Predição de Fraudes em Operações Financeiras

Aluno: Gustavo Rodrigues Carvalho
Email: 10175838@mackenzista.com.br
RA: 10175838

3. RESUMO

Com o avanço da tecnologia em diferentes áreas, uma delas sendo a financeira, gerando assim um maior número de usuários que fazem uso dos meios digitais para realizar suas transações, junto com o aumento de clientes legítimos que usam as ferramentas de forma legal e ordeira, há também aqueles que se utilizam da facilidade e conveniência da facilitação que o meio digital trouxe para as operações financeiras e buscam cometer fraudes, há também um aumento no número de fraudes em operações financeiras. Tendo isso em mente, este trabalho tem como objetivo identificar as características de uma operação fraudulenta, desenvolver modelos de aprendizado de máquina e descobrir qual modelo tem maior precisão ao prever uma operação financeira fraudulenta.

4. INTRODUÇÃO

Contextualização

Segundo o jornal E-Investidor (2023), houve mais de 2,8 mil tentativas de fraudes financeiras por minuto em canais eletrônicos financeiros. No mesmo período, aproximadamente 365 milhões de tentativas de golpes foram registradas, indicando que cerca de 1,73% das transações digitais no sistema financeiro do país tiveram intenções criminosas.

Essa crescente incidência de fraudes pode ser explicada pela maior vulnerabilidade do consumidor no ambiente online. Conforme apontado por Sampaio (2023), *"A facilitação do acesso à internet evidenciou a situação da vulnerabilidade agravada do consumidor dentro do ambiente online. Nessa conjuntura, criminosos deram nova roupagem a crimes já conhecidos e transferiram a sua prática também para o ambiente digital"*.

Justificativa

A mudança do ambiente físico para o digital fez com que a forma como lidamos com fraudes financeiras evoluísse. Antes, a detecção de fraudes dependia predominantemente da capacidade humana. Hoje, devido ao volume massivo de dados e à velocidade com que são gerados, há a necessidade do envolvimento de tecnologia avançada. Lokanan (2024) destaca que essa transição tornou a detecção de fraudes um desafio que exige o uso de inteligência artificial e aprendizado de máquina.

Dado o crescimento contínuo dessas atividades criminosas, torna-se essencial aprimorar métodos automatizados para detectar operações fraudulentas com maior precisão, reduzindo prejuízos financeiros e impactos negativos sobre consumidores e instituições.

Objetivo

O presente estudo tem como objetivo utilizar algoritmos de aprendizado de máquina para classificar operações financeiras fraudulentas. Os resultados obtidos serão comparados com aqueles

de estudos anteriores sobre detecção de fraudes financeiras, nos quais foram disponibilizados tanto os modelos aplicados quanto as bases de dados analisadas. Além disso, será realizada uma análise comparativa com indicadores financeiros.

Opção do Projeto

Este projeto se enquadra na **Opção Framework**, empregando técnicas de **Machine Learning** para resolver um problema de **classificação**. O estudo se baseará na aplicação de algoritmos de aprendizado de máquina para prever fraudes em transações financeiras, utilizando um conjunto de dados apropriado para o treinamento e validação dos modelos.

5. DESCRIÇÃO DO PROBLEMA

Antes de prosseguirmos é necessário introduzir alguns conceitos, dentre eles um interessante de se tomar conhecimento é o de fraude financeira. Segundo Wells(2014) "No sentido mais lato, a fraude pode englobar qualquer crime com fins lucrativos que utilize o engano como principal modus operandi. Das três formas de retirar ilegalmente dinheiro a uma vítima - força, artifício ou furto - todos os crimes que recorrem a artifícios são fraudes" (p. 8).

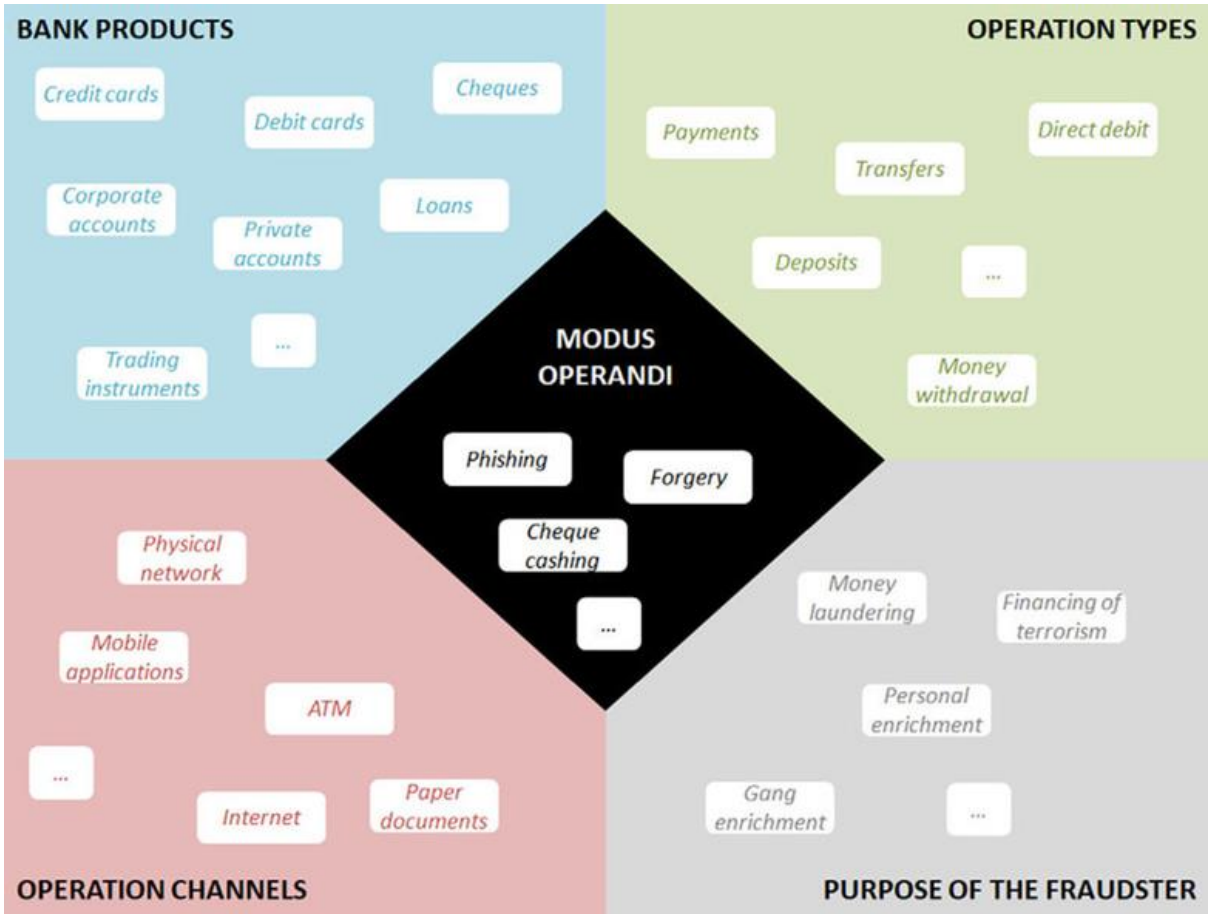


Figura 1. Illustration of fraud dimensions.

Somado ao conceito de fraudes financeiras temos o conceito de operações financeiras segundo (Praxis 2024) as operações financeiras são transações do mercado financeiro, que envolvem a compra

e venda de ações e títulos em moedas estrangeiras, transações essas que ocorrem entre investidores, empresas e instituições financeiras (ver Figura 1).

6. DISCUTIR A RESPEITO DOS ASPECTOS ÉTICOS DO USO DA IA E SUA RESPONSABILIDADE NO DESENVOLVIMENTO DA SOLUÇÃO.

Um dos dilemas éticos, que podem ser enfrentados é com relação ao viés da análise, por conta do pequeno número, de casos de fraudes financeiras em relação ao grande número de transações não fraudulentas, o que pode interferir nos resultado das predições, gerando falsos positivos e falsos negativos.

Outro ponto que deve ser levado em consideração, é a sensibilidade dos dados financeiros de pessoas, para isso o dataset selecionado

7. DATASET

O conjunto de dados utilizado neste projeto é uma representação sintética de transações financeiras móveis, gerada pelo **PaySim**, um simulador que utiliza dados agregados de registros financeiros reais de um serviço de dinheiro móvel em um país africano. Esse dataset foi desenvolvido para fins de pesquisa e tem como objetivo permitir estudos de detecção de fraudes em transações eletrônicas.

Origem e Estrutura do Dataset
O PaySim simula transações financeiras móveis com base em um mês de registros financeiros reais, disponibilizados por uma empresa multinacional que opera esse serviço em mais de 14 países. Esse dataset foi reduzido para **1/4 do tamanho original** para facilitar seu uso e análise na plataforma Kaggle.

8. METODOLOGIA

8.1. Coleta e Preparação dos Dados

O dataset utilizado é o **PaySim**, que simula transações financeiras móveis e inclui atividades fraudulentas. Antes de aplicar algoritmos de aprendizado de máquina, realizaremos:

- Análise Exploratória de Dados (EDA)
  - Tratamento de dados desbalanceados,
  - Remoção de colunas inviáveis
  - Transformação e normalização de dados
  - Pré-processamento**.**

8.2. Seleção e Treinamento do Modelo

Serão testados diferentes algoritmos de *Machine Learning* para verificar qual melhor se adapta ao problema de detecção de fraudes:

- **Modelos de aprendizado supervisionado:**

- Regressão Logística.
- KNN.
- Decision Tree.
- Random Forest.
- Gradient Boosting.

Os modelos serão avaliados com **validação cruzada** e métricas adequadas ao problema de classificação desbalanceada, como **AUC-ROC, F1-score, precisão e recall**.

8.3. Ferramentas/frameworks utilizados.

Para Bibliotecas e Frameworks python

- **NumPy e Pandas**: Para manipulação de dados, será utilizado, numpy e pandas, pois permite manipular e realizar operações matemáticas assim como manipular e analisar os conjuntos de dados.
- **Seaborn e Matplotlib**: Para visualização de dados, será utilizado seaborn e matiplotlib, para visualização de dados estatísticos, e gráficos.
- **Scikit-learn (sklearn)**: Para pré-processamento de dados, normalização e segregação do conjunto sklearn.
- **sklearn.metrics**: E sklearn metrics, para avaliação do desempenho do modelo.

8.4. Avaliação de Performance e Ajustes

Com base nos resultados iniciais, ajustaremos os hiperparâmetros dos modelos para otimizar seu desempenho, e utilizaremos métricas financeiras para detectar fraudes como f1-score.

8.5. Implementação e Resultados Esperados

O modelo final será implementado e avaliado em um ambiente de teste para verificar sua capacidade de identificar fraudes em novas transações. Os resultados esperados incluem, uma baixa taca de detecção de fraudes, com uma baixa taxa de falsos negativos, e positivos e reduzindo o tempo de resposta para possivelmente ser usado em tempo real.

8.6. Pipeline adotada

1. Pré-processamento dos dados:
   a. Codificação de variáveis categóricas.
   b. Normalização dos dados.
   c. Remoção de duplicatas e tratamento de valores ausentes.
2. Divisão dos dados:
3. Treinamento:
4. Validação e avaliação
5. Comparação dos resultados:

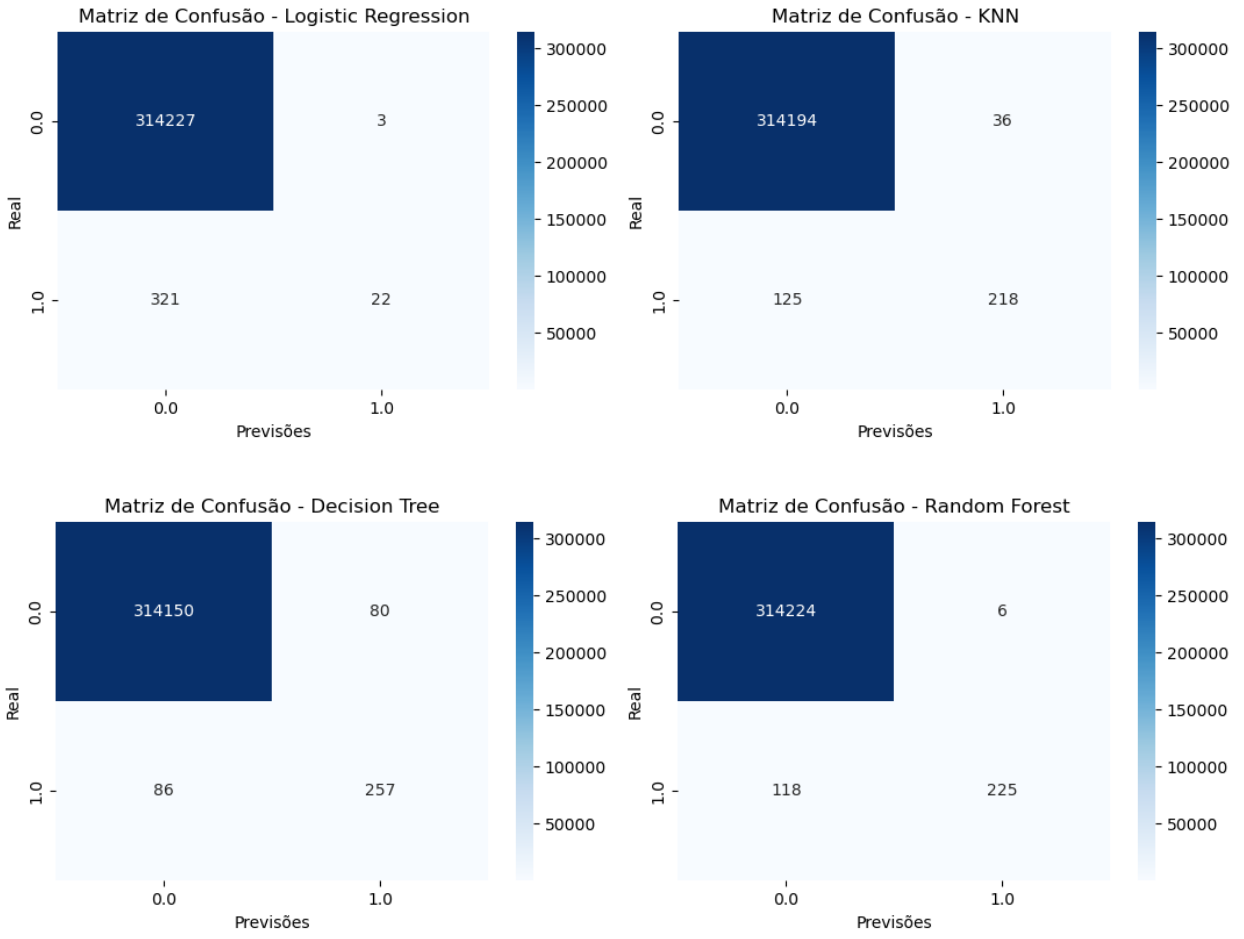
9. RESULTADOS

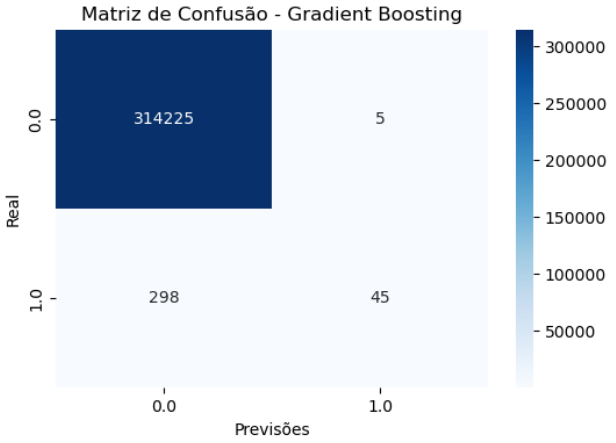
9.1. Métricas de desempenho

Model	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Logistic Regression	99,897%	99,898%	99,999%	99,948%	88,000%	6,414%	11,957%
KNN	99,949%	99,960%	99,989%	99,974%	85,827%	63,557%	73,032%
Decision Tree	99,946%	99,973%	99,973%	99,973%	75,291%	75,510%	75,400%
Random Forest	99,960%	99,962%	99,997%	99,980%	96,567%	65,598%	78,125%
Gradient Boosting	99,904%	99,905%	99,998%	99,952%	90,000%	13,120%	22,901%

Precision (Macro)	Recall (Macro)	F1-Score (Macro)	Precision (Weighted)	Recall (Weighted)	F1-Score (Weighted)
93,949%	53,207%	55,952%	99,885%	99,897%	99,853%
92,894%	81,773%	86,503%	99,945%	99,949%	99,945%
87,632%	87,742%	87,687%	99,946%	99,946%	99,946%
98,264%	82,798%	89,052%	99,959%	99,960%	99,956%
94,953%	56,559%	61,426%	99,894%	99,904%	99,868%

9.2. Matrizes de confusão





9.3. Interpretações

Métricas para 1 e 0

Accuracy (Acurácia) : Com base nos resultados, conseguimos identificar que todos os modelos possuem uma alta acurácia (acima dos 99%), sendo assim entendemos que ele acerta a maioria das previsões, mas vindo de um conjunto de dados desbalanceado, essa métrica não é muito confiável, pois a classe 0 é substancialmente maior que a classe 1, portanto fica muito mais fácil identificar a classe 0 como sendo não fraudulenta.

Precisão (Precision): Com base nos resultados de precisão, vemos todas as que as classificações dividindo todas as classificações corretas, e vemos que dois modelos se destacam com maior precisão, elas sendo (Random Forest e Gradient Boosting com resultados acima de 90%)

Recall (Revocação): Para a classe 0, as taxas são altas indicando que os modelos tem identificado essas classes corretamente, já para a classe 1 alguns modelos tem falhado em identificar casos reais de fraude, dentre eles a regressão logística e o gradiente boosting que por sua vez tiveram taxas menores que 20%.

F1-Score: É a média harmônica entre precision e recall, novamente para as classes 0 a taxa está alta, porém para os casos da classe 1 existe uma grande variação Random Forest com a maior taxa 78,125% e a Logistic Regression com 11,957%, devido ao baixo recall e precisão.

Métricas para macro

Não leva em consideração, o peso das classes apenas e trata ambas como iguais, tirando a média simples das classes para cada métrica Precision, Recall e F1-Score.

Métricas Weighted (Ponderada):

É a média ponderada das métricas analisadas (Precision, Recall, F1-Score) para as classes Precision, Recall, F1-Score.

10. Conclusão

Com base no resultado apresentado pela métricas analisadas chegamos a conclusão de que a que possui melhor desempenho é Random Forest, possui um maior recall, o que significa que dos modelos é o que melhor consegue identificar fraudes. Além disso possui um maior equilíbrio entre recall e precision, o que gera menor ocorrência de falsos negativos e de falsos positivos.

Em adição acredito que seja válido uma menção honrosa ao KNN, que ficou para trás, porém por pouco e assim como o Random Forest, possui bons resultas, mas o Random Forest nos fornece uma superioridade na precision e no recall.

Os resultados atenderam ás expectativas?

Sim, os resultados obtidos forma satisfatórios, ao término do trabalho saio com um algoritmo que realiza a análise exploratória de dados em um outro arquivo, e um arquivo onde é realizado o pré-processamento e preparação dos dados para o treinamento e execução dos modelos.

O problema foi resolvido?

Sim, foi possível testar e comparar os resultados de diferentes modelos machine learning para classificar as fraudes financeiras e encontrar dentre eles um modelo com bons resultados e com janela de melhorias no desempenho dele.

O que é possível de melhorar para uma versão 2.0

Acredito que para uma rodada 2.0 seria interessante realizar um técnicas de balanceamento de classes como undersampling ou oversampling, assim como uma atribuição de pesos às classes para verificar a possibilidade de otimização dos resultados.

11. Links

- **Youtube:** https://youtu.be/TADKhwj2Q7U
- **GitHub:** https://github.com/grcarvalho1032/projeto-fraud-detection-IA-2025/tree/main

12. Referências

SAMPAIO, Marília de Ávila e Silva. Responsabilidade civil das instituições financeiras nas fraudes eletrônicas. *Tribunal de Justiça do Distrito Federal e dos Territórios*. Disponível em: https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/artigos-discursos-e-entrevistas/artigos/2023/responsabilidade-civil-das-instituicoes-financeiras-nas-fraudes-eletronicas

LOKANAN, M. E. Predicting Money Laundering Using Machine Learning and Artificial Neural Networks Algorithms in Banks. ***Journal of Applied Security Research***, v. 19, n. 1, p. 20-44, 2024. DOI: 10.1080/19361610.2022.2114744. Disponível em: https://doi.org/10.1080/19361610.2022.2114744. Acesso em: 12 Set 2024.

GARETH, J.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. ***An Introduction to Statistical Learning with Applications in R***. 2. ed. Nova York: Springer, 2013. Capítulos 4 e 5.
SILVA, A. S.; PERES, S. M.; BOSCARIOLI, C. ***Introdução a Mineração de Dados – Com Aplicações em R.*** 1. ed. São Paulo: Elsevier, 2016. Capítulos 1, 2 e 3.

HAN, J.; KAMBER, M.; PEI, J. ***Data Mining: Concepts and Techniques***. 3. ed. San Francisco: Morgan Kaufmann, 2012. Capítulos 6, 8, 9, 12 e 13.

WITTEN, I. H.; FRANK, E. ***Data Mining: Practical Machine Learning Tools and Techniques*.** 2. ed. San Francisco: Morgan Kaufmann, 2005. Capítulos 5, 6 e 8.

WELLS, Joseph T. ***Principles of Fraud Examination***. 4. ed. Hoboken: John Wiley & Sons, 2014. Capítulo 1.

HEWAPATHIRANA, I. Utilizing Prediction Intervals for Unsupervised Detection of Fraudulent Transactions: A Case Study. ***Asian Journal of Engineering and Applied Technology***, v. 11, n. 2, p. 1-10, 2022. DOI: https://doi.org/10.51983/ajeat-2022.11.2.3348. Disponível em: www.trp.org.in. Acesso em: 11 Set 2024.

XU, T.; LIU, J.; CHENG, H. Predicting Fraud in U.S. - Listed Chinese Companies: An Empirical Analysis Based on M-Score and F-Score Models. In: ***Proceedings of the 2023 International Conference on Management Research and Economic Development***. DOI: 10.54254/2754-1169/20/20230177. Acesso em: 12 Set. 2024.

DALAL, S.; SETH, B.; RADULESCU, M.; SECARA, C.; TOLEA, C. Predicting Fraud in Financial Payment Services through Optimized Hyper-Parameter-Tuned XGBoost Model. ***Mathematics***, v. 10, n. 24, p. 4679, 2022. DOI: https://doi.org/10.3390/math10244679. Acesso em: 12 Set. 2024.

GAO, J. X.; ZHOU, Z. R.; AI, J. S.; XIA, B. X.; COGGESHALL, S. Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms**. *Journal of Intelligent Learning Systems and Applications*,** v. 11, p. 33-63, 2019. DOI: https://doi.org/10.4236/jilsa.2019.113003. Recebido em: 6 abr. 2019. Aceito em: 11 ago. 2019. Publicado em: 14 ago. 2019. Disponível em: http://www.scirp.org/journal/jilsa. Acesso em: 16 Set 2024.

NESVIJEVSKAIA, A.; OUILLADE, S.; GUILMIN, P.; ZUCKER, J.-D. The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy*, v. 3, e12, 2021. DOI: 10.1017/dap.2021.3. Disponível em: https://doi.org/10.1017/dap.2021.3. Acesso em: 16 Set. 2024.

SAMPAIO, Marília de Ávila e Silva. Responsabilidade civil das instituições financeiras nas fraudes eletrônicas. *Tribunal de Justiça do Distrito Federal e dos Territórios*. Disponível em: https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/artigos-discursos-e-entrevistas/artigos/2023/responsabilidade-civil-das-instituicoes-financeiras-nas-fraudes-eletronicas. Acesso em: 23 set. 2024.

E-INvestidor. Brasil sofre 2,8 mil tentativas de fraudes financeiras por minuto; saiba como se proteger. *Estadão*, 27 jun. 2023. Disponível em: https://einvestidor.estadao.com.br/ultimas/brasil-dados-tentativas-fraude-dicas-se-proteger/. Acesso em: 23 set. 2024.

SEJA PRAXIS. O que é: Operações Financeiras. Disponível em: https://www.sejapraxis.com.br/glossario/o-que-e-operacoes-financeiras/. Acesso em: 23 set. 2024.

NEHA, S. V. S. T.; YADAV, Yogesh; GOYAL, Yashika. *Introduction to Machine Learning*. **International Journal of Advanced Research in Science, Communication and Technology** (IJARSCT), v. 4, n. 2, p. 100-108, mar. 2024. DOI: 10.48175/IJARSCT-15723. Disponível em: www.ijarsct.co.in. Acesso em: 28 set. 2024.

WANG, Le; HAN, Meng; LI, Xiaojuan; ZHANG, Ni; CHENG, Haodong. *Review of classification methods on unbalanced data sets*. IEEE Access, v. 9, p. 48152-48171, 2021. DOI: 10.1109/ACCESS.2021.3074243.