**APRIMORANDO A DETECÇÃO DE FRAUDES FINANCEIRAS EM INSTITUIÇÕES BANCÁRIAS: UMA ABORDAGEM DE MACHINE LEARNING COM ÁRVORES DE DECISÃO E RANDOM FOREST**

Amanda Rocha Evaristo¹\*;Wagner Wilson Bortoletto2 ]

1 Itaú Unibanco. Licenciatura em Física. Rua Sônia Maria de Moraes Angel Jones, Pq Oziel; 13049-060 Campinas, SP, Brasil

2 PECEGE. Mestre em Engenharia de Produção e Manufatura. Rua Antônio Menardi, Residencial Victor D’Andrea; 13481-344 Limeira, SP, Brasil

\*autor correspondente: rochae.amanda@gmail.com

**APRIMORANDO A DETECÇÃO DE FRAUDES FINANCEIRAS EM INSTITUIÇÕES BANCÁRIAS: UMA ABORDAGEM DE MACHINE LEARNING COM ÁRVORES DE DECISÃO E RANDOM FOREST**

**Resumo**

Este trabalho tem por objetivo explorar a eficácia de técnicas avançadas de Machine Learning, especificamente árvores de decisão e Random Forest, na detecção de fraudes financeiras em instituições bancárias. Por meio de uma revisão sistemática da literatura, identificamos as principais abordagens e desafios na implementação dessas tecnologias no contexto bancário, incluindo a seleção de variáveis, o impacto do tamanho e do balanceamento dos conjuntos de dados, e a importância das técnicas ensemble para melhorar a detecção de fraudes. O estudo destacou a necessidade de equilibrar precisão e eficiência operacional, considerando o tempo de treinamento e inferência dos modelos, e enfatizou a relevância de soluções escaláveis que possam se adaptar ao crescimento do volume de transações e à evolução das técnicas fraudulentas. Os resultados sugerem que, apesar dos desafios, a aplicação de Machine Learning no combate a fraudes financeiras oferece um potencial significativo para aumentar a segurança e eficiência das operações bancárias. A pesquisa aponta para a importância de avanços contínuos na área, visando aprimorar as capacidades de detecção e responder de maneira eficaz às ameaças em constante evolução.

**Palavras-chave:** Machine Learning, Detecção De Fraudes Financeiras, Árvores De Decisão, Random Forest.

**Abstract**

This work explored the efficacy of advanced Machine Learning techniques, specifically decision trees and Random Forest, in detecting financial fraud within banking institutions. Through a systematic literature review, we identified the main approaches and challenges in implementing these technologies in the banking context, including variable selection, the impact of dataset size and balancing, and the importance of ensemble techniques for improving fraud detection. The study highlighted the need to balance accuracy and operational efficiency, considering the training and inference time of the models, and emphasized the relevance of scalable solutions that can adapt to transaction volume growth and the evolution of fraudulent techniques. The results suggest that, despite the challenges, the application of Machine Learning in combating financial fraud holds significant potential to enhance the security and efficiency of banking operations. The research points to the importance of continuous advancements in the field, aiming to improve detection capabilities and effectively respond to constantly evolving threats.

**Keywords:** Machine Learning, Financial Fraud Detection, Decision Trees, Random Forest.

**Introdução**

Com o advento do Big Data e suas tecnologias, houve um aumento significativo nas transações de comércio eletrônico e no uso do cartão de crédito, como observado em Boutaher, N. et al. (2020). Esse crescimento tem sido acompanhado por um aumento correspondente nas fraudes e problemas no setor bancário, afetando os sistemas de controle e a transparência das transações online. A detecção eficiente dessas fraudes em instituições bancárias é um desafio crítico, conforme destacado por Al-Hashedi e Magalingam (2021), tornando a implementação de métodos avançados de detecção de fraudes, como técnicas baseadas em inteligência artificial e mineração de dados, crucial para mitigar os riscos associados.

À medida que as técnicas fraudulentas se tornam mais complexas, é essencial desenvolver e implementar modelos avançados de detecção de fraudes para proteger os ativos dos clientes e preservar a integridade do sistema financeiro. Nesse cenário, a aplicação de técnicas de Data Analytics e Machine Learning, conforme demonstrado em um estudo recente sobre detecção de fraudes em transações com cartão de crédito Khatri et al. (2020), tem se destacado como uma abordagem altamente promissora para identificar padrões suspeitos e comportamentos fraudulentos.

Este estudo utiliza a base de dados "Credit Card Fraud Detection" disponível publicamente em Mishra5001. Essa base contém informações sobre transações financeiras de cartões de crédito, incluindo valor, tipo e hora da transação. Os dados serão coletados diretamente desse conjunto, passando por pré-processamento e análise para identificar padrões relevantes para a detecção de fraudes. Reina, D. et al. (2008) definem "Red flags" como indicadores de alerta e os destaca como um instrumento relevante na prevenção e detecção de fraudes. Medrado (2016) ressalta a importância das "red flags" como um meio eficaz de avaliar riscos e detectar possíveis distorções nas demonstrações financeiras das empresas.

Além disso, de acordo com Boutaher, N. et al. (2020), pesquisas recentes têm enfatizado o potencial das técnicas de Machine Learning na detecção de fraudes financeiras, reforçando a necessidade de abordagens avançadas na avaliação de riscos e na prevenção de perdas financeiras. A partir do estudo conduzido por Patil, S. et al. (2018), nota-se a importância da modelagem de Random Forest na detecção de fraudes em transações de cartão de crédito. Assim, diante da disponibilidade desses dados e da necessidade crescente de mitigar os riscos associados às atividades bancárias, este trabalho se propõe a realizar uma análise abrangente e detalhada da base de dados "Credit Card Fraud Detection", explorando técnicas avançadas de análise de dados e Machine Learning para identificar padrões suspeitos e aprimorar a detecção de fraudes financeiras em instituições bancárias.

A justificativa para a adoção de técnicas avançadas como árvores de decisão e Random Forest na detecção de fraudes financeiras em instituições bancárias reside na complexidade e na sofisticação crescente dos métodos empregados por fraudadores. Com a digitalização das operações financeiras, os bancos enfrentam desafios sem precedentes na proteção dos ativos e na manutenção da confiança de seus clientes. A capacidade dessas técnicas de Machine Learning de analisar grandes volumes de dados, aprender com eles e identificar padrões sutis e complexos que possam indicar comportamento fraudulento é crucial. A necessidade de abordagens mais eficazes é evidenciada pelo aumento constante das perdas financeiras devido a fraudes, o que impacta não apenas as instituições financeiras, mas também seus clientes. Portanto, explorar e implementar soluções inovadoras em detecção de fraudes não é apenas uma necessidade operacional, mas também uma responsabilidade social das instituições bancárias.

O principal objetivo deste trabalho é investigar a eficácia das árvores de decisão e do método Random Forest na detecção de fraudes financeiras em instituições bancárias, focando na aplicação dessas técnicas de Data Analytics e Machine Learning. Pretende-se analisar a capacidade desses modelos em processar e aprender com grandes conjuntos de dados financeiros, identificando com precisão transações fraudulentas, minimizando falsos positivos e melhorando a segurança das operações bancárias. Além disso, almeja-se contribuir para o corpo de conhecimento existente, fornecendo insights práticos e teóricos que possam orientar o desenvolvimento de estratégias mais robustas e eficientes de prevenção a fraudes, beneficiando assim o setor bancário e seus clientes por meio da redução de perdas financeiras e do fortalecimento da confiança nas transações eletrônicas.

**Material e Métodos**

A condução deste estudo baseia-se na utilização da base de dados "Credit Card Fraud Detection", amplamente reconhecida por profissionais e pesquisadores como uma fonte confiável para desenvolver e testar modelos de análise de dados. Fornecida por Mishra5001, ela contém uma variedade de atributos cruciais, como valor da transação, tipo de transação e hora da transação, bem como variáveis relacionadas ao comportamento do cliente. Este conjunto de dados representa fielmente o cenário real de negócios e é frequentemente empregado em estudos de análise de risco no setor bancário e financeiro. A análise detalhada desses dados visa identificar padrões e características relevantes para a detecção de fraudes, contribuindo assim para uma compreensão mais profunda dos processos envolvidos.

Para a análise dos dados e desenvolvimento dos modelos de detecção de fraudes, serão utilizadas técnicas de Machine Learning, com foco principal nas abordagens de Árvores de Decisão e Random Forest. Estas escolhas são fundamentadas em estudos prévios que destacam a eficácia dessas técnicas na detecção de fraudes financeiras. O estudo conduzido por Afriyie et al. (2023) ressalta a importância da modelagem de Random Forest na detecção de fraudes em transações de cartão de crédito, demonstrando a eficácia dessa técnica na identificação de padrões complexos e não lineares em grandes conjuntos de dados financeiros.

Adicionalmente, conforme citado em Husejinovic (2020), a árvore de decisão é uma técnica adequada para este tipo de análise. De maneira complementar, de acordo com as análises encontradas em Alarfaj, F et al. (2022), foi possível avaliar a eficácia dos modelos a partir do uso de métricas de desempenho e técnicas de análise exploratória de dados. Boutaher, N. et al. (2020) também destacam a importância da Análise Exploratória de Dados na identificação de padrões suspeitos e comportamentos fraudulentos em grandes conjuntos de dados. A metodologia de tratamento dos dados segue um passo a passo estruturado, conforme descrito a seguir:

**Visão Geral dos dados**

O dataset selecionado apresenta um total de 122 colunas que incluem uma variedade de atributos importantes para a análise de fraudes em transações de cartões de crédito. Inicialmente realizou-se a seleção dos atributos considerados relevantes para a detecção de fraudes, pelo fato de apresentarem dados que possibilitam a identificação de padrões comportamentais que indicam atividades fraudulentas, como renda, valores de crédito e histórico de aplicações. A seguir encontra-se a descrição dos atributos selecionados:

Tabela 1. Variáveis selecionadas

| Variável | Tipo | Descrição |
| --- | --- | --- |
| AMT\_CREDIT | float | Montante total do crédito concedido. |
| AMT\_INCOME\_TOTAL | float | Renda anual total do cliente. |
| CNT\_CHILDREN | int | Número de filhos do cliente. |
| CNT\_FAM\_MEMBERS | float | Número de membros da família do cliente. |
| DAYS\_BIRTH | int | Idade do cliente em dias. |
| DAYS\_EMPLOYED | int | Tempo de emprego em dias. |
| EXT\_SOURCE\_2 | float | Pontuações externas que representam a probabilidade de inadimplência do cliente. |
| AMT\_ANNUITY | float | Valor anual da anuidade do empréstimo. |
| AMT\_GOODS\_PRICE | float | Preço dos bens para os quais o crédito foi solicitado. |
| DAYS\_REGISTRATION | float | Quantidade de dias desde o registro. |
| FLAG\_PHONE | int | Indica se o cliente possui telefone. |
| REGION\_RATING\_CLIENT | int | Classificação da região de residência do cliente. |
| TARGET | int | Indicador binário do status do empréstimo (0 = empréstimo reembolsado, 1 = empréstimo com dificuldades) |
| FLAG\_OWN\_CAR | string | Indicador se o cliente possui carro (Y = Sim, N = Não). |
| FLAG\_OWN\_REALTY | string | Indicador se o cliente possui imóvel (Y = Sim, N = Não). |
| NAME\_CONTRACT\_TYPE | string | Tipo de contrato do empréstimo (ex: Cash loans, Revolving loans). |
| CODE\_GENDER | string | Descrição do sexo do cliente. |

Fonte: Dados originais da pesquisa.

**Limpeza e Pré-processamento dos Dados**

Na sequência, a partir dos atributos selecionados, realizou-se a análise detalhada dos atributos numéricos, incluindo estatísticas descritivas e a remoção dos valores nulos. Este passo garantiu a integridade e a consistência dos dados numéricos. De forma análoga, analisou-se os atributos categóricos, considerando a distribuição de valores e o manejo de dados ausentes. Foi realizada a codificação das variáveis categóricas através de técnicas como codificação de variáveis dummy, a fim de garantir que os dados estejam adequados para a modelagem estatística. Inicialmente, o dataset continha 307.511 linhas. Após a remoção das linhas nulas, o número de linhas foi reduzido para 306.562, resultando na exclusão de 949 linhas, o que representa apenas 0,31% do total de dados.

**Análise Estatística dos Valores Numéricos**

De forma complementar, realizou-se a avaliação detalhada das colunas numéricas após o processo de limpeza, incluindo identificação de dados discretos e contínuos relevantes para a detecção de padrões de fraude. Foram realizados testes estatísticos para entender a distribuição dos dados e suas características. A seguir encontra-se a relação dos resultados obtidos. Vale destacar que algumas colunas apresentaram comportamentos que não se enquadraram em nenhuma distribuição estatística específica, indicando a complexidade e a singularidade dos dados analisados.

Tabela 2. Relação das variáveis e suas distribuições estatísticas

| Variável | Distribuição estatística | p-valor |
| --- | --- | --- |
| amt\_credit | - | - |
| amt\_income\_total | - | - |
| cnt\_children | Binomial | 1.0 |
| cnt\_fam\_members | Binomial | 1.0 |
| days\_birth | Log-normal | 1.0 |
| days\_employed | Log-normal | 1.0 |
| ext\_source\_2 | - | - |
| amt\_annuity | Log-normal | 1.0 |
| amt\_goods\_price | - | - |
| days\_registration | - | - |
| flag\_phone | Binomial | 1.0 |
| region\_rating\_client | Binomial | 1.0 |
| target | Binomial | 1.0 |

Fonte: Dados originais da pesquisa.

**Árvores de Decisão e Random Forest**

Na sequência, realizou-se a avaliação do desempenho de modelos de Árvores de Decisão utilizando duas versões da base tratada: com e sem variáveis dummy. Além disso, foi realizada a seleção de colunas com base em análise estatística para otimizar o desempenho dos modelos.

As ferramentas de análise utilizadas incluirão bibliotecas de Machine Learning em Python, como Scikit-learn e Pandas, para o desenvolvimento e treinamento dos modelos. Técnicas de validação cruzada e métricas de avaliação de desempenho, como precisão, recall, F1-score e matriz de confusão, serão empregadas para avaliar a eficácia dos modelos desenvolvidos, seguindo as práticas recomendadas na literatura (referências).

**Resultados Preliminares**

O título da seção Resultados Preliminares deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos, seguindo a de acordo com a descrição feita no item 1.1 Formato e margens e apresentados na mesma ordem da seção Material e Métodos. Nesta seção devem ser apresentados os resultados parciais obtidos na pesquisa, ou seja, os resultados obtidos até o momento.

**Atenção:** antes de enviar o arquivo para o Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

Vamos analisar e comparar os resultados das duas árvores de decisão e das duas florestas aleatórias (random forests) com base na acurácia, área sob a curva ROC (ROC AUC) e nas matrizes de confusão.

### Comparação das Árvores de Decisão

#### Primeira Árvore de Decisão

- \*\*Accuracy\*\*: 0.8492

- \*\*ROC AUC\*\*: 0.5279

- \*\*Matriz de Confusão\*\*:

```

[[51349 5045]

[ 4204 715]]

```

- Verdadeiros Negativos (TN): 51349

- Falsos Positivos (FP): 5045

- Falsos Negativos (FN): 4204

- Verdadeiros Positivos (TP): 715

#### Segunda Árvore de Decisão

- \*\*Accuracy\*\*: 0.8480

- \*\*ROC AUC\*\*: 0.5299

- \*\*Matriz de Confusão\*\*:

```

[[51248 5146]

[ 4176 743]]

```

- Verdadeiros Negativos (TN): 51248

- Falsos Positivos (FP): 5146

- Falsos Negativos (FN): 4176

- Verdadeiros Positivos (TP): 743

\*\*Análise Comparativa\*\*:

- Ambas as árvores têm acurácias e ROC AUC similares.

- A segunda árvore de decisão tem uma leve vantagem em termos de ROC AUC (0.5299 vs 0.5279) e possui menos falsos negativos (4176 vs 4204), sugerindo que detecta ligeiramente mais fraudes.

- No entanto, a primeira árvore de decisão possui um pouco menos de falsos positivos (5045 vs 5146).

### Comparação das Florestas Aleatórias (Random Forests)

#### Primeira Random Forest

- \*\*Accuracy\*\*: 0.9196

- \*\*ROC AUC\*\*: 0.6780

- \*\*Matriz de Confusão\*\*:

```

[[56367 27]

[ 4903 16]]

```

- Verdadeiros Negativos (TN): 56367

- Falsos Positivos (FP): 27

- Falsos Negativos (FN): 4903

- Verdadeiros Positivos (TP): 16

#### Segunda Random Forest

- \*\*Accuracy\*\*: 0.9197

- \*\*ROC AUC\*\*: 0.6712

- \*\*Matriz de Confusão\*\*:

```

[[56362 32]

[ 4892 27]]

```

- Verdadeiros Negativos (TN): 56362

- Falsos Positivos (FP): 32

- Falsos Negativos (FN): 4892

- Verdadeiros Positivos (TP): 27

\*\*Análise Comparativa\*\*:

- Ambas as random forests têm acurácias muito próximas (0.9196 vs 0.9197).

- A primeira random forest tem um ROC AUC ligeiramente melhor (0.6780 vs 0.6712).

- A segunda random forest tem menos falsos negativos (4892 vs 4903), indicando que detecta mais fraudes.

- A primeira random forest tem menos falsos positivos (27 vs 32).

### Conclusão

- \*\*Árvores de Decisão\*\*: A segunda árvore de decisão é ligeiramente melhor na detecção de fraudes, com uma ROC AUC um pouco maior e menos falsos negativos. Contudo, a diferença é pequena.

- \*\*Florestas Aleatórias (Random Forests)\*\*: A segunda random forest detecta mais fraudes (mais verdadeiros positivos e menos falsos negativos), embora tenha uma ligeira desvantagem no ROC AUC comparado à primeira. Considerando que a prioridade é detectar fraudes, a segunda random forest pode ser considerada melhor.

### Recomendações

- \*\*Para Detecção de Fraudes\*\*: Considere a segunda árvore de decisão e a segunda random forest, devido ao melhor desempenho na detecção de casos positivos (fraudes).

- \*\*Para Menor Taxa de Falsos Positivos\*\*: A primeira random forest é ligeiramente melhor, mas a diferença é mínima.

Essas análises podem ser usadas para justificar a escolha do modelo no seu TCC, destacando a importância do balanceamento entre a detecção de fraudes e a manutenção de baixas taxas de falsos positivos.

**Conclusão(ões) ou Considerações Finais**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares. A seção deve conter frases curtas, apresentando as conclusões e inferências elaboradas a partir da discussão dos resultados obtidos até o momento. Os autores não devem, em hipótese alguma, mencionar, citar ou reproduzir resultados de outros estudos na(s) Conclusão(ões) ou Considerações Finais do TCC. Além disso, esta seção não deve conter tabelas ou figuras, sendo redigida de forma sucinta.

**Atenção:** antes de enviar o arquivo para o Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

Conclusoes:

As probabilidades calculadas confirmam a distribuição do número de filhos: A maior parte dos dados está concentrada em 0 a 2 filhos. Conforme o número de filhos aumenta, a probabilidade diminui drasticamente, o que é consistente com a distribuição esperada em uma população.

As probabilidades calculadas confirmam a distribuição dos membros da família: A maior parte dos dados está concentrada entre 1 a 4 membros da família.Conforme o número de membros da família aumenta, a probabilidade diminui drasticamente, o que é consistente com a distribuição esperada em uma população.

### Conclusão

O estudo demonstrou que a aplicação de técnicas avançadas de Machine Learning, como Árvores de Decisão e Random Forest, na detecção de fraudes financeiras, é promissora. A abordagem sistemática para a seleção e pré-processamento das variáveis, aliada à utilização de técnicas ensemble, mostrou-se eficaz na melhoria da precisão e eficiência dos modelos de detecção de fraudes. A pesquisa destaca a importância da contínua evolução das técnicas de análise e modelagem para enfrentar as ameaças emergentes no setor financeiro, garantindo assim a segurança e a eficiência das operações bancárias.

**Referências**

Boutaher, N., Elomri, A., Abghour, N., Moussaid, K., & Rida, M., 2020. A review of credit card fraud detection using machine learning techniques. In 2020 5th International Conference on cloud computing and artificial intelligence: technologies and applications (CloudTech) (pp. 1-5). IEEE.

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402.

Khatri, S., Arora, A., & Agrawal, A. P. (2020, January). Supervised machine learning algorithms for credit card fraud detection: a comparison. In *2020 10th international conference on cloud computing, data science & engineering (confluence)* (pp. 680-683). IEEE.

Mishra5001. (s/d). Credit Card Fraud Detection. Kaggle. [Disponível em: <](https://www.kaggle.com/datasets/mishra5001/credit-card)https://www.kaggle.com/datasets/mishra5001/credit-card?select=application\_data.csv>. Acesso em: 20 jun. 2024.

Reina, D., do Nascimento, S., & Reina, D. R. M. (2008). The perception of the auditors regarding the use of red flags in the major audit firms in Brazildoi: 10.4025/enfoque. v27i2. 6172. *Enfoque*, *27*(2), 71.

Medrado, F. C. (2017). Indicadores econômico-financeiros como red flags de riscos de fraudes ou manipulação contábil.

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, *132*, 385-395.

Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access*, *10*, 39700-39715.

**Apêndice ou Anexo** (opcional)

Apêndices são textos e/ou documentos que foram elaborados pelo autor e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para figuras e tabelas.

O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).

**Atenção:** antes de enviar o arquivo para o Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.