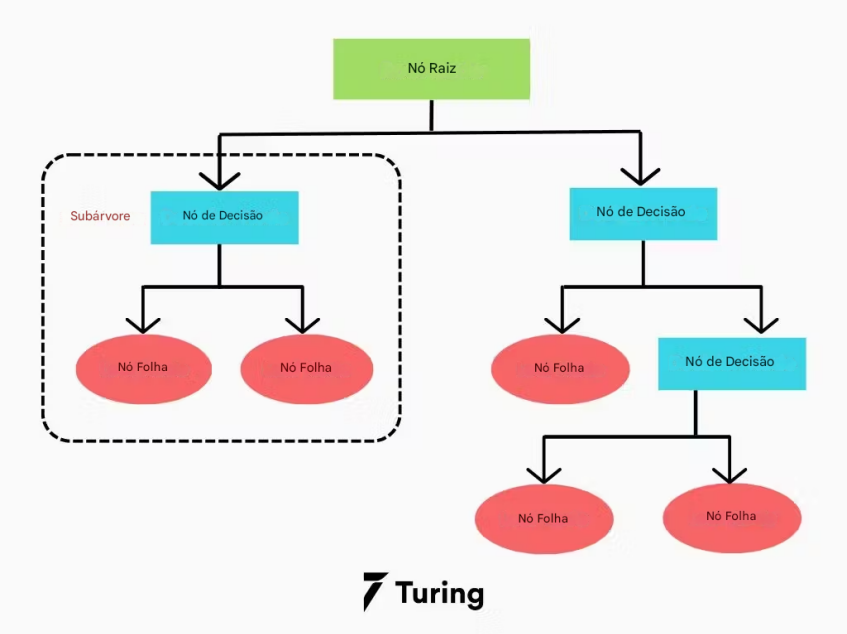
> A árvore de decisão é um algoritmo de aprendizado de máquina supervisionado usado para tarefas de classificação e regressão. É uma técnica poderosa e amplamente utilizada, especialmente por sua interpretabilidade e facilidade de uso. A ideia fundamental por trás da árvore de decisão é criar um modelo preditivo em forma de árvore, na qual cada nó interno representa uma decisão com base em um atributo específico, e cada folha representa uma classe ou valor de saída. (SANTOS, 2023, p. 42)

Ele também explica seu funcionamento, como:

> O algoritmo busca dividir o conjunto de dados de forma recursiva com base nos atributos disponíveis. O objetivo é encontrar as divisões que maximizam a homogeneidade dos dados dentro de cada subconjunto resultante. Uma vez que a árvore de decisão tenha sido construída, ela pode ser usada para fazer previsões em novos exemplos.
>
> A árvore é percorrida de cima para baixo, seguindo os ramos correspondentes às decisões baseadas nos valores dos atributos. No final, a previsão é feita com base na classe ou valor associado à folha alcançada.
>
> Essa técnica possui algumas vantagens, incluindo a capacidade de lidar com dados numéricos e categóricos, além disso, ela é relativamente rápida de construir e pode lidar com grandes volumes de dados. (SANTOS, 2023, p. 43)

Considerando a explicação de maneira teórica sobre o funcionamento do algoritmo, apresento uma outra ótica do seu processo de funcionamento, através da Figura 2 abaixo, representada de maneira visual, a árvore inicia seu processo de classificação pelo nó raiz baseando-se em suas decisões de classificação por atributos até finalizar o conjunto de dados e formar uma árvore completa.

Figura 3 – Representação do Funcionamento do Modelo de Árvore.



Fonte: Adaptado de Turing (2024).

2.3.2 FLORESTA ALEATÓRIA

O Autor Santos (2023) Define o algoritmo de floresta aleatória como:

> A Floresta Aleatória (Random Forest) é um algoritmo que combina várias árvores de decisão individuais para realizar tarefas de classificação e regressão. Ela é conhecida por sua capacidade de lidar com problemas complexos e produzir resultados precisos. (SANTOS, 2023, pag. 47)

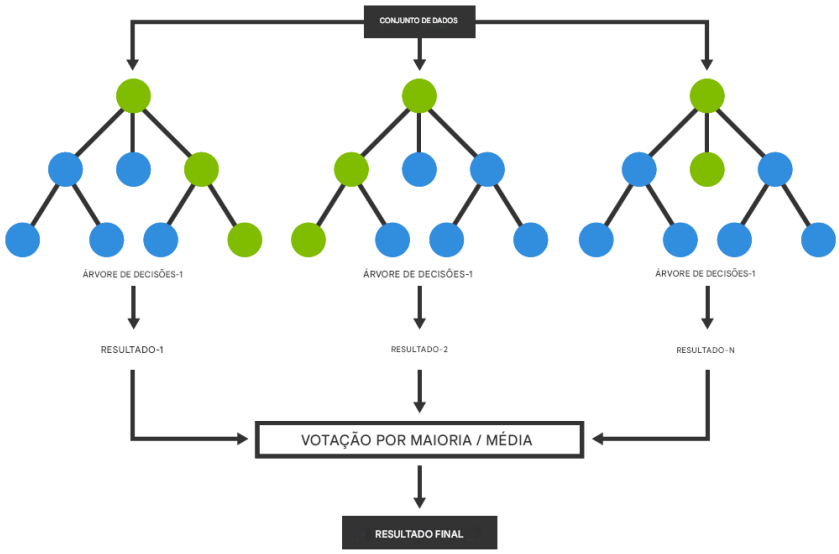
Ele também explica seu funcionamento, como:

> O algoritmo da Floresta Aleatória funciona construindo múltiplas árvores de decisão independentes entre si. Cada árvore é treinada em uma amostra aleatória do conjunto de dados de treinamento, chamada de amostragem de bootstrap, na qual exemplos são selecionados com substituição.
>
> Durante a fase de treinamento, cada árvore é expandida até sua capacidade máxima ou até que um critério de parada seja atingido. Em problemas de classificação, a árvore toma decisões com base na maioria das classes dos exemplos de treinamento nas folhas.
>
> Para fazer uma previsão com uma Floresta Aleatória treinada, cada árvore individual na floresta é percorrida e sua previsão é computada. No caso de classificação, a classe mais frequente entre as árvores é escolhida como a previsão final. (SANTOS, 2023, pag. 48)

Considerando a explicação de maneira teórica sobre o funcionamento do algoritmo, apresento uma outra ótica do seu processo de funcionamento, através da Figura 4 abaixo, representada de maneira visual, a floresta se inicia pela construção de diversas árvores (Conceituadas na seção anterior), construídas de maneira aleatória, devido ao treinamento realizado com amostras aleatórias do conjunto de dados, ademais, estas árvores são expandidas até sua capacidade

máxima. Por fim, a árvore que melhor representar o conjunto de dados ou obter o melhor valor de previsão, é a árvore escolhida no fim do treinamento.

Figura 4 – Representação do Funcionamento do Modelo de Floresta Aleatória.



Fonte: Adaptado de Gunay (2023).

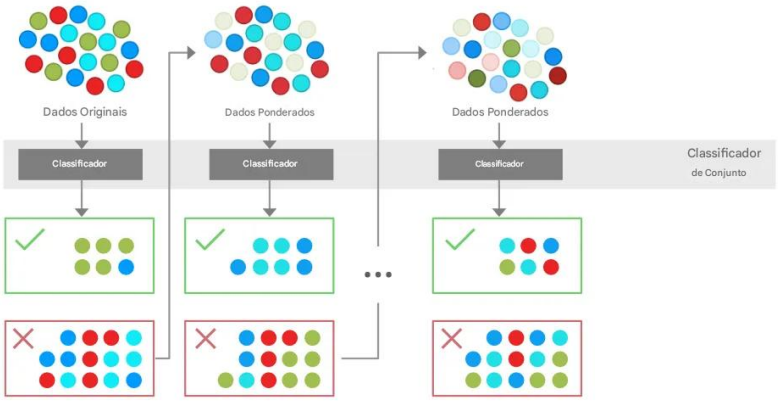
2.3.3 XGBOOST (Extreme Gradient Boosting) - Reforço de Gradiente Extremo

O XGBoost é um algoritmo de aprendizado de máquina, pertencente à família de algoritmos de boosting. Segundo Santos (2023), estes algoritmos funcionam de forma iterativa e a cada iteração do modelo gerado busca corrigir os erros do modelo gerado anteriormente, buscando melhorar seu desempenho e reduzindo a taxa de erro a cada iteração através de uma nova versão do modelo, gerada a partir de uma combinação de vários modelos individuais previstos anteriormente.

Ainda sim, Santos (2023) explica que o XGBoost, em especial, utiliza o conceito de regularização para tratamento eficiente dos recursos. Este conceito permite evitar o overfitting ao classificar, aumenta sua taxa de aprendizado, números de sub-árvores. Além disso, este modelo garante liberdade ao cientista de dados ajustar seus hiperparâmetros a fim de melhorar seu desempenho de classificação ou regressão, permitindo ao cientista ajustar o algoritmo ao seu problema.

Considerando a explicação de maneira teórica sobre o funcionamento do algoritmo, apresento uma outra ótica do seu processo de funcionamento, através da Figura 5 abaixo, representada de maneira visual. O algoritmo de XGBoost, realiza a classificação de forma iterativa, ele realiza uma primeira classificação, a partir dos erros da classificação inicial, ele utiliza estes dados como entrada para a próxima árvore gerada, com objetivo de minimizar a taxa de erro do modelo, executando este processo até uma condição de parada ou percorrer toda a base de dados.

Figura 5 – Representação do Modelo de XGBoost (Extreme Gradient Boosting).



Fonte: Adaptado de Verma (2022).

2.4 CIÊNCIA DE DADOS

Segundo artigo de Engineering e Sciences (2024), a ciência de dados pode ser definida como uma área de estudo que utiliza processos, técnicas e métodos científicos para transformar dados brutos em conhecimento e *insights*. Neste sentido, o artigo ainda destaca que a ciência de dados é uma área interdisciplinar que combina conhecimentos de estatística, ciência da computação e conhecimento de domínio. Isso a torna muito versátil, com aplicações que abrangem desde à saúde até pesquisas ambientais.

Já o autor Cao (2017) em seu artigo, traz uma definição formal e prática acerca da ciência de dados, ele a define como um novo campo interdisciplinar que sintetiza e se baseia a partir das áreas de estatística, informática, computação, comunicação, gestão e sociologia para possibilitar o estudo dos dados e seu ambiente, acrescidos de outros domínios e aspectos contextuais, como os organizacionais e sociais. Além disso, ele expõe que a ciência de dados tem por objetivo a transformação dos dados em *insights* e decisões, a partir de um pensamento e metodologia dos dados. Ela é dada pela fórmula:

Figura 6 – Fórmula elaborada pelo autor Cao (2017) sobre ciência de dados.

$$\begin{aligned} \text{Ciência de Dados} = {} & \text{Estatística} + \text{Informática} + \text{Computação} && (1) \\ & + \text{Comunicação} + \text{Sociologia} + \text{Gestão} && (2) \\ & \mid \text{dados} + \text{ambiente} + \text{pensamento} && (3) \end{aligned}$$

Onde $\mid$ significa “condicional a”.

2.4.1 OS PILARES DA CIÊNCIA DE DADOS

Como definido anteriormente, a Figura 7 abaixo, representa os quatro pilares fundamentais para a ciência de dados, eles referem-se às suas áreas de conhecimento envolvidas. Dentre as quais, se referem ao ”conhecimento de domínio”, ”habilidades em matemática”, ”ciência da computação”e ”habilidade em comunicação”.

O autor Rout (2025), detalha cada um destes quatro pilares básicos como:

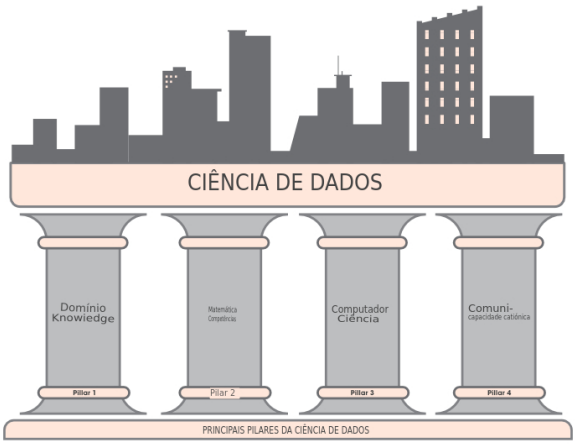
Conhecimento de domínio: O conhecimento de domínio é referente ao lado comercial da empresa, seu modelo de negócios e a relação dos dados com a área de atuação daquela empresa. Se faz necessário o estudo e uso de ferramentas para obtenção de informações sobre o negócio da empresa.

Habilidades em Matemática: A relação entre a ciência de dados e a matemática se faz presente no uso dos algoritmos de aprendizado de máquina, modelagem matemática do problema e construção destes algoritmos complexos. Seus conhecimentos incluem, álgebra linear, cálculo, técnicas de otimização, por fim, estatística e probabilidade.

Ciência da Computação: A relação entre a ciência da computação e a ciência de dados envolve a compreensão e modelagem computacional dos modelos, além de conhecimentos em programação, banco de dados, aprendizado de máquina e computação distribuída.

Habilidades de Comunicação: A apresentação dos resultados, seja de maneira verbal ou escrita, é crucial para um projeto em ciência de dados, especialmente quando esta apresentação não dialoga diretamente com o time de tecnologia ou engenharia. Devida sua multidisciplinaridade e forte relação com o time de negócios da organização, utilizar técnicas de apresentação e habilidades de comunicação é muito importante.

Figura 7 – Os 4 Pilares da Ciência de Dados.



Fonte: Adaptado de Rout (2025).

3 METODOLOGIA

Neste capítulo, será apresentado o fluxo de desenvolvimento do presente trabalho. A metodologia utilizada neste trabalho, chamada de ciclo da ciência de dados ou CRISP-DM (Cross-Industry Standard Process for Data Mining) [Processo padrão intersetorial para mineração de dados], em resumo, o CRISP surgiu para aplicações em mineração de dados, porém, seu funcionamento é amplamente utilizado em projetos de ciência de dados.

3.1 CRISP-DM

Segundo o autor Rivo, Fuente e others. (2012), o projeto CRISP-DM foi desenvolvido em 1996 por três players importantes do recém iniciado mercado de mineração de dados, os quais são: Daimler-Chrysler (Daimler-Benz), SPSS (ISL) e NCR. Este projeto foi desenvolvido e validado como um modelo em forma de etapas sequenciais do processo de mineração de dados. Ele pode ser aplicado em uma ampla gama de setores da economia, o Crisp garante o desenvolvimento de projetos de mineração de dados mais rápidos, baratos e confiáveis.

Conforme expresso na Figura 8 abaixo, o ciclo de desenvolvimento CRISP-DM, também conhecido como ciclo da ciência de dados é acompanhado pelas etapas de compreensão do problema e negócio, compreensão dos dados, preparação e organização dos dados, análise e modelagem computacional, validação e apresentação ou implantação. Cada uma destas etapas, serão detalhadas a seguir.

Figura 8 – Representação do Modelo CRISP. Fonte: Adaptado de Souza (2023).

3.1.1 ENTENDIMENTO DO NEGÓCIO

O autor Wirth e Hipp (2000) em seu artigo, define a etapa de entendimento do negócio, como uma fase inicial do processo que tem por objetivo realizar a compreensão dos

objetivos e requisitos do projeto, partindo de um ponto de vista empresarial, convertendo-o em uma definição de problema.

3.1.2 ENTENDIMENTO DOS DADOS

Segundo o autor Wirth e Hipp (2000), a etapa de entendimento dos dados inicia com a coleta inicial dos dados, seguido de atividades cujo objetivo é a familiarização com os dados, identificando seus problemas, como, qualidade destes dados, ausência e tipos dos dados. Ao realizar estas etapas, é identificado as informações iniciais sobre aqueles dados.

3.1.3 PREPARAÇÃO DOS DADOS

Ainda sim, o autor Wirth e Hipp (2000), informa que a etapa de preparação dos dados abrange todas as atividades para construir um conjunto de dados final (dados que serão alimentados nas ferramentas de modelagens). Ele, ainda sim, informa que tarefas como, seleção de tabelas, registros e atributos, limpeza dos dados, construção de novos atributos e transformação sobre os dados são necessárias para minimização de erros na etapa de modelagem.

3.1.4 ANÁLISE E MODELAGEM

O autor Wirth e Hipp (2000), informa que a etapa de modelagem inclui a utilização de diversas técnicas, elas são selecionadas e aplicadas de acordo com a necessidade de cada projeto, os parâmetros de cada técnica devem ser calibrados para seus valores ideais. Ele ainda diz que de modo geral, existem diversas técnicas para modelagem do mesmo tipo de problema, dentre elas, algumas exigem determinada padronização ou formatação dos dados. Ainda sim, Wirth (2000) comenta que a etapa de preparação dos dados é estreitamente relacionada com a de modelagem.

3.1.5 VALIDAÇÃO DO MODELO

O autor Wirth e Hipp (2000), explica que a etapa de validação é uma fase do projeto que permite a avaliação dos modelos criados na etapa anterior, ele ainda explica que nesta etapa, é muito importante realizar a re-avaliação e revisão detalhada do desempenho do modelo, antes de implantá-lo. Ele ainda explica que além da etapa técnica de modelagem, é importante verificar questões de negócios que possam não ter sido consideradas.

3.1.6 APRESENTAÇÃO DO RESULTADO

Por fim, mas não menos importante, Wirth e Hipp (2000) explica que a criação do modelo não implica (na maioria das vezes) no fim do projeto, este conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente seja capaz de interpretá-lo e utilizá-lo. Sendo assim, a fase de implantação resulta em um simples relatório ou implementação de

um novo recurso, vale ressaltar que esta escolha dependerá do processo e necessidade de cada projeto.

4 RESULTADOS

Como se trata de um dataset sobre fraude de cartão de crédito, é essencial preservar os dados originais para garantir a segurança da informação. No entanto, dadas as especificidades do dataset utilizado, não foi possível obter os melhores resultados apenas por meio da análise exploratória. Além disso, é importante frisar que se trata de um conjunto de dados que contém os registros de transações com cartão de crédito por titulares europeus ao longo do ano de 2023. Para isso, é utilizado o conceito de anonimidade dos dados, ou seja, para preservar os dados, os dados foram processados por uma função que permite sua representação anonimizada, garantindo a privacidade e segurança dos dados.

O dataset, criado por Elgiriyewithana (2023) possui aproximadamente 570 mil registros e 31 colunas, sendo elas: ID, V1-V28, Montante e Classe. As colunas V1 a V28 são as colunas anonimizadas, nelas não existem nomes e os dados são todos do tipo float que variam de -70 à +220. Além disso, elas representam dados como nome, sobrenome, endereço, números de telefone, hora, localização da transação, etc. A coluna Montante representa o valor da transação e a Classe representa se ela é classificada em Fraude ou Não Fraude.

Apesar dos desafios práticos, foi realizada a análise exploratória completa, mesmo sem encontrar os melhores resultados, esta etapa é fundamental no ciclo da ciência de dados, é nela que elaboramos nossas hipóteses acerca do problema. Se na análise exploratória, nossos resultados foram complexos, quando partimos para os resultados dos modelos, eles foram muito promissores e se mostraram bastante eficientes para construção de uma solução ao tema abordado.

4.1 RESULTADOS DA ANÁLISE EXPLORATÓRIA

A análise exploratória deste problema apresentou desafios devido à natureza dos dados, apesar de seguir os princípios da estatística descritiva e tentando extrair o máximo de informações do dataset, devido à natureza destes dados, não foi possível alcançar uma solução perfeita, mas que mesmo assim, não houve grande impacto sobre o resultado dos modelos.

4.1.1 CONHECENDO OS DADOS

Antes de iniciarmos a análise exploratória e a aplicação da estatística descritiva, é importante entender sobre os dados, seus tipos, seus valores e uma representação da base de dados trabalhada. Neste sentido, o primeiro comando que executamos é uma visualização do *dataframe* (uma estrutura de dados da biblioteca de tratamento e manipulação de dados do python pandas). Esta estrutura de *dataframe*, permite que os dados sejam armazenados de maneira bidimensional e tabular, ou seja, são armazenados em formas de linhas e colunas, semelhante a uma tabela excel. O *dataframe* facilita a manipulação e análise dos dados de maneira eficiente, com isso, podemos ver na imagem abaixo, uma apresentação breve da base de dados contendo suas primeiras e últimas 5 linhas, além de uma amostra próxima do número de colunas do dataframe.

A partir de agora, a base de dados será chamada de dataframe. Como podemos ver na Tabela 1 abaixo.

Tabela 1 – Análise Geral do Dataset

id	V1	V2	V3	V4	V5	…	V28	Amount	Class
0	-0.260648	-0.469648	2.496266	-0.08374	0.129681	…	-1.051045	17982.10	0
1	0.985100	-0.356045	0.558506	-0.429654	0.277104	…	-0.654515	6531.37	0
2	-0.260272	-0.949385	1.728538	-0.457966	0.074062	…	-2.244718	2513.54	0
3	-0.152152	-0.508959	1.746840	-1.090178	0.249486	…	0.048424	5384.44	0
4	-0.206820	-0.165280	1.527053	-0.448293	0.106125	…	0.419117	14278.97	0

Além desta visualização do dataframe, outro comando importante do pandas que permite analisar a base de dados é o info, ele permite identificar dados gerais do dataframe, como o número de colunas, número de registros e valores das colunas do dataset (Base de Dados).

A Tabela 2 nos permite observar informações sobre as 31 colunas do *dataframe*, incluindo a quantidade de valores não nulos e o tipo de dado de cada uma. A verificação da quantidade de valores não nulos é fundamental para o tratamento e o pré-processamento dos dados, pois colunas com muitos valores ausente podem comprometer o desempenho de modelos de aprendizado de máquina, que não lidam bem com dados faltantes. Nesses casos, seria necessário aplicar técnicas de manipulação ou imputação de dados. No entanto, no nosso caso específico, isso não é um problema, uma vez que todas as colunas possuem valores completos.

Tabela 2 – Análise Info Dataset

#	Coluna	Não Nulo	Tipo
0	id	568630	int64
1	V1	568630	float64
2	V2	568630	float64
3	V3	568630	float64
4	V4	568630	float64
5	V5	568630	float64
6	V6	568630	float64
7	V7	568630	float64
8	V8	568630	float64
9	V9	568630	float64
10	V10	568630	float64
11	V11	568630	float64
12	V12	568630	float64
13	V13	568630	float64
14	V14	568630	float64
15	V15	568630	float64
16	V16	568630	float64
17	V17	568630	float64
18	V18	568630	float64
19	V19	568630	float64
20	V20	568630	float64
21	V21	568630	float64
22	V22	568630	float64
23	V23	568630	float64
24	V24	568630	float64
25	V25	568630	float64
26	V26	568630	float64
27	V27	568630	float64
28	V28	568630	float64
29	Amount	568630	float64
30	Class	568630	int64

4.1.2 VERIFICAR BALANCEAMENTO DO DATASET

O próximo passo na análise exploratória, depois de entender o Dataset, é verificar se ele está balanceado. Segundo Castro e Braga (2011), o balanceamento de um dataset acontece quando a distribuição das classes está proporcional, e o desbalanceamento acontece quando existe uma desigualdade entre essas proporções. Um exemplo simples é quando se estuda uma doença rara em uma população, onde a quantidade de pessoas sadias será muito maior do que a quantidade de pessoas doentes, representando assim um desbalanceamento entre as classes.

Ou seja, a classe preditora (doença sobre uma população) vai ter uma diferença grande entre os valores 0 [FALSE], que seriam as pessoas sadias, e os valores 1 [TRUE], que seriam as pessoas doentes. Como a proporção de pessoas doentes é menor, isso mostra claramente

um desbalanceamento. Este tipo de situação causa problema nos algoritmos de aprendizado de máquina, porque os modelos acabam aprendendo mais a classe majoritária e cometem erros nas predições da classe minoritária, favorecendo sempre a que tem maior número de registros.

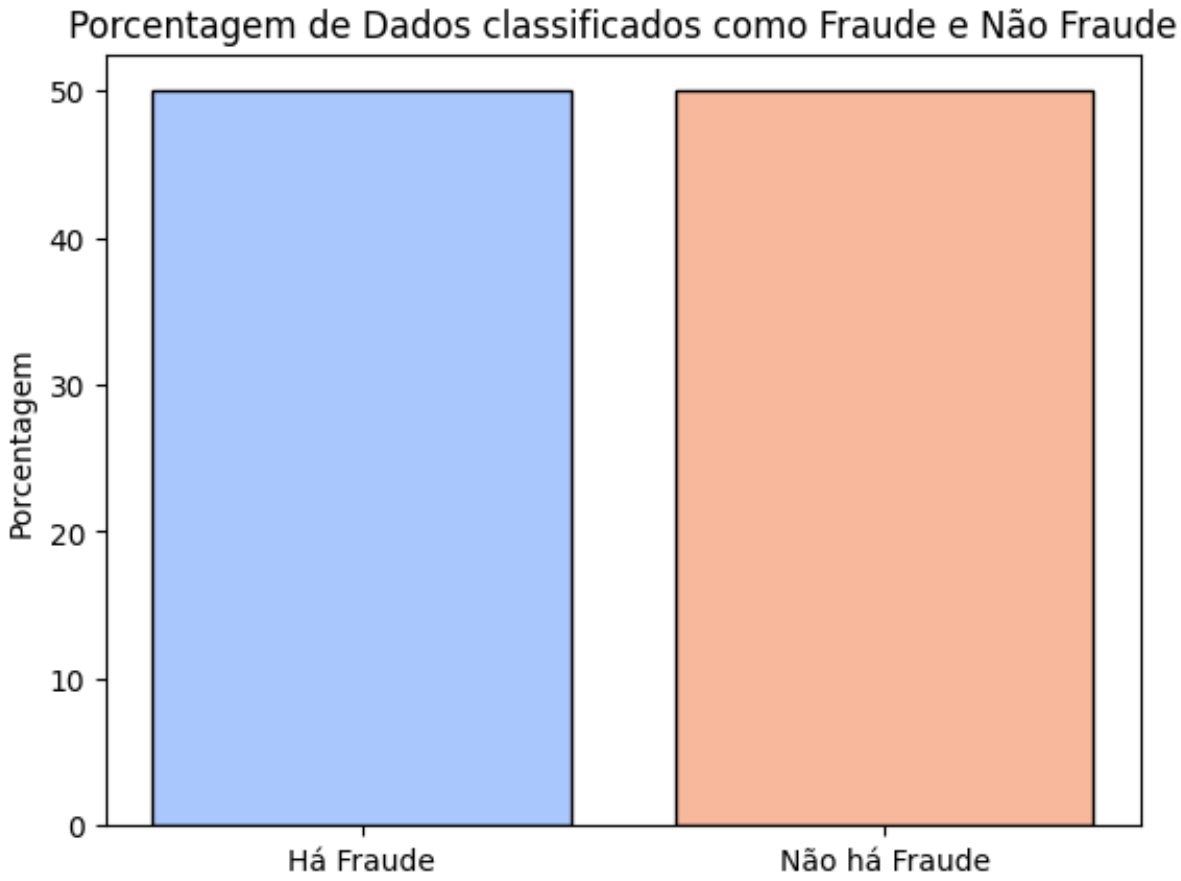
Por isso, realizar esta análise no dataset é essencial, já que dependendo do resultado pode ser necessário aplicar ajustes para equilibrar as classes e evitar classificações equivocadas. No caso deste trabalho, o Dataset estava balanceado, não exigindo nenhum tipo de ajuste. Como é possível observar no Gráfico 1, o dataset possui exatamente 50% das classes, classificadas como “Fraude” e 50% das classes classificadas como “Não Fraude”, o que confirma que o dataset está balanceado.

Quando um dataset não está balanceado, é preciso aplicar técnicas que permitam corrigir essa desigualdade para que os algoritmos consigam aprender de forma mais justa. Segundo Castro e Braga (2011), este tipo de problema pode ser tratado através de métodos de balanceamento de dados, que buscam aproximar a proporção entre a classe majoritária e a classe minoritária.

Um dos métodos mais conhecidos é o Oversampling, que aumenta a quantidade de exemplos da classe minoritária, replicando ou criando novos registros artificiais para que ela fique mais próxima da classe majoritária. Já o Undersampling faz o contrário, reduzindo os exemplos da classe majoritária para diminuir a diferença de proporção. Também existe a possibilidade de combinar os dois métodos, o que em muitos casos pode trazer um resultado mais equilibrado.

Sem esse tipo de correção, os modelos de aprendizado de máquina acabam favorecendo a classe majoritária, gerando predições equivocadas e pouco úteis. Por isso, quando identificado o desbalanceamento, aplicar técnicas de Oversampling e Undersampling se torna essencial para ajustar as classes e melhorar a qualidade da classificação, evitando que apenas uma classe seja priorizada pelo algoritmo.

Gráfico 1 – Análise Dataset Balanceado



Fonte: Autoria Própria

4.1.3 ESTATÍSTICA DESCRITIVA

A estatística descritiva é um dos pilares básicos para o processo de análise de dados. Guedes *et al.* (2005) define a estatística descritiva como, a área da estatística que se preocupa com a descrição dos dados. Além disso, tem por objetivo principal, sintetizar uma série de valores de mesma natureza, garantindo dessa forma que se tenha uma visão global da variação desses valores, organizar e descrever os dados em três maneiras: Tabelas, Gráficos e Medidas Descritivas. A forma de tabela resume o conjunto de observações, enquanto os gráficos são uma forma de apresentação dos dados.

Nesse contexto, dividiremos o estudo da estatística descritiva entre os seus principais conceitos básicos: Mìnima e Máxima, Medidas de Tendências Centrais e Medidas de Dispersão.

4.1.3.1 VALORES MÍNIMOS E MÁXIMOS

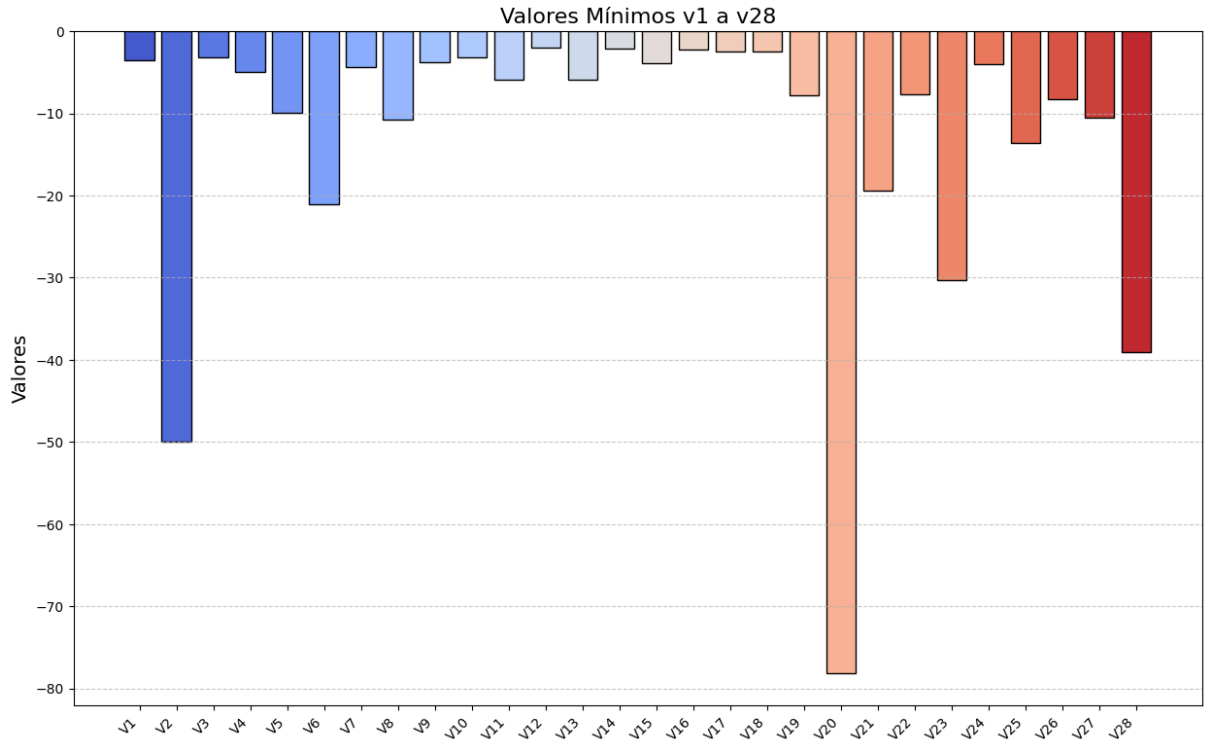
O conceito de mínimo e máximo, envolve a ordenação de valores numéricos. Organizados de maneira crescente, fornecendo informações preciosas. O termo “mínimo e máximo” representa o valor extremo dos conjuntos numéricos ordenados, ou seja, o menor e maior valor de uma ordenação numérica. A partir disso, foi dividido os gráficos dentre as principais colunas

e por uma questão de escala e distribuição, foram divididos das colunas V1 a V28 e a coluna Montante, de forma separada. Assim, foi feito para todos os gráficos da estatística descritiva.

Mínimos:

A partir do Gráfico 2 e da Tabela 3 abaixo, podemos perceber que referente aos valores mínimos no Gráfico 2 sobre as colunas V1-V28, é difícil ter definição clara sobre algo, especialmente por se tratar de uma base de dados com 570 mil registros. Já a Tabela 3 sobre a coluna montante obteve um valor que faça sentido para análise, o valor de 50.01 pode ser representado tanto como apenas uma transação comum de um item de "baixo" valor, como um valor de uma possível fraude.

Gráfico 2 – Valores Mínimos das Colunas V1 - V28



Fonte: Autoria Própria

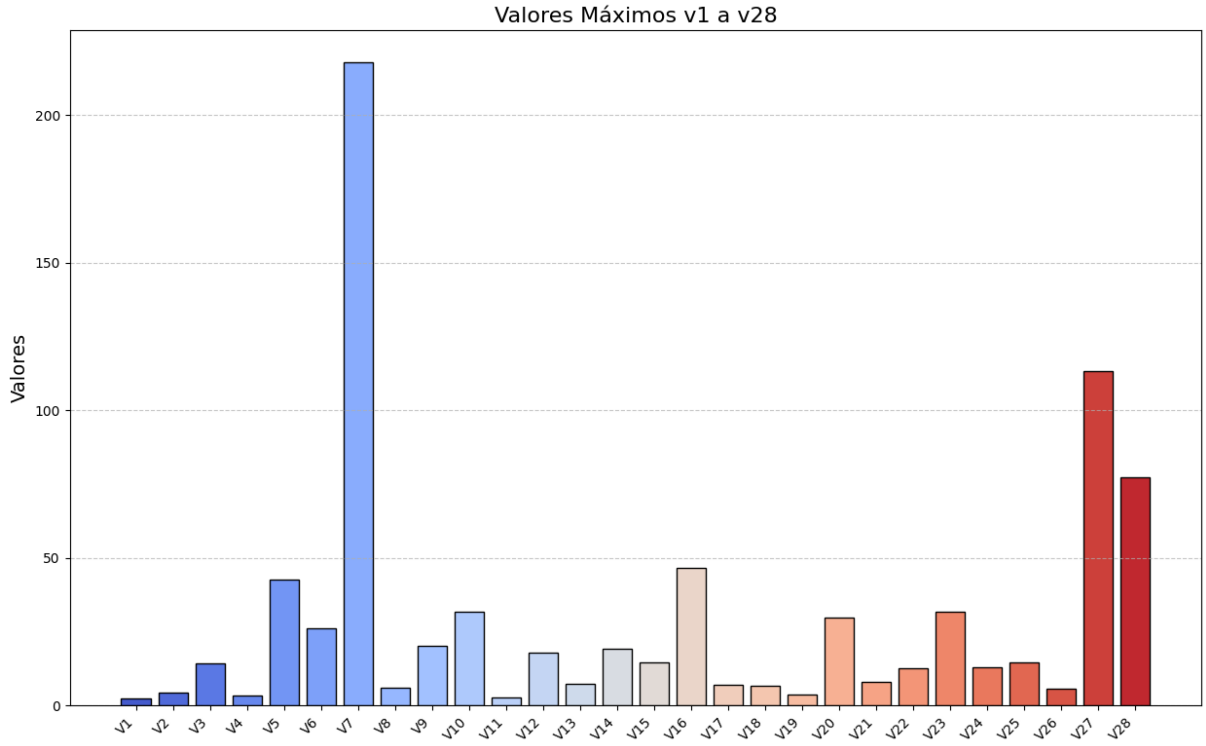
Tabela 3 – Valor Mínimo da Variável *Amount*

Estatística	Valor
Mínimo	50,01

Máximos:

A análise do Gráfico 3 e Tabela 5 abaixo sobre o valor máximo se assemelha bastante com a análise do valor mínimo, com a dificuldade de analisar o Gráfico 3 sobre as colunas V1-V28, e a tabela 4 sobre a coluna montante representando a única coluna de possível interpretação.

Gráfico 3 – Valores Máximos das Colunas V1-V28



Fonte: Autoria Própria

Tabela 4 – Valor Máximo da Variável *Amount*

Estatística	Valor
Máximo	24.039,93

4.1.3.2 MEDIDAS DE TENDÊNCIA CENTRAL

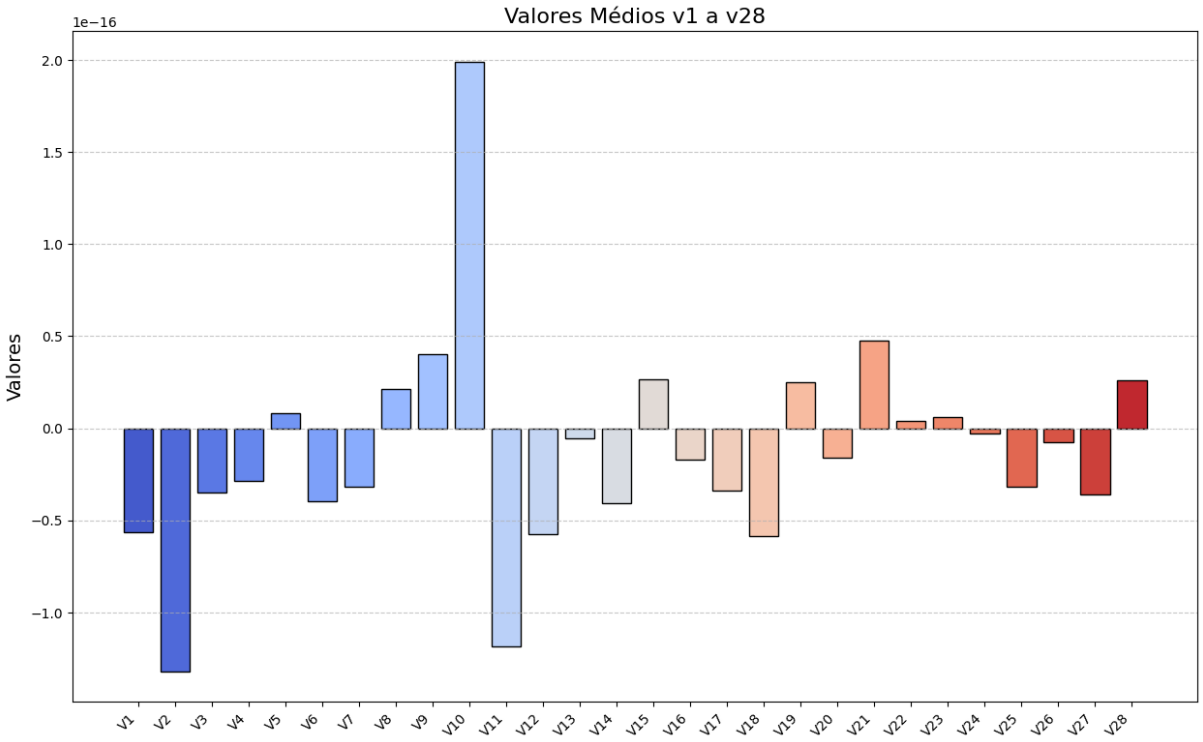
Os conceitos referentes a estas medidas é buscar alguns valores representativos ou recorrentes de um determinado conjunto de dados, dentre estas medidas estão contidos os conceitos de média, mediana e moda.

4.1.3.2.1 MÉDIA

A autora Guedes *et al.* (2005) define a média aritmética como, a soma de todos os valores observados da variável dividida pelo número total de observação. A média de um conjunto numérico representa o ponto de equilíbrio deste conjunto, é a medida de tendência central mais utilizada para representar a massa dos dados. É importante destacar que a média é altamente influenciada pelos outliers, não representando de maneira exata, o conjunto de dados.

Sobre a média, a sequência das métricas anteriores se perduram, com dificuldades de avaliar o Gráfico 4 sobre as colunas V1 - V28. Já a Tabela 5, referente à coluna Montante, permanece como a única a ser avaliada. Ainda assim, considerando a grande quantidade de registros no dataset e o fato de seus valores não serem elevados, observa-se a ausência de outliers. Como podemos ver no Gráfico 4 e Tabela 5 abaixo.

Gráfico 4 – Valores Médios das Colunas V1-V28



Fonte: Autoria Própria

Tabela 5 – Média da Variável *Amount*

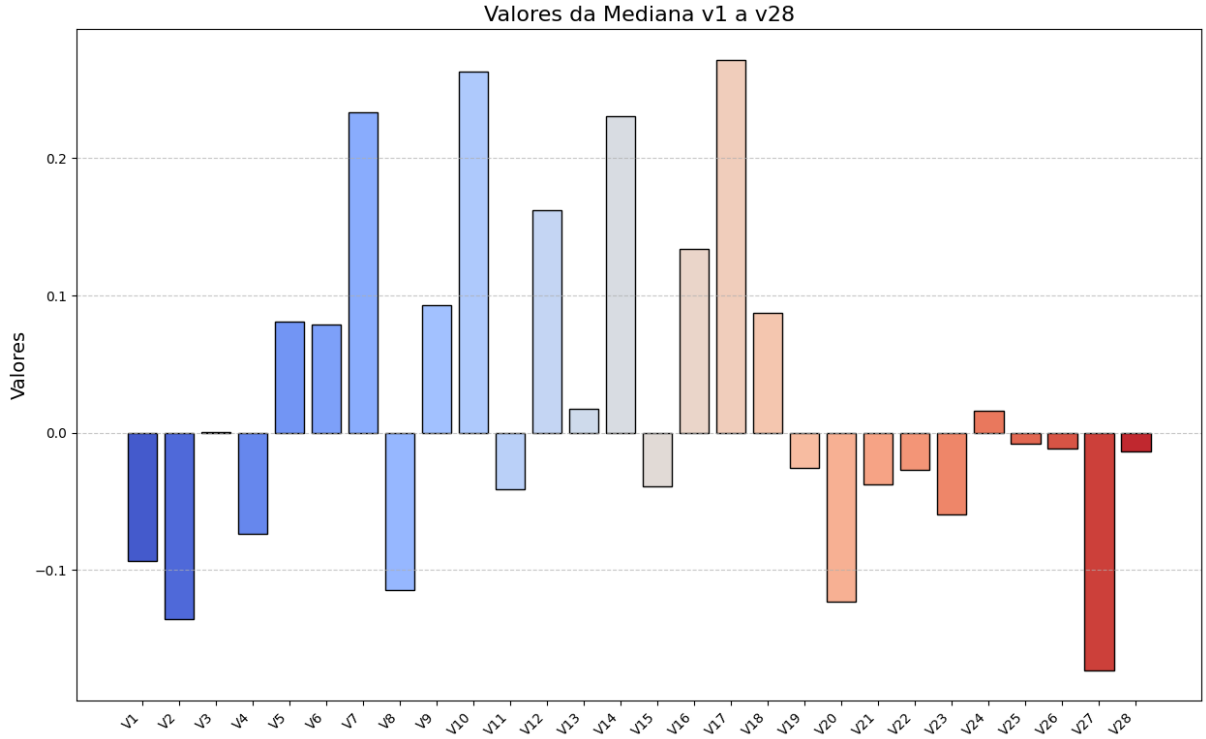
Estatística	Valor
Média	12.041,96

4.1.3.2.2 MEDIANA

A autora Guedes *et al.* (2005) define a mediana como o valor que ocupa a posição central da série de observação de uma variável, dividindo o conjunto em duas partes iguais, ou seja, a quantidade de valores inferiores à mediana é igual à quantidade de valores superiores a ela. É importante ressaltar que a mediana não é influenciada pelos outliers, representando de maneira mais fiel, o conjunto de dados.

Sobre a Mediana, podemos encontrar uma análise semelhante à da média, inclusive com a mediana muito próxima da média, o que representa que na base de dados não existem muitos outliers, pois caso exista um grande número de outlier, a média seria mais alta que a mediana, já que a mediana lida melhor com outlier. Como podemos ver no Gráfico 5 e Tabela 6 abaixo.

Gráfico 5 – Valores Mediana das Colunas V1-V28



Fonte: Autoria Própria

Tabela 6 – Mediana da Variável *Amount*

Estatística	Valor
Mediana	12.030,15

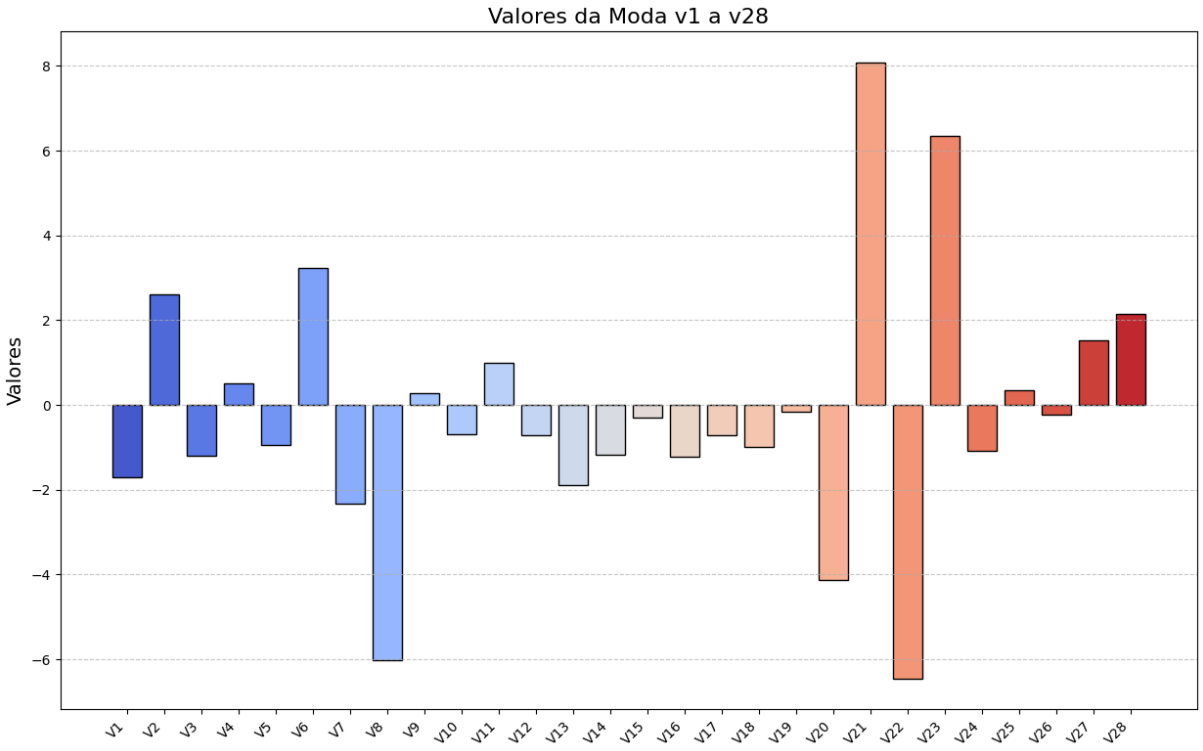
4.1.3.2.3 MODA

Por fim, Guedes *et al.* (2005) define a moda como o valor que apresenta maior frequência da variável entre os valores observados. Ou seja, o valor que mais se repete naquele conjunto de dados. A moda pode ser determinada imediatamente observando a frequência absoluta dos dados.

Sobre a moda, na Tabela 7 temos um número interessante de valores que mais se repetem, o primeiro ponto é que os 6 primeiros números que mais se repetem, estão com valores

muito próximos variando de $11.374,95 à $23.840,23, uma variação baixa para valores que se repetem, confirmando os valores e análise da média e mediana que o número de outliers seriam baixos. Como podemos ver na Tabela 7 e Gráfico 6 abaixo.

Gráfico 6 – Valores Moda das Colunas V1-V28



Fonte: Autoria Própria

Tabela 7 – Moda da Variável *Amount*

Estatística	Valores
Moda	70,64; 282,35; 9.178,59; 11.374,95; 11.505,10; 14.264,39; 14.773,74; 15.842,27; 18.953,05; 19.073,22; 19.600,59; 22.239,17; 22.481,52; 23.840,23

4.1.3.3 MEDIDAS DE DISPERSÃO

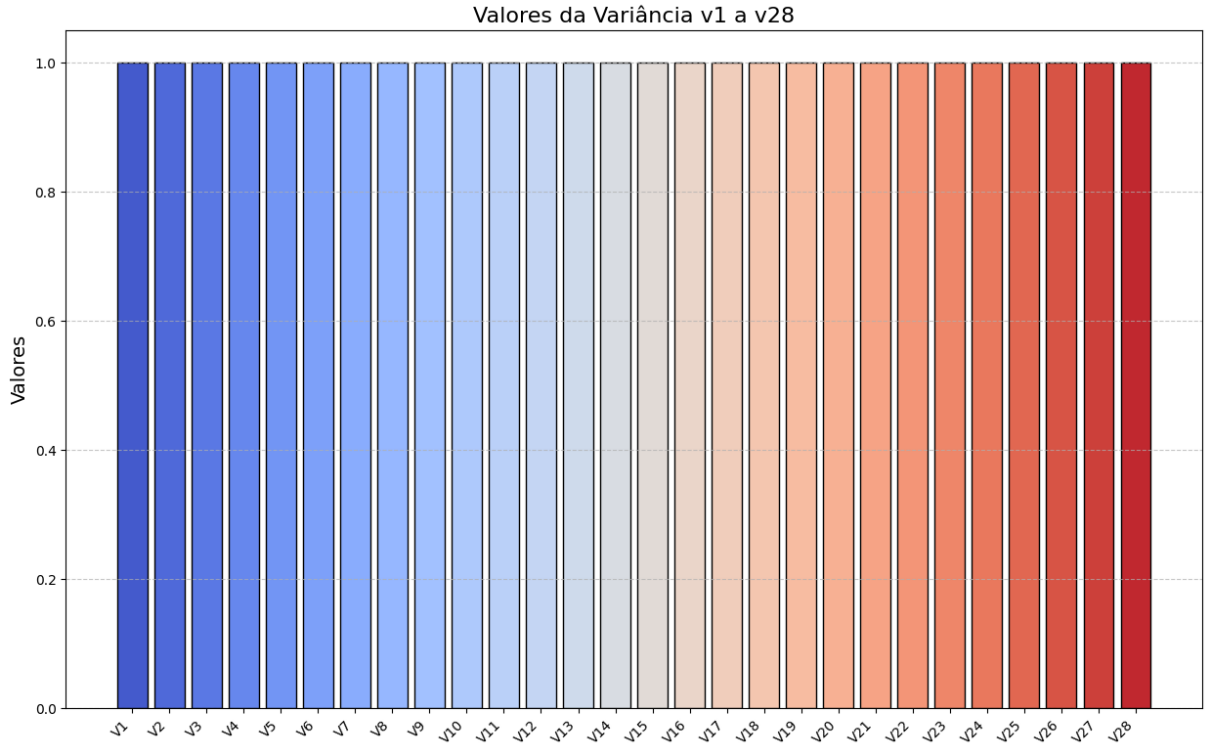
As técnicas que são abordadas nestas medidas buscam alguns valores que permitam quantificar um valor de diferença entre si. Dentre os conceitos, estão: Variância e Desvio Padrão.

4.1.3.3.1 VARIÂNCIA

O autor Granato e Kist (2018) define o conceito de variância como um valor que indica a extensão com que os resultados de uma base de dados se diferem, uns dos outros, sig-

nificando que quanto maior a variância, maior será a dispersão de valores que compõem este valor. Além disso, a variância é uma medida populacional e o Granato complementa que por se tratar de uma medida populacional, sua implementação não é comum, tornando-a um valor interessante para se utilizar em experimentos. Como podemos ver no Gráfico 7 e na Tabela 8 abaixo.

Gráfico 7 – Valores Variância das Colunas V1-V28



Fonte: Autoria Própria

Tabela 8 – Variância da Variável *Amount*

Estatística	Valor
Variância	47.881.479,31

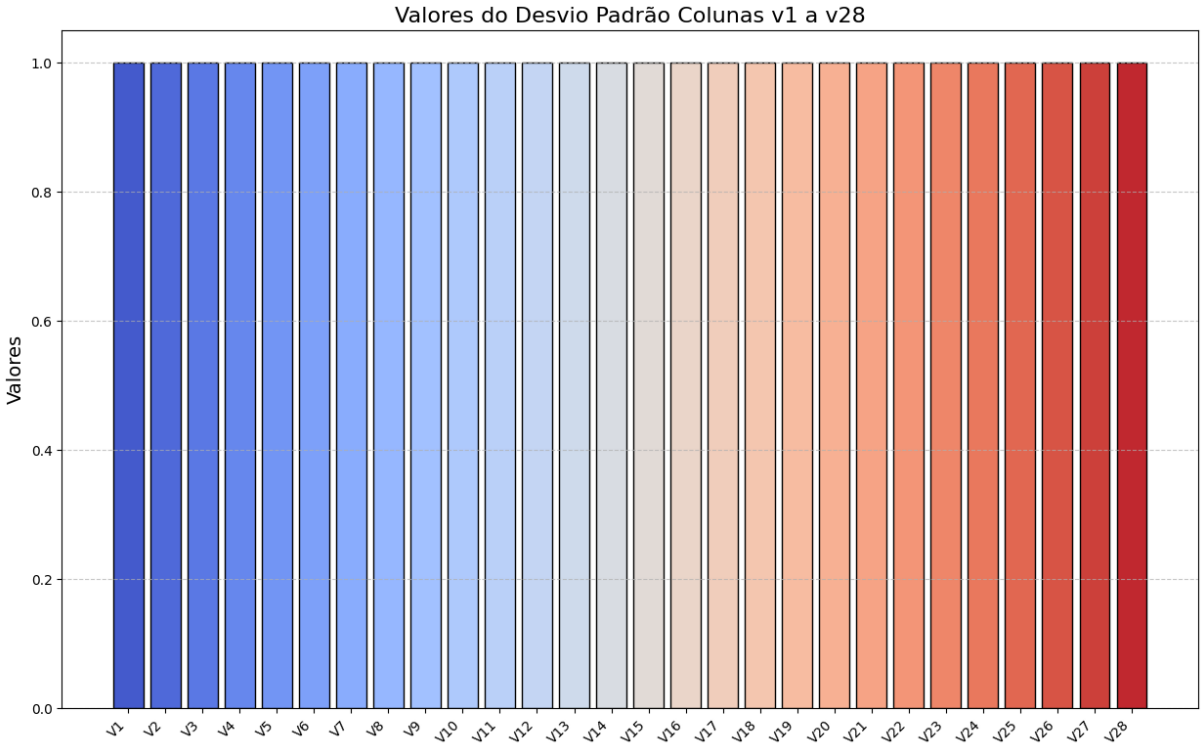
4.1.3.3.2 DESVIO PADRÃO

O autor Gouveia (s.d.) define o desvio padrão como uma medida que expressa o grau de dispersão do conjunto de dados. Ainda sim, ela complementa informando que quanto mais próximo de 0 é o desvio padrão, mais homogêneos são os dados do conjunto. Ela ainda complementa definindo o conceito de variância como uma medida de dispersão, também utilizada para expressar o quanto um conjunto de dados se desvia da média, a variância está relacionada com

o desvio padrão, sendo o desvio padrão definido como a raiz quadrada da variância. Seu uso é facilitado pela expressão na mesma unidade de medida dos dados.

Sobre a métrica de desvio padrão, nós temos um resultado que comprova toda esta análise, em especial, que o desvio padrão de todas as colunas do dataset, não são nada homogêneos, o que mais uma vez, auxilia e comprova toda a análise. Assim como a medida de variância, comprovando a heterogeneidade do conjunto de dados. Como podemos ver no Gráfico 8 e na Tabela 9 abaixo.

Gráfico 8 – Valores Desvio Padrão das Colunas V1-V28



Fonte: Autoria Própria

Tabela 9 – Desvio Padrão da Variável *Amount*

Estatística	Valor
Desvio Padrão	6.919,64

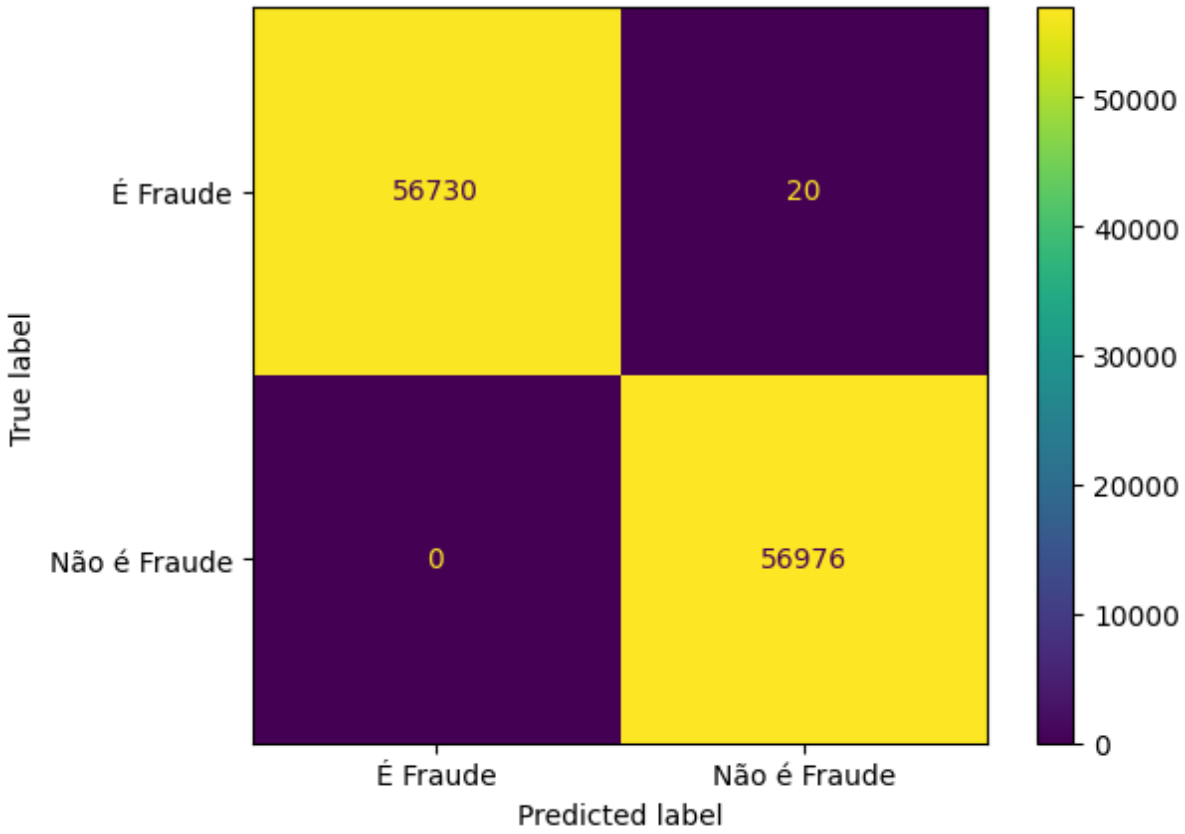
4.1.4 MATRIZ DE CORRELAÇÃO

Segundo Roque (s.d.), A medida de correlação pode ser definida como um tipo de medida que se usa quando se deseja saber a relação entre duas ou mais variáveis de um conjunto de dados, de maneira que quando uma varia, a outra varia também. Este conceito permite adquirir conhecimento sobre a análise do comportamento das variáveis do conjunto de dados, de maneira que a partir do valor de uma variável, é possível prever o valor da outra variável, de

acordo com a sua relação. A correlação pode ser positiva, igual a 0 e negativa. Dizemos que a correlação é positiva, quando uma variável tende a aumentar, a outra variável relacionada a ela, aumenta também. Já a correlação negativa, é quando uma variável tende a diminuir e a outra variável relacionada a ela tende a aumentar. Já a correlação igual a 0, indica que a variação entre uma das variáveis (aumento ou diminuição) não influencia a outra variável, relacionada à ela.

Sobre a Matriz de Correlação, representada pelo Gráfico **??** abaixo, podemos ver que algumas variáveis possuem maior conexão, ou uma conexão mais forte, umas com as outras, do que outras. Em um dataset com 31 colunas, ter este controle de correlação é extremamente importante para preparar o modelo apenas com variáveis relevantes ou com forte conexão, ou relação entre si. Como as variáveis: 'V1', 'V2', 'V3', 'V4', 'V6', 'V7', 'V9', 'V10', 'V11', 'V12', 'V14', 'V16', 'V17'.

Gráfico 9 – Correlação entre variáveis do Dataset



Fonte: Autoria Própria

4.2 RESULTADOS DO MODELO

Os algoritmos de aprendizado de máquina utilizados foram o de árvore de decisão (Decision Tree), floresta aleatória (Random Forest) e XGBoost (Extreme Gradient Boosting). O processo de escolha destes algoritmos foi o resultado de uma análise dentre as opções de algoritmos disponíveis e que melhor se ajusta ao problema ou base de dados. Todas as opções

escolhidas, são algoritmos modernos, eficientes e muito utilizados nos trabalhos acadêmicos para solucionar problemas de classificação, em especial, os algoritmos de Árvore e Floresta, são amplamente utilizados, corroborando para esta escolha.

Na etapa de modelagem, foram escolhidos 5 algoritmos de aprendizado de máquina, inicialmente foi utilizado 2 variações do algoritmo de decision tree, após, utilizamos sua versão aprimorada, a random forest, por fim, foram utilizados 2 variações do algoritmo XGBoost que é uma versão aprimorada da random forest.

O algoritmo de Support Vector Machine (SVM) foi testado como uma experiência exploratória, mesmo sabendo de suas limitações em relação às características da base de dados, como o alto número de registros e variáveis. A execução do modelo confirmou a expectativa de inviabilidade: após 30 horas de processamento, o treinamento ainda não havia sido concluído, evidenciando que o SVM não é uma abordagem adequada para este problema.

Vale ressaltar que todos os algoritmos e suas respectivas variações foram submetidos ao mesmo processo de pré-processamento, treinamento, predição e validação dos resultados. Se diferenciando apenas nas variações do algoritmo de árvore e do XGBoost.

No XGBoost, foram utilizado um algoritmo padrão, tal qual, se encontra sugerido na documentação da biblioteca, porém, este algoritmo permite ao longo do processo de treinamento e predição realizar ajustes ou sobre-ajustes em hiperparâmetros internos que otimiza e melhora seu desempenho, neste sentido, foi utilizado uma versão default e outra com esta "Modificação". Utilizar esta configuração com ajustes nos hiperparâmetros tem por objetivo buscar o melhor desempenho do modelo, que já era ótimo.

Já no algoritmo de árvore, foram utilizados 2 versões do mesmo algoritmo, se diferenciando apenas no que diz respeito a base de dados utilizada para treinamento, em uma versão, foram utilizadas todas as colunas do dataset para realizar seu treinamento e predição. Na outra versão, foi utilizado apenas as colunas relevantes de acordo com o gráfico de correlação, sua utilização não surtiu tanta diferença, por opção de baixa diferença nos resultados, a partir deste teste, foi utilizado o dataset com todas as colunas para treinamento e predição dos outros modelos utilizados. Porém, estes resultados serão explicados com maiores detalhes nos tópicos a seguir.

4.2.1 DECISION TREE

Este algoritmo, inicialmente, como parte fundamental do ciclo da ciência de dados, foi idealizado 2 variações de uso, para identificar qual se comportaria de melhor maneira ou teria melhor adaptação para a base de dados.

A primeira versão utiliza para treinamento e validação através do método de predição, todas as colunas úteis da base de dados, ou seja, de V1 a V28 + Montante, como as variáveis dependentes do modelo e a Classe, como variável independente, através disso, foi criado uma estrutura de data frame contendo todas estas colunas e facilitando todo o processo de manipulação e treinamento do modelo.

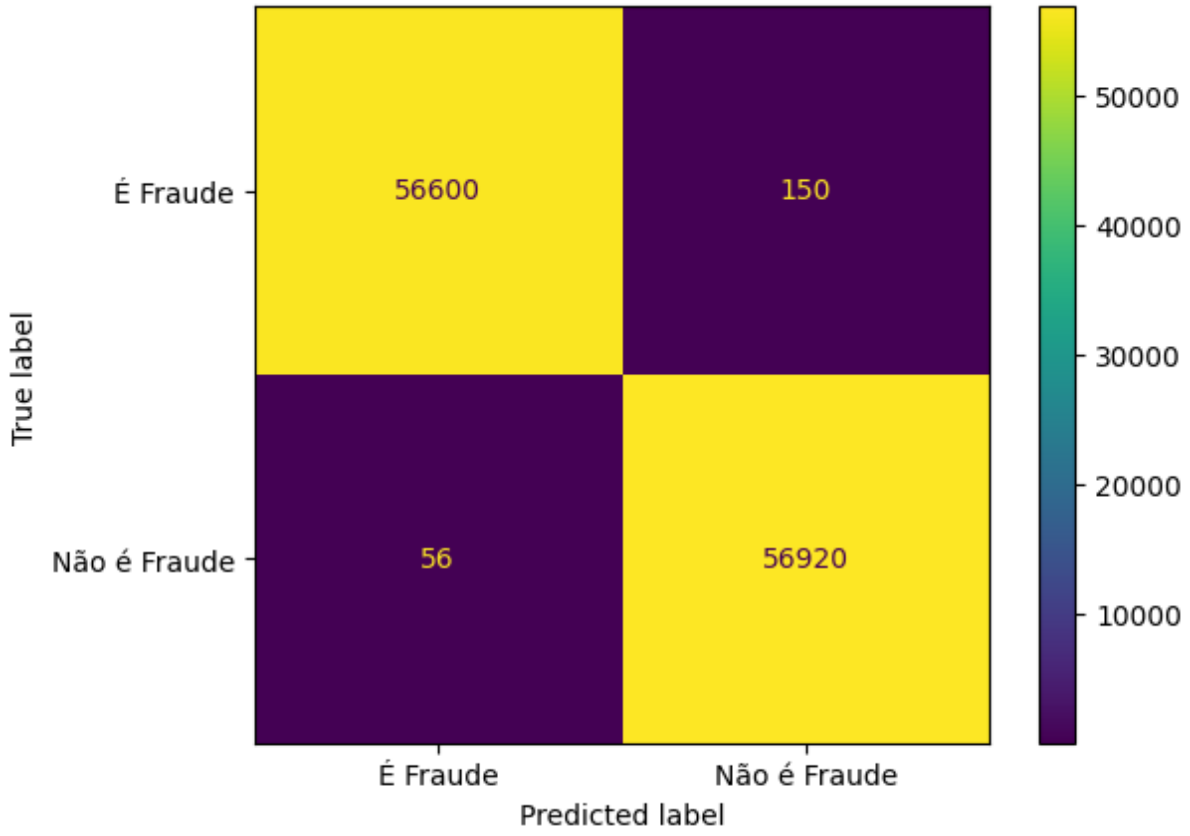
A segunda versão, utiliza para treinamento e validação do método de predição, as principais colunas que possuem grande valor de correlação entre si, reduzindo de maneira significativa as variáveis do modelo. Desse modo, foram utilizadas as colunas 'V1', 'V2', 'V3', 'V4', 'V6', 'V7', 'V9', 'V10', 'V11', 'V12', 'V14', 'V16', 'V17', 'Amount', como variáveis dependentes do modelo e a variável Classe como variável independente. A partir dessas colunas, foi criado uma estrutura de data frame, contendo todos estes dados e facilitando todo o processo de manipulação e treinamento do modelo.

Neste sentido, ambos algoritmos obtiveram desempenhos ótimos em todas as métricas avaliativas, se mostrando excelentes opções para o desenvolvimento de algoritmos e soluções para identificar e detectar fraudes de cartão de crédito. As métricas avaliadas são acurácia, precisão, recall, f1-score, curva roc-auc, matriz de confusão e a validação cruzada. Sendo representada, seus respectivos resultados abaixo:

Matriz de Confusão:

De acordo com a Matriz de Confusão apresentada no Gráfico 10, tendo o modelo classificado corretamente 56.596 dados como "Fraude", chamado verdadeiro positivo. Ele também classificou 56.920 dados como "Não é Fraude" corretamente, chamado de verdadeiro negativo. Porém, ele também classificou 154 dados erroneamente, como "Fraude", chamado de falso negativo. Por fim, ele classificou 56 dados erroneamente, como "Não é Fraude", chamado falso positivo.

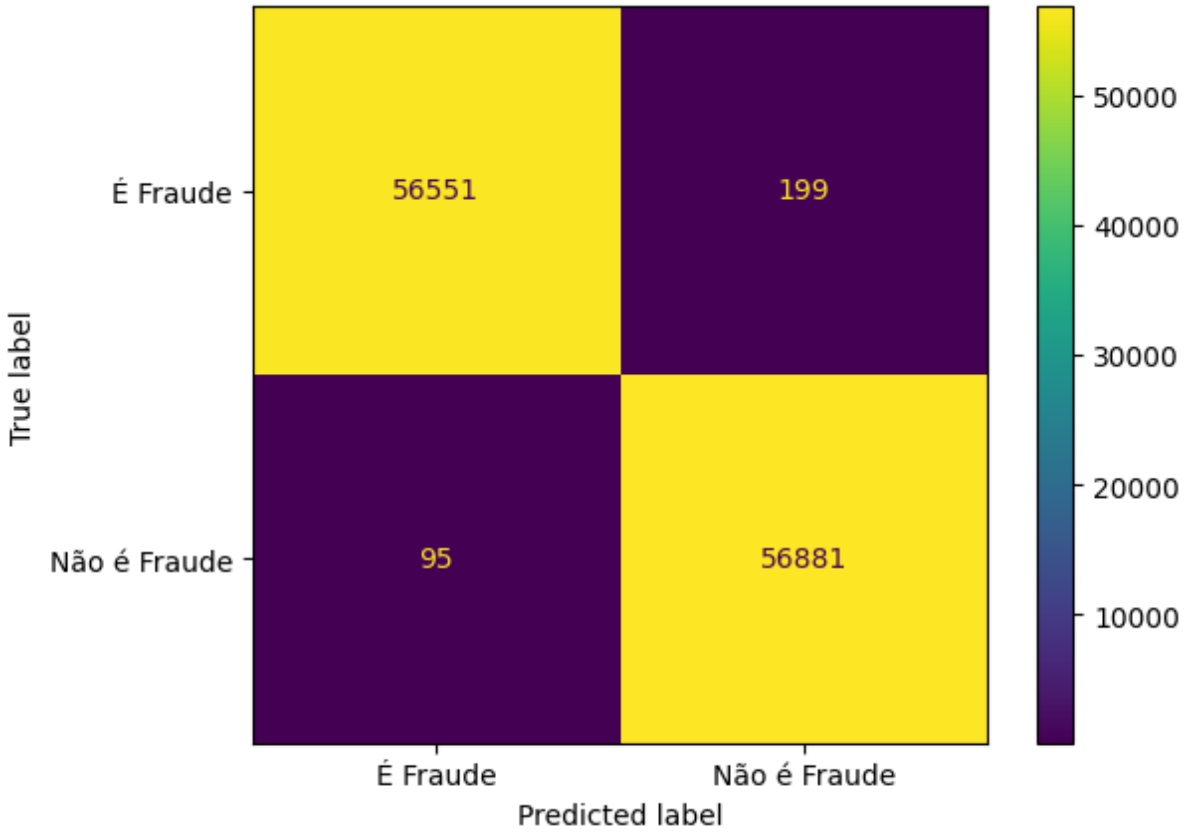
Gráfico 10 – Representação da Matriz de Confusão do Modelo com Todas Variáveis



Fonte: Autoria Própria

De acordo com a Matriz de Confusão apresentada no Gráfico 11, tendo o modelo classificado corretamente 56.561 dados como “Fraude”, chamado verdadeiro positivo. Ele também classificou 56.890 dados como “Não é Fraude” corretamente, chamado de verdadeiro negativo. Porém, ele também classificou 189 dados erroneamente, como “Fraude”, chamado de falso negativo. Por fim, ele classificou 86 dados erroneamente, como “Não é Fraude”, chamado falso positivo. Representando um desempenho inferior ao modelo com todas as variáveis, o que significa que mesmo as variáveis com baixa correlação, são importantes para o modelo realizar uma melhor qualidade de classificação.

Gráfico 11 – Representação da Matriz de Confusão do Modelo com as Variáveis Principais



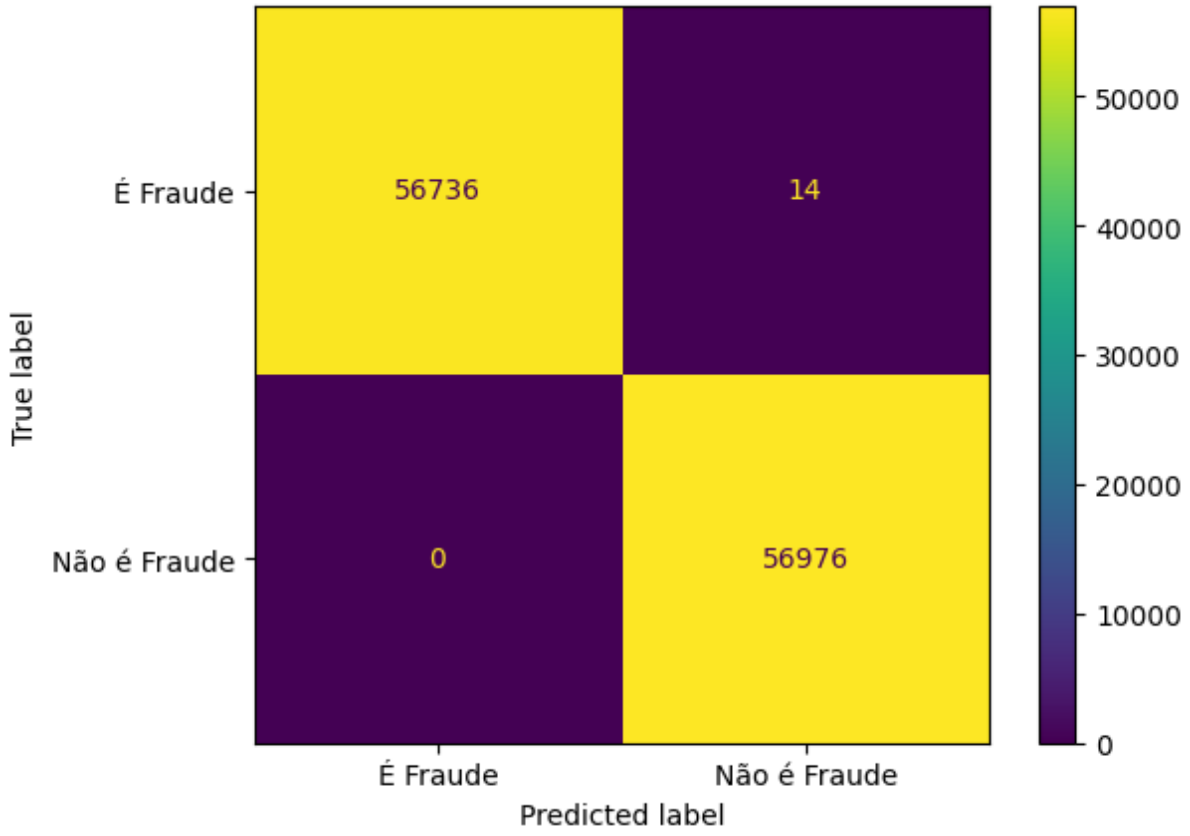
Fonte: Autoria Própria

4.2.2 RANDOM FOREST

O uso deste algoritmo no projeto, a partir do experimento do modelo de árvore com todas variáveis se mostrar mais eficiente e com melhor resultado, a partir dele, todos os modelos de aprendizado de máquina utilizarão o data frame com todas variáveis, pois acredito ser o de melhor resultado e adaptado à este problema.

De acordo com a Matriz de Confusão apresentada no Gráfico 12, é possível observar que a qualidade de classificação do modelo para este problema e conjunto de dados é bastante elevada, tendo o modelo classificado corretamente 56.736 registros como “Fraude”, também chamado de verdadeiro positivo. O modelo também classificou corretamente 56.976 registros como “Não é Fraude”, também chamado de verdadeiro negativo. O modelo classificou de maneira errônea, apenas 14 registros como “Não é fraude”, também chamado de falso negativo. O modelo não obteve classificação errada de falso positivo. Considerando os erros do modelo pela matriz de confusão, são erros mínimos, o que explica seu desempenho excelente para esta base de dados e problema proposto.

Gráfico 12 – Representação da Matriz de Confusão do Modelo



Fonte: Autoria Própria

4.2.3 XGBoost (Extreme Gradient Boosting)

No projeto, foram utilizadas duas variações do modelo, uma com os valores de hiperparâmetros default, outra com ajustes nos hiperparâmetros para tentar identificar qual é o limite de capacidade de classificação do modelo, buscando o mais próximo possível de erro igual a zero.

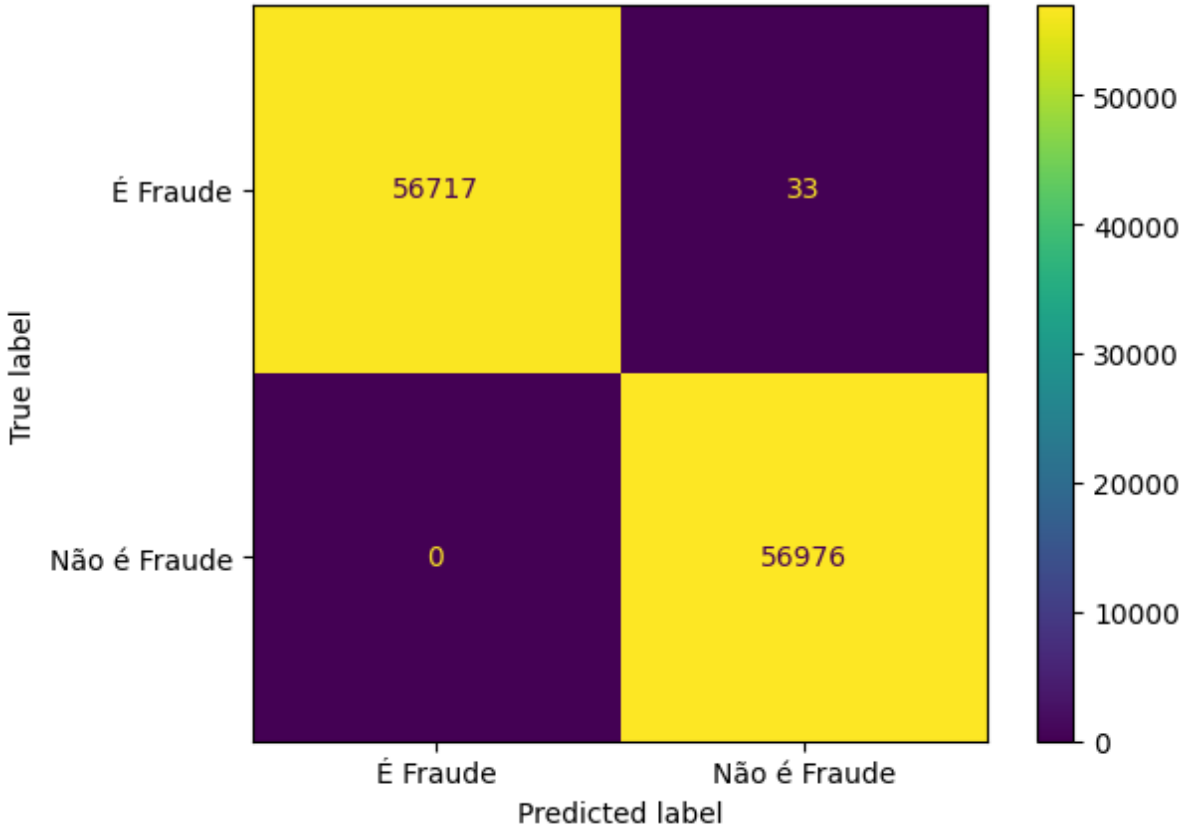
XGBoost Versão Padrão:

Após conceituado e explicado de maneira geral o funcionamento do algoritmo de aprendizado de máquina (XGBoost). Vamos ao uso deste algoritmo no projeto, apenas para efeito de comparação e análise de desempenho, esta versão do algoritmo utiliza a mesma implementação presente na documentação do modelo.

De acordo com a Matriz de Confusão apresentada no Gráfico **??**, é possível observar que a qualidade de classificação do modelo para este problema e conjunto de dados é bastante elevada, tendo o modelo classificado corretamente 56.717 registros como “Fraude”, também chamado de verdadeiro positivo. O modelo também classificou corretamente 56.976 registros como “Não é Fraude”, também chamado de verdadeiro negativo. O modelo classificou de maneira errônea, apenas 33 registros como “Não é fraude”, também chamado de falso negativo. O

modelo não obteve classificação errada de falso positivo. Considerando os erros do modelo pela matriz de confusão, são erros mínimos.

Gráfico 13 – Representação da Matriz de Confusão do Modelo Padrão



Fonte: Autoria Própria

XGBoost Versão Modificado:

No uso deste algoritmo no projeto, apenas para efeito de comparação e análise de desempenho, esta versão do algoritmo utiliza a mesma implementação presente na documentação do modelo, mas com seus hiperparâmetros modificados, representado pelo Algoritmo 1 abaixo.

Algoritmo 1 – Os hiperparâmetros modificados, são:

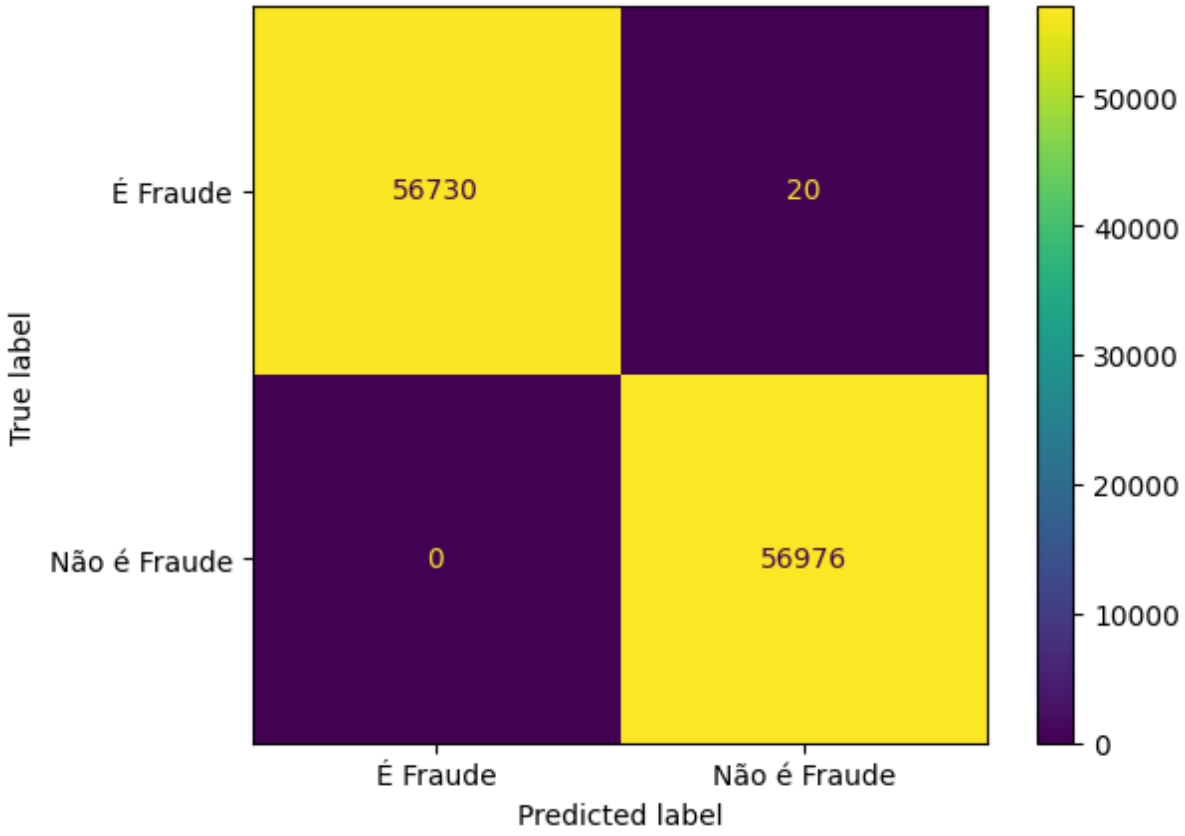
```
        # Parâmetros do modelo
        n_rounds = 5000 # Número máximo de iterações (árvores) no boosting
        base_Score = 0.5 # Taxa inicial de aprendizado

        params = {
            "max_depth": 10,  # Número máximo de ramificações
            "eta": 0.2,  # Taxa de aprendizado
            "objective": "binary:logistic",
            "eval_metric": "auc",
            "tree_method": "hist",
            "base_score": base_Score,
            "subsample": 0.8,  # Usa apenas 80% dos dados por árvore
            "colsample_bytree": 0.8  # Usa apenas 80% das colunas por
    árvore
        }

```

Fonte: Autoria Própria

É importante destacar que este ajuste nos hiperparâmetros do modelo lhe permite um ganho de desempenho, porém a forma com que ele é treinado não permite que seja realizado a etapa de validação cruzada dos dados, pois, este modelo não retorna o modelo.fit(), ele retorna um modelo do tipo boost que já é um modelo treinado. Devido a esta incompatibilidade, não é possível que seja realizada a validação cruzada. Sendo o único modelo deste estudo a não passar pela etapa de validação cruzada.

De acordo com a Matriz de Confusão apresentada no Gráfico **??**, é possível observar que a qualidade de classificação do modelo para este problema e conjunto de dados é bastante elevada, tendo o modelo classificado corretamente 56.730 registros como “Fraude”, também chamado de verdadeiro positivo. O modelo também classificou corretamente 56.976 registros como “Não é Fraude”, também chamado de verdadeiro negativo. O modelo classificou de maneira errônea, apenas 20 registros como “Não é fraude”, também chamado de falso negativo. O modelo não obteve classificação errada de falso positivo.

Gráfico 14 – Representação da Matriz Confusão do Modelo Modificado



Fonte: Autoria Própria

4.2.4 MÉTRICAS DE AVALIAÇÃO DOS MODELOS

A partir da matriz de confusão, são geradas as métricas de acurácia, precisão, recall, f1-score e curva AUC ROC e Validação Cruzada. Explicado sobre a matriz de confusão de cada modelo, vamos agora realizar uma análise gráfica sobre as métricas aplicadas a todos os modelos, a fim de identificar qual é o melhor modelo. Mas antes, os conceitos acerca das métricas anteriores serão explicadas com maiores detalhes, abaixo:

Acurácia:

O autor Junior *et al.* (2022) em seu artigo explica a acurácia, como uma medida de avaliação dos modelos de aprendizado de máquina que busca identificar e quantificar o percentual de casos verdadeiros classificados pelo modelo, a partir da matriz de confusão, em relação a todos os resultados classificados pelo modelo. Ele ainda destaca que ela é uma medida que nem sempre indica precisão do modelo.

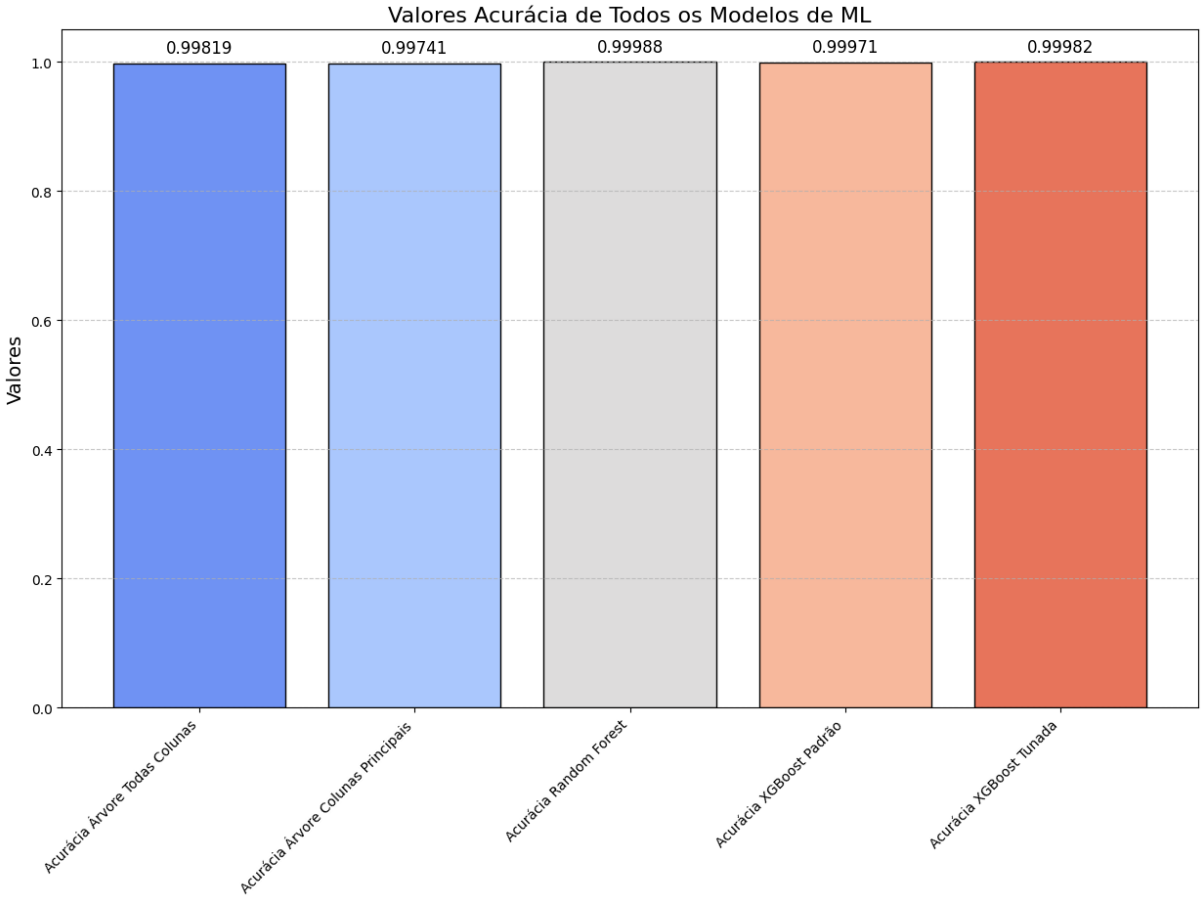
Analisando o Gráfico 15 plotado com o resultado da acurácia de todos os modelos utilizados, podemos observar alguns pontos interessantes, o primeiro e que complementa os resultados das matrizes de confusão, é o excelente resultado de todos os modelos, todos os

modelos ficaram pelo menos acima de 99%, se diferenciando apenas por algumas casas decimais entre eles.

Outro ponto interessante de se observar, também observado na matriz de confusão, foi que o algoritmo com todas as variáveis obteve um desempenho superior ao modelo com as principais variáveis, escolhidos a partir do gráfico de correlação. Além disso, podemos observar que mesmo realizando a modificação de hiperparâmetros no modelo de XGBoost, seu ganho de desempenho não foi tão elevado, se comparado ao modelo padrão, em especial, pensando no custo computacional da modificação, é um experimento que valeu para aprendizado e curiosidade, porém, se avaliar seu ganho real, não vale a pena o custo.

Por fim, o modelo que obteve melhor resultado e chamou atenção pelo custo computacional X desempenho foi o modelo de floresta aleatória, sem dúvidas um modelo muito eficiente e que obteve o melhor desempenho e com um custo computacional satisfatório.

Gráfico 15 – Gráfico Comparativo Acurácia de Todos Modelos Propostos



Fonte: Autoria Própria

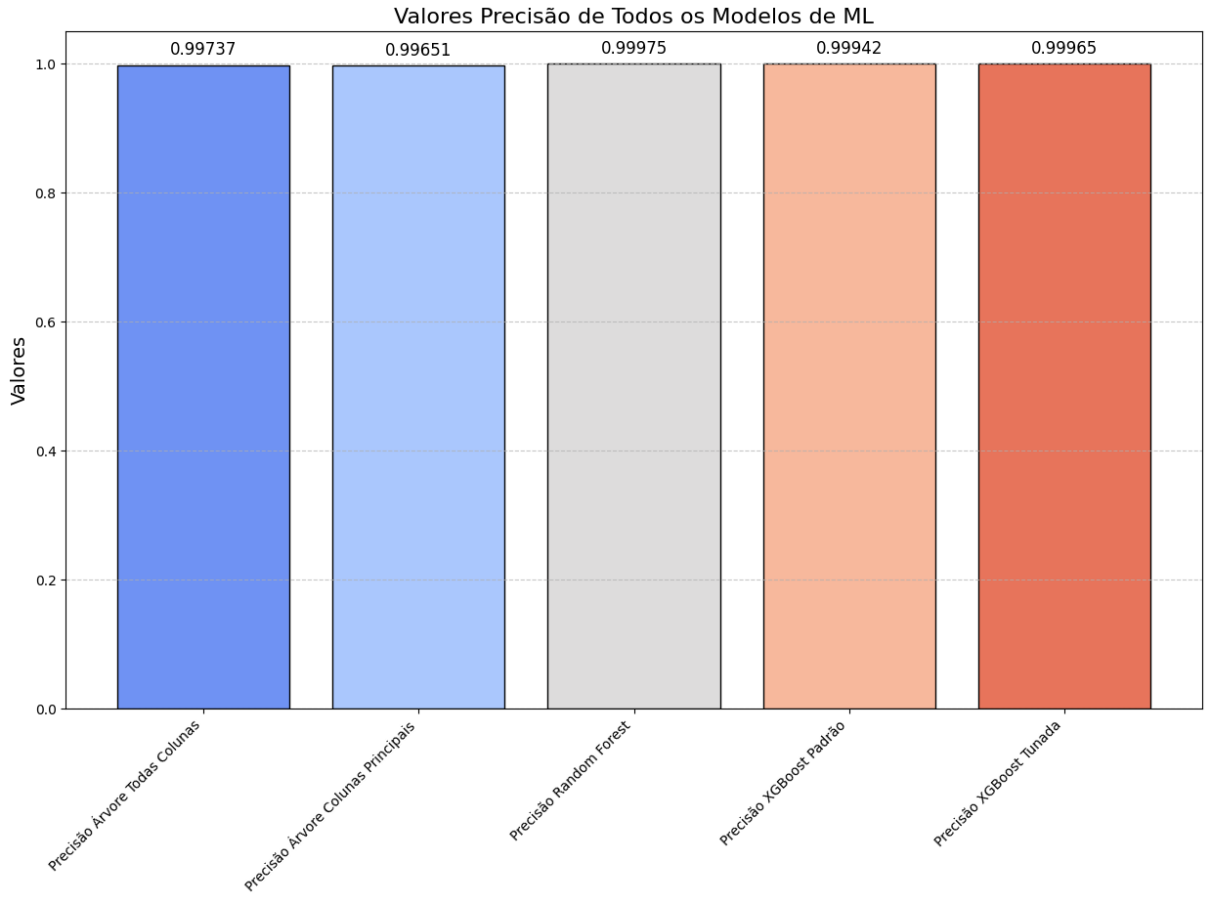
Precisão:

O autor Junior *et al.* (2022) também explica a precisão, como uma medida de avaliação que busca identificar e quantificar dentre a porcentagem de verdadeiros positivos prevista

pelo modelo, a partir da matriz de confusão, está de acordo com a realidade exata. Ele ainda explica no seu artigo, que caso um modelo possua um falso positivo alto, sua precisão é reduzida.

É interessante ver que assim como na análise proposta na etapa de acurácia, foi uma comprovação dos resultados das matrizes de confusão, a etapa de avaliação da precisão não foi diferente. Como representado no Gráfico 16 abaixo.

Gráfico 16 – Gráfico Comparativo Precisão de Todos Modelos Propostos



Fonte: Autoria Própria

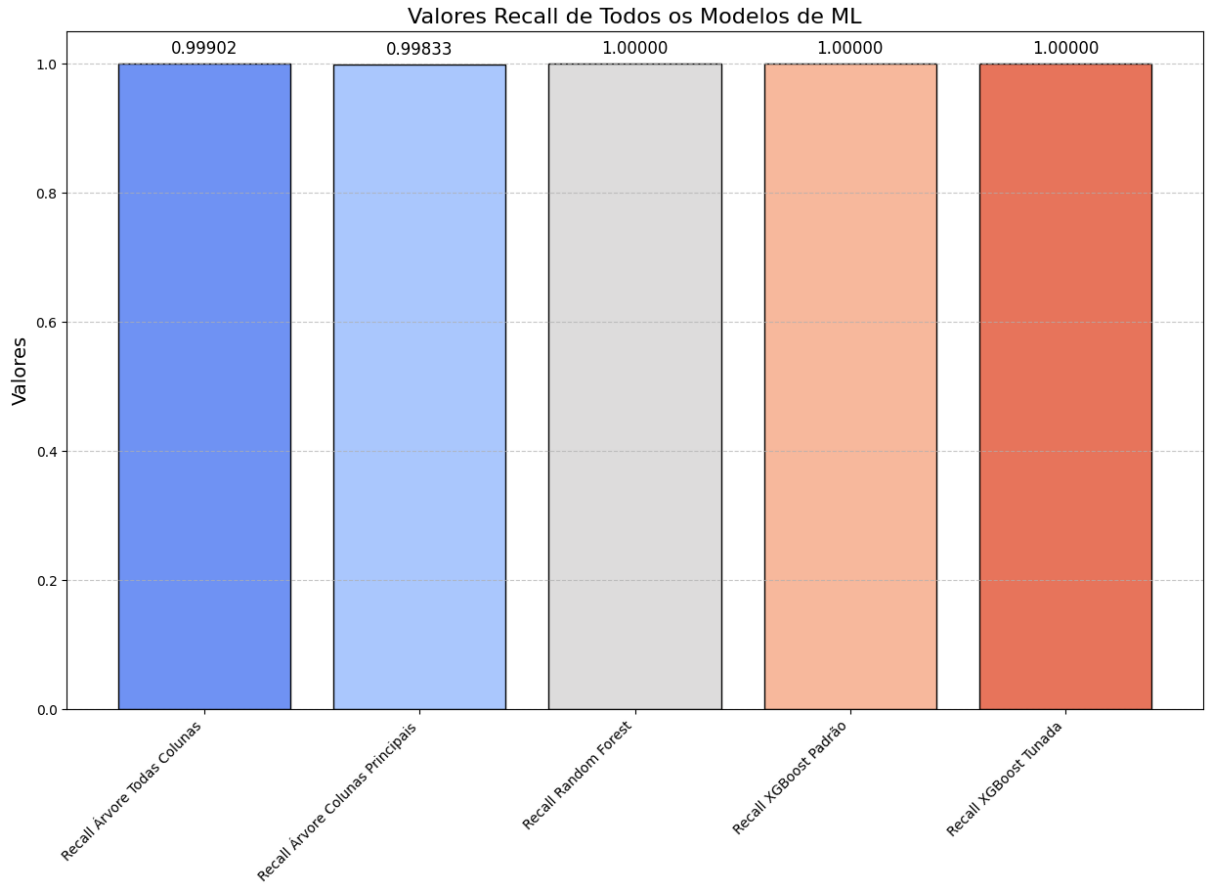
Retorno:

Ainda assim, o Autor Junior *et al.* (2022) em seu artigo, explica que a métrica do retorno é uma medida de avaliação que busca responder qual a porcentagem de todos os verdadeiros positivos que foram previstos corretamente pelo modelo. Essa métrica permite ao classificador identificar os falsos negativos.

Quando avaliamos a métrica de retorno, nós obtemos um resultado interessante, dos 5 tipos e variações de modelos propostos, abaixo de 100%, foram apenas os modelos de árvores. Mas o resultado deles se mantiveram o mesmo da acurácia, precisão e agora no retorno, foi o modelo com todas variáveis, já o modelo com as principais variáveis ficou abaixo, mas ainda com um excelente resultado.

Por fim, mas não menos importante, ambos modelos de floresta aleatória e as 2 variações do XGBoost obtiveram resultado igual a 100%, sendo o melhor resultado possível. Como representado no Gráfico 17 abaixo.

Gráfico 17 – Gráfico Comparativo Retorno de Todos Modelos Propostos



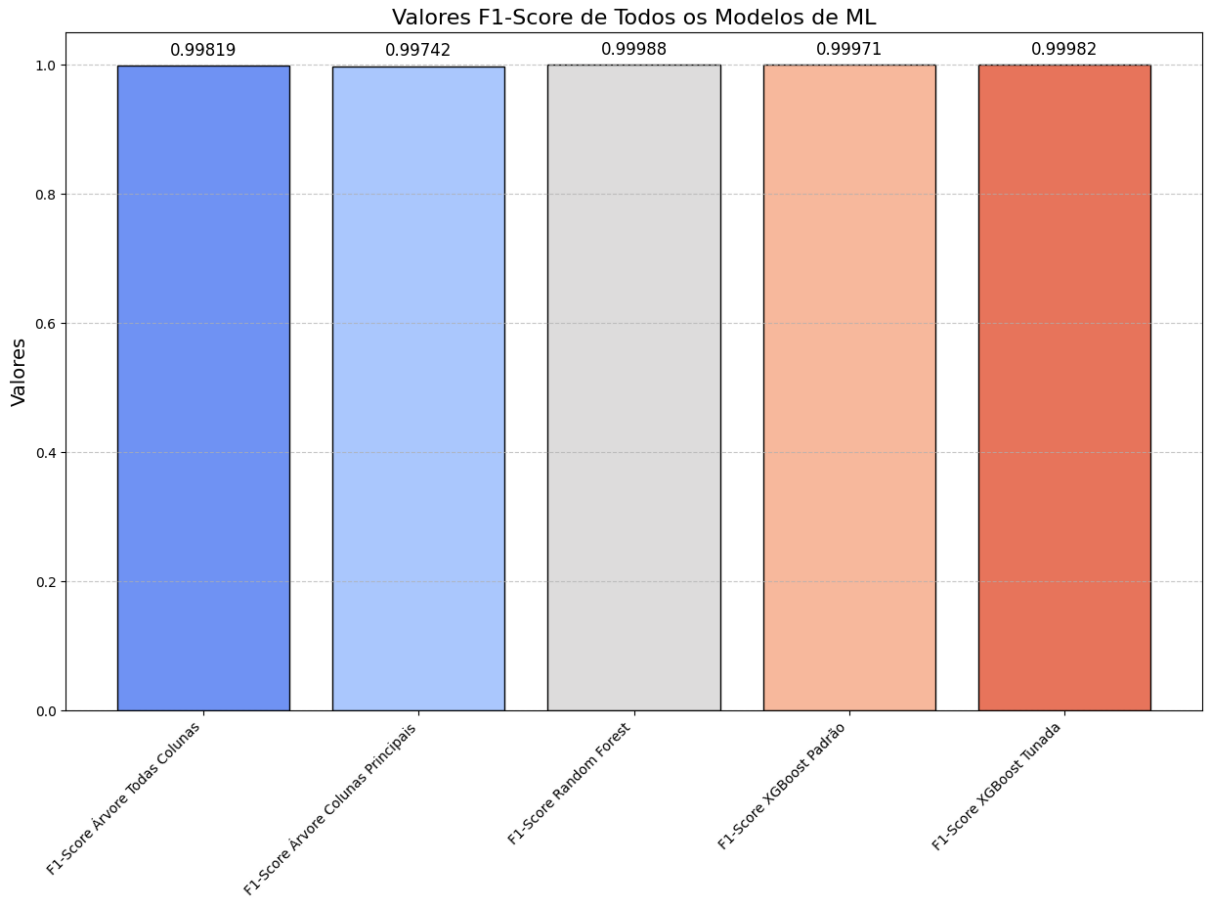
Fonte: Autoria Própria

F1-Score:

Segundo o autor Junior *et al.* (2022) explica que a métrica F1-Score permite associar os valores preditivos negativos do modelo, junto com a sensibilidade em classificadores binários de aprendizado de máquina. Sendo assim, esta métrica é obtida a partir da média harmônica entre ambas, resultando em uma medida de classificação que evidencia a qualidade dos valores preditivos negativos e sensibilidade, sendo o impacto real da capacidade de predição do modelo, permitindo também a avaliação da relação de equilíbrio entre precisão e sensibilidade.

Os resultados da métrica de avaliação F1-Score seguem o padrão já observados através das matrizes de confusão, métrica de acurácia e precisão. Com o melhor resultado sendo da floresta aleatória, das árvores, o modelo com todas variáveis é o com melhor desempenho e dentre os XGBoost, o modelo com modificação de parâmetros tem um leve ganho de desempenho, que não justifica o custo computacional. Como representado no Gráfico 18 abaixo.

Gráfico 18 – Gráfico Comparativo F1-Score de Todos Modelos Propostos



Fonte: Autoria Própria

Curva AUC ROC:

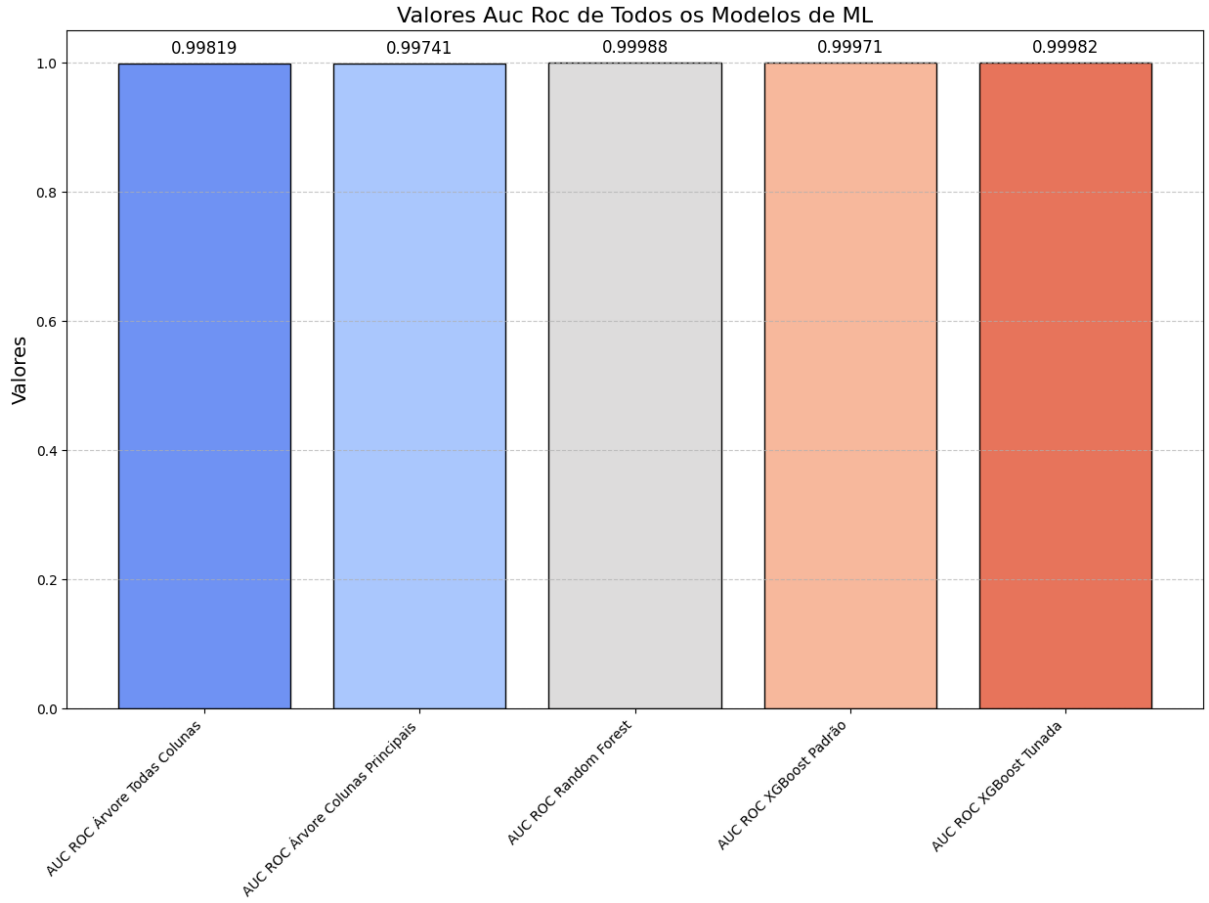
Por fim, o autor Junior *et al.* (2022) define a curva AUC ROC como um gráfico de sensibilidade em função da especificidade, este gráfico busca a avaliação do modelo em forma gráfica, traçando um limite para a predição do modelo, se destacando em 3 casos especiais.

1. Valores menores no eixo X do gráfico, indicam uma quantidade de falso positivo menor e uma quantidade de verdadeiro negativo em maior quantidade.
2. Valores maiores no eixo y, indicam maior quantidade de verdadeiro positivo e menor quantidade de falso negativo.
3. Valores altos no eixo y e baixos no eixo x, indicam um bom modelo.

A área sob a curva ROC também é definida por ele como uma métrica de separabilidade, informando o quanto o modelo é capaz de distinguir suas classes, quanto maior esta área sobre a curva, maior é a capacidade do modelo em predizer corretamente sua classificação.

Quando avaliamos a métrica de curva AUC - ROC, o mesmo padrão visto anteriormente se repete, por isso, não tem sentido detalhar sua explicação dos resultados. Como representado no Gráfico 19 abaixo.

Gráfico 19 – Gráfico Comparativo Curva AUC - ROC de Todos Modelos Propostos



Fonte: Autoria Própria

Validação Cruzada:

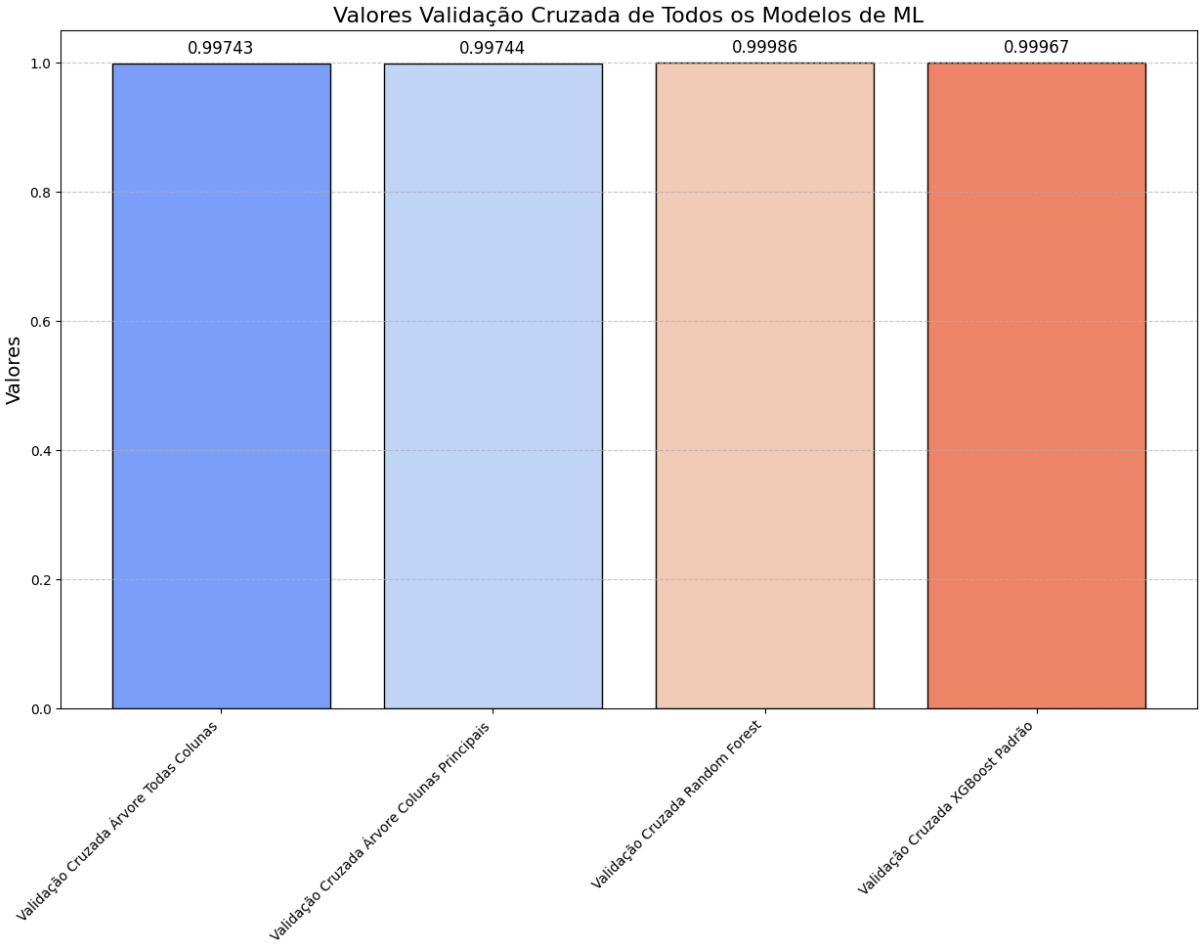
O autor Savietto (s.d.) explica a validação cruzada como um processo de validação dos modelos de aprendizado de máquina que permite dividir a base de dados de maneira aleatória em “K” grupos, quando o grupo é separado para teste, os outros são utilizados para treinamento, durante o treinamento do modelo, o modelo é treinado com os dados de treinamento, após, é testado com os dados de teste, é salvo o valor da métrica e é descartado o modelo. Após, o próximo grupo é selecionado e se repete o processo. Este método é chamado de k-fold cross-validation.

Quando avaliamos a métrica de validação cruzada, acontece uma mudança nos testes vistos anteriormente. Ao olhar para o gráfico acima, percebemos que só existe uma variação do algoritmo XGBoost, isso se deve pois, como dito anteriormente ao explicar sobre o modelo, a versão com os hiperparâmetros modificada não permite a realização da validação cruzada pela forma com que o modelo é executado, pois o retorno do modelo de treinamento é um arquivo do tipo boost já treinado, para execução da validação cruzada, é necessário que o modelo realize o treinamento por meio do modelo.fit(). Devido a esta incompatibilidade de arquivos e retorno, seu uso foi inviabilizado neste teste, o que é uma pena, pois a validação cruzada, como dito

anteriormente, permite avaliar a capacidade do modelo a se adaptar aos novos dados. Neste sentido, o gráfico acima representa a acurácia, após validação dos modelos propostos.

Podemos observar que mesmo após a validação cruzada, os modelos se mantiveram com um excelente desempenho, alcançando um resultado ótimo, mesmo considerando que a base dados possui aproximadamente 570 mil registros, a redução da acurácia dos modelos foi mínima. Isso significa que o modelo já estava muito bem treinado e preparado para trabalhar com diferentes dados se tornando uma opção viável e com um custo de processamento satisfatório para implementação deste tipo de técnica para tentar inibir os problemas abordados neste presente trabalho. Como representado no Gráfico 20 abaixo.

Gráfico 20 – Gráfico Comparativo Validação de Todos Modelos Propostos Exceto o XGBoost Modificado



Fonte: Autoria Própria

4.2.5 TABELA COMPARATIVA DOS MODELOS E MÉTRICAS

A fim de deixar mais claro todos os resultados coletados, foi criado a Tabela 10 e a Tabela 11 comparando cada algoritmo com suas respectivas métricas de avaliação coletadas e expressas nos resultados acima. É válido destacar que por se tratar de muitas métricas e algoritmos, a tabela foi particionada em 3 versões. A primeira versão com os algoritmos de árvore e

Random Forest, a segunda com os de XGBoost e por fim o melhor resultado desta comparação e o algoritmo respectivo por este resultado.

Tabela 10 – Parte 1: Comparação de algoritmos (métricas básicas)

Métrica	Árv. (Todas)	Árv. (Red.)	R. Forest	XGB Padrão
Acurácia (Pré)	0.99814	0.99750	0.99988	0.99971
Precisão	0.99727	0.99664	0.99975	0.99942
Recall	0.99902	0.99839	1.00000	1.00000
F1-Score	0.99814	0.99751	0.99988	0.99971
AUC ROC	0.99813	0.99750	0.99988	0.99971
Acurácia (Pós)	0.99738	0.99737	0.99986	0.99986

Tabela 11 – Parte 2: Resultados do XGBoost Tunado e melhores resultados

Métrica	XGB Tunado	Melhor	Algoritmo
Acurácia (Pré)	0.99982	0.99988	R. Forest
Precisão	0.99965	0.99975	R. Forest
Recall	1.00000	1.00000	R.F./XGB
F1-Score	0.99982	0.99988	R. Forest
AUC ROC	0.99982	0.99988	R. Forest
Acurácia (Pós)	0.00000	0.99986	R. Forest

5 CONCLUSÃO

Este trabalho teve como proposta a comparação e a avaliação de modelos de Aprendizado de Máquina para a classificação e detecção de fraude bancária Para isso, foi realizado um levantamento bibliográfico que embasou a escolha das técnicas de ciência de dados, algoritmos de classificação e avaliação dos modelos, respectivamente, seus resultados. Dessa forma, contribuindo para o entendimento e aperfeiçoamento do trabalho.

Na metodologia, foi-se aplicado o ciclo da ciência de dados, iniciando da aquisição e entendimento dos dados, utilização de métodos estatísticos para realizar a etapa de análise exploratória, após, foram escolhidos e implementados os algoritmos de aprendizado de máquina supervisionados, após, foi realizado a avaliação dos resultados destes modelos.

Os resultados experimentais demonstraram que o uso do algoritmo *Random Forest*, e passando uma avaliação robusta de treino, validação e teste, foi eficaz para a tarefa de classificação de fraude, atingindo uma precisão de 99% na base de teste, para a classe fraude. Esse desempenho confirma a capacidade do modelo em identificar corretamente comentários desse contexto, minimizando o ruído de dados menos relevantes. Vale ressaltar que dentre os algoritmos utilizados neste experimento ele obteve o melhor resultado e com um custo de processamento e de tempo processado relativamente baixo se comparado ao XGBoost, em ambas versões, por exemplo. Se tornando uma excelente opção para identificação de fraudes.

Para futuros trabalhos, sugere-se o desenvolvimento de aplicações com o uso do modelo, como *dashboards* que coletam os dados de transações e aplicam o modelo para filtrar e classificar as fraudes. Além disso, é recomendado o aperfeiçoamento das técnicas, como a utilização de outros algoritmos, como de Redes Neurais Artificiais modernos, outro ponto é aplicar o processo com um banco de dados mais específico, para evitar possíveis vieses dos modelos.

REFERÊNCIAS

ALI, M. **Aprendizado de máquina supervisionado**. 2024. Acesso em: 13 maio 2025. Disponível em: <https://www.datacamp.com/pt/blog/supervised-machine-learning>.

AMORIM, M. J. V. **Técnicas de aprendizado de máquina aplicadas na previsão de produtividade de operadores de centros de teleatendimento**. Dissertação (Mestrado) — Universidade João Pessoa, Salvador, 2008. Acesso em: 13 maio 2025. Disponível em: <https://doi.org/10.17648/SBAI-2019-111211>.

Banco Pan. **Fraude bancária: conheça os principais golpes, seus direitos e como se proteger**. 2022. Acesso em: 16 ago. 2024. Disponível em: <https://www.bancopan.com.br/blog/seguranca/fraude-bancaria-conheca-os-principais-golpes-seus-direitos-e-como-se-proteger>.

BigDataCorp. **Impacto das fraudes: conheça os dados do mercado**. 2023. Acesso em: 20 maio 2025. Disponível em: <https://blog.bigdatacorp.com.br/impacto-das-fraudes-conheca-os-dados-do-mercado/>.

CAO, L. *Data Science: A Comprehensive Overview*. **ACM Computing Surveys**, v. 50, n. 3, p. 1–42, 2017. Acesso em: 22 set. 2025.

CASTRO, C.; BRAGA, A. Aprendizado supervisionado com conjuntos de dados desbalanceados. **SBA: Controle Automação**, v. 22, n. 5, p. 501–512, 2011. Acesso em: 29 abr. 2025. Disponível em: <https://doi.org/10.1590/S0103-17592011000500002>.

ClearSale. **Fraude no cartão de crédito: principais tipos e como proteger seu varejo**. 2024. Acesso em: 20 maio 2025. Disponível em: <https://br.clear.sale/blog/fraude-cartao-de-credito-principais-tipos-e-como-proteger-seu-varejo>.

ELGIRIYEWITHANA, N. **Credit card fraud detection dataset 2023**. Kaggle, 2023. Acesso em: 30 mar. 2025. Disponível em: <https://doi.org/10.34740/KAGGLE/DSV/6492730>.

ENGINEERING, H. J. A. P. S. of; SCIENCES, A. **"What is data science? Definition, skills, applications & more"**. 2024. Acesso em: 22 set. 2025. Disponível em: <https://seas.harvard.edu/news/what-data-science-definition-skills-applications-more>.

Federação Brasileira de Bancos. **"Febraban e bancos lançarão Selo de Prevenção a Fraudes das Instituições Financeiras"**. 2023. Acesso em: 20 maio 2025. Disponível em: <https://portal.febraban.org.br/noticia/3996/pt-br/>.

FURLAN, M. **Golpe do cartão de crédito: conheça as fraudes mais comuns**. 2023. Acesso em: 12 jun. 2024. Disponível em: <https://www.serasa.com.br/premium/blog/conheca-as-fraudes-mais-comuns/>.

GOUVEIA, R. **Desvio padrão**. s.d. Acesso em: 21 ago. 2024. Disponível em: <https://www.todamateria.com.br/desvio-padrao/>.

GRANATO, D.; KIST, A. **Análise estatística descritiva aplicada à ciência e tecnologia de alimentos usando programas estatísticos**. São Paulo: IAL, 2018. 120 p.

GUEDES, T. A. *et al.* **Estatística descritiva**. Projeto de ensino Aprender Fazendo Estatística, 2005. 1–49 p. Acesso em: 21 ago. 2024. Disponível em: <https://www.ime.usp.br/~rvicente/Guedes_etal_Estatistica_Descritiva.pdf>.

GUNAY, D. **”Random Forest”**. 2023. Acesso em: 20 maio 2025. Disponível em: <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e>.

JUNIOR, G. d. B. V. *et al.* Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. **Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida**, v. 14, n. 2, 2022. Acesso em: 3 abr. 2025. Disponível em: <https://doi.org/10.36692/v14n2-01R>.

LEONARDIS, R. W. J. **Detecção de fraudes em transações com cartão de crédito: uma comparação do desempenho de técnicas inteligentes com base na avaliação da função de custo**. 75 p. Dissertação (Dissertação (Mestrado em Informática e Gestão do Conhecimento)) — Universidade Nove de Julho, São Paulo, 2023. Acesso em: 16 ago. 2024. Disponível em: <http://bibliotecatede.uninove.br/handle/tede/3242>.

LONGO, L. **Brasil registra R$ 3,5 bi em tentativas de fraude com cartão de crédito, aponta ‘Mapa da Fraude’**. 2024. Acesso em: 12 jun. 2024. Disponível em: <https://valorinveste.globo.com/produtos/servicos-financeiros/noticia/2024/04/18/brasil-registra-r-35-bi-em-tentativas-de-fraude-com-cartao-de-credito-aponta-mapa-da-fraude.ghtml>.

MathWorks. **Machine Learning in MATLAB**. 2024. Acesso em: 20 maio 2025. Disponível em: <https://www.mathworks.com/help/stats/machine-learning-in-matlab.html>.

PAIXãO, E. **Golpe do cartão de crédito: quais são os tipos mais comuns e como evitar?** 2023. Acesso em: 12 jun. 2024. Disponível em: <https://blog.nubank.com.br/golpe-cartao-de-credito/>.

RIVO, E.; FUENTE, D. L.; OTHERS. *Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management*. **Clinical and Translational Oncology**, v. 14, n. 1, p. 73–79, 2012. Acesso em: 29 set. 2025. Disponível em: <https://doi.org/10.1007/s12094-012-0764-8>.

ROQUE, A. **Estatística aplicada à educação: aula 8**. s.d. Acesso em: 22 ago. 2024. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/5209106/mod_resource/content/1/aula%208.pdf>.

ROUT, A. R. **”4 Key Pillars of Data Science”**. 2025. Acesso em: 22 set. 2025. Disponível em: <https://www.geeksforgeeks.org/data-science/key-pillars-of-data-science/>.

SANTOS, A. C. d. **Aprendizado de máquina aplicado na detecção de fraudes em cartão de crédito**. 42 p. Monografia (Graduação em Estatística) — Universidade Federal de Ouro Preto, Ouro Preto, 2023. Acesso em: 5 abr. 2025. Disponível em: <http://www.monografias.ufop.br/handle/35400000/6454>.

SANTOS, M. K. *et al.* Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards precision medicine. **Radiologia Brasileira**, v. 52, n. 6, p. 387–396, 2019. Acesso em: 11 mai. 2025. Disponível em: <https://doi.org/10.1590/0100-3984.2019.0049>.

SARKER, I. H. Machine learning: algorithms, real-world applications and research directions. **SN Computer Science**, v. 2, p. 160, 2021. Acesso em: 11 mai. 2025. Disponível em: <https://doi.org/10.1007/s42979-021-00592-x>.

SAVIETTO, J. V. **Machine learning: métricas, validação cruzada, bias e variância**. s.d. Acesso em: 5 abr. 2025. Disponível em: <https://medium.com/@jvsavietto6/machine-learning-m%C3%A9tricas-valida%C3%A7%C3%A3o-cruzada-bias-e-vari%C3%A2ncia-380513d97c95>.

Serasa. **Pesquisa da Serasa aponta que consumidores têm quatro ou mais cartões**. 2022. Acesso em: 20 maio 2025. Disponível em: <https://www.serasa.com.br/imprensa/pesquisa-da-serasa-aponta-que-consumidores-tem-quatro-ou-mais-cartoes/>.

SICA, N. **Pesquisa: impressões digitais e sua relação com as pessoas e as empresas**. 2022. Acesso em: 12 jun. 2024. Disponível em: <https://www.kaspersky.com.br/blog/pesquisa-impressoes-digitais/18906/>.

SILVA, G. F. d. S. **Fraude de cartão de crédito: como a estatística e o machine learning se conversam**. Dissertação (Dissertação (Mestrado em Estatística)) — Universidade de São Paulo, Piracicaba, 2020. Acesso em: 12 abr. 2025. Disponível em: <https://www.teses.usp.br/teses/disponiveis/11/11134/tde-14012021-200042/>.

SOUZA, A. **Metodologia CRISP-DM: Uma Abordagem Abrangente para Projetos de Dados**. 2023. Acesso em: 22 set. 2025. Disponível em: <https://medium.com/blog-do-zouza/metodologia-crisp-dm-uma-abordagem-abrangente-para-projetos-de-dados-d7e7135b907e>.

TURING. **Importance of Decision Trees in Machine Learning**. 2024. Acesso em: 20 maio 2025. Disponível em: <https://www.turing.com/kb/importance-of-decision-trees-in-machine-learning>.

VERMA, N. **"XGBoost Algorithm Explained in Less Than 5 Minutes"**. 2022. Acesso em: 20 maio 2025. Disponível em: <https://medium.com/@techynilesh/xgboost-algorithm-explained-in-less-than-5-minutes-b561dcc1ccee>.

WIRTH, R.; HIPP, J. *CRISP-DM: Towards a Standard Process Model for Data Mining*. **Procedings / Practical Applications of Knowledge Discovery and Data Mining**, 2000. Acesso em: 29 set. 2025. Disponível em: <https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>.


UFVJM