UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

RHUAN LOPES ASSIS

DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO UTILIZANDO MÉTODOS DE BASEADOS EM ÁRVORES DE DECISÃO

PATO BRANCO

2023

RHUAN LOPES ASSIS

DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO UTILIZANDO MÉTODOS DE BASEADOS EM ÁRVORES DE DECISÃO

DETECTION OF CREDIT CARD FRAUD USING DECISION TREE-BASED METHODS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia da Computação do Curso de Bacharelado em Engenharia da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Professor Doutor Jefferson Tales Oliva

Coorientador: Professor Doutor Dalcimar Casanova

PATO BRANCO

2023

RHUAN LOPES ASSIS

DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO UTILIZANDO MÉTODOS DE BASEADOS EM ÁRVORES DE DECISÃO

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia da Computação do Curso de Bacharelado em Engenharia da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 19/junho/2023

Jefferson Tales Oliva
Doutor em Ciências da Computação e Matemática Computacional
Universidade Tecnológica Federal do Paraná

Dalcimar Casanova
Doutor em Física Computacional
Universidade Tecnológica Federal do Paraná

Kelly Lais Wiggers
Doutora em Informática
Universidade Tecnológica Federal do Paraná

Fábio Favarim
Doutor em Engenharia Elétrica
Universidade Tecnológica Federal do Paraná

PATO BRANCO

2023

Dedico este trabalho a minha mãe Valmira Veronice de Assis e meu pai José Lopes Barbosa, duas pessoas que além de me ensinarem valores que se levam para além da vida, deram o sangue e suor para que eu conseguisse o conhecimento que resultou nestre trabalho, em que finalizo o curso .

AGRADECIMENTOS

Agradeço aos meus amigos Joshua, Augusto, Thiago, Matheus, Gustavo, Renato, Ralf, Vinícius, Felipe, Gabriel, André, Elioenai, Alexsander, Nathalia, Carlos H., Igor, Rader, Giuliano, Thomaz, Pedro, Arthur e Nathan, pessoas excelentes que tive o prazer de compartilhar lágrimas e conquistas ao meu lado.

Não posso deixar de agradecer a Bettão, eterno líder, que foi uma pessoa essencial durante um momento de minha vida, trazendo alegria e motivação para seguir em frente, e seguir feliz.

Este trabalho também é fruto da motivação que tive de ter um orientador competente! Não existem palavras para mensurar a gratidão que tenho a meu professor orientador, Jefferson, que confiou em mim e me direcionou a buscar o melhor.

Também agradeço a Karine, por ter sido a melhor companheira de trabalho que tive, tendo me ensinado e mostrado skills essenciais para minha carreira, bem como sendo amiga e conselheira em dias difíceis.

Por fim, agradeço a Franciéli Faccin, por me mostrar lados bons da vida que eu não conhecia até então, por ser uma pessoa incrível ao mostrar seu brilho, resiliência, parceria e cultura às pessoas ao seu redor.

As circunstâncias do nascimento de alguém
são irrelevantes. É o que você faz com o dom
da vida que determina quem você é.
(MEWTWO, 1998)

RESUMO

Em 2020, o mundo foi surpreendido com o surgimento de uma pandemia. Devido ao isolamento social causado por esta, as compras pela Internet tiveram uma aumento significativo, e, junto a elas, também aumentaram os casos de tentativas de fraude, especialmente em compras com cartão de crédito. Apesar do aumento de fraudes, o número de transações legítimas continua expressivamente maior, o que dificulta sua detecção. Tendo isso em mente, sabe-se que transações fraudulentas são consideradas anomalias perante as legítimas. Este trabalho tem como objetivo a utilização de algoritmos de aprendizado de máquina baseados em árvores de decisão, que são modelos facilmente interpretáveis por humanos e podem ser utilizados para detectar potenciais fraudes. A base foi separada em treino, teste e validação, de forma estratificada. O conjunto de teste foi gerado utilizando a maior fração dos dados, com o objetivo de criar modelos representativos, ou seja, que utilizasse uma menor parcela dos dados para o treinamento. Durante a construção de modelos, os dados foram padronizados via *StandardScaler* antes das próximas fases, de otimização de hiperparâmetros com o *BayesianSearch*, treinamento dos modelos *Isolation Forest* (IF) e *ExtraTree* (ET) com os hiperparâmetros encontrados, predição e validação cruzada *K-fold*. Por fim, para a comparação dos modelos, foi aplicado o teste estatístico de hipótese de Friedman considerando o nível de significância de 95%. Como foi constatada diferença estatística extremamente significativa, o pós-teste de Nemenyi foi aplicado para verificar quais pares de modelos tiveram diferença estatisticamente significativa. Como resultado, conclui-se que, com 95% de certeza, que os modelos ETs tiveram desempenho superior em comparação com o modelo IF-Matt. Por fim, o modelo supervisionado obteve melhores medidas de classificação de transações legítimas, enquanto o não supervisionado foi o melhor classificando fraudes. Pode-se também notar que o coeficiente de correlação de Matthews era maior em modelos com sensitividade maior.

Palavras-chave: fraudes em cartão de crédito; detecção de anomalias; aprendizado de máquina; árvores de decisão; .

ABSTRACT

In 2020, the world was surprised by the arise of the pandemic. Due the social isolation issued by this, Internet purchases had a significant increase, and, with them, fraud attempts has increased as well, especially in credit card purchases. Although the fraud increased, the legit transactions are still expressively bigger, which difficult its detection. With this in mind, is known that fraudulent transactions are considered outliers towards the legit ones. This work aims to utilize tree-based machine learning algorithms, which are easily interpretable models for humans and can be used to detect potential frauds. The dataset was split into training, testing, and validation sets in a stratified manner. The testing set was generated using the largest fraction of the data to create representative models, meaning that a smaller portion of the data was used for training. During the model construction, the data was standardized using the StandardScaler before proceeding to the next steps, which were hyperparameter optimization with BayesianSearch, training the Isolation Forest and ExtraTree models with the found hyperparameters, prediction, and K-fold cross-validation. Finally, to compare the models, the Friedman statistical hypothesis test was applied with a significance level of 95%. Since an extremely significant statistical difference was found, the Nemenyi post-test was applied to determine which pairs of models had a statistically significant difference. As a result, it can be concluded with 95% certainty that the ET models performed better compared to the IF-Matt model. Additionally, the supervised model achieved better classification measures for legitimate transactions, while the unsupervised model excelled in classifying frauds. It can also be observed that models with higher sensitivity had a higher Matthews correlation coefficient.
Keywords: credit card fraud; anomaly detection; machine learning; decision trees; .

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

AM	Aprendizado de máquina
GD	*GridSearch*
HTTPS	*Hypertext Tranfers Protocol Secure*
IA	Inteligência Artificial
IF	*Isolation Forest*
ET	*Extra Tree*
BAS	*Beetle Antennae Search*
KNN	K-ésimo vizinho mais próximo
LDA	*Linear Discriminant Analysis*
LOF	*Local Outlier Factor*
MC	Matrizes de Confusão
PCA	*Principal Component Analysis*
RB	Redes Baesyanas
RNA	Redes Neurais Artificiais
ROC	*Receiver Operating Curve*
SMOTE	*Synthetic Minority Oversampling Technique*
SSL	*Secure Sockets Layer*
SVM	Máquina de Vetores de Suporte
TCC	Trabalho de Conclusão de Curso
TLS	*Transport Layer Security*
VP	Verdadeiros positivos
VN	Verdadeiros negativos
FP	Falsos Positivos

FN	Falsos Negativos
ESP	Especificidade
TFN	Taxa de falsos negativos
TFP	Taxa de falsos positivos
RL	Regressão Logística
RF	*Random Forests*
MCC	Coeficiente de correlação de Matthews
ACC	Acurácia
PREC	Precisão
SENS	Sensitividade
F1	*F1-Score*
EFI	Eficiência
ROC	*Receiver Operating Curve*

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais

Dados são componentes fundamentais no contexto do mundo contemporâneo, visto que, através destes, pode-se otimizar praticamente tudo o que os dão origem, através, por exemplo, da analise de tendências. Dada a sua importância, há quem diga que dados são o novo petróleo, matéria prima para inovação. O desafio é trabalhar os dados de forma inteligente e, a partir da sua combinação, gerar informação com resultados satisfatórios e de alto desempenho, potencializando resultados, provendo segurança para tomadas de decisões (DINIZ, 2021).

A detecção de pequenas fraudes, por exemplo, é um problema desafiador para as seguradoras, já que cada vez mais aumenta o volume desse tipo de fraude, bem como o custo para detectá-las. Apesar disso, sem o uso de ferramentas e sistemas tecnológicos, uma pequena fração da atividade fraudulenta pode escalar rapidamente, gerando em grandes prejuízos. Para evitar essas perdas, o uso da inteligência artificial (IA) vem ganhando relevância na detecção de fraudes (INSURTALKS, 2022).

Na detecção de fraudes os dados podem ter o seu processamento através de *pipelines*, que consistem de uma série de etapas de processamento de dados. Estes são sensíveis a mudanças, que ocorrem, por exemplo, na fase de extração dos dados, que pode trazer valores nulos, podendo ocasionar em dados ruidosos. Este estudo busca uma solução para identificar alterações de comportamentos dos dados (*outliers*) através do emprego de aprendizado de máquina (AM), cujos algoritmos são aplicados na base de dados para identificar anomalias (SOUZA, 2021).

Outliers são valores atípicos, podendo causar anomalias nos resultados dos modelos preditivos. Entender e identificar os *outliers* é de grande importância, visto que eles podem gerar um viés negativo nos resultados, ou, como neste trabalho, ser o objeto de estudo (HOPPEN; PRATES, 2017).

Também, como a cada ano cresce o uso dos cartões de crédito como forma de pagamento, uma das consequências é a atenção de criminosos que estão em busca de maneiras de fraudar as transações. Para diminuir o prejuízo monetário, as prestadoras do serviço de cartão de crédito estão desenvolvendo novos métodos a fim de combater esses crimes (VIEIRA, 2019). A detecção de *Outliers* pode ser uma das melhores opções para a detecção e combate de fraudes, visto que há um número ínfimo de transações fraudulentas em comparação às legítimas.

1.2 Problema de Pesquisa

Uma das principais questões no tema de detecção de fraude seria encontrar a melhor alternativa para resolver este problema, visto que, em relação às transações legítimas, existe um número muito baixo de fraudes.

1.3 Hipótese

Dada a flexibilidade e adaptabilidade de modelos baseados em árvores de decisão, visto que possuem pontos de corte destinados a separar as classes, mostram-se potenciais candidatos à detecção de anomalias.

1.4 Objetivos

1.4.1 Objetivo geral

Desenvolver e comparar modelos baseados em árvores utilizando uma otimização Bayesiana de hiperparâmetros para auxiliar na detecção de fraudes de cartões de crédito.

1.4.2 Objetivos específicos

- Pré-processamento dos dados obtidos.
- Obtenção de uma amostragem representativa para treinamento.
- Utilizar algoritmos de aprendizado de máquina direcionados à detecção de anomalias para a construção de modelos preditivos.
- Avaliação dos modelos preditivos.
- Comparação dos resultados com os obtidos em trabalhos relacionados.

2 REFERENCIAL TEÓRICO

2.1 Cartão de Crédito: O que é e como funciona

De maneira direta, o cartão de crédito é uma forma de empréstimo monetário, no qual o cliente tem até 40 dias para efetuar o pagamento da fatura sem que sejam cobradas taxas adicionais. É disponibilizado por bancos e instituições financeiras, nas quais o solicitante, se aprovado, recebe um limite de crédito que pode ser utilizado em compras ou serviços (CARDOSO; VIEIRA, 2019).

2.1.1 Fraudes em transações de crédito

No primeiro semestre de 2021, o Brasil sofreu um aumento de quase 33% nas tentativas de transações fraudulentas em relação ao mesmo período em 2020, isso pode ser observado na Figura 1, onde as barras em rosa apresentam a quantidade de transações em milhões e, a linha em azul apresenta a quantidade de tentativas de fraude. O destaque foi na área do *e-commerce*, pois o isolamento social devido à pandemia, fez com seu uso fosse bem amplificado, já que as pessoas passaram a comprar mais pela Internet (SUTTO, 2021).

Figura 1 – Transações e tentativas de fraude em 2020 e 2021 no Brasil

Fonte: Adaptado de e-commercebrasil (2021).

Devido à facilidade, comodidade e constantes novas funcionalidades, o uso do crédito se popularizou de maneira globalizada. Junto a ele, vieram atacantes que buscam utilizar transações fraudulentas, já que podem ser feitas de maneira rápida e com uma alta quantia de dinheiro (MAES *et al.*, 2002).

Existem diversas formas de se fazer uma transação fraudulenta, que ocorre devido ao roubo de dados. Os dados de uma pessoa podem ser obtidos devido a um vazamento em algum site. Por exemplo, o atacante pode utilizar os dados vazados para tentar obter acesso ao cartão de uma vítima (CARDOSO; VIEIRA, 2019). Outro engano comum é a inserção de dados em uma página de *phishing*, normalmente simulando uma loja real,com o objetivo de confundir o usuário que pensa estar em um site legítimo devido ao cadeado fechado (via certificado SSL/TSL) e, no fim, é enganado tendo seus dados roubados (GONÇALVES, 2022).

2.2 Inteligência artificial e aprendizado de máquina

Uma das possíveis definições de inteligência artificial (IA) é considerá-la uma espécie de mente elaborada pelo homem a fim de fazer com que alguma máquina simule uma habilidade natural humana (SILVA, 2005).

A sociedade contemporânea sente cada vez mais o crescimento da IA (por exemplo, em seu uso na medicina). É estimado que até 2030 seu impacto econômico mundial chegue a 30 trilhões de dólares (GIL, 2021).

O aprendizado de máquina (AM) foi definido como área de estudo em que se dá aos computadores a capacidade de aprender sem que estes tenham sido programados para isso. Quando se cria um modelo a ser treinado, o computador não tem conhecimento algum sobre o mesmo, e a partir de uma perspectiva estatística/matemática, podem vir a identificar padrões no modelo treinado (DUARTE, 2021).

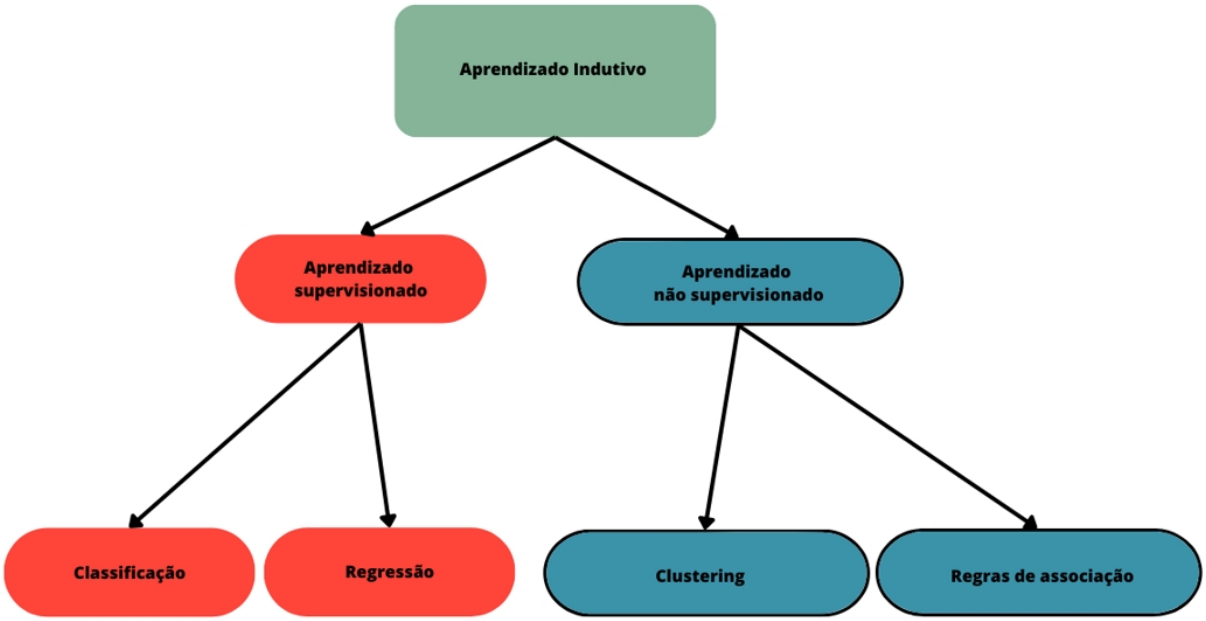
Os principais métodos de AM estão dentro do aprendizado indutivo, como mostra a Figura 2, ao apresentar os aprendizados supervisionado e não supervisionado e seus respectivos casos de uso.

2.2.1 Aprendizado supervisionado

Quando a base de dados possui as variáveis descritivas (independentes) e também as variáveis alvo (resultado/classificação), pode-se utilizar um algoritmo supervisionado. Observando o comportamento das variáveis independentes (características) e dependentes (alvo) utilizando uma base de treino, os modelos supervisionados avaliam matematicamente esta para estimar os rótulos de uma base de teste. Existem dois tipos de técnicas supervisionadas:

- Classificação: A variável dependente é categórica, ou seja, determina o pertencimento a uma classe (TACONELI, 2022). O método de avaliação se baseia na quantidade de classificações corretas ou incorretas, um exemplo de medida é a acurácia. Algumas técnicas de classificação supervisionadas são: K-vizinhos mais próximos, *random forests* e máquina de vetores de suporte(HONDA; FACURE; YAOHAO, 2017).

Figura 2 – Hierarquia do aprendizado indutivo.



Fonte: Autoria própria (2023).

- Regressão: A variável dependente é numérica (TACONELI, 2022). Para avaliar este tipo de modelo, deve-se calcular o erro entre o valor calculado e o valor real, e uma das medidas de avaliação pode ser o *mean squared error* (MSE). Algumas técnicas de regressão supervisionadas são: *Regression trees*, *Ordinary least squares estimation* e a regressão logística (RL) (LOTERMAN *et al.*, 2012).

2.2.2 Aprendizado não supervisionado

Este se trata de um treinamento feito a partir de dados não classificados. Os algoritmos buscam reconhecer os padrões implícitos nos dados, seja mostrando semelhanças ou diferenças.

Abaixo são apresentados exemplos de abordagens não supervisionadas:

- Clusterização: corresponde ao agrupamento dos dados não rotulados baseados em suas semelhanças ou diferenças, ou seja, busca-se particionar os dados em subgrupos em que seus membros possuem características similares. Podem ser subdivididos em exclusivos, hierárquicos, sobrepostos e probabilísticos. Alguns dos algoritmos são o *K-Means* e o *Hierarchical clustering* (FERREIRA; ZEVIANI, 2013).

- Regras de associação: são utilizada para descobrir características que descrevem grandes parcelas de dados. Consistem em uma expressão $X \rightarrow Y$, que são os conjuntos de itens os quais existe a tendencia de dados que contêm X conterem Y (JUNIOR, 2022). Alguns dos algoritmos mais utilizados são o *Eclat* e o *FP-Growth* (SILVA, 2021).

São muito utilizadas em sistemas de mercado com autocaixa, onde a empresa busca entender as preferências com base em dados de produtos comprados.

2.2.3 Overfitting e Underfitting

Após o treinamento do modelo, existem 2 características que podem indicar que este não obteve resultados confiáveis, sendo elas o *Overfitting* e o *Underfitting*.

- *Underfitting*: É evidenciado quando os resultados (como acurácia, precisão, etc) do modelo são ruins ainda na fase de treino. Costuma ocorrer devido a este não conseguir identificar as relações entre as entradas e a saída (KOEHRSEN, 2018).
- *Overfitting*: É destacado quando o modelo consegue bom desempenho com os dados de treino, mas não consegue com os dados de teste. O motivo deste evento é o modelo conseguir aprender tão bem o que ocorre com os dados de treino que não consegue generalizar com amostras não vistas durante o processo de treinamento (AWS, 2016).

2.3 Processo de Descoberta de Conhecimento

Dado que o trabalho tem como objetivo a detecção de anomalias através do treinamento de modelos preditivos, algumas etapas do processo de descoberta de conhecimento podem ser realizadas, conforme apresentadas nas próximas subseções.

2.3.1 Pré Processamento dos dados

Diversos aspectos podem influenciar no desempenho de um sistema de aprendizado devido à qualidade dos dados. Grande parte dos algoritmos de aprendizado de máquina são sensíveis à qualidade da entrada de dados. Nesse sentido, pode ser necessário o pré-processamento da base de dados para a preparação e estruturação das informações para um formato apropriado (JIAWEI; MICHELINE; JIAN, 2016). Devido a isso, algumas medidas que podem ser aplicadas serão listadas a seguir:

- Limpeza de dados: é comum achar dados incompletos, ruidosos, enfim, com inconsistências que podem atrapalhar os resultados. Os procedimentos a serem utilizados abordam as seguintes técnicas:
  - Retirada/ajuste de dados: na base de dados podem ser encontrados registros parcialmente nulos, ou com valores redundantes, e sua limitação tende a diminuir a qualidade dos resultados. Com isso, pode ser interessante a retiradas desses registros, visto que deixariam de prejudicar a qualidade dos dados.

  - Agregação em valores nulos: alguns registros podem vir com poucos valores nulos, e pode ser que agregar valores em tais situações pode trazer ser mais vantajoso que retirar o registro em questão, já que poderiam ser mantidos valores relevantes para a base de dados. Algumas técnicas como o preenchimento utilizando medidas estatísticas, como a média ou a moda, podem ser utilizadas.

- Normalização: Esta operação consiste em ajustar a escala dos valores dos atributos de maneira que os mesmos sejam mapeados para valores restritos a um determinado intervalo (normalmente entre 0 e 1). A motivação é evitar que alguns atributos, por apresentarem escalas diferentes (como o peso em quilos ou em libras), influenciem de forma tendenciosa os métodos de mineração de dados. Um exemplo é a utilização da função Min-Max (Equação 1), em que X' é o novo valor de X, $Xmin$ é o valor mínimo de X e $Xmax$ é o valor máximo valor de X (GOLDSCHMIDT; PASSOS, 2005).

$$X' = \frac{X - Xmin}{Xmax - Xmin} \tag{1}$$

- Padronização: Com a mesma ideia da normalização, a padronização faz com que os valores tenham média 0 e desvio padrão 1. A padronização foi realizada neste trabalho, utilizando o método StandardScaler (HORKY; PROKEŠ; HUBÁČEK, 2022), que segue a Equação 2.

$$z' = \frac{X - u}{s} \tag{2}$$

  Na qual z' é o valor padronizado, X é o valor a ser padronizado, u é a média e s é o desvio padrão.

- Redução de dimensionalidade: Os algoritmos de AM são capazes de extrair informações dos conjuntos de dados ricos em características, quer sejam tabelas extensas ou imagens com milhões de pixeis (SOFTTEK, 2021). Analisar os dados e os transformar em informações úteis pode ser uma tarefa difícil. É comum que os dados tenham um grande número de registros, bem com um grande número de atributos (dimensões), o que dificulta sua compreensão (RAZENTE; JUNIOR, 2004). A técnica PCA, do inglês *Principal Component Analysis*, reduz a dimensão dos dados, mantendo suas informações e características, como os valores de variância (YAMADA *et al.*, 2021). O objetivo do PCA é identificar a base mais significativa para reestruturar a base de dados, revelando uma estrutura oculta e filtrando ruidos na mesma (KURITA, 2019).

2.3.2 Construção de modelos: detecção de anomalias

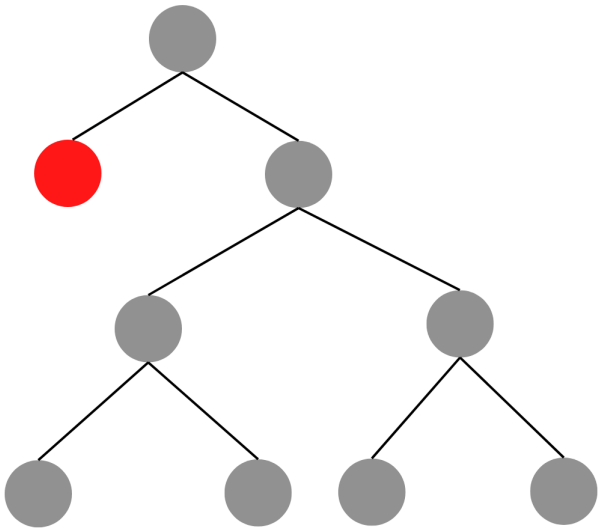
A construção de modelos para a detecção de anomalias pode ser realizada por meio de diferentes métodos de AM. Uma das abordagens não supervisionadas é a *Isolation Forest*, enquanto uma supervisionada seria a *ExtraTree*.

2.3.2.1 Isolation Forest

A ferramenta isola observações, dado que, aleatoriamente, escolhe uma característica, e então, novamente de maneira aleatoria, seleciona um valor entre os valores máximo e mínimo da característica escolhida (LIU; TING; ZHOU, 2008).

O particionamento é feito de maneira recursiva, então é representado por uma estrutura de árvore. O número de divisões utilizados para isolar uma amostra se define pelo comprimento desta. Como as anomalias são diferentes e estão em menor número, como ilustrado na Figura 3, possuem um caminho mais curto, como observado no nó marcado em vermelho. Então, quando uma floresta produz de forma comunitária comprimentos mais curtos para determinadas amostras, existe uma grande possibilidade de considera-las como anomalias (LIU; TING; ZHOU, 2008).

Figura 3 – Anomalia em uma *Isolation Forest*



Fonte: Autoria própria (2023).

2.3.2.2 Extra Tree

Extra Trees são baseadas nas *Random Forests* (RF) (MAIER *et al.*, 2015), que são uma combinação de preditores no formato de árvore, em que cada árvore depende dos valores de um vetor aleatório que é amostrado de forma independente e com distribuição igual para todas as árvores da floresta (BREIMAN, 2001).

Esses algoritmos têm as seguintes vantagens:

- *White-box model*: Apresentam a lógica, estrutura interna e os parâmetros de forma clara, podendo assim, serem interpretáveis por humanos e avaliados de forma transparente (PINTELAS; LIVIERIS; PINTELAS, 2020).
- Rápidos tanto no treino quanto na predição (BREIMAN, 2017).

As ET se diferenciam por adicionar uma camada extra de aleatoriedade às RF: agora não procuram pelo ponto de corte ótimo em cada nó (sendo este, o que divide os dados em duas novas ramificações, baseadas em uma característica), mas um valor aleatório é selecionado para cada característica, o que reduz o espaço de busca (MAIER *et al.*, 2015).

Por outro lado, o tamanho/profundidade da floresta é aumentado ao se introduzir cortes subótimos. Teoricamente, essa medida aborda de maneira não intuitiva o problema de classificação, aumentando a chance de saída de mínimos locais, porém ao custo de ter cortes subótimos (MAIER *et al.*, 2015).

2.3.3 Otimização de hiperparâmetros

Durante a construção de modelos, a escolha de hiperparâmetros afeta diretamente o desempenho de seus respectivos algoritmos. Para isso, existem ferramentas de otimização que são utilizadas visando melhores resultados. Um exemplo didático seria o clássico algoritmo de K-vizinhos mais próximos, em que a escolha correta do valor K tende a melhorar o desempenho do modelo.

Um dos métodos de otimização é o *GridSearchCV* (SYARIF; PRUGEL-BENNETT; WILLS, 2016), o qual consiste em avaliar todas as combinações possíveis de hiperparâmetros e, dada uma medida de avaliação, retornar a combinação que obteve o melhor desempenho. A maior vantagem deste método é a possibilidade de mostrar o melhor resultado, em contraponto tem custo elevado, já que testa todas as combinações possíveis. Outro método é o *Bayesian Search CV*, cujo objetivo é encontrar o ótimo global de uma função *Black-Box* (NGUYEN *et al.*, 2019), já que se possui uma entrada e um resultado, porém, se desconhece sua expressão matemática (TERAYAMA *et al.*, 2021). Esse método constrói um modelo probabilístico para a função, e então o utiliza para tomar decisões sobre a próxima avaliação (NGUYEN *et al.*, 2019). Sendo assim, pode-se escolher o numero de iterações, e em cada uma é sempre esperada uma

melhor pontuação. Este foi o motivo da escolha da otimização *Bayesiana*, visto que pode-se avaliar as melhoras nas pontuações após um número escolhido de iterações ao mesmo tempo em que pode-se controlar o custo (tempo), dado que possui condição de parada.

2.3.4 Validação e avaliação do modelo

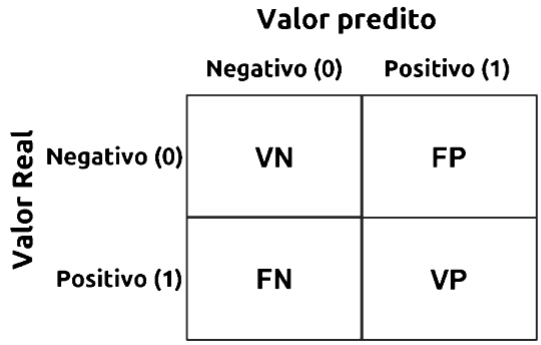
Para garantir a confiabilidade de um modelo, uma prática comum ao realizar um experimento de AM é a realização da validação cruzada (SKLEARN, 2020).

A validação do modelo será feita através da técnica de validação cruzada K-fold, que consiste em dividir o *dataset* de forma aleatória em K amostras, com tamanhos semelhantes. O processo terá K iterações, onde em cada passo é utilizada uma amostra para teste e as $k - 1$ amostras restantes, para treinamento. Após cada iteração, é gerada uma pontuação de um índice de desempenho, que fica retida em uma variável, e então o modelo é descartado. Ao fim, as pontuações serão utilizadas para estimar o desempenho preditivo do modelo (BROWNLEE, 2018). Tendo os resultados, as medidas a serem utilizadas para a avaliação tanto da validação cruzada, quanto da aplicação do modelo treinado na base de teste, são calculadas a partir da matriz de confusão.

2.3.4.1 Matriz de confusão e índices de avaliação

Matrizes de confusão (MC) são utilizadas na avaliação do relacionamento entre variáveis, revelando o pertencimento ou não a uma determinada classe (OLIVA, 2019). Indica as porcentagens de erros e acertos do modelo, conforme ilustrado na Figura 4, e se divide em 4 categorias:

Figura 4 – Matriz de confusão



Fonte: Adaptado de Juliana Scudilio (2020).

- Verdadeiros positivos (VP): Exemplos da classe positiva classificados corretamente.

- Verdadeiros negativos (VN): Exemplos da classe negativa classificados corretamente.
- Falsos positivos (FP): Exemplos da classe positiva classificados de maneira incorreta.
- Falsos negativos (FN): Exemplos da classe negativa classificados de maneira incorreta.

Construída a matriz de confusão, pode-se extrair algumas medidas de avaliação da qualidade do modelo. As utilizadas neste estudo estarão descritas abaixo.

- Acurácia (Acc): De maneira geral, indica o desempenho do modelo, ou seja, a taxa de acertos do modelo (OLIVA, 2019). Essa medida pode ser obtida pela Equação 3:

$$Ac = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

- Precisão (Prec): Dentre todas as classificações positivas, a porcentagem que foi classificada correta (OLIVA, 2019). A Equação 4 demonstra o cálculo da precisão.

$$Pc = \frac{VP}{VP + FP} \tag{4}$$

- Revocação/Sensitividade (Sens): A técnica consiste em avaliar a taxa de verdadeiros positivos em relação a todos os valores positivos esperados (OLIVA, 2019). A Equação 5 representa o cálculo da sensitividade.

$$Sens = \frac{VP}{VP + FN} \tag{5}$$

- *F1-Score* (F1): É calculada a partir de uma média harmônica entre a precisão e a revocação (OLIVA, 2019). Seu cálculo é dado pela Equação 6.

$$F1 = \frac{2 * Pc * Sens}{Pc + Sens} \tag{6}$$

- Especificidade (Esp): Indica a quantidade de instancias negativas rotuladas corretamente em relação a todos valores negativos esperados.(OLIVA, 2019). É calculada pela Equação 7.

$$Esp = \frac{VN}{VN + FP} \tag{7}$$

- Taxa de falsos negativos (TFN): Representada pela Equação 8, mostra a quantidade de falsos negativos em relação às classificações positivas (OLIVA, 2019).

$$TFN = \frac{FN}{VP + FN} \tag{8}$$

- Taxa de falsos positivos (TFP): Mostra a quantidade de falsos positivos em relação ao total de classificações negativas esperadas (OLIVA, 2019). Seu cálculo é dado pela Equação 9.

$$TFN = \frac{FP}{VN + FP} \tag{9}$$

- Área abaixo da *Receiver Operating Curve* (ROC): medida utilizada para avaliar o quão bem um modelo pode discriminar duas diferentes classes. Seus valores vão de 0,5 (não discrimina) a 1 (discrimina perfeitamente) (COOK, 2007).

- Coeficiente de correlação de Matthews (MCC): Avalia a correlação entre as classificações binárias observadas e previstas (JUNIOR *et al.*, 2022). É representada pela Equação 10.

$$MCC = \frac{(VP * VN) - (FP * FN)}{\sqrt{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}} \tag{10}$$

- Eficiência (Efi): Média entre a Sensitividade e a Especificidade (OLIVA, 2019). Seu valor é obtido da Equação 11.

$$Efi = \frac{Sens + Esp}{2} \tag{11}$$

- Média Geométrica (*G-Mean*): Média geométrica entre as acurácias específicas de cada classe (MULLICK *et al.*, 2020). Ao lidar com problemas de classificação binária, são utilizadas a sensitividade e a especificidade para o cálculo, que é dado pela Equação 12.

$$G - Mean = \sqrt{Sens * Esp} \tag{12}$$

Para complementar a avaliação desses modelos, os mesmo podem ser comparados por meio de testes estatísticos de hipótese, considerando um nível de significância (e.g. 95%).

Para a escolha do teste adequado,são considerados 2 aspectos essenciais (OLIVEIRA, 2008):

- O primeiro é a caracterização da variável alvo, considerando sua natureza (qualitativa ou quantitativa), distribuição (normal ou não), continuidade (contínua ou discreta) e instabilidade (alta ou baixa).

- O segundo aspecto é a existência de dois erros reconhecidos pela teoria estatística: o erro tipo I (concluir significância quando não há) e o erro tipo II (concluir equivalência quando há diferença significativa).

Alguns exemplos de testes são o teste de Friedman (SILVA; KORITIAKI; MELEM, 2021), o teste de Kruskal-Wallis o teste de Nemenyi (ELLIOTT; HYNAN, 2011).

3 TRABALHOS RELACIONADOS

Maes *et al.* (2002) apresentou duas abordagens diferentes, utilizando redes neurais artificiais (RNA) e redes *baesyanas* (RB). Para medir o desempenho, a medida utilizada foi a área abaixo da *Receiver Operating Curve* (AUROC). O objetivo desse trabalho era observar quantas classificações fraudulentas foram classificadas como legítimas, bem como quantas classificações legítimas foram definidas fraudulentas. Na abordagem com RNA, são realizadas duas abordagens, que revelam a importância do pré-processamento dos dados, visto que através dele foi feita uma análise de correlação, em que se descobre que uma característica tem forte correlação com a maioria das outras. Já na abordagem utilizando RB foram utilizadas 4 características, junto ao rótulo. Foi mostrado que duas das características influenciavam diretamente na classificação de fraude, bem como a classificação influenciava seus valores. Na comparação de resultados, foi observada a medida de taxa de verdadeiros e positivos, com o melhor resultado de 70% para a RNA e 68% para a RB, enquanto os menores resultados foram de 47% para RNA e 58% para RB. Outra comparação foi a do tempo de aprendizado, onde a RNA levou horas, enquanto a RB terminou em 20 minutos.

Awoyemi, Adetunmbi e Oluwadare (2017), com a mesma base de dados do presente trabalho, utilizou técnicas para reduzir a quantidade de dados legítimos e ao mesmo tempo em que aumentará o de fraudes, para serem comparados aos resultados da base original. Este processo foi realizado para criar dois diferentes *datasets*, com proporções de 10%/90% e 34%/66%, de transações fraudulentas e legítimas, respectivamente. Foram escolhidos 3 algoritmos, novamente, supervisionados, para o treinamento dos modelos, sendo estes o Nayve Bayes, o K-Vizinho mais próximo (KNN) e a Regressão Logística. As medidas de avaliação escolhidas foram a acurácia, sensitividade, especificidade, precisão, coeficiente de correlação de Matthews e eficiencia. Na comparação de resultados, nota-se que a RL teve o melhor desempenho apenas na base não amostrada, não tendo desempenhado bem nas outras bases segundo Awoyemi, Adetunmbi e Oluwadare (2017). O coeficiente de correlação de Matthew foi o que teve maior desvio padrão dentre os resultados, já que os valores oscilam de 0,007 a 0,9535, obtidos respectivamente pela RL e KNN. Já o KNN obteve as melhores pontuações nas outras bases, obtendo 100% nas medidas de precisão e especificidade.

Em outro trabalho, Cardoso e Vieira (2019), a base de dados utilizada é a mesma do presente trabalho. Foi feita uma análise utilizando os seguintes métodos supervisionados: Máquina de vetores de suporte (SVM), *Naive Bayes* e K-ésimo vizinho mais próximo (KNN). O trabalho se divide em quatro avaliações experimentais, nas quais é alterada a base de dados:

- Utilização da base completa: Nesse experimento, o KNN teve o menor desempenho, acertando apenas 1,7% da classificação minoritária. O SVM obteve 35,83% de acerto. Por fim, o *Naive Bayes* teve uma taxa de 63,33%, resultado considerado promissor.

- Retirada de 150000 registros da classe majoritária: Neste, o KNN continua com baixo desempenho, acertando apenas 1,69% da classificação minoritária. O SVM obteve 31,35% de acerto, tendo uma queda no desempenho. Já o *Naive Bayes* teve aumento, com taxa de 64,40%.

- Retirada de 15 colunas da base: o KNN continua com o pior desempenho, registrando 0,83% de taxa de acerto para a classificação minoritária. Já o SVM obteve 15,83% de acerto, apresentando outra queda no desempenho. O *Naive Bayes* desta vez teve uma queda, com taxa de 46,66%.

- Redução de dimensionalidade para 5 colunas utilizando o PCA: Neste caso, o KNN não consegue acertos da classificação minoritária, errando todas as predições, enquanto o SVM obteve um resultado semelhante ao anterior, com 16,66% de acerto. Por fim, o *Naive Bayes* teve novamente uma queda, com acerto de 35,83%.

Já em Varmedja *et al.* (2019), que utilizou a mesma base de dados que o presente trabalho, fez uso da SMOTE para balancear o número de transações fraudulentas e legítimas na mesma base de dados que foi utilizada neste trabalho. O conjunto de dados foi divido entre 80% para treino e 20% para teste. Em seguida, as seguintes abordagens de aprendizado supervisionado foram aplicadas: Regressão Logística (RL), *Naïve bayes* (NB), *Random Forest* (RF) e Redes Neurais Artificiais (ANN). As medidas de avaliação utilizadas foram a acurácia, especificidade e precisão da classe minoritária, que obteve melhores resultados de acurácia e precisão com a RF, com 99,96% e 96,38% respectivamente, enquanto a melhor especificidade foi obtida com a RL, com 91,84%. Enfim, Varmedja *et al.* (2019) concluiu que a técnica SMOTE é eficaz na melhoria da detecção de fraude, bem como considera que o algoritmo RF obteve o melhor desempenho em seus experimentos dados os resultados. Em um comparativo das matrizes de confusão com as do presente trabalho, nas abordagens supervisionadas que este utiliza, pode-se observar uma semelhança nas taxas de falsos negativos, o que tende a significar que abordagens supervisionadas costumam classificar corretamente a classe majoritária.

Jain *et al.* (2019) fez uma comparação de diversas abordagens, utilizando RNA, árvores de decisão, sistema baseado em lógica *Fuzzy*, SVM, redes *Bayesiana*, KNN e RL. As medida de avaliação foram a acurácia, precisão e a taxa de falsos positivos (TFP). Como resultado, o modelo RNA atingiu 99,71% de acurácia, 99,68% de precisão e 0,12% de TFP, seguida das redes *bayesianas*, com 97,52%, 97,04% e 2,5% de acurácia, precisão e TFP, respectivamente. Por outro lado, a RL teve o pior desempenho em todas as medidas, com acurácia de 94,7%, precisão 77,8% e 2,9% de TFP. Jain *et al.* (2019) também comenta que as abordagens com melhor desempenho também possuem maior custo de treino, em comparação o pior desempenho da regressão logística, que possui um custo não tão caro para o treino.

O trabalho de Khan *et al.* (2022) utiliza uma abordagem com o algoritmo de *Beetle Antennae Search* (BAS), que é uma metaheurística que simula o comportamento dos besouros

em sua busca por comida na natureza. O objetivo principal desse trabalho é a otimização para minimizar a função de perda. A fim de comparação, Khan *et al.* (2022) também são utilizados três outros algoritmos: *logit*, *Support Vector Machine* e *RUSBoost*. Para lidar com o grande desbalanceamento da base de dados, foram aplicados quatro técnicas:

- *Under-sampling*: A técnica reduz a quantidade da classe majoritária a mesma quantidade da classe minoritária. Sua única vantagem é a eficiencia computacional, visto que serão perdidos muitos dados, o que afeta de forma negativa a acurácia.
- Sobre-amostragem de dados com a classe minoritária: A abordagem iguala o número de dados da classe minoritária com a classe majoritária. Isso resolve o problema do desbalanceamento, porém, os dados continuam sem diversidade. É esperado que o modelo aprenda os dados da classe minoritária junto a um *overfitting*.
- Dividir e sobreamostrar: É uma técnica onde se divide a classe majoritária em N *subsets*, e, então, a classe minoritária é sobreamostrada em cada um. O objetivo é criar um comitê, onde o resultado é decidido pela maioria de votos dos N modelos. É uma abordagem computacionalmente cara e demorada.
- Sobreamostragem com a técnica *SMOTE*: A classe minoritária é sobreamostrada com a técnica *SMOTE* (*Synthetic Minority Oversampling Technique*). A *SMOTE* pega k amostras da classe minoritária e utiliza o KNN para computar a nova amostra. Isso é feito de forma iterativa, e termina quando se igualam o número de amostras de cada classe. A maior vantagem é a diversidade obtida na base de dados.

Os seis parâmetros utilizados para avaliação do desempenho foram: acurácia, precisão, sensitividade, ganho cumulativo descontado normalizado, AUROC e o tempo. Foi observado que o algoritmo BAS obteve o melhor resultado em todas as medidas, comparado aos outros algoritmos, visto que atingiu 84,9% de acurácia, 5,07% de sensitividade, 8,01% de precisão, 0,181% de ganho cumulativo descontado normalizado, 0,774 de valor para a AUROC e levou 91,52 segundos.

Ahmad *et al.* (2023) abordou a detecção de fraudes com uma técnica de agrupamento e seleção baseada em similaridade, utilizando o *Fuzzy C-means* em conjunto com a *Random Under-Sampling* (RUS), técnica que remove dados da classe majoritária de maneira aleatória. Como o RUS atua de maneira randomica, existe a possibilidade de excluir dados críticos, o que pode gerar uma queda de desempenho. O framework proposto utilizando o *fuzzy C-means* não só conserta os problemas do RUS, como também garante similaridade e integridade nas características dos dados.

O trabalho consistiu nas seguintes etapas:

- Pre-processamento e seleção de características.

- Agrupamento da base de dados com o *fuzzy C-means*.
- Balanceamento e distribuição dos dados (foram utilizados os dados com 50% e 50% (classe A), 34% e 66% (classe B), 25% e 75% (classe C) de transações fraudulentas e legítimas, respectivamente).
- Geração dos dados de treino e teste.
- Treino dos modelos utilizando os algoritmos ANN, LR, NB e KNN.
- Análise do desempenho dos modelos.

As medidas de avaliação utilizadas em Ahmad *et al.* (2023) foram a acurácia, precisão, *F1-Score*, sensitividade, especificidade e AUROC. Comparando o desempenho dos algoritmos, o ANN obteve as melhores classificações, com uma acurácia de 94,3%, 95,17% e 96,6% para as classes A, B e C, respectivamente. Em termos de precisão na identificação de registros de fraude, o kNN os menores resultados, com 89,1%, 89,6% e 89,4%, enquanto o NB obteve a maior precisão, nas classe A e B 97,5% e 94,5% na classe C. Todos os algoritmos mostraram bom desempenho na especificidade, com o ANN alcançando 98,9%, o LR com 98,6%, o NB com 98,4% e o kNN com 96,6%. Também é destacado que a distribuição 34% e 66% de fraudes e legítimas, respectivamente, mostra um aumento significativo nas medidas, principalmente na especificidade, o que mostra que o método proposto se mostra eficiente na identificação de transações legítimas.

Pode-se observar que a RL é uma abordagem utilizada em diversos trabalhos, e obtendo desempenho semelhante na maioria, tendendo sempre a ficar entre os métodos com as menores medidas de avaliação.

Existem outros diversos trabalhos relacionados que abordam diferentes contextos de fraude, bem como variados métodos para detecção das mesmas (ZHANG *et al.*, 2022), (ILEBERI; SUN; WANG, 2022), (ESENOGHO *et al.*, 2022), (SETTIPALLI; GANGADHARAN, 2023), (SISODIA; SISODIA, 2023).

Este trabalho, diferente dos citados, aborda uma comparação entre algoritmos baseados em árvore, sendo um supervisionado e outro não. Embora a metodologia seja diferente, o *BayesianSearch* foi utilizado, visto que enriqueceu os resultados, trazendo melhores pontuações nos índices de avaliação. Outra diferença em relação a trabalhos observados foi a divisão em treino, teste e validação, visto que o presente trabalho utiliza 50% dos dados para teste, com objetivo de obter uma base representativa utilizando uma menor amostragem. Também foram realizados testes estatísticos visando mostrar se existem diferenças significativas entre os modelos.

4 MATERIAIS E MÉTODOS

4.1 Materiais

- Python 3.11[1]: é uma linguagem interpretada e de alto nível.
- Numpy[2]: é uma biblioteca Python onde se trabalha com computação numérica.
- Pandas[3]: é uma biblioteca Python utilizada para análise e manipulação de dados.
- Scikit-learn[4]: é uma biblioteca Python voltada para algoritmos de AM.
- Matplotlib[5]: é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python.
- Visual Studio Code[6]: é um IDE utilizada para programação em geral.
- Notebook Acer Nitro 5
- Processador AMD Ryzen 7
- Placa de vídeo GTX 1650
- 8 *gigabytes* de memória RAM

4.1.1 Base de dados

A base de dados *Credit Card Fraud Detection* foi obtida no website Kaggle (2018), e sua extração foi feita via *API* (*Application Programming Interface*). Ela possui dados obtidos de setembro de 2013 de transações de cidadãos europeus, sendo o conteúdo obtido em 2 dias. O número total de transações e suas classificações estão descritos na tabela 1.

Classificação	Número	Porcentagem (%)
Legítimas	284315	99,828%
Fraudes	492	0.172%
Total	284807	100%

Tabela 1 – Distribuição das classificações dos dados

A base contém apenas atributos numéricos, dos quais, a maioria foi resultado da aplicação do método de redução de dimensionalidade denominado PCA. O PCA foi feito por questões

[1] https://www.python.org/
[2] https://numpy.org/
[3] https://pandas.pydata.org/
[4] https://scikit-learn.org/stable/
[5] https://matplotlib.org/
[6] https://code.visualstudio.com/

de anonimização, então as colunas possuem nomes enumerados de V1 a V28, e apenas 3 características permaneceram como na base original:

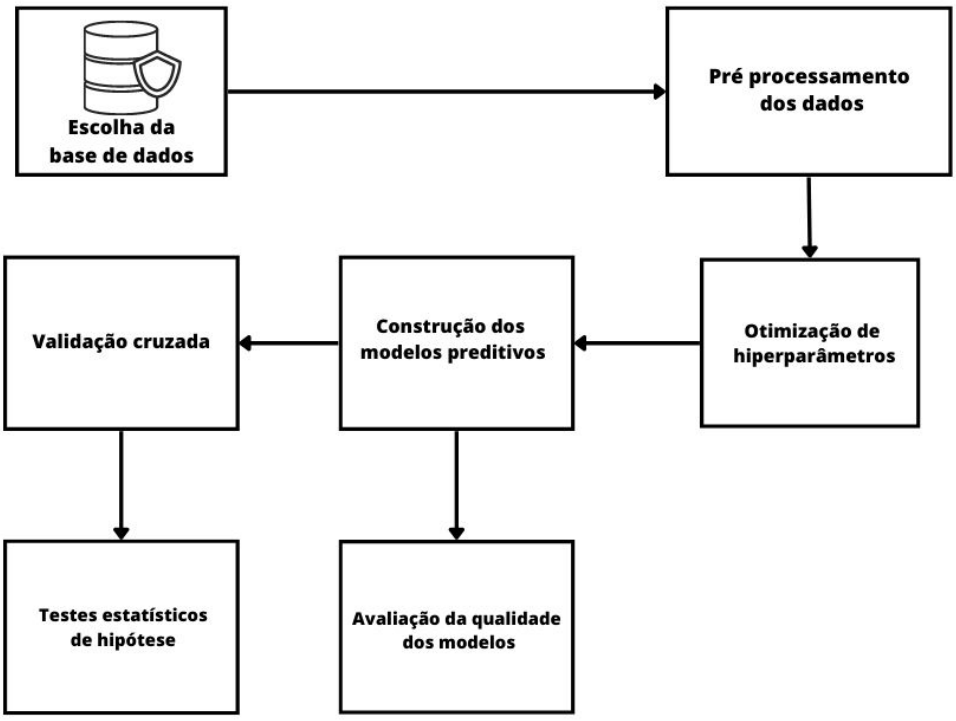
- *time*: representa o tempo decorrido tendo como referência a primeira transação da base.
- *amount*: é o valor da transação em questão.
- *class*: coluna em que se rotulam as transações como legítimas (0) ou fraudulentas (1).

A base de dados foi lida e transformada em um *dataframe* através da biblioteca Pandas.

4.2 Métodos

A Figura 5 representa as etapas que serão realizadas para a realização deste trabalho.

Figura 5 – Fluxograma das etapas do trabalho.



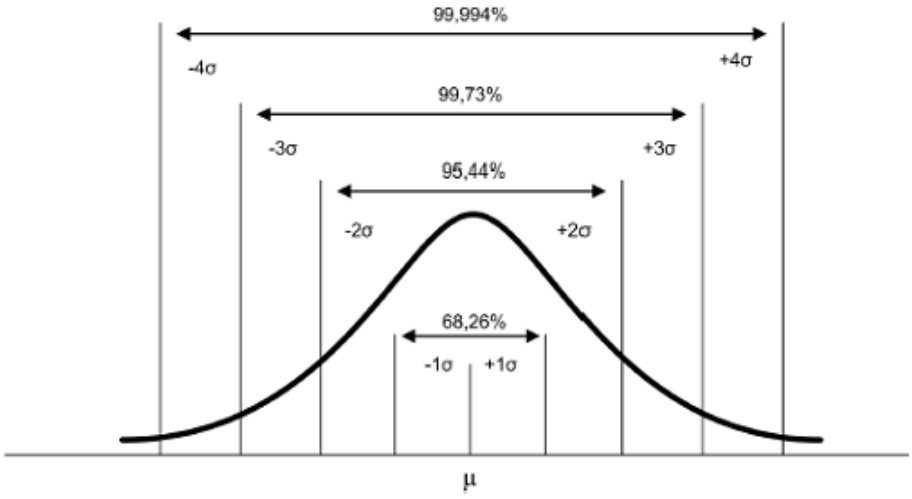
Fonte: Autoria própria (2023).

4.2.1 Pré Processamento dos dados

Na primeira etapa da avaliação experimental, a coluna de classe, que classifica as transações, é remapeada. Como os algoritmos utilizados para o trabalho classificam transações legítimas como 1, e fraudes como -1, foi utilizado um dicionário, que é representado por um par chave-valor, para fazer um de-para, e a função map para remodelar a coluna conforme este.

Em seguida, a técnica de padronização deverá ser aplicada na matriz de características para deixar as médias (μ) das colunas com valor 0, bem como seus desvio-padrões (σ) como 1. Para isso, foi utilizada a ferramenta *Standardscaler*, que realiza tal procedimento . Ao utilizar o método, obtem-se para todas as colunas um comportamento de curva gaussiana. A partir da origem (μ), uma gaussiana, como observado na Figura 6, possui cerca de 68,2% dos dados em uma distância de até 1(σ), aproximadamente 95,4% em até 2σ, e praticamente 99,7% em até 3σ (LARSON; FARBER; PATARRA, 2004). Como o presente trabalho busca observar outliers, é esperado que as anomalias encontradas estejam acima de 3σ.

Figura 6 – Curva gaussiana



Fonte: Wikimedia Commons (2016).

4.2.2 Construção dos modelos

Após a etapa anterior (pré-processamento), é feita uma separação dos dados em treino (25%), validação (25%) e teste (50%), sendo a divisão estratificada para manter a proporção de cada classe em todas as amostras (também foram realizados experimentos utilizando outras partições). Em todos os experimentos, a *seed* do *random_state* foi a mesma. O motivo dessa separação seria encontrar uma amostra representativa que gerasse bons resultados na aplicação do modelo na base de teste.

As ferramentas escolhidas são *Isolation Forest* (IF) e o *Extra Tree* (ET). Ambas as alternativas são algoritmos baseados em árvore, o que as dá a vantagem de possuirem baixo custo computacional para o treinamento dos modelos, e, dada a diferença de uma ser supervisionada e a outra não, foi feita uma comparação do desempenho de ambas.

Durante a construção de modelos é realizada a otimização de hiperparâmetros com o objetivo de melhorar o desempenho do modelo.

4.2.2.1 Otimização de hiperparâmetros

Nesta fase, foi utilizado o *BayesianSearchCV*, em que os principais hiperparâmetros são o *search_space* e o modelo em si. Os dados utilizados para encontrar os melhores hiperparâmetros foram os dados de validação. Para todos os experimentos, os hiperparâmetros gerais foram configurados da seguinte maneira:

- *estimator*: Recebe o modelo. Nos experimentos com a IF, recebe *random_state = 87*; nos com ET, recebe *class_weight=balanced* (necessário devido ao desbalanceamento de classes).
- *search_spaces*: Recebe o espaço de busca de cada algoritmo (um para IF, outro para ET).
- *n_iter*: Recebe o número de iterações, que teve valor 20 para todos os experimentos.
- cv: Validação Cruzada *k-fold*, onde foram usados 10 folds em todos os experimentos, visto que, em geral, é o valor recomendado para analisar a capacidade de generalização do modelo (SINGH; PANDA *et al.*, 2011).
- *n_jobs*: Número de processadores a serem utilizados, com valor -1 para utilização de todos da máquina.
- *scoring*: Medida de avaliação que o otimizador deseja melhorar. Foram feitos experimentações para otimização da eficiência (acurácia balanceada), área abaixo da curva ROC e coeficiente de correlação de Matthews.

Para a IF e a ET foram otimizados 3 pontuações no *BayesianSearch*: a eficiência (acurácia balanceada), a área abaixo da curva ROC e o coeficiente de correlação de Matthews. Os 6 modelos foram treinados utilizando a base de treino com os hiperparâmetros encontrados, e, posteriormente, foi feita a predição em todas as bases (treino, teste e validação), a fim de verificar *underfit* ou *overfit*.

4.2.3 Validação e avaliação dos modelos

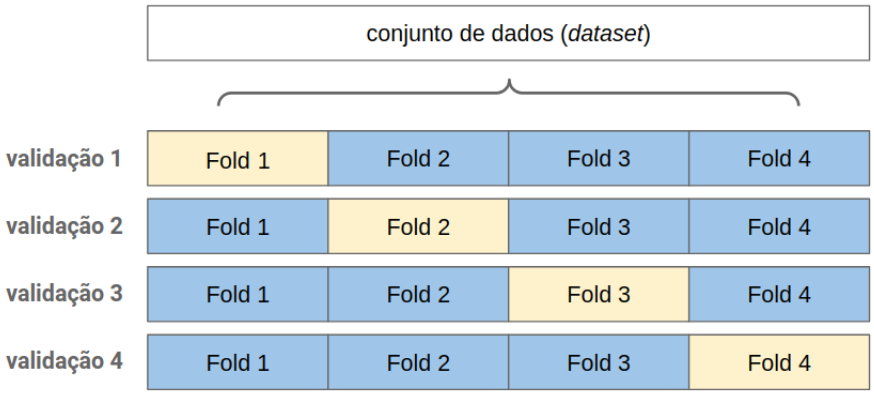
Após o treinamento do modelo, é necessário validar a qualidade e confiabilidade dos resultados. Para isso, as medidas utilizadas são extraídas de uma matriz de confusão, da qual serão obtidas:

- Acurácia.
- Precisão.
- Sensitividade.
- *F1-Score*.
- Especificidade.
- Taxa de falsos negativos.
- Taxa de falsos positivos.
- Área abaixo da *Receiver Operating Curve* (ROCAUC).
- Coeficiente de correlação de matthews.
- Eficiência.
- Média Geométrica.

Para a validação dos modelos foi aplicada a validação cruzada *k-fold* utilizando 10 *folds*. Este método utiliza os seguintes passos (BROWNLEE, 2018):

- Reordenação da base de dados de forma aleatória;
- Separação da base em K grupos de subamostras;
- Para cada grupo, este será utilizado como base de teste, enquanto todo o restante será utilizado na base de treino, como ilustra a Figura 7 (ou seja, cada amostra estará no modelo de teste 1 vez, enquanto K-1 vezes no de treino);
- Após o treino, avalia-se o modelo com a base de teste, se armazena o resultado em uma variável vetorial, e o modelo é descartado;
- Os resultados serão utilizados para avaliar a qualidade do treinamento.

Figura 7 – Validação cruzada *K-fold*



Fonte: Wikimedia Commons (2021).

Após o término do *K-fold*, os vetores da eficiência obtidos para cada modelo foram comparados entre si utilizando um teste estatístico não paramétrico, o teste de Friedman, que visa apontar se existe diferença significativa entre os modelos, o que é apontado caso o p-valor seja menor que 0.05. O motivo da escolha da eficiência para o teste se da pela razão de seu cálculo focar tanto para o acerto da classe majoritária (sensitividade) como acerto da classe minoritária (especificidade). Caso o teste de Friedman resultar em diferença estatisticamente significativa, o pós-teste de Nemenyi deverá ser executado com o propósito de encontrar os pares de modelos em que houve diferença estatisticamente significativa.

5 RESULTADOS E DISCUSSÕES

Para ambas, seja IF (*isolation forest*) ou ET (*extra trees*), foram feitos 3 experimentos, e cada um visou a otimização de uma medida de avaliação via *BayesianSearch*, sendo estas medidas:

- Eficiência (ou acurácia balanceada)
- Área abaixo da curva ROC
- Coeficiente de correlação de Matthews

Após encontrar os melhores hiperparâmetros após 20 iterações, o modelo foi treinado com os hiperparâmetros encontrados utilizando a base de treino.

Após a fase de treinamento, as três amostragens - treino, teste e validação - foram aplicadas ao método de predição. Em seguida, a base de teste foi utilizada em uma validação cruzada *K-fold* para obter as médias e desvios-padrões de cada *fold*. Para a matriz de confusão, considera-se 0 para transações fraudulentas, e 1 para transações legítimas. Os aquivos relacionados aos experimentos se encontram em (ASSIS, 2023).

5.1 *Isolation Forest*

5.1.1 Experimento 1 - Otimizando Eficiência

Após a otimização, o BayesianSearch atingiu uma Eficiência de $90{,}32\%$ para a IF.

Ao aplicar o método de predição à base de treino, foi obtida a seguinte matriz de confusão, representada pela Tabela 2.

		Valor Real	
		0	1
Predição	0	110	13
	1	8124	62954

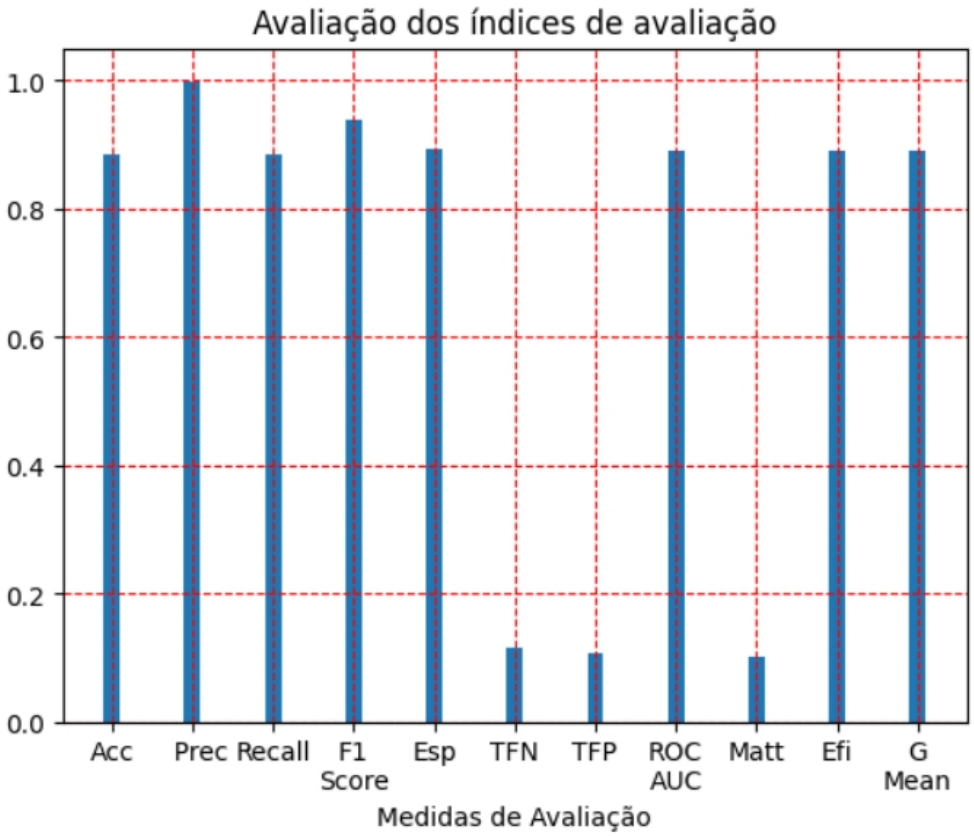
Tabela 2 – Matriz de confusão - IF e Eficiência - dados de treino

Onde pode-se observar que a taxa de acerto da classe fraudulenta, em proporção, parece maior que a da classe legítima, o que se confirma observando o gráfico, representado pela Figura 8.

Já na aplicação à base de validação, que foi utilizada para encontrar os melhores hiperparâmetros, é a matriz ilustrada pela Tabela 3.

A qual teve comportamento semelhante ao da predição em treino, como observado no seguinte gráfico da Figura 9.

Figura 8 – Medidas de avaliação em gráfico - IF e Eficiência - Treino



Fonte: Autoria própria (2023).

		Valor Real	
		0	1
Predição	0	113	10
	1	8560	62519

Tabela 3 – Matriz de confusão - IF e Eficiência - dados de validação

Por fim, ao aplicar a base de teste, que possui dados totalmente desconhecidos, a matriz de confusão é ilustrada pela Tabela 4. Pode-se notar que os valores praticamente foram duplicados, o que indica performance semelhante.

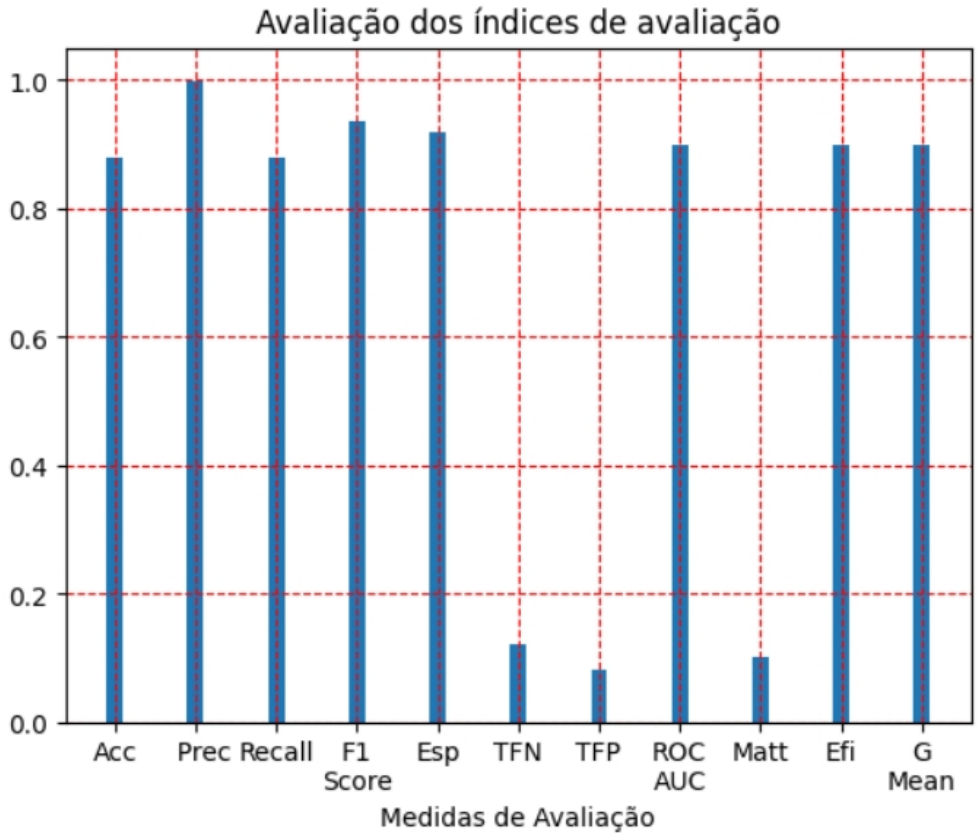
		Valor Real	
		0	1
Predição	0	223	23
	1	17254	124904

Tabela 4 – Matriz de confusão - IF e Eficiência - dados de teste

E esta gerou os resultados ilustrados no gráfico da Figura 10. Observando os gráficos, nota-se que em todas as predições o desempenho foi semelhante.

Observando a Tabela 5 e a Tabela 6, pode-se observar que para todas as amostragens, os resultados tiveram baixa diferença entre seus valores.

Figura 9 – Medidas de avaliação em gráfico - IF e Eficiência - Validação



Fonte: Autoria própria (2023).

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	88,57%	99,98%	88,57%	93,93%	89,43%	89,00%	10,13%	89,00%	89,00%
Validação	87,96%	99,98%	87,96%	93,59%	91,87%	89,91%	10,14%	89,91%	89,89%
Teste	87,87%	99,98%	87,86%	93,53%	90,65%	89,26%	9,94%	89,26%	89,25%

Tabela 5 – Medidas de avaliação - IF e Eficiência

Amostra	TFN	TFP
Treino	11,43%	10,57%
Validação	12,04%	8,13%
Teste	12,14%	9,35%

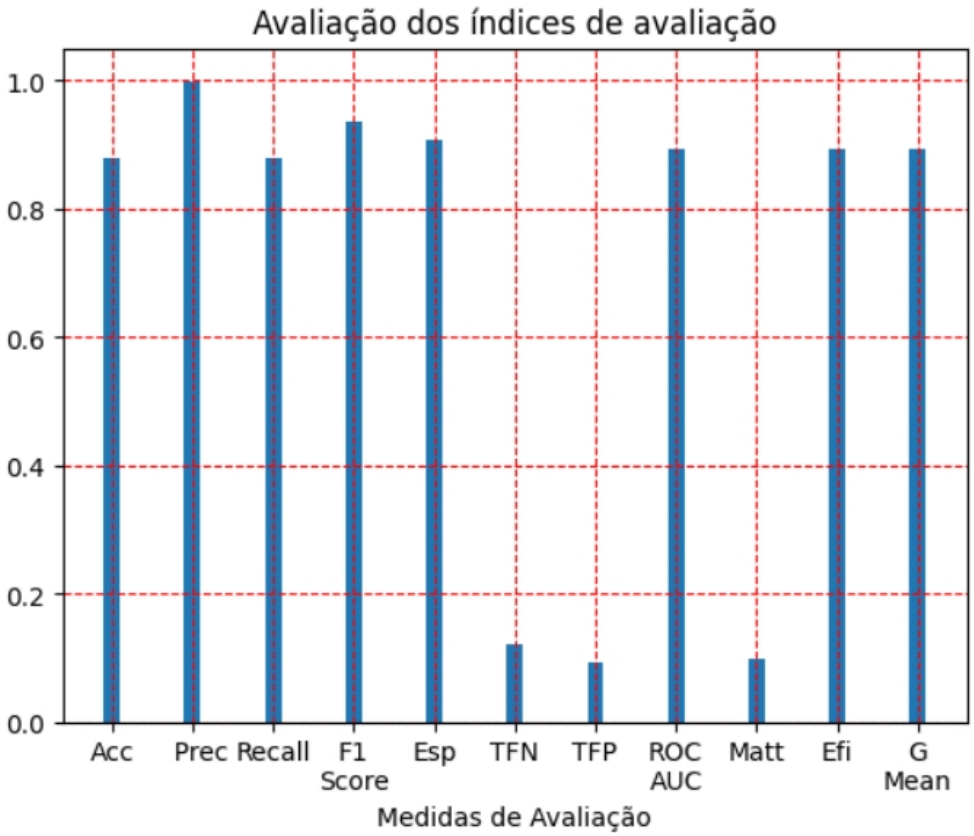
Tabela 6 – Taxas de falsos negativos e positivos - IF e Eficiência.

Na validação cruzada, obtem-se a média e desvio padrão dos índices de avaliação, que estão contidos nas Tabela 7 e Tabela 8. Pode-se observar que os maiores desvios-padrões estão na especificidade e taxa de falsos positivos. Além disso, nota-se que os valores são próximos aos das predições nas bases de treino, validação e teste.

Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	88,57%	99,98%	88,57%	93,93%	90,28%	89,43%	10,23%	89,43%	89,39%
Desvio-Padrão	0,29%	0,01%	0,29%	0,16%	4,89%	2,45%	1,38%	2,45%	2,44%

Tabela 7 – Resuldados da validação cruzada - IF e Eficiência

Figura 10 – Medidas de avaliação em gráfico - IF e Eficiência - Teste



Fonte: Autoria própria (2023).

Medida-estatística	TFN	TFP
Média	11,43%	9,72%
Desvio-Padrão	0,29%	4,89%

Tabela 8 – Taxas de falsos negativos e positivos - Validação cruzada - IF e Eficiência.

Na Tabela 5 pode-se notar que os valores de especificidade são maiores que os de sensitividade, o que mostra que o modelo diferencia melhor fraudes à transações legítimas. O Coeficiente de Matthews foi o único que não teve desempenho promissor. Dado que a base de teste é maior que ambas as bases de treino e validação, pode-se inferir que, dados os resultados, a amostragem foi representativa neste experimento.

5.1.2 Experimento 2 - Otimizando a área abaixo da curva ROC

Para a otimização da área abaixo da curva ROC na IF foi obtido o valor de $95{,}34\%$.

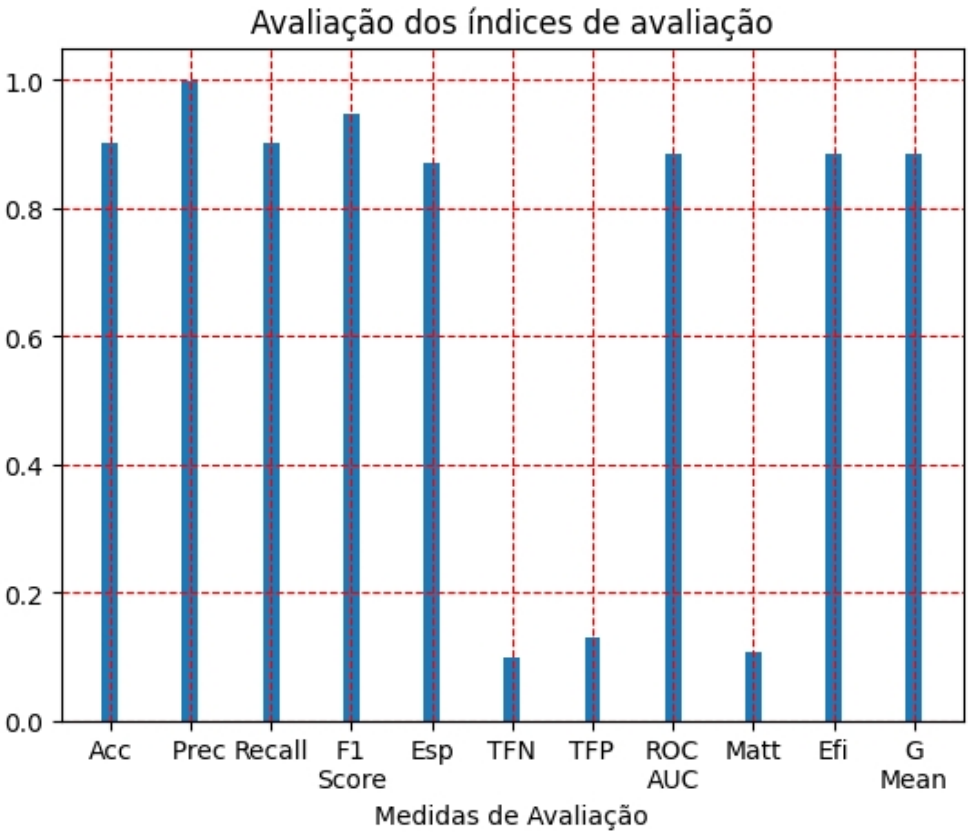
Aplicando a predição à base de treino, foi obtida a matriz de confusão ilustrada na Tabela 9. Observa-se que o acerto em treino foi ligeiramente inferior ao do experimento anterior.

Dada a matriz, foi o gráfico mostrado pela Figura 11 é criado. desta vez, a sensitividade é maior que a especificidade.

		Valor Real	
		0	1
Predição	0	107	16
	1	6943	64135

Tabela 9 – Matriz de confusão - IF e AUROC - dados de treino

Figura 11 – Medidas de avaliação em gráfico - IF e AUROC - Treino



Fonte: Autoria própria (2023).

Ao aplicar a predição aos dados de validação, foi obtida a seguinte matriz de confusão, mostrada na Tabela 10. Pode-se observar que a classe de fraudes teve um acerto ligeiramente maior.

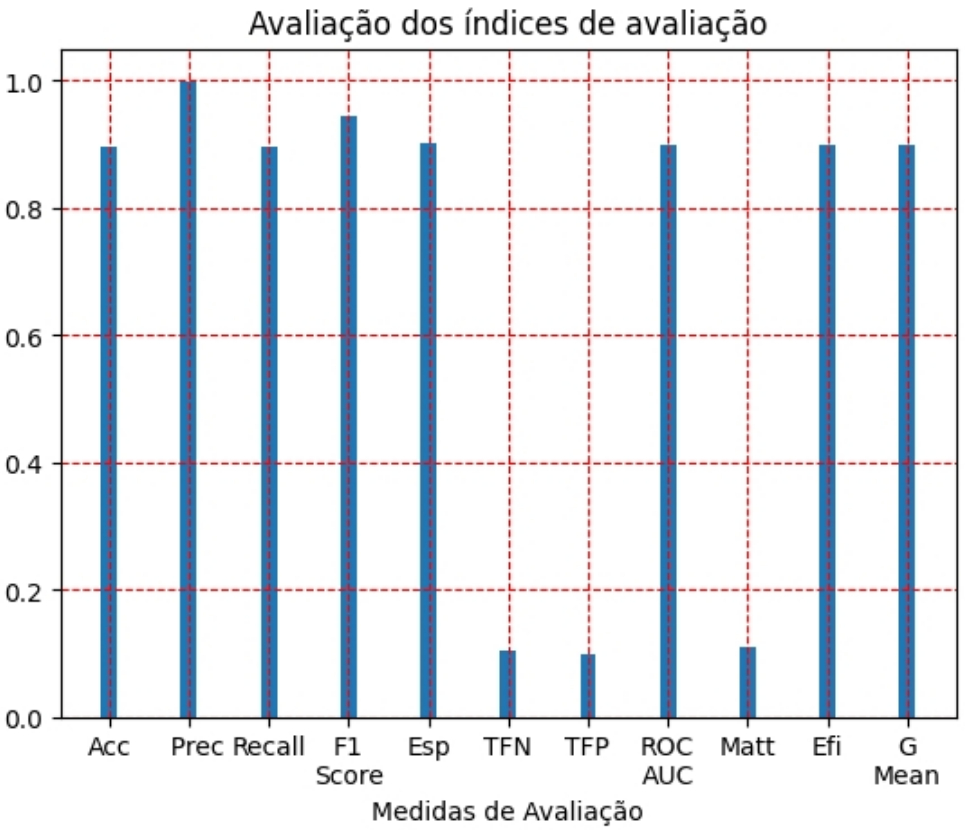
		Valor Real	
		0	1
Predição	0	111	12
	1	7364	63715

Tabela 10 – Matriz de confusão - IF e AUROC - dados de validação

Então, é criado o gráfico ilustrado pela Figura 12. Neste, dado o maior acerto da classe negativa, pode-se observar que a especificidade passou a ser maior que a sensitividade.

Por fim, para a matriz dos dados de teste, tem-se a Tabela 11. Nota-se que, apesar da queda de cerca de 14% da taxa de falsos negativos, a primeira linha da matriz é identica a primeira linha da Tabela 4, o que mostra que obtiveram a mesma especificidade.

Figura 12 – Medidas de avaliação em gráfico - IF e AUROC - Validação



Fonte: Autoria própria (2023).

		Valor Real	
		0	1
Predição	0	223	13
	1	14784	127374

Tabela 11 – Matriz de confusão - IF e AUROC - dados de teste

A partir da Tabela 11 foi gerado o gráfico ilustrado pela figura Figura 13. Pode-se notar que a especificidade se manteve maior que a sensitividade na predição de teste.

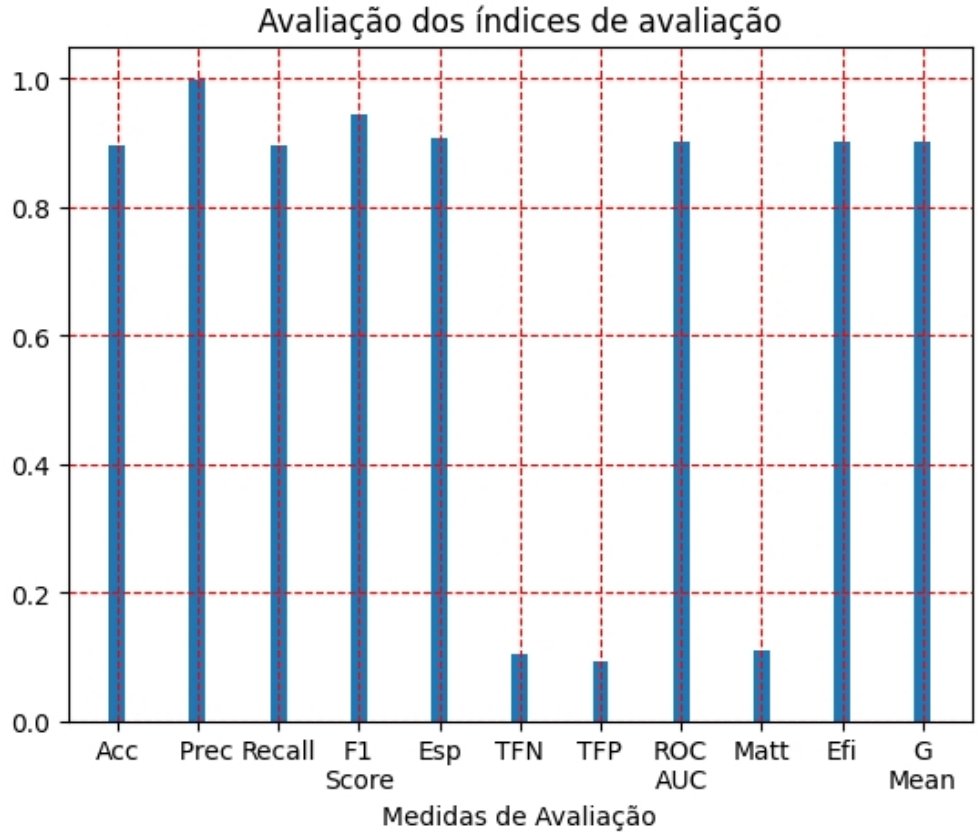
Novamente, ao observar a Tabela 12, nota-se que os resultados não apresentam grandes diferenças entre treino, teste e validação. O único ponto diferente, como mostra a Tabela 13, é a taxa de falsos positivos superar a taxa de falsos negativos.

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	90,23%	99,98%	90,23%	94,85%	86,99%	88,61%	10,74%	88,61%	88,60%
Validação	89,64%	99,98%	89,64%	94,53%	90,24%	89,94%	10,82%	89,94%	89,94%
Teste	89,60%	99,98%	89,60%	94,51%	90,65%	90,13%	10,85%	90,13%	90,12%

Tabela 12 – Medidas de avaliação - IF e AUROC

Observando os resultados da validação cruzada, obtem-se a média e desvio padrão para os índices de avaliação deste experimento, que estão nas Tabela 14 e Tabela 15. Pode-se observar que os maiores desvios-padrões estão, novamente, na especificidade e taxa de falsos

Figura 13 – Medidas de avaliação em gráfico - IF e AUROC - Teste



Fonte: Autoria própria (2023).

Amostra	TFN	TFP
Treino	9,77%	13,01%
Validação	10,36%	9,76%
Teste	10,40%	9,35%

Tabela 13 – Taxas de falsos negativos e positivos - IF e AUROC.

positivos, como também a proximidade dos valores aos resultados das predições nas bases de treino, validação e teste.

Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	90,25%	99,98%	90,25%	94,87%	90,28%	90,27%	11,19%	90,27%	90,23%
Desvio-Padrão	0,25%	0,01%	0,24%	0,14%	4,89%	2,46%	1,48%	2,46%	2,47%

Tabela 14 – Resuldados da validação cruzada - IF e AUROC

Medida-estatística	TFN	TFP
Média	9,75%	9,72%
Desvio-Padrão	0,24%	4,89%

Tabela 15 – Taxas de falsos negativos e positivos - Validação cruzada - IF e AUROC.

5.1.3 Experimento 3 - Otimizando o coeficiente de correlação de Matthews

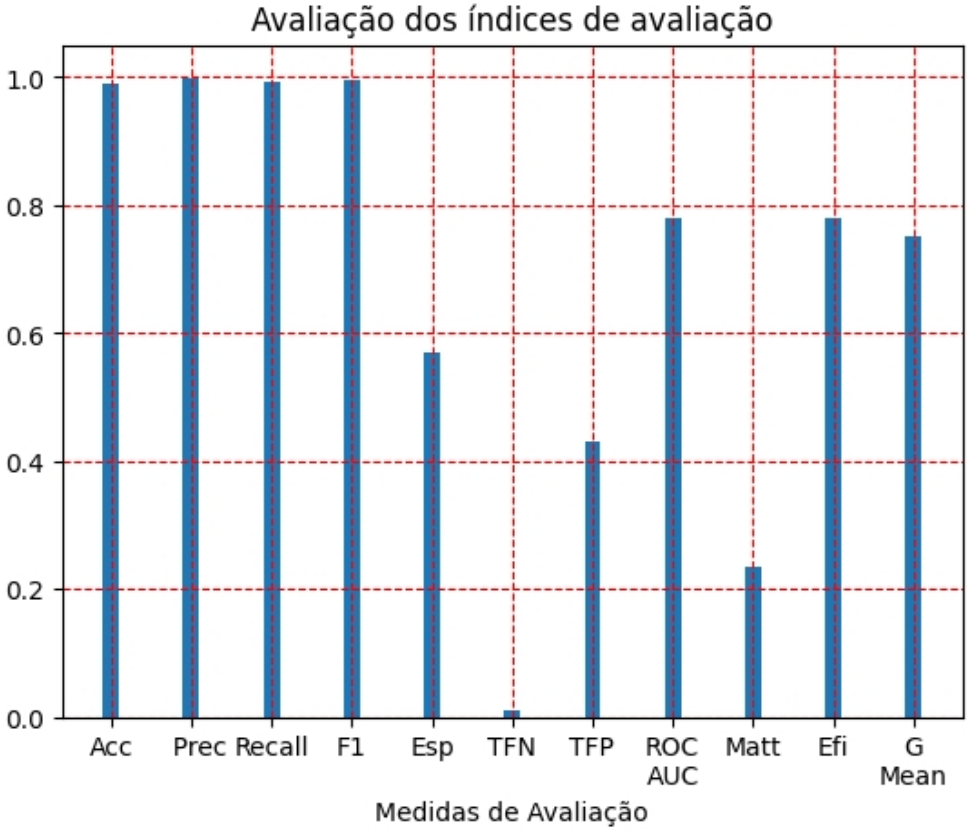
Na otimização do coeficiente de correlação de Matthews para a IF, o resultado obtido foi de $23{,}15\%$.

Na aplicação da predição na base de treino, foi obtida a matriz de confusão da Tabela 16, que gerou a Figura 14, que ilustra o gráfico com as medidas de avaliação. Pode-se observar, tanto na matriz como no gráfico, que houve queda e aumento significativos na especificidade e sensitividade, respectivamente.s

		Valor Real	
		0	1
Predição	0	70	53
	1	642	70436

Tabela 16 – Matriz de confusão - IF e Matt - dados de treino

Figura 14 – Medidas de avaliação em gráfico - IF e Matt - Treino



Fonte: Autoria própria (2023).

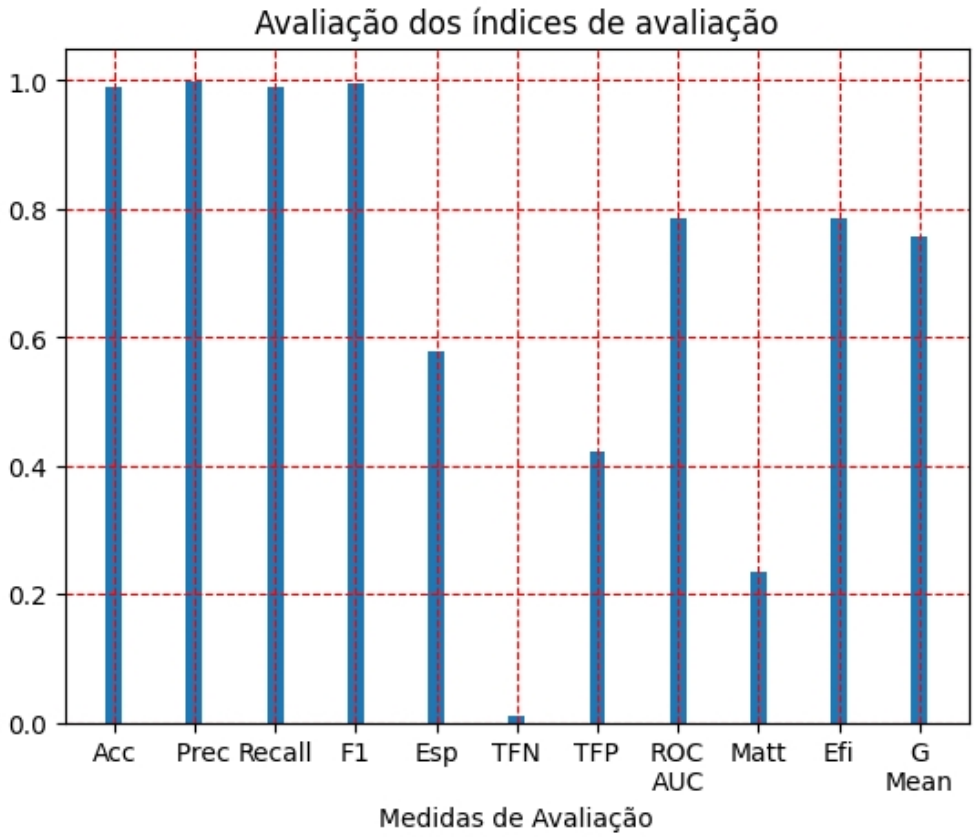
Já nos dados de validação, pode-se observar a seguinte matriz de confusão, mostrada na Tabela 17. O gráfico da Figura 15 mostra as medidas geradas a partir desta. Observa-se que não houveram grandes diferenças em relação a predição de treino.

Por fim, a seguinte matriz, presente na Tabela 18, é obtida da predição nos dados de teste. Em seguida pode-se observar as medidas no gráfico da Figura 16. Aqui se confirma a

		Valor Real	
		0	1
Predição	0	71	52
	1	657	70422

Tabela 17 – Matriz de confusão - IF e Matt - dados de validação

Figura 15 – Medidas de avaliação em gráfico - IF e Matt - Validação



Fonte: Autoria própria (2023).

diferença deste experimento em relação aos experimentos 1 e 2, visto que, seguindo as predições em treino e validação, a especificidade manteve a queda apresentada, bem como a sensitividade também esteve maior. Observa-se que o valor do Matt, embora aumentou consideravelmente, não teve um resultado promissor.

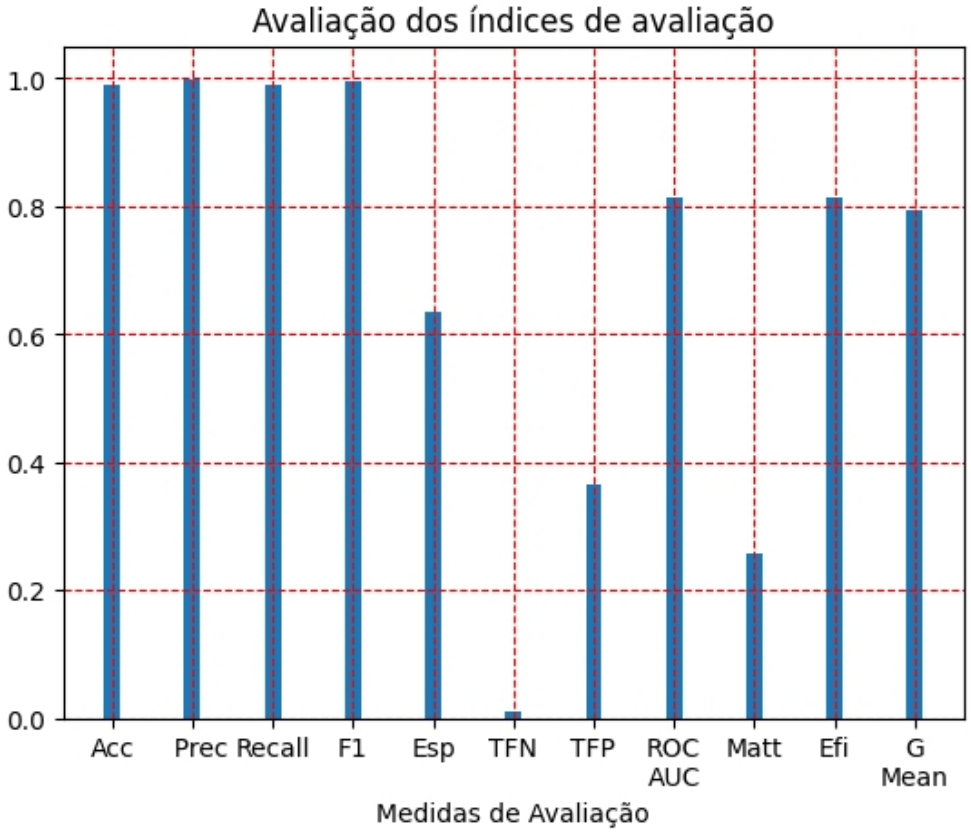
		Valor Real	
		0	1
Predição	0	156	90
	1	1302	140856

Tabela 18 – Matriz de confusão - IF e Matt - dados de teste

Os resultados deste experimento são apresentados nas Tabela 19 e Tabela 20.

Nesta etapa, a validação cruzada também obtve valores de média e desvio padrão semelhantes aos das medidas de avaliação deste experimento. Tais valores estão nas Tabela 21

Figura 16 – Medidas de avaliação em gráfico - IF e Matt - Teste



Fonte: Autoria própria (2023).

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	99,02%	99,92%	99,10%	99,51%	56,91%	78,00%	23,38%	78,00%	75,10%
Validação	99,00%	99,93%	99,08%	99,50%	57,72%	78,40%	23,45%	78,40%	75,62%
Teste	99,02%	99,94%	99,08%	99,51%	63,41%	81,25%	25,78%	81,25%	79,27%

Tabela 19 – Medidas de avaliação - IF e Matt

Amostra	TFN	TFP
Treino	0,90%	43,09%
Validação	0,92%	42,28%
Teste	0,92%	36,59%

Tabela 20 – Taxas de falsos negativos e positivos - IF e Matt.

e Tabela 22. Novamente, ambas especificidade e taxa de falsos positivos obtiveram o maior desvio padrão, bem como apresentaram os menores desempenhos deste trabalho.

Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	99,04%	99,93%	99,11%	99,52%	61,83%	80,47%	25,45%	80,47%	78,18%
Desvio-Padrão	0,07%	0,01%	0,07%	0,04%	6,46%	3,23%	3,92%	3,23%	4,04%

Tabela 21 – Resultados da validação cruzada - IF e Matt

Medida-estatística	TFN	TFP
Média	0,89%	0,07%
Desvio-Padrão	38,17%	6,46%

Tabela 22 – Taxas de falsos negativos e positivos - Validação cruzada - IF e Matt.

5.2 *Extra Tree*

5.2.1 Experimento 1 - Otimizando Eficiência

A otimização da eficiência para ET obteve o valor de $93{,}29\%$.

Na aplicação da predição à base de treino, foi obtida a matriz de confusão da Tabela 23, para a qual se destaca pontuação de 100% de especificidade, que pode ser evidência de um possível *overfit*.

		Valor Real	
		0	1
Predição	0	123	0
	1	153	70925

Tabela 23 – Matriz de confusão - ET e Eficiência - dados de treino

Foram gerados resultados a partir desta, apresentados pelo gráfico representado pela Figura 17. Nela, pode-se observar um Matt superior a todos os obtidos nos experimentos com a IF, bem como desempenho de quase 100% na medida ROC.

Na Tabela 24, tem-se a matriz de confusão da predição dos dados de validação, em que se pode notar que, desta vez, houve erros na classificação de fraudes. Esta gerou a Figura 18, onde nota-se que a especificidade caiu em cerca de 14%.

		Valor Real	
		0	1
Predição	0	106	17
	1	170	70909

Tabela 24 – Matriz de confusão - ET e Eficiência - dados de validação

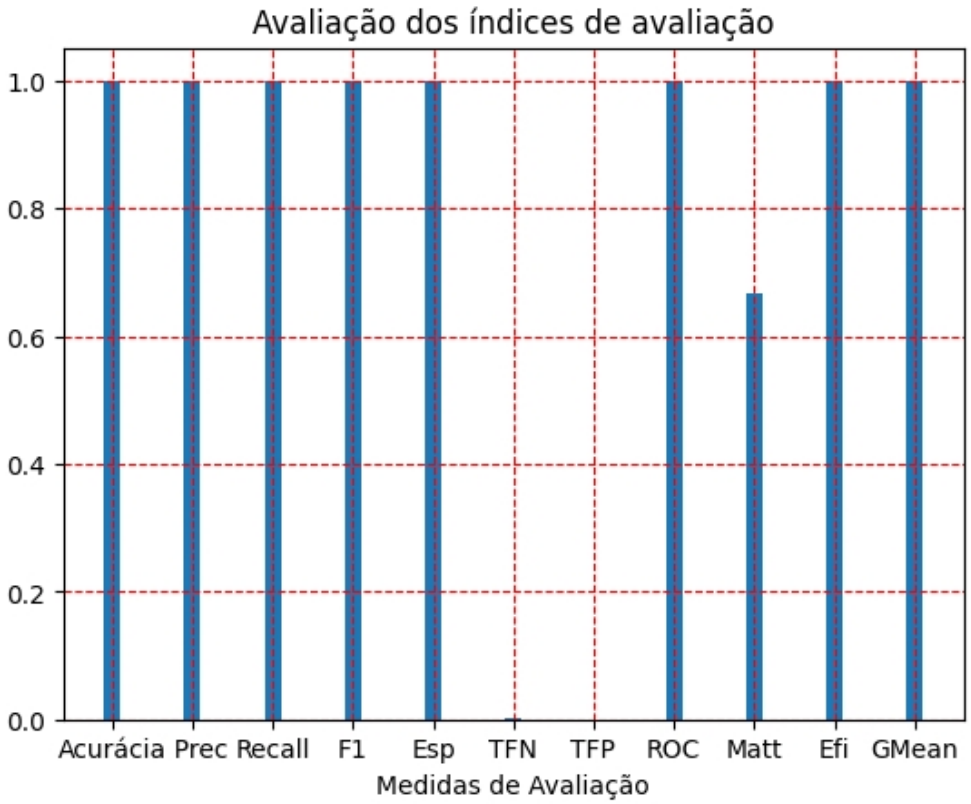
Para o teste, a matriz se ilustra na Tabela 25. Seus valores nas predições negativas são praticamente o dobro da Tabela 24. Seu gráfico é apresentado pela Figura 19. Nota-se que este é semelhante ao gráfico da Figura 18.

		Valor Real	
		0	1
Predição	0	211	35
	1	308	141850

Tabela 25 – Matriz de confusão - ET e Eficiência - dados de teste

Apresentam-se os resultados, de forma numérica, nas Tabela 26 e Tabela 27. Destaca-se que houve diferença notável dentre as especificidades, eficiências e *G-means*. Isso pode ser

Figura 17 – Medidas de avaliação em gráfico - ET e Eficiência - Treino



Fonte: Autoria própria (2023).

um indicador de overfit. Porém, apesar da queda, dada a semelhança dos dados de validação e teste, os resultados parecem promissores.

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	99,79%	100,00%	99,78%	99,89%	100,00%	99,89%	66,69%	99,89%	99,89%
Validação	99,74%	99,98%	99,76%	99,87%	86,18%	92,97%	57,43%	92,97%	92,72%
Teste	99,76%	99,98%	99,78%	99,88%	85,77%	92,78%	58,96%	92,78%	92,51%

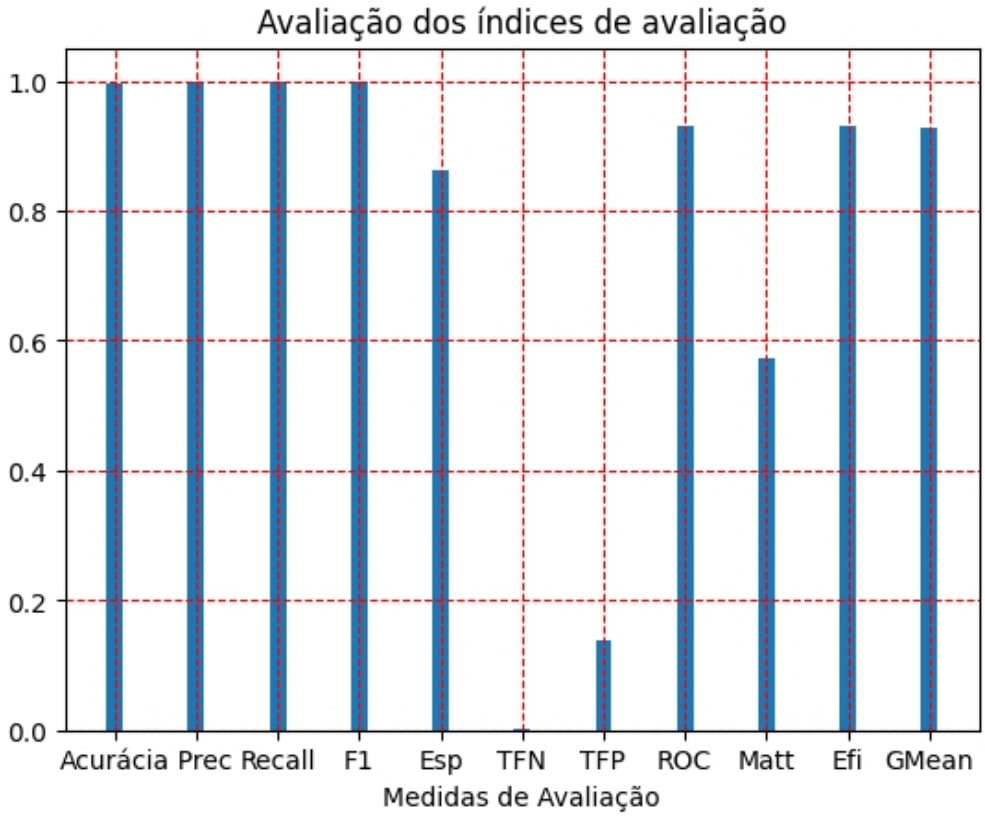
Tabela 26 – Medidas de avaliação - ET e Eficiência

Amostra	TFN	TFP
Treino	0,22%	0,00%
Validação	0,24%	13,82%
Teste	0,22%	14,23%

Tabela 27 – Taxas de falsos negativos e positivos - ET e Eficiência.

Na validação cruzada, obtem-se a média e desvio padrão das medidas de avaliação escolhidas, contidos nas Tabela 28 e Tabela 29. Observa-se que o modelo consegue classificar bem a classe positiva, visto que possui alta sensitividade. O modelo também superou todos modelos de IF nos coeficientes de correlação de Matthews.

Figura 18 – Medidas de avaliação em gráfico - ET e Eficiência - validação



Fonte: Autoria própria (2023).

Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	99,73%	99,98%	99,75%	99,86%	86,74%	93,25%	56,83%	93,25%	92,98%
Desvio-Padrão	0,02%	0,01%	0,03%	0,01%	4,81%	2,40%	4,59%	2,40%	2,55%

Tabela 28 – Resuldados da validação cruzada - ET e Eficiência

Medida-estatística	TFN	TFP
Média	0,25%	13,26%
Desvio-Padrão	0,03%	4,81%

Tabela 29 – Taxas de falsos negativos e positivos - Validação cruzada - IF e Eficiência.

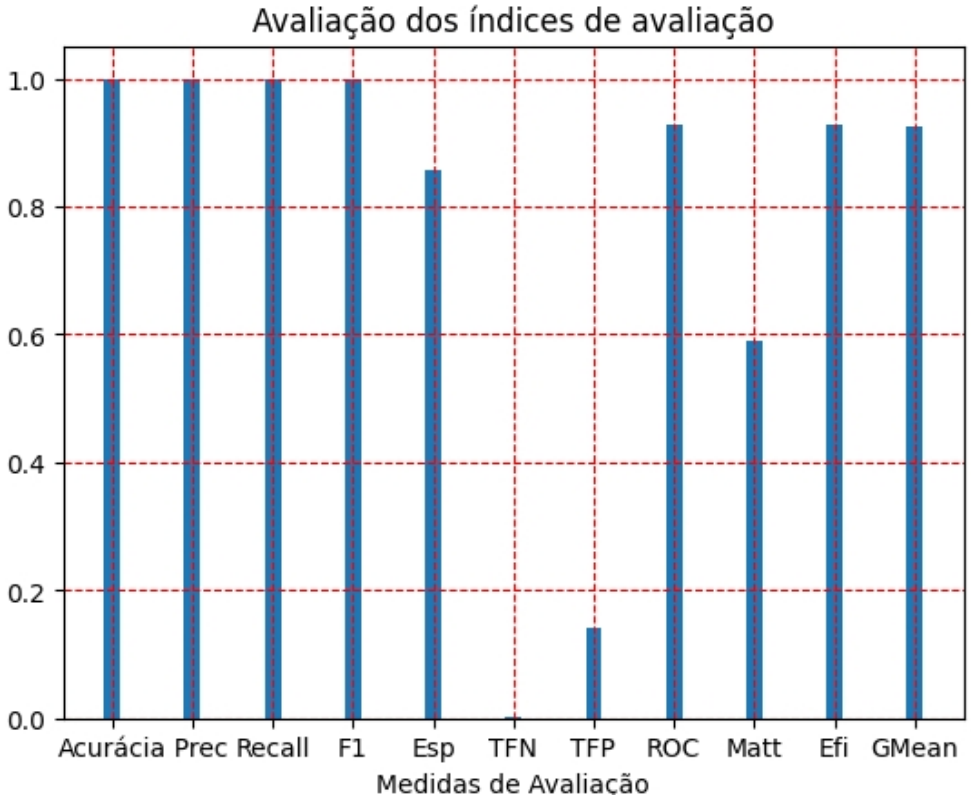
5.2.2 Experimento 2 - Otimizando a área abaixo da curva ROC

Para este experimento, a pontuação da área abaixo da curva ROC obtida pelo otimizador foi de $97{,}69\%$.

Aplicando a predição à base de treino, se obteve a matriz de confusão da Tabela 30. Diferente da Tabela 23, observa-se que ocorreram erros na predição de fraudes.

A partir desta, apresentam-se os resultados pelo gráfico representado pela Figura 20. Embora este experimento buscou a otimização da área abaixo da curva ROC, pode-se evidenciar que teve desempenho inferior ao gráfico ilustrado pela Figura 17, visto que existiu o indício de um possível *overfit* dada especificidade de 100%.

Figura 19 – Medidas de avaliação em gráfico - ET e Eficiência - Teste



Fonte: Autoria própria (2023).

		Valor Real	
		0	1
Predição	0	107	16
	1	115	70963

Tabela 30 – Matriz de confusão - ET e AUROC - dados de treino

Em seguida, para a base de validação, a Tabela 31 mostra a matriz de confusão gerada. Esta mostra resultados semelhantes à Tabela 30. Enquanto a Figura 21 apresenta o gráfico com seus resultados, semelhantes aos da Figura 20.

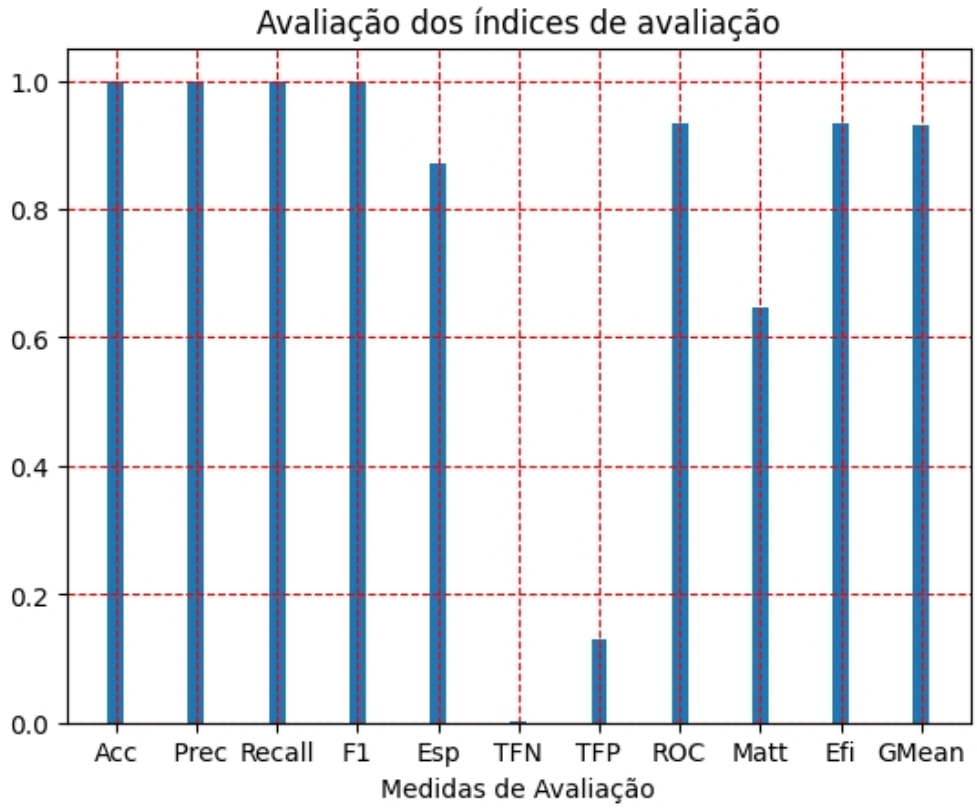
		Valor Real	
		0	1
Predição	0	104	19
	1	131	70948

Tabela 31 – Matriz de confusão - ET e AUROC - dados de validação

Já nos dados de teste, foi gerada a Tabela 32, em que observa-se que a linha da predição de fraudes obteve exatamente o dobro dos valores da Tabela 31. Seu gráfico é ilustrado pela Figura 22, com os respectivos resultados, em que se nota que não houve grande diferença em relação as predições em treino e validação.

Nas Tabela 33 e Tabela 34 pode-se observar o os valores em cada predição. Diferente do experimento 1 com a ET, e conforme observado nas matrizes e gráficos deste experimento, os

Figura 20 – Medidas de avaliação em gráfico - ET e AUROC - Treino



Fonte: Autoria própria (2023).

		Valor Real	
		0	1
Predição	0	208	38
	1	223	141935

Tabela 32 – Matriz de confusão - ET e AUROC - dados de teste

valores não apresentaram grandes diferenças entre as predições nas diferentes amostragens, o que indica que não houve overfit.

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	99,82%	99,98%	99,84%	99,91%	86,99%	93,42%	64,68%	93,42%	93,19%
Validação	99,79%	99,97%	99,82%	99,89%	84,55%	92,18%	61,09%	92,18%	91,87%
Teste	99,82%	99,97%	99,84%	99,91%	84,55%	92,20%	63,80%	92,20%	91,88%

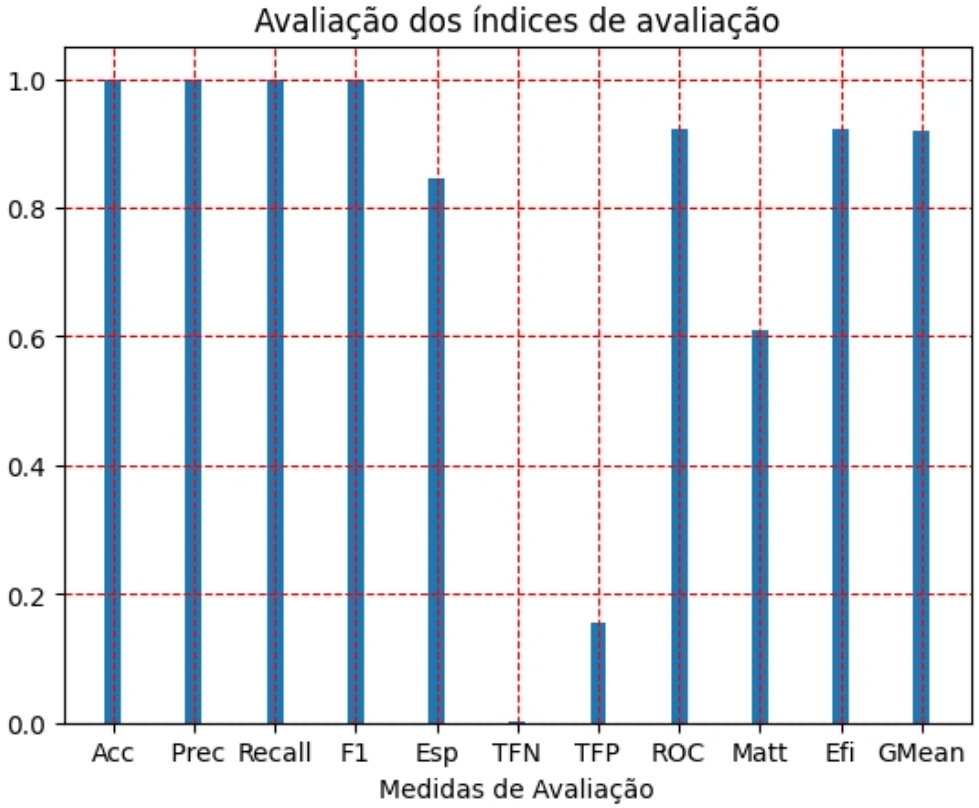
Tabela 33 – Medidas de avaliação - ET e AUROC

Amostra	TFN	TFP
Treino	0,16%	13,01%
Validação	0,18%	15,45%
Teste	0,16%	15,45%

Tabela 34 – Taxas de falsos negativos e positivos - ET e AUROC.

Na fase de validação cruzada, média e desvio padrão das medidas de avaliação escolhidas são obtidas, estando nas Tabela 35 e Tabela 36. Observa-se, novamente, que o modelo

Figura 21 – Medidas de avaliação em gráfico - ET e AUROC - Validação



Fonte: Autoria própria (2023).

consegue classificar bem a classe positiva, visto que possui alta sensitividade. O modelo novamente também superou todos a IF em todos os coeficientes de correlação de Matthews, por outro lado, esteve com o segundo maior desvio padrão.

Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	99,82%	99,97%	99,84%	99,91%	84,70%	92,27%	63,84%	92,27%	91,91%
Desvio-Padrão	0,03%	0,01%	0,03%	0,01%	5,77%	2,89%	4,94%	2,89%	3,14%

Tabela 35 – Resuldados da validação cruzada - ET e AUROC

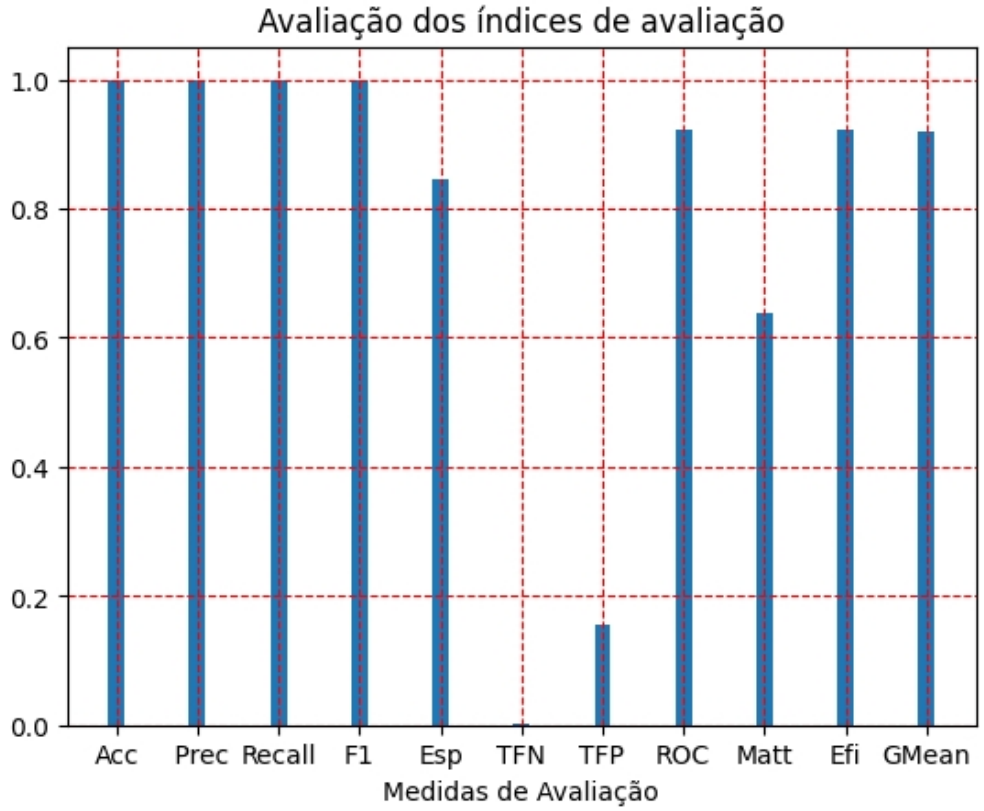
Medida-estatística	TFN	TFP
Média	0,16%	15,30%
Desvio-Padrão	0,03%	5,77%

Tabela 36 – Taxas de falsos negativos e positivos - Validação cruzada - ET e AUROC.

5.2.3 Experimento 3 - Otimizando o coeficiente de correlação de Matthews

Por fim, neste experimento, o valor do coeficiente de correlação de Matthews atingido pelo otimizador foi de $76,42\%$.

Figura 22 – Medidas de avaliação em gráfico - ET e AUROC - Teste



Fonte: Autoria própria (2023).

Aplicando a predição à base de treino, foi obtida a matriz de confusão da Tabela 37. Um destaque, novamente, a pontuação perfeita de especificidade, indicando um possível overfit.

		Valor Real	
		0	1
Predição	0	123	0
	1	31	71047

Tabela 37 – Matriz de confusão - ET e Matt - dados de treino

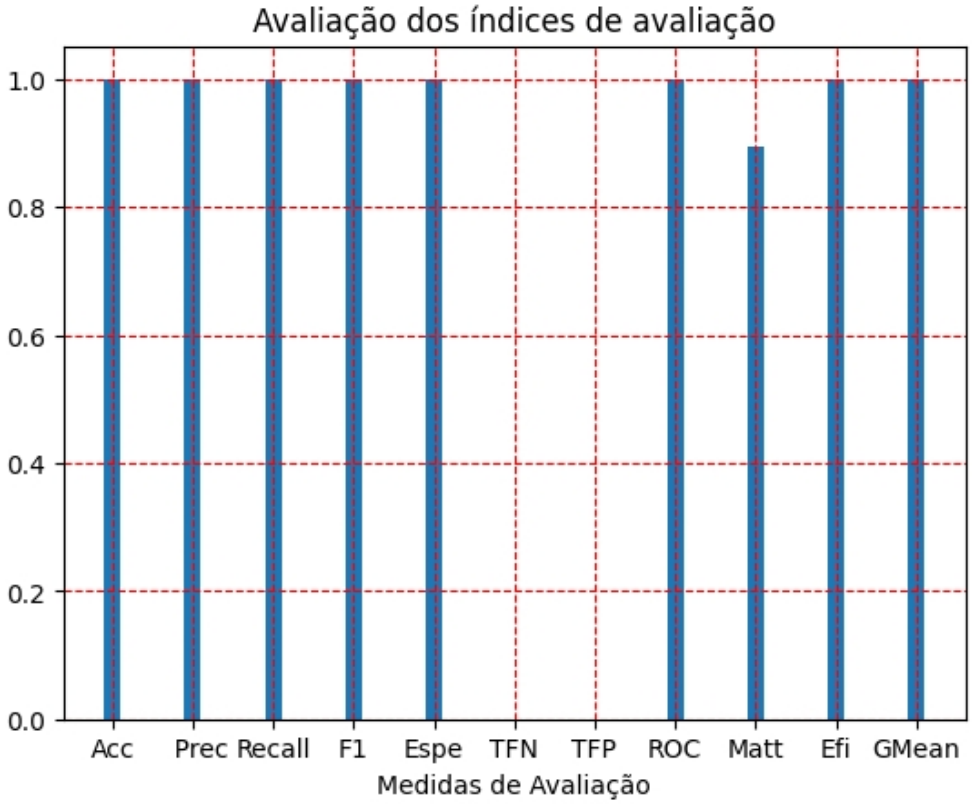
Os resultados da Tabela 37 são apresentados pelo gráfico da Figura 23. Novamente, pode-se observar o desempenho de quase 100% do indice ROC.

Já na predição da validação, tem-se a Tabela 38 representando a matriz de confusão, e suas medidas explanadas no gráfico da Figura 24. Observando-as, semelhante ao experimento em que se otimiza eficiência, nota-se novamente que houve erro nas predições de fraude, bem como uma queda na especificidade.

		Valor Real	
		0	1
Predição	0	103	20
	1	50	71029

Tabela 38 – Matriz de confusão - ET e Matt - dados de validação

Figura 23 – Medidas de avaliação em gráfico - ET e Matt - Treino



Fonte: Autoria própria (2023).

Por fim, para a base de teste, tem-se a matriz de confusão ilustrada pela Tabela 39 e seus resultados apresentados pela Figura 25. Nestes, observa-se um comportamento parecido com o experimento que otimizou eficiência, visto que os resultados de teste seguem proporção semelhante ao experimento de validação.

		Valor Real	
		0	1
Predição	0	208	38
	1	65	142093

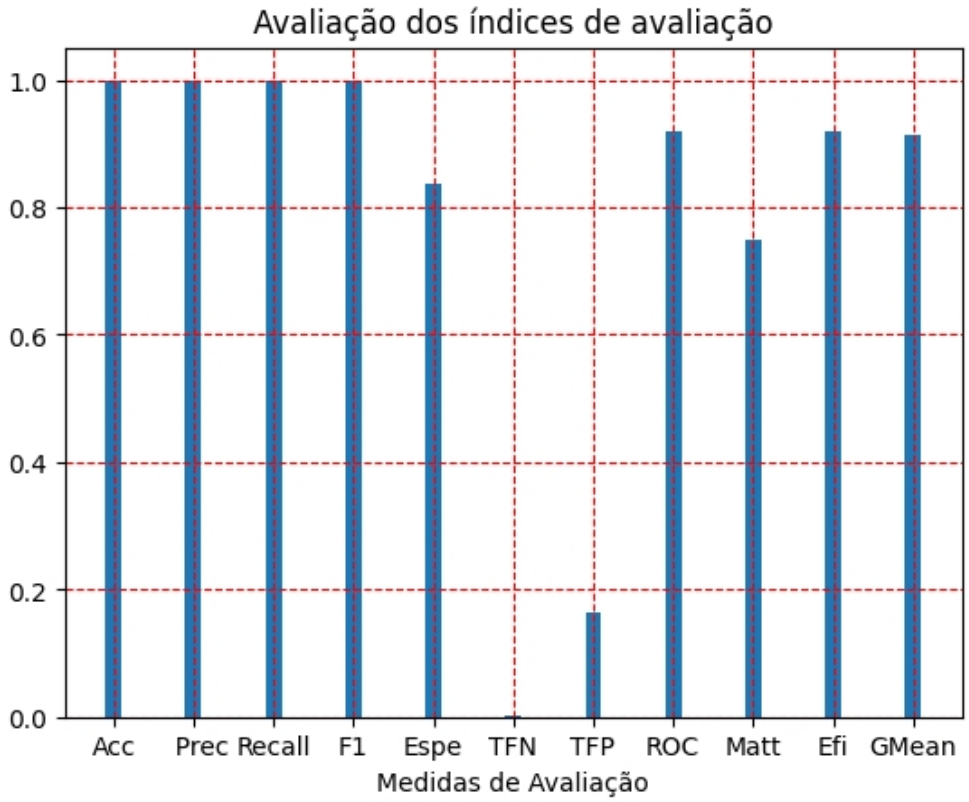
Tabela 39 – Matriz de confusão - ET e Matt - dados de teste

Nas Tabela 40 e Tabela 41 pode-se observar o os valores em cada predição. Da mesma forma que o experimento 1 com a ET, os valores de treino apresentaram uma notável diferença nas medidas de especificidade e taxa de falsos positivos, o que indica um possível overfit.

Amostra	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Treino	99,96%	100,00%	99,96%	99,98%	100,00%	99,98%	89,35%	99,98%	99,98 %
Validação	99,90%	99,97%	99,93%	99,95%	83,74%	91,83%	75,04%	91,83%	91,48%
Teste	99,93%	99,97%	99,95%	99,96%	84,55%	92,25%	80,23%	92,25%	91,93%

Tabela 40 – Medidas de avaliação - ET e Matt

Figura 24 – Medidas de avaliação em gráfico - ET e Matt - Validação



Fonte: Autoria própria (2023).

Amostra	TFN	TFP
Treino	0,04%	0,00%
Validação	0,07%	16,26%
Teste	0,05%	15,45%

Tabela 41 – Taxas de falsos negativos e positivos - ET e Matt.

Por fim, na fase de validação cruzada, obtem-se a média e desvio padrão das medidas de avaliação escolhidas, estando nas Tabela 42 e Tabela 43. Novamente pode-se notar a boa classificação da classe positiva, dada a alta sensitividade.

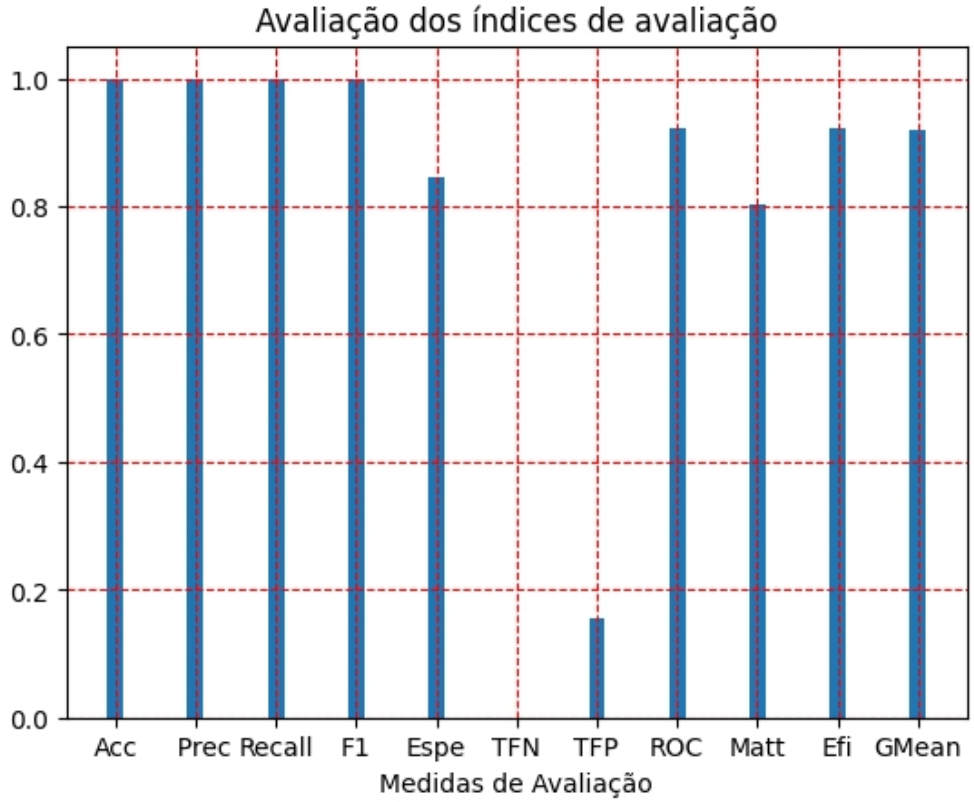
Medida-estatística	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
Média	99.91%	99,98%	99.93%	99.95%	85.95%	92.94%	77.11%	92.94%	92.65%
Desvio-Padrão	0.02%	0,01%	0,02%	0,01%	4.47%	2.23%	3.84%	2.23%	2.40%

Tabela 42 – Resuldados da validação cruzada - ET e Matt

Medida-estatística	TFN	TFP
Média	0.07%	14.05%
Desvio-Padrão	0,02%	4.47%

Tabela 43 – Taxas de falsos negativos e positivos - Validação cruzada - ET e Matt.

Figura 25 – Medidas de avaliação em gráfico - ET e Matt - Teste



Fonte: Autoria própria (2023).

5.3 Comparações entre os modelos

Cada um dos modelos foi otimizado visando a melhora de uma medida de avaliação. As tabelas a seguir apresentam comparações utilizando as bases de teste, visto que foram dados não utilizados para otimização de hiperparâmetros ou para treino (ou seja, dados não vistos).

Na Tabela 44, é notável a diferença entre os valores de sensitividade e especificidade. Observa-se que a IF aprendeu melhor a classificar fraudes, em proporção menor ao aprendizado da ET sobre as transações legítimas, que quase obteve pontuação máxima na sensitividade. Também nota-se uma grande diferença nos coeficientes de correlação de Matthews, onde o método supervisionado obteve um desempenho muito maior ao não supervisionado.

Método	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
IF	87,87%	99,98%	87,86%	93,53%	90,65%	89,26%	9,94%	89,26%	89,25%
ET	99,76%	99,98%	99,78%	99,88%	85,77%	92,78%	58,96%	92,78%	92,51%

Tabela 44 – Comparação sobre a otimização da eficiência

Já na Tabela 45 observa-se o mesmo cenário, em que as maiores diferenças estão na sensitividade, especificidade e coeficiente de correlação de Matthews.

Na Tabela 46, é mostrado que, diferente das comparações das Tabela 44 e Tabela 45, a IF teve desempenho semelhante a ET na medida de sensitividade. Por outro lado, foi o único

Método	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
IF	89,60%	99,98%	89,60%	94,51%	90,65%	90,13%	10,85%	90,13%	90,12%
ET	99,82%	99,97%	99,84%	99,91%	84,55%	92,20%	63,80%	92,20%	91,88%

Tabela 45 – Comparação sobre a otimização da área abaixo da curva ROC

modelo de IF que teve desempenho inferior ao da ET na medida de especificidade. Apesar da maior pontuação no coeficiente de correlação de Matthews entre as IF, foi o modelo com pior desempenho.

Método	Acc	Prec	Sens	F1	Esp	ROC	Matt	Efi	Gmean
IF	99,02%	99,94%	99,08%	99,51%	63,41%	81,25%	25,78%	81,25%	79,27%
ET	99,93%	99,97%	99,95%	99,96%	84,55%	92,25%	80,23%	92,25%	91,93%

Tabela 46 – Comparação sobre a otimização do coeficiente de correlação de Matthews

5.4 Testes estatísticos

Para verificar se houve diferença estatisticamente significativa entre modelos, foi aplicado o teste de Friedman, considerando o nível de significância de 95%, nos resultados da validação cruzada. Como resultado, foi obtido um p-valor menor que 0,0001, constatando diferença extremamente significativa entre os modelos.

Para verificar quais pares de modelos houveram diferença estatisticamente significativa, foi realizado o pós-teste de Nemenyi, onde, de acordo com a Tabela 47, foi constatada diferença estatisticamente significativa entre o modelo IF otimizado pelo coeficiente de correlação de Matthews com todos os modelos criados com a ET, dado que seus valores estão abaixo de 0.05. Dessa forma, pode-se com concluir, com 95% de certeza, que os modelos ET tiveram desempenho superior em comparação com o modelo IF-Matt.

Modelo	ET-ROC	ET-Matt	ET-Efi	IF-ROC	IF-Matt	IF-Efi
ET-ROC	1,0000	0,9969	0,9966	0,9189	0,0039	0,7086
ET-Matt	0,9969	1,0000	1,0000	0,6741	0,0004	0,3858
ET-Efi	0,9966	1,0000	1,0000	0,6671	0,0004	0,3790
IF-ROC	0,9189	0,6741	0,6671	1,0000	0,1195	0,9982
IF-Matt	0,0039	0,0004	0,0004	0,1195	1,0000	0,3081
IF-Efi	0,7086	0,3858	0,3790	0,9982	0,3081	1,0000

Tabela 47 – Teste de Nemenyi

5.5 Comparações com trabalhos relacionados

Nesta seção, os resultados deste trabalho são comparados com pesquisas relacionadas que utilizaram a mesma base de dados. A medida de comparação foi a especificidade, visto que foi encontrada em todos.

Awoyemi, Adetunmbi e Oluwadare (2017) utilizou técnicas de amostragem para criar dados sintéticos e modificar as proporções de transações legítimas e fraudulentas. Embora não tenha utilizado algoritmos não supervisionados e nem baseados em árvores, pode-se dizer que seu desempenho foi superior ao do presente trabalho, visto que encontrou especificidades de 100% em 2 experimentos, enquanto neste, a melhor teve valor de $90{,}65\%$.

No trabalho de Cardoso e Vieira (2019), o melhor desempenho dos experimentos foi de $64{,}4\%$ para especificidade, com o algoritmo *Naive Bayes*, modelo supervisionado, que esteve apenas $1{,}01\%$ acima do pior desempenho deste trabalho na medida em questão.

Varmedja *et al.* (2019) utilizou a técnica SMOTE para balancear a quantidade de cada classe, considerando a proporção de 50% para cada. Em seguida, foi feita uma separação de 80% de dados para treino e 20% para teste. Embora com o pior desempenho em relação aos outros algoritmos, dada a baixa precisão, a RL obteve a maior especificidade, atingindo $91{,}84\%$, que superou a maior pontuação deste trabalho nesta medida. Varmedja *et al.* (2019) considera que a RF obteve o melhor desempenho, com especificidade de $81{,}63\%$, valor superado por 5 dos experimentos feitos neste trabalho.

É válido lembrar que as comparações são feitas meramente entre valores, visto que as abordagens de detecção de fraude entre os trabalhos foram diferentes. Pode-se apontar também o uso de algoritmos baseados em árvore, o que torna os modelos mais facilmente interpretáveis por humanos. Embora este trabalho não necessariamente atingiu os maiores índices de avaliação, deve-se frisar que seu objetivo era encontrar uma amostra representativa para a base de treino, o qual teve 50% da base de dados reservada. Essa abordagem se difere de outros trabalhos em geral, onde normalmente o teste tem uma amostragem de menor fração.

6 CONCLUSÃO

A detecção de fraudes em cartões de crédito detém um papel crucial na segurança financeira e proteção dos consumidores. Ao longo deste trabalho, foram explorados métodos baseados em árvores de decisão junto à otimização *bayesiana* para potencializar a detecção de fraudes.

As anomalias são consideradas de grande valor em diversos estudos para a aplicação de métodos de aprendizado de máquina, e conseguir entender o seu comportamento e detectá-las é um processo que gera grande valor para as mais diversas áreas, já que sua detecção pode gerar uma ação preventiva. Por fim, o presente trabalho teve êxito nos experimentos e mostrou resultados promissores para detecção de fraudes.

6.0.1 Principais Dificuldades do Trabalho

Dado o desbalanceamento da base de dados, é evidente que um dos maiores desafios seria equilibrar a sensitividade e a especificidade (medidas que avaliam a correta classificação de cada classe). Dada a proporção de fraudes, um modelo que simplesmente classifica todas as transações como legítimas teria alta sensitividade, ao custo de deixar passar todas as fraudes. Por outro lado, deve-se ter cuidado ao treinar o modelo para classificar os dados apenas como fraudes, pois embora seja melhor classificar uma transação legítima como fraude à uma fraudulenta como legítima, deve-se ter cuidado com a proporção dessa ocorrência. Outra adversidade foi a realização dos experimentos em um notebook, visto que em outro ambiente poderiam ser realizados experimentos que demandam de mais recursos computacionais.

6.0.2 Principais Contribuições

Ao observar o desempenho dos experimentos, pode-se considerar os resultados promissores, visto que são apresentados valores de desempenho competitivos para a classificação de ambas as transações. Vale destacar que os dados foram separados em 25% para validação, 25% para treino e 50% para teste, visando buscar uma base representativa e que seja escalável na predição de novos dados desconhecidos.

Pode-se notar que a abordagem supervisionada teve melhor aprendizado das transações legítimas, dada a alta sensitividade, que também observa-se ser maior com um maior valor do coeficiente de correlação de Matthews (MCC). Observando as pontuações do MCC nas aplicações da IF, modelo não supervisionado, nota-se seu baixo desempenho.

Por fim, é importante ressaltar que a IF, modelo não supervisionado focado em detecção de anomalias, atingiu maior especificidade que o modelo supervisionado utilizado, o que mostra o potencial do modelo na detecção de fraudes.

6.0.3 Trabalhos Futuros

Para futuros experimentos, técnicas de amostragem mostram ter grande valor, visto que também estão presentes em trabalhos correlatos.

Embora a base de dados já tenha passado por um processo de anonimização, a utilização de métodos de redução de dimensionalidade pode ser avaliada.

A criação de um comitê de classificadores também pode ser interessante para melhorar o desempenho preditivo, visto que alguns modelos tendem a classificar melhor cada classe, então, as fraquezas de cada modelo podem ser supridas nas fortalezas de outros.

REFERÊNCIAS

AHMAD, H. *et al.* Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (sbs). **International Journal of Information Technology**, Springer, v. 15, n. 1, p. 325–333, 2023.

ASSIS, R. L. **TCC Models**. 2023. https://github.com/im-not-rhuan/TCC_Models.

AWOYEMI, J. O.; ADETUNMBI, A. O.; OLUWADARE, S. A. Credit card fraud detection using machine learning techniques: A comparative analysis. *In*: IEEE. **2017 international conference on computing networking and informatics (ICCNI)**. [*S.l.*], 2017. p. 1–9.

AWS. **Amazon Machine Learning: Developer Guide**. 2016. https://docs.aws.amazon.com/pdfs/machine-learning/latest/dg/machinelearning-dg.pdf#model-fit-underfitting-vs-overfitting. Accessed: 2022-11-10.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BREIMAN, L. **Classification and regression trees**. [*S.l.*]: Routledge, 2017.

BROWNLEE, J. A gentle introduction to k-fold cross-validation. **Machine learning mastery**, v. 2019, 2018.

CARDOSO, M. d. C.; VIEIRA, H. F. Detecção de fraude no uso de cartões de crédito utilizando técnicas supervisionadas de machine learning. jun. 2019.

COOK, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. **Circulation**, Am Heart Assoc, v. 115, n. 7, p. 928–935, 2007.

DINIZ, L. **Sim, dados são o novo petróleo!** 2021. https://www.industria40.ind.br/artigo/20949-sim-dados-sao-novo-petroleo. Accessed: 2022-08-24.

DUARTE, N. S. A dinâmica ética das máquinas em inteligência artificial e o novo papel do professor de filosofia. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 7, n. 6, p. 968–995, 2021.

ELLIOTT, A. C.; HYNAN, L. S. A sas® macro implementation of a multiple comparison post hoc test for a kruskal–wallis analysis. **Computer methods and programs in biomedicine**, Elsevier, v. 102, n. 1, p. 75–80, 2011.

ESENOGHO, E. *et al.* A neural network ensemble with feature engineering for improved credit card fraud detection. **IEEE Access**, IEEE, v. 10, p. 16400–16407, 2022.

FERREIRA, E. V.; ZEVIANI, W. M. Aprendizado não supervisionado. **Universidade Federal do Paraná (UFPR), entre**, 2013.

GIL, M. A. **IA vai movimentar 30 trilhões de dólares em 2030", diz especialista**. 2021. https://epocanegocios.globo.com/Upload/noticia/2021/08/ia-vai-movimentar-30-trilhoes-de-dolares-em-2030-diz-especialista.html. Accessed: 2022-11-13.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining**. [*S.l.*]: Gulf Professional Publishing, 2005.

GONÇALVES, A. **O que é SSL / TLS e HTTPS?** 2022. https://www.hostinger.com.br/tutoriais/o-que-e-ssl-tls-https. Accessed: 2022-10-14.

HONDA, H.; FACURE, M.; YAOHAO, P. **Os Três Tipos de Aprendizado de Máquina**. 2017. https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/. Accessed: 2022-11-10.

HOPPEN, J.; PRATES, W. **Outliers, o que são e como tratá-los em uma análise de dados?** 2017. https://www.aquare.la/o-que-sao-outliers-e-como-trata-los-em-uma-analise-de-dados/. Accessed: 2022-09-01.

HORKY, P.; PROKEŠ, A.; HUBÁČEK, P. Unsupervised time series pattern recognition for purpose of electronic surveillance. *In*: IEEE. **2022 24th International Microwave and Radar Conference (MIKON)**. [*S.l.*], 2022. p. 1–5.

ILEBERI, E.; SUN, Y.; WANG, Z. A machine learning based credit card fraud detection using the ga algorithm for feature selection. **Journal of Big Data**, SpringerOpen, v. 9, n. 1, p. 1–17, 2022.

INSURTALKS. **Inteligência Artificial é capaz de detectar fraudes em reclamações de baixo valor e grandes volumes em seguros**. 2022. https://www.insurtalks.com.br/posts/inteligencia-artificial-e-capaz-de-detectar-fraudes-em-reclamacoes-de-baixo-valor-e-grandes-volumes-em- Accessed: 2022-08-24.

JAIN, Y. *et al.* A comparative analysis of various credit card fraud detection techniques. **Int J Recent Technol Eng**, v. 7, n. 5S2, p. 402–407, 2019.

JIAWEI, H.; MICHELINE, K.; JIAN, P. **Data mining concepts and techniques**. 2016.

JUNIOR, G. d. B. V. *et al.* Importância do índice fowlkes-mallows (fmi), do coeficiente de correlação de matthews (mcc) e do índice youden (iy) nos classificadores de inteligência artificial na área da saúde. **Revista CPAQV–Centro de Pesquisas Avançadas em Qualidade de Vida| Vol**, v. 14, n. 3, p. 2, 2022.

JUNIOR, S. C. L. **Aplicação de regras de associação para o desenvolvimento de sistemas de recomendação de materiais bibliográficos da UFRN**. 2022. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2022.

KAGGLE. **Credit Card Fraud Detection**. 2018. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Accessed: 2022-08-25.

KHAN, A. T. *et al.* Fraud detection in publicly traded us firms using beetle antennae search: A machine learning approach. **Expert Systems with Applications**, Elsevier, v. 191, p. 116148, 2022.

KOEHRSEN, W. Overfitting vs. underfitting: A complete example. **Towards Data Science**, 2018.

KURITA, T. Principal component analysis (pca). **Computer Vision: A Reference Guide**, Springer, p. 1–4, 2019.

LARSON, R.; FARBER, B.; PATARRA, C. tradução técnica. **Estatística aplicada**. [*S.l.*]: Prentice Hall, 2004.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. *In*: IEEE. **2008 eighth ieee international conference on data mining**. [*S.l.*], 2008. p. 413–422.

LOTERMAN, G. *et al.* Benchmarking regression algorithms for loss given default modeling. **International Journal of Forecasting**, Elsevier, v. 28, n. 1, p. 161–170, 2012.

MAES, S. *et al.* Credit card fraud detection using bayesian and neural networks. ago. 2002.

MAIER, O. *et al.* Extra tree forests for sub-acute ischemic stroke lesion segmentation in mr sequences. **Journal of neuroscience methods**, Elsevier, v. 240, p. 89–100, 2015.

MEWTWO. **Pokémon: Mewtwo Contra-ataca**. 1998.

MULLICK, S. S. *et al.* Appropriateness of performance indices for imbalanced data classification: An analysis. **arXiv e-prints**, p. arXiv–2008, 2020.

NGUYEN, V. *et al.* Filtering bayesian optimization approach in weakly specified search space. **Knowledge and Information Systems**, Springer, v. 60, p. 385–413, 2019.

OLIVA, J. T. **Geração automática de laudos médicos para o diagnóstico de epilepsia por meio do processamento de eletroencefalogramas utilizando aprendizado de máquina**. 2019. Tese (Doutorado) — Universidade de São Paulo, 2019.

OLIVEIRA, A. F. Testes estatísticos para comparação de médias. **Revista Eletrônica Nutritime**, v. 5, n. 6, p. 777–788, 2008.

PINTELAS, E.; LIVIERIS, I. E.; PINTELAS, P. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. **Algorithms**, MDPI, v. 13, n. 1, p. 17, 2020.

RAZENTE, H. L.; JUNIOR, C. T. Análise visual em processos de redução de dimensionalidade para mineração em sistemas de bases de dados. 2004.

SETTIPALLI, L.; GANGADHARAN, G. Wmtdbc: An unsupervised multivariate analysis model for fraud detection in health insurance claims. **Expert Systems with Applications**, Elsevier, v. 215, p. 119259, 2023.

SILVA, B. M. d. Inteligência artificial, aprendizado de máquina. 2005.

SILVA, E. H. N. da; KORITIAKI, N. A.; MELEM, V. M. Exemplos de aplicação do teste de friedman. **Anais do Pró-Ensino: Mostra Anual de Atividades de Ensino da UEL**, n. 3, p. 94–94, 2021.

SILVA, L. N. M. **Tipos de aprendizado de máquina e algumas aplicações**. 2021. http://www2.decom.ufop.br/terralab/tipos-de-aprendizado-de-maquina-e-algumas-aplicacoes/. Accessed: 2022-11-12.

SINGH, G.; PANDA, R. K. *et al.* Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: a small agricultural watershed, kapgari, india. **Int. J. Earth Sci. Eng**, v. 4, n. 06, p. 443–450, 2011.

SISODIA, D.; SISODIA, D. S. A hybrid data-level sampling approach in learning from skewed user-click data for click fraud detection in online advertising. **Expert Systems**, Wiley Online Library, v. 40, n. 2, p. e13147, 2023.

SKLEARN. **Cross-validation: evaluating estimator performance**. 2020. https://scikit-learn.org/stable/modules/cross_validation.html. Accessed: 2022-08-25.

SOFTTEK. **Redução da dimensionalidade na Machine Learning**. 2021. Accessed: 2022-08-25.

SOUZA, J. O. d. **Detecção de outliers em pipelines de dados**. 2021. Monografia (Especialização em em Ciência de Dados) — Universidade Tecnológica Federal do Paraná, Parana, 2021.

SUTTO, G. **Brasil teve alta de quase 33% nas tentativas de fraude com cartão de crédito no 1° semestre, mostra estudo**. 2021. https://www.infomoney.com.br/minhas-financas/

brasil-teve-alta-de-quase-33-nas-tentativas-de-fraude-com-cartao-de-credito-no-1-semestre-mostra-estud
Accessed: 2022-10-13.

SYARIF, I.; PRUGEL-BENNETT, A.; WILLS, G. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. **TELKOMNIKA (Telecommunication Computing Electronics and Control)**, v. 14, n. 4, p. 1502–1509, 2016.

TACONELI, C. Uma introdução aos modelos uni e multivariados de classificação e regressão por árvores. 12 2022.

TERAYAMA, K. *et al.* Black-box optimization for automated discovery. **Accounts of Chemical Research**, ACS Publications, v. 54, n. 6, p. 1334–1346, 2021.

VARMEDJA, D. *et al.* Credit card fraud detection-machine learning methods. *In*: IEEE. **2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)**. [*S.l.*], 2019. p. 1–5.

VIEIRA, H. F. **DETECÇÃO DE FRAUDE NO USO DE CARTÕES DE CRÉDITO UTILIZANDO TÉCNICAS SUPERVISIONADAS DE MACHINE LEARNING**. 2019. Monografia (Bacharelado em Engenharia de Computação) — Centro Universitário de Anápolis, Goiás, 2019.

YAMADA, F. T. *et al.* **Um estudo sobre a influência de métodos imparciais de análise de componentes principais em problemas de classificação**. 2021. Tese (Doutorado) — [sn], 2021.

ZHANG, G. *et al.* efraudcom: An e-commerce fraud detection system via competitive graph neural networks. **ACM Transactions on Information Systems (TOIS)**, ACM New York, NY, v. 40, n. 3, p. 1–29, 2022.