UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELECOMUNICAÇÃO

CURSO DE GRADUAÇÃO EM ENGENHARIA COMPUTAÇÃO

ANNA VITHORIA MENDES DE SOUSA

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA FRAUDES DE CARTÃO DE CRÉDITO

FORTALEZA

2023

ANNA VITHORIA MENDES DE SOUSA

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA FRAUDES DE CARTÃO DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Computação.

Orientador: Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx

FORTALEZA

2023

ANNA VITHORIA MENDES DE SOUSA

ANÁLISE DE DESEMPENHO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA FRAUDES DE CARTÃO DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Dois (SIGLA)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Três (SIGLA)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Quatro (SIGLA)

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

O uso de cartões de crédito e transações financeiras eletrônicas é algo que esta cada vez mais popularizado, essa nova relação com o dinheiro também nos deixa mais suscetíveis em alguns pontos, tendo em vista que diariamente uma forma nova de golpe é pensada ou colocada em prática, oferecer uma proteção a esses novos perigos é torna algo muito necessário.

Uma possibilidade de identificar essa possível fraude é por meio de algoritmos de Aprendizagem de Máquina, ter algoritmos treinados que sejam capazes de classificar uma transação como inadequado ou não e, com grandes capacidade assertiva com certeza, é um grande reforço. É esse o objetivo deste trabalho analisar como é o desempenho de algoritmos como Logistic Regression, Decision Tree e Random Forest não resolução desse desafio.

Os algoritmos foram treinados usando um dataset de fraudes de cartão de crédito, passando por um processo de validação cruzada, buscando os parâmetros que oferecem os melhores resultados, para então treinar os modelos novamente, mas dessa vez com os hiperparâmetros escolhidos. Analisando métricas como acurácia, precisão, recall e f1-score, além do tempo de desempenho, para medir como cada um dos modelos desempenha.

Palavras-chave: Fraude. Cartões de Crédito. Aprendizagem de Máquina

LISTA DE FIGURAS

Figura 1 – Distribuição dos Dados 14
Figura 2 – Correlações de Pearson e Spearman. 15
Figura 3 – Acurácia dos Modelos 20
Figura 4 – Acurácia Média dos Modelos 21
Figura 5 – Prcisão dos Modelos 21
Figura 6 – Precisão Média dos Modelos 22
Figura 7 – Recall dos Modelos 22
Figura 8 – Precisão Média dos Modelos 23
Figura 9 – Fscore dos Modelos 24
Figura 10 – Fscore Média dos Modelos 24
Figura 11 – Tempo de Execução dos Modelos 25
Figura 12 – Tempo Médio de Execução dos Modelos 26

LISTA DE TABELAS

Tabela 1 – Médias Resultados da Validação Cruzada para RL 18
Tabela 2 – Desvio Padrão Resultados da Validação Cruzada para RL 18
Tabela 3 – Médias Resultados da Validação Cruzada para DT 18
Tabela 4 – Desvio Padrão Resultados da Validação Cruzada para DT 19
Tabela 5 – Médias Resultados da Validação Cruzada para RF 19
Tabela 6 – Desvio Padrão Resultados da Validação Cruzada para RF 19
Tabela 7 – Acurácia dos testes 20
Tabela 8 – Estatísticas das Acurácias 20
Tabela 9 – Precisão dos testes 21
Tabela 10 – Estatísticas das Precisões 22
Tabela 11 – Recall dos testes 22
Tabela 12 – Estatísticas de Recall 23
Tabela 13 – F1-score dos testes 23
Tabela 14 – Estatísticas de Fscore 24
Tabela 15 – Tempo de Execução dos testes 25
Tabela 16 – Estatísticas de Tempo de Execução 25

SUMÁRIO

1 INTRODUÇÃO 9
2 FUNDAMENTOS TEÓRICOS 10
2.1 Algoritmos de Machine Learning 10
2.1.1 Regressão Logística 10
2.1.2 Árvore de Decisão 11
2.1.3 Floresta Aleatória 12
2.2 Métricas 12
2.2.1 Acurácia 12
2.2.2 Precisão 13
2.2.3 Recall 13
2.2.4 F1-score 13
2.2.5 Tempo de Execução 13
3 METODOLOGIA 14
3.1 Base de Dados 14
3.2 Pré-processamento 15
3.3 Experimento 16
3.4 Testes Estatísticos 16
4 RESULTADOS 18
4.1 Validação Cruzada da Regressão Logística 18
4.2 Validação Cruzada da Árvore de Decisão 18
4.3 Validação Cruzada da Floresta Aleatória 19
4.4 Resultado dos Testes 19
4.4.1 Acurácias 19
4.4.2 Precisão 20
4.4.3 Recall 21
4.4.4 F1-score 23
4.4.5 Tempo de Execução 24
4.5 Testes Estatísticos 25
5 CONCLUSÕES E TRABALHOS FUTUROS 28
REFERÊNCIAS 29

ANEXOS 29

1 INTRODUÇÃO

Os constantes avanços tecnológicos têm provocado diversas mudanças nos hábitos das pessoas, não seria diferente quando se trata da forma de lidar com o consumo e finanças. Assim é cada vez mais comum o uso de transações financeiras eletrônicas, logo fraudes em cartões de crédito surgem como uma enorme preocupação para consumidores e instituições financeiras.

Segundo o Banco Central (BC, 2023), somente no Brasil o número de cartões de credito no pais até junho de 2022 era de 190,8 milhões, esse número representa quase o dobro da população economicamente ativa no Brasil. Além disso, segundo um levantamento feito pela Visa,(VALOR. . . , 2020), o Brasil está em segundo lugar no ranking de países da América Latina com maior numero de fraude no cartão em compras no comercio online

Diante deste desafio, a implementação de estratégias eficazes para detecção e prevenção de fraudes torna-se imperativa. Nesse contexto, a utilização de Algoritmos de Aprendizagem de Máquina surge como uma forma proteção, explorando a capacidade desses sistemas em analisar vastos conjuntos de dados para identificar possíveis fraudes.

Sendo assim, esse trabalho propõe analisar como os Algoritmos de Regressão Logística, Decision Tree e Random Forest desempenham a detecção dessas transações perigosas.

2 FUNDAMENTOS TEÓRICOS

2.1 Algoritmos de Machine Learning

Machine Learning é uma subárea da Inteligência Artificial, que busca fazer o computador aprender por meio de dados. Reconhecendo padrões e aprendendo com eles, além de realizar tarefas específicas sem uma programação explícita. Objetivo fundamental do aprendizado de máquina é capacitar sistemas a adquirirem conhecimento e aprimorar seu desempenho por meio da experiência, oferecendo soluções eficientes para uma ampla gama de problemas complexos em diversas áreas, como reconhecimento de padrões, processamento de linguagem natural, e, como abordado aqui, detecção de fraudes em transações financeiras.

Esses algoritmos de Aprendizagem de Máquina podem resolver algoritmos de Regressão ou de Classificação. Os problemas de Regressão requer como saída um número real. Pode ser qualquer valor dentro de um intervalo contínuo, casos como o de previsão de vendas e disponibilidades de crédito são aplicados esses algoritmos. Já os de Classificação são usados com o objetivo de atribuir uma instância a uma categoria ou classe específica. Tendo como saída uma classe ou rótulo discreto, representando a categoria à qual a instância pertence. Esses algoritmos são aplicados em problemas médicos, se o pacientes está saudável ou não, ou em casos de fraude como é o caso deste trabalho.

Entre esses modelos há os supervisionados e não supervisionados. No caso dos modelos não supervisionados são usados com dados não rotulados, enquanto que os supervisionados são usados para dados rotulados, em que se entende como os dados vão ser classificados, ou seja, a variável resposta ou dependente é conhecida. Como este é o caso dos dados analisados no trabalho, isto é, o dataset é rotulado, os modelos aplicados serão supervisionados, Regressão Logística (Logistic Regression), Árvore de Decisão (Decision Tree) e Floresta Aleatória (Random Forest).

2.1.1 Regressão Logística

É um algoritmo usado para classificação, em que é estimada a probabilidade de uma instância pertencer a uma determinada classe. Em que a probabilidade maior que 50% a amostras pertence uma classe, no caso deste trabalho, a amostra é considerada fraudulenta, já com uma porcentagem menor de 50% a instância pertence a classe 0, isto é, uma não fraude.

A Regressão Logística funciona como uma soma ponderada das características de

entrada, gerando uma logística dos resultados. A logística é uma função sigmóide que mostra um número entre 0 e 1.

Um dos principais hiperparâmetros de RL é o parâmetro Regularização, ou "penalty"(penalidade), e a Força de Regularização, o parâmetro C, o qual controla a força da penalidade. Essa penalidade ou regularização é útil para evitar que o modelo se ajuste muito aos dados de treinamento. Por sua vez, o C é o inverso da força de regularização, assim valores menores significam uma regularização mais forte.

2.1.2 Árvore de Decisão

São algoritmos que podem ser usados para Regressão ou Classificação. Uma árvore de decisão é construída dividindo recursivamente o conjunto de dados em uma sequência de subconjuntos com base em perguntas do tipo "se-então". O algoritmo baseia-se na ideia de divisão dos dados em grupos homogêneos. Cada nó interno da árvore representa uma pergunta sobre uma característica, e cada ramo representa uma possível resposta para essa pergunta. Os nós folha contém os rótulos ou valores previstos.

O processo de construção da árvore envolve selecionar a melhor característica para dividir os dados em cada etapa. A característica considerada como melhor é aquela que maximiza a pureza dos subconjuntos resultantes, tornando os grupos mais homogêneos em relação às suas classes. Este processo é repetido de forma recursiva para cada subconjunto até que determinados critérios de parada sejam atendidos.

Entre os hiperparâmetros de uma DT temos, o “criterion” ou Divisão de Critério, o qual, é um parâmetro que determina a função usada para medir a qualidade de uma divisão, podendo ser “gini”, Índice de Gini ou Impureza, e a “entropy”, Entropia. O Índice de Gini diz respeito a impureza de um nó, medindo a quantidade de vezes que um elemento escolhido aleatoriamente do conjunto de dados seria rotulado de maneira incorreta se fosse rotulado aleatoriamente de acordo com a distribuição de rótulos do subconjunto. Por sua vez a Entropia representa a falta de uniformidade ou uma medida de aleatoriedade nos dados. Quanto mais alta a entropia, mais caótico e misturados estão os dados e quanto menor a entropia, mais uniforme e homogênea está o conjunto de dados.

Outro parâmetro é o “max-depth”, Profundidade Máxima, a profundidade de um árvore se refere ao caminho mais longo, indo da raiz até uma folha. O hiperparâmetro limita a profundidade máxima da árvore, controlando assim a complexidade do modelo, ajudando a

evitar o overfitting.

2.1.3 Floresta Aleatória

O algoritmo de Floresta Aleatória combina múltiplas Árvores de Decisão, permitindo um modelo mais robusto, isso porque um problema das árvores de Decisão é a alta variância,isto é, elas são muito sensíveis às variações nos dados de treinamento.

Uma Floresta Aleatória constrói várias árvores de decisão independentes durante o treinamento. Cada árvore é treinada em uma amostra aleatória do conjunto de treinamento, e as amostras são geralmente selecionadas com reposição (chamado de amostragem com substituição). Assim, ao treinar cada árvore em uma amostra aleatória, a Floresta Aleatória promove a diversidade entre as árvores, reduzindo assim a probabilidade de overfitting. Ao final, cada uma das árvores que compõem a floresta é utilizada para determinar e escolher o melhor resultado.

Assim como nas Árvores de Decisão, os hiperparâmetros de Divisão de Critério e Máxima profundidade, também aparecem nos algoritmos de Random Forest. Um outro fator importante é o “n-estimators”, Número de Árvores, em que se determina o número de árvores que compõem a floresta. Vale dizer que um número maior de árvores geralmente leva a um desempenho mais robusto, mas também aumenta o tempo de treinamento.

2.2 Métricas

A avaliação desses modelos se deu por meio de diferentes métricas, como :

- Acurácia (accuracy)
- Precisão (precision)
- Recall
- F1 -score
- Tempo de processamento

2.2.1 Acurácia

"Accuracy"representa a porcentagem de previsões corretas feitas por um modelo em relação ao total de previsões. Essa métrica fornece uma medida direta da precisão global do modelo, indicando a proporção de instâncias classificadas corretamente em relação ao número total de instâncias avaliadas. Quanto maior a precisão, melhor o desempenho do modelo na

tarefa de previsão.

2.2.2 Precisão

O “Precision” mede a capacidade de um modelo de machine learning em classificar corretamente as instâncias positivas, ou seja, a capacidade de encontrar somente as amostras relevantes. A precisão fornece uma medida da qualidade das previsões positivas do modelo. É especialmente útil em situações em que o custo de um falso positivo é alto, ou seja, quando é prejudicial rotular incorretamente uma instância como positiva.

2.2.3 Recall

O recall, também conhecido como sensibilidade ou taxa de verdadeiros positivos, é uma métrica de desempenho em problemas de classificação que mede a capacidade de um modelo encontrar todas as amostras positivas. Em outras palavras, o recall indica a proporção de instâncias positivas que foram corretamente identificadas pelo modelo em relação ao total de instâncias positivas presentes nos dados.

2.2.4 F1-score

A pontuação F1 (ou F1-score) é uma métrica que combina precisão e recall em uma única medida, fornecendo uma avaliação mais equilibrada do desempenho de um modelo de classificação. A média harmônica é utilizada em vez da média aritmética para dar mais peso às pontuações mais baixas. Isso significa que a F1-score será mais influenciada pelo menor valor entre precisão e recall, tornando-se uma métrica útil quando você deseja encontrar um equilíbrio entre essas duas métricas. A F1-score varia de 0 a 1, onde 1 indica um desempenho perfeito (precisão e recall perfeitos) e 0 indica um desempenho pobre (ou precisão ou recall muito baixos).

2.2.5 Tempo de Execução

O tempo de desempenho de um algoritmo de machine learning é uma consideração crítica que afeta a eficiência e a utilidade prática do modelo. Esse aspecto refere-se ao tempo que um algoritmo leva para treinar um modelo e fazer previsões com base nos dados.

3 METODOLOGIA

O trabalho foi feito em linguagem Python, por meio do Jupyter Notebook, usando bibliotecas Panda, Matplotlib, Sklearn e Numpy. A fonte dos dados foi o Kangle,(DADOS, 2023).

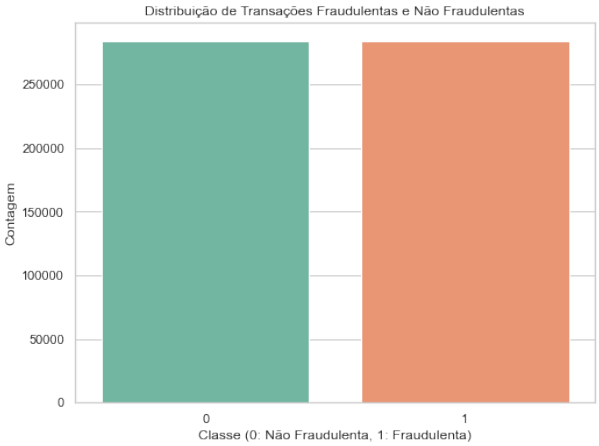
3.1 Base de Dados

Este conjunto de dados contém transações de cartão de crédito realizadas por titulares de cartões europeus no ano de 2023, vale ressaltar que um problema comum em dados de fraude cartão é o desbalanceamento, algo que não ocorreu aqui, ao ser coletado não é deixado claro se os dados foram balanceados de forma artificial ou não. De qualquer forma, os dados foram anonimizados para proteger as identidades dos titulares dos cartões. Ele é composto 568630 linhas e 31 colunas, entre elas temos :

- id : Identificador
- Coluna V1 a V28: São os atributos anonimizados
- Amount: O valor da transação
- Class: Rótulo binário indicando se a transação é fraudulenta (1) ou não (0)

Os dados eram compostos por 284315 amostras não fraudulentas e 284315 fraudulentas. Já em relação aos valores a média de valores das transações não fraudulentas foi de 12026.31, não ficando claro a unidade monetária, já para as fraudulentas foi de 12057.60.

Figura 1 – Distribuição dos Dados



Fonte: elaborada pelo autor.

Para o treinamento dos modelos foram usadas as colunas de maior correlação com a coluna "Class", sendo assim foram escolhidas as colunas "V2", "V4","V8", "V11", "V21","V27"e

"Amount". Pois apresentvam valores maiores de correlação de Pearson e Spearman.

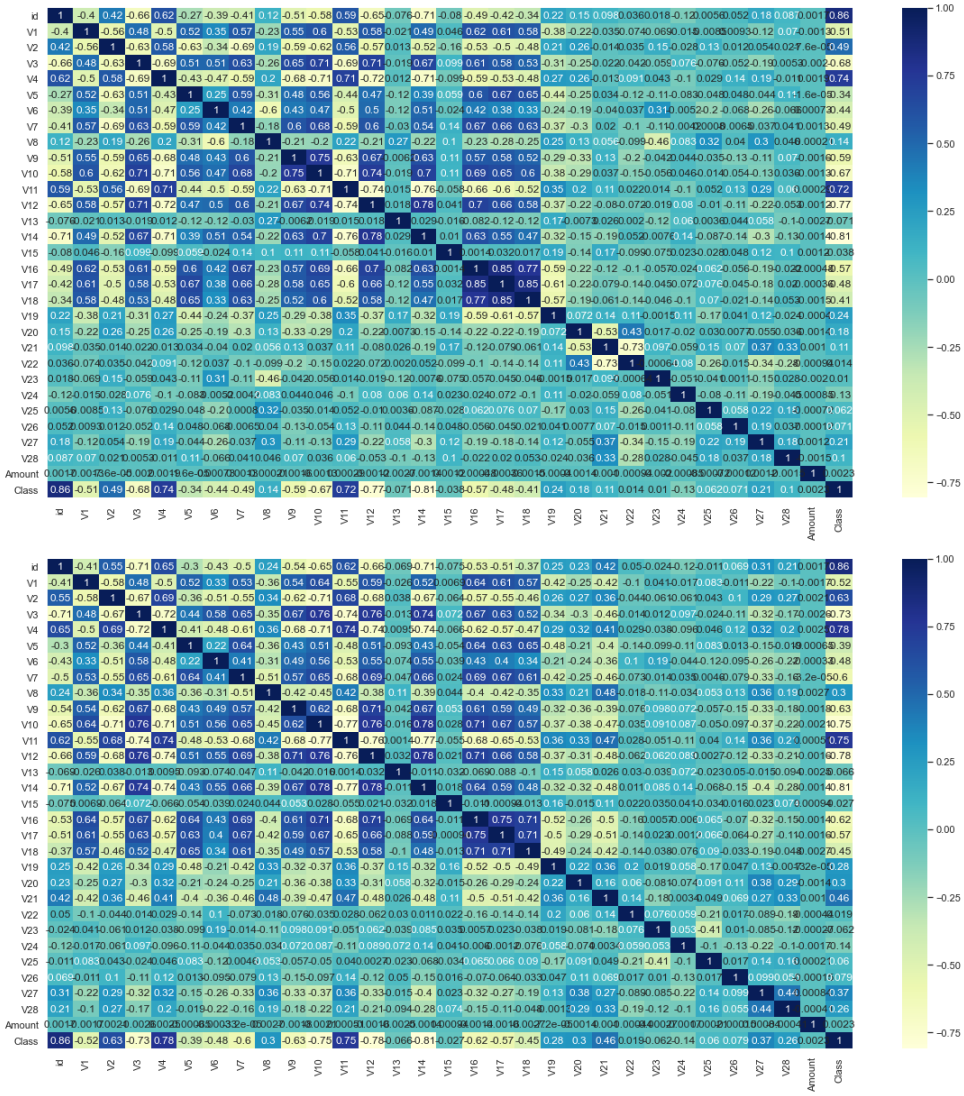


Figura 2 – Correlações de Pearson e Spearman.

3.2 Pré-processamento

O processo de limpeza dos dados teve início na busca de dados faltantes e duplicados, os quais não constavam no dataset. Os dados passaram pela divisão treino e teste, em que 80% dos dados foram usados para treino e 20% para teste.

Os dados de treino passaram por uma padronização, StandardScaler, em que a média é removida e a variação é unitaria, isto é, cada coluna deve ter um valor de média igual a zero e um desvio-padrão igual a 1.

3.3 Experimento

Inicialmente, cada um dos algoritmos foi treinado usando variações de parâmetros,(BORGNE *et al.*, 2022), em busca do que seria a melhor combinação. Quando a melhor combinação foi encontrada, isto é, a combinação com melhor score, média da pontuação da validação cruzada, no caso a métrica foi o F1 score, por ser a média harmônica entre a precisão e o recall, eram realizados mais testes com esse arranjo de parâmetros, no total de 5 repetições para cada um dos algoritmos.

O método usado para realizar as combinações dos parâmetros foi o Grid Search CV, em que é realizada uma busca em um grid de combinações de hiperparâmetro que é definida previamente, após os ajuste do modelo ocorre a avaliação por meio da validação cruzada, neste caso foram usado 5 k folds, ao final obtém-se como resposta os fatores que melhor desempenho.

Para a Regressão Logística foram analisadas 4 possibilidades para o parâmetro C, isto é, foram testados quatro valores diferentes para o parâmetro, são eles 0.1, 1, 10 e 100.

Por sua vez a Decision Tree teve seus parâmetros de Criterion e Profundidade mudados, o no caso do primeiro foi usado a Entropia e para o segundo os valores aplicados foram 2, 5, 10 e 20, assim são 5 combinações de parâmetros para esse algoritmo.

Enquanto que para os parâmetros do Random Forest , o Criterion também foi Entropia, o estimador foi igual a 100 e os valores de profundidade variaram None, 2, 5 e 10, sendo None um valor default da biblioteca. Os valores escolhidos foram 10 para o parâmetro C da RL, já as profundidade da Decision Tree e Random Forest são, respectivamente, 20 e None.

3.4 Testes Estatísticos

Apos a obtenção dos resultados dos testes para os três modelos será feita uma analise estatística dos resultados. O objetivo é realizar um experimento a fim de confirmar ou rejeitar uma hipótese. O Teste de Hipótese contava com a Hipótese nula, isto é, a hipótese cuja possibilidade é provada e a hipótese alternativa, um contraponto da nula (o que se espera provar). Foram usadas duas hipóteses nulas uma para as métricas de recall, precisão e fscore, por serem bem relevantes em um contexto de fraude de cartões, em que uma complementa a outra. E um teste de hipótese relacionado ao tempo de processamento dos modelos.

O nível de significância (α), isto é, a probabilidade de rejeitar a hipótese nula quando ela é verdadeira, usado foi de 0.05, por ser um valor frequentemente aplicado. O teste usado foi

o ANOVA (Análise de Variância), em que se faz a comparação entre as médias de mais de duas amostras. Assim se o P-value for menor que o nível de significância a hipótese nula é rejeitada.

Quando o teste de ANOVA indicar uma diferença significativa, considerar testes de comparação pós-hoc, como o teste de Tukey, para identificar quais grupos diferem entre si.

4 RESULTADOS

4.1 Validação Cruzada da Regressão Logística

Os resultados do grid para Regressão Logística mostraram que os resultados mais adequados se deram para o parâmetro C recebendo o valor 10. A métrica de escolha foi a média do F1-score das 5 repetições da validação cruzada, sendo assim f1 é igual 0.957.

Tabela 1 – Médias Resultados da Validação Cruzada para RL

Params	Avg Accuracy	Avg Recall	Avg Precision	Avg F1 Score
'C': 0.1	0.955	0.930	0.979	0.954
'C': 1	0.957	0.934	0.979	0.956
'C': 10	0.958	0.937	0.978	0.957
'C': 100	0.957	0.935	0.978	0.956

Fonte: elaborada pelo autor.

Tabela 2 – Desvio Padrão Resultados da Validação Cruzada para RL

Params	Std Accuracy	Std Recall	Std Precision	Std F1 Score
'C': 0.1	0.005	0.010	0.001	0.005
'C': 1	0.001	0.001	0.002	0.001
'C': 10	0.001	0.002	0.002	0.001
'C': 100	0.001	0.004	0.002	0.001

Fonte: elaborada pelo autor.

4.2 Validação Cruzada da Árvore de Decisão

A cross-validation para a Árvore de Decisão teve seu melhor resultado com a aplicação da entropia e a profundidade de 20, em que f1 é igual a 0.990.

Tabela 3 – Médias Resultados da Validação Cruzada para DT

Params	Avg Accuracy	Avg Recall	Avg Precision	Avg F1 Score
'max depth': 2	0.903	0.927	0.885	0.906
max depth': 5	0.930	0.912	0.946	0.929
'max depth': 10	0.957	0.942	0.971	0.956
'max depth': 20	0.990	0.993	0.987	0.990

Fonte: elaborada pelo autor.

Tabela 4 – Desvio Padrão Resultados da Validação Cruzada para DT

Params	Std Accuracy	Std Recall	Std Precision	Std F1 Score
'max depth': 2	0.001	0.001	0.001	0.001
'max depth': 5	0.001	0.010	0.008	0.002
'max depth': 10	0.001	0.002	0.002	0.001
'max depth': 20	0.000	0.001	0.000	0.000

Fonte: elaborada pelo autor.

4.3 Validação Cruzada da Floresta Aleatória

A realização do grid para a Floresta Aleatória obteve menlhores resultados com o estimador de 100, entropia e a profundidade None, isto é, entre os valores testados o default foi considerado como mais adequado, pois obteve-se um F1 de 0.997.

Tabela 5 – Médias Resultados da Validação Cruzada para RF

Params	Avg Accuracy	Avg Recall	Avg Precision	Avg F1 Score
'max depth': None	0.997	0.998	0.997	0.997
max depth': 2	0.911	0.856	0.962	0.906
'max depth': 5	0.938	0.902	0.972	0.936
'max depth': 10	0.963	0.941	0.984	0.962

Fonte: elaborada pelo autor.

Tabela 6 – Desvio Padrão Resultados da Validação Cruzada para RF

Params	Std Accuracy	Std Recall	Std Precision	Std F1 Score
'max depth': None	0.000	0.000	0.000	0.000
'max depth': 2	0.002	0.006	0.002	0.002
'max depth': 5	0.002	0.004	0.001	0.002
'max depth': 10	0.001	0.001	0.001	0.000

Fonte: elaborada pelo autor.

4.4 Resultado dos Testes

4.4.1 Acurácias

Os valores de acurácia para cada um dos testes realizados foi observado que os melhores resultados obtidos foi para o Random Forest, seguido pelo Decision Tree e Logistic Regresion. O primeiro conseguiu uma acurácia de 0.999, enquanto os outros tinham 0.998 e 0.958, em valores arredondados. Isso significa que os modelo desempenharam bem os seus

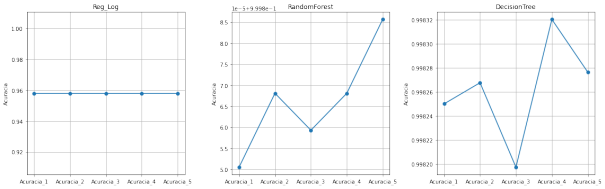
objetivos, no entanto é valido dizer que somente a acurácia não é suficiente para afirmar qual modelo se comporta melhor.

Tabela 7 – Acurácia dos testes

Modelo	Acuracia 1	Acuracia 2	Acuracia 3	Acuracia 4	Acuracia 5
Reg Logistica	0.958	0.958	0.958	0.958	0.958
Arv Decisão	0.998	0.998	0.998	0.998	0.998
Flr Aleatoria	0.999	0.999	0.999	0.999	0.999

Fonte: elaborada pelo autor.

Figura 3 – Acurácia dos Modelos



Fonte: elaborada pelo autor.

Como o desvio padrão era muito próximo a 0.0, mostrando que os valores não apresentavam grandes variações, a média aritmética entre as acurácias das 5 repetições foi considerada como adequada. . Essa escolha se mostrou apropriada, uma vez que a estabilidade nos resultados sugere consistência nas avaliações e reforça a confiabilidade da média como uma representação válida do desempenho médio dos modelos.

Tabela 8 – Estatísticas das Acurácias

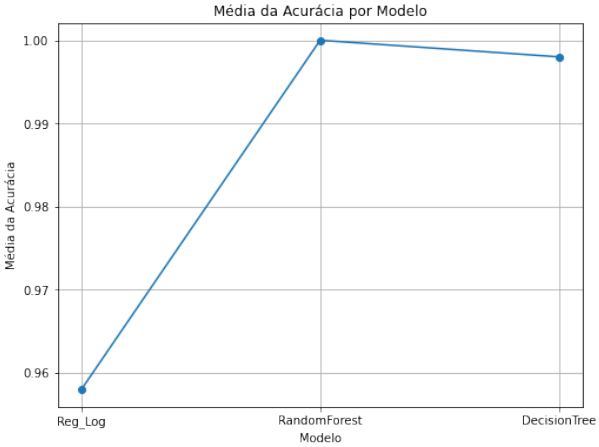
Modelo	Media	Std	Mediana
Reg Logistica	0.958	0.0	0.958
Arv Decisão	0.998	0.0	0.998
Flr Aleatoria	0.999	0.0	0.999

Fonte: elaborada pelo autor.

4.4.2 Precisão

A precisão dos modelos foram valores muito próximos, para a Regressão Logística a precisão foi de 0.979, por sua vez a Árvore de Decisão e Floresta Aleatória tiveram resultados iguais a 0.997 e a 0.999, respectivamente. Indicando um bom desempenho para os três casos, sem uma analise aprofundada é possível considerar uma robustez maior do terceiro modelo, tendo em vista que a precisão é a proporção de instâncias relevantes recuperadas pelo modelo

Figura 4 – Acurácia Média dos Modelos



Fonte: elaborada pelo autor.

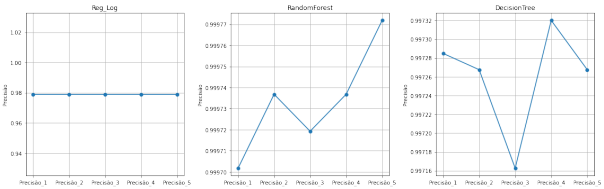
que são realmente relevantes, os resultados são bem satisfatórios. Mas a Random Forest possui resultados mais precisos, seguidos por Decision Tree e Logistic Regression.

Tabela 9 – Precisão dos testes

Modelo	Precisão 1	Precisão 2	Precisão 3	Precisão 4	Precisão 5
Reg Logistica	0.979	0.979	0.9798	0.979	0.979
Arv Decisão	0.997	0.997	0.997	0.997	0.997
Flr Aleatoria	0.999	0.999	0.999	0.999	0.999

Fonte: elaborada pelo autor.

Figura 5 – Prcisão dos Modelos



Fonte: elaborada pelo autor.

As estatísticas revelam não apenas altos níveis médios de precisão para cada modelo, mas também uma notável estabilidade e consistência em suas repetições. Esses resultados indicam um desempenho robusto e confiável dos modelos, reforçando a confiança em suas capacidades de classificação de instâncias positivas.

4.4.3 Recall

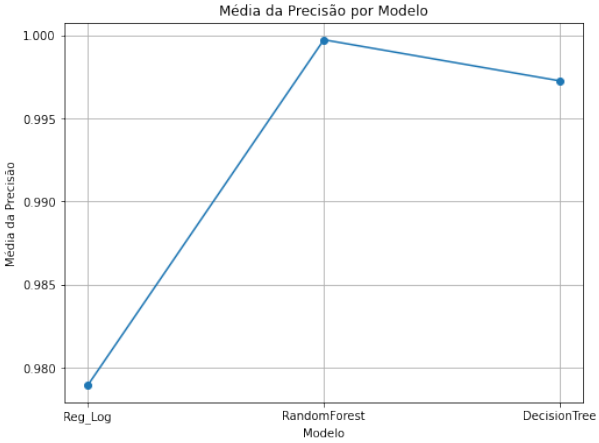
Os valores de recall unidos com os de precisão podem permitir uma avaliação mais completa dos modelos. Pois um modelo pode ter alta precisão, mas um recall baixo, o que

Tabela 10 – Estatísticas das Precisões

Modelo	Media	Std	Mediana
Reg Logistica	0.979	0.0	0.979
Arv Decisão	0.997	0.0	0.997
Flr Aleatoria	0.999	0.0	0.999

Fonte: elaborada pelo autor.

Figura 6 – Precisão Média dos Modelos



Fonte: elaborada pelo autor.

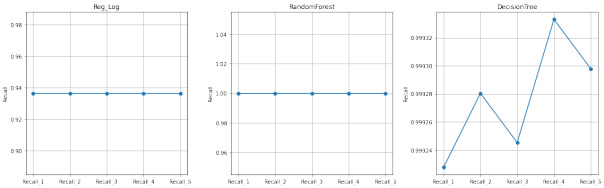
significa que é preciso ao classificar instâncias positivas, mas perde algumas instâncias relevantes. Por outro lado, um modelo com alto recall pode capturar a maioria das instâncias positivas, mas pode ter uma precisão menor, classificando incorretamente algumas instâncias negativas.

Tabela 11 – Recall dos testes

Modelo	Recall 1	Recall 2	Recall 3	Recall 4	Recall 5
Reg Logistica	0.936	0.936	0.9368	0.936	0.936
Arv Decisão	0.999	0.999	0.999	0.999	0.999
Flr Aleatoria	1.000	1.000	1.000	1.000	1.000

Fonte: elaborada pelo autor.

Figura 7 – Recall dos Modelos



Fonte: elaborada pelo autor.

Novamente, os resultados são bem satisfatórios, tendo em vista que os modelos geraram altos índices de recall, os quais são bem próximos uns dos outros, isso porque para a Regressão Logística obteve-se 0.936, que é um diferença pequena com os resultados da Árvore

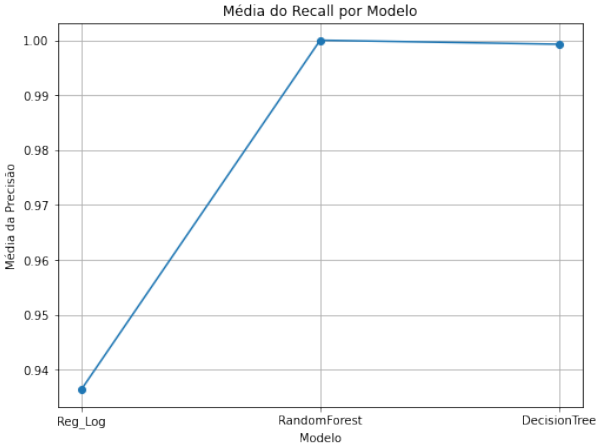
de Decisão que foram 0.999, que também é uma diferença mínima para o Random Forest que tinha recall de 1.00.

Tabela 12 – Estatísticas de Recall

Modelo	Media	Std	Mediana
Reg Logistica	0.936	0.0	0.936
Arv Decisão	0.999	0.0	0.999
Flr Aleatoria	1.00	0.0	1.00

Fonte: elaborada pelo autor.

Figura 8 – Precisão Média dos Modelos



Fonte: elaborada pelo autor.

4.4.4 F1-score

O F-score para a Regressão Logística permaneceu constante em todas as repetições, mantendo-se em 0.957, algo que também aconteceu para os outros dois algoritmos, para o Decision Tree o resultado obtido foi de 0.998, enquanto que para o Random Forest o F1 foi igual a 0.999.

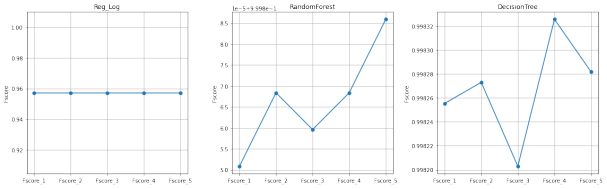
Tabela 13 – F1-score dos testes

Modelo	F-score 1	F-score 2	F-score 3	F-score 4	F-score 5
Reg Logistica	0.957	0.957	0.9578	0.957	0.957
Arv Decisão	0.998	0.998	0.998	0.998	0.998
Flr Aleatoria	0.999	0.999	0.999	0.999	0.999

Fonte: elaborada pelo autor.

Essa consistência destaca a habilidade do modelo em realizar previsões precisas, mantendo um equilíbrio entre a capacidade de classificar instâncias positivas e negativas. Assim

Figura 9 – Fscore dos Modelos



Fonte: elaborada pelo autor.

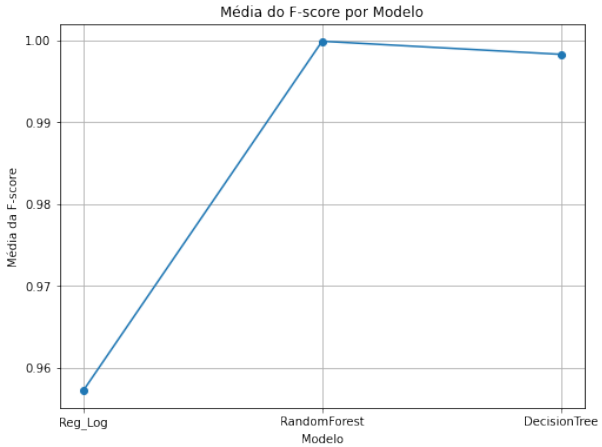
os resultados dos F-scores apontam para um desempenho robusto e equilibrado em todos os modelos avaliados, indicando que esses modelos são eficazes na tarefa em questão.

Tabela 14 – Estatísticas de Fscore

Modelo	Media	Std	Mediana
Reg Logistica	0.957	0.0	0.957
Arv Decisão	0.998	0.0	0.998
Flr Aleatoria	0.999	0.0	0.999

Fonte: elaborada pelo autor.

Figura 10 – Fscore Média dos Modelos



Fonte: elaborada pelo autor.

4.4.5 Tempo de Execução

O modelo de Regressão Logística demonstrou tempos de execução relativamente curtos, variando entre 8.734 e 15.501 segundos nas diferentes repetições. Essa eficiência computacional sugere uma rápida capacidade de processamento do modelo, tornando-o uma opção eficaz para a tarefa.

A RandomForest, por outro lado, exibiu tempos de execução mais substanciais, oscilando entre 1145.561 e 1677.391 segundos. Embora essa abordagem possa demandar mais

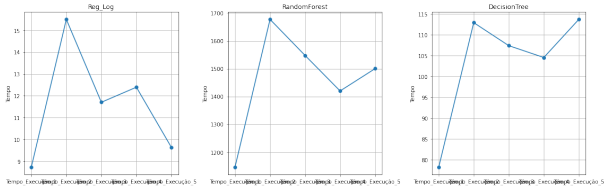
tempo computacional, seu desempenho geral robusto pode justificar o investimento temporal em cenários onde a precisão e robustez são prioridades.

Tabela 15 – Tempo de Execução dos testes

Modelo	Tempo 1	Tempo 2	Tempo 3	Tempo 4	Tempo 5
Reg Logistica	8.734	15.501 711.706	12.396	9.639	
Arv Decisão	78.159	112.914	107.382	104.537	113.706
Flr Aleatoria	1145.561	1677.391	1547.582	1419.845	1500.216

Fonte: elaborada pelo autor.

Figura 11 – Tempo de Execução dos Modelos



Fonte: elaborada pelo autor.

A DecisionTree, comparativamente, apresentou tempos de execução intermediários, variando de 78.159 a 113.706 segundos. Essa abordagem equilibrada em termos de tempo sugere uma eficácia computacional considerável, oferecendo uma alternativa que equilibra desempenho e eficiência.

A média de tempo de execução é de aproximadamente para os três modelos, com um desvio padrão relativamente baixos e mediana, próxima à média, sugere uma distribuição simétrica dos tempos de execução, com uma variação moderada.

Tabela 16 – Estatísticas de Tempo de Execução

Modelo	Media	Std	Mediana
Reg Logistica	11.595	2.642	11.706
Arv Decisão	103.340	14.586	107.382 Flr Aleatoria
1458.119	198.125	1500.216	

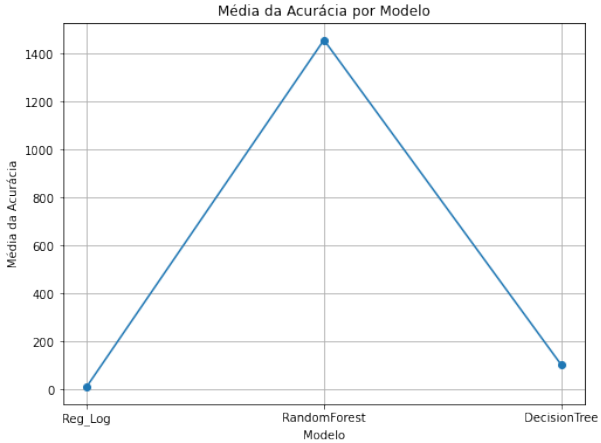
Fonte: elaborada pelo autor.

4.5 Testes Estatísticos

A aplicação do teste de ANOVA a partir das hipóteses para as métricas de precisão, recall e fscore:

Hipótese Nula (H0): Não há diferença estatisticamente significativa nas métricas de desempenho (precisão, recall, F-score) entre os modelos de detecção de fraude.

Figura 12 – Tempo Médio de Execução dos Modelos



Fonte: elaborada pelo autor.

Hipótese Alternativa (H1): Existe diferença estatisticamente significativa nas métricas de desempenho entre pelo menos dois dos modelos de detecção de fraude

A hipótese nula foi rejeitada, já que valor p foi menor que 0.05, indicando que há evidências de diferença estatisticamente significativa nas métricas de desempenho entre os modelos.

Assim ao se realizar o Teste de Tukey a diferença média (meandiff) entre a precisão da Decision Tree e Random Forest é de 0.0022, já para o recall e o fscore a meandiff foi de 0.0016 para os dois. O intervalo de confiança (lower, upper) está entre 0.0018 e 0.0025, para a precissão e 0.001 0.0016 são os limites para recall e fscore, enquanto o p-valor ajustado (p-adj) é 0.001, além da rejeição da hipótese nula (reject=True), para todas a métricas, indicando que há uma diferença significativa na precisão entre DecisionTree e RandomForest.

A diferença média entre a precisão da Decision Tree e Logistic Regression foi -0.018, para o recall e fscore foi -0.0411, por sua vez os limites são iguais para as metricas de recall e fscore, -0.0411 e -0.041, enquanto a precisão tinha -0.0183 e-0.0177 como limites. Além de ter a hipótese nula rejeitada em todos os casos, indica que há uma diferença significativa nas métricas dos modelos.

Já entre Logistic Regression e Random Forest a diferença média foi de -0.0202 e -0.0427, o primeiro para precisão e o segundo para o recall e fscore. No que diz respeito aos intervalos de confiança, para a precisão o lower foi igual -0.0205 e para as outras métricas igual a -0.0427, enquanto o upper foi de -0.0198 para a precisão e -0.0426 para os outros, em que mais uma vez a hipótese nula foi rejeitada.

A analise com base no tempo de execução foi feita com base na Hipótese Nula de

que não há diferença significativa nos tempos de execução entre os modelos. Com a rejeição dessa hipótese pelo Teste de ANOVA e usando o teste de Tukey. A diferença média entre o tempo de execução da DecisionTree e RandomForest é de 1354.7794 segundos. O intervalo de confiança está entre 1161.329 e 1548.2298 segundos. o p-valor ajustado (p-adj) é 0.001 e a rejeição da hipótese nula indica que há uma diferença significativa no tempo de execução entre DecisionTree e RandomForest.

Entre DecisionTree e Logistic Regression, a diferença média no tempo de execução é de -91.7444 segundos. O intervalo de confiança está entre -285.1948 e 101.706 segundos. O p-valor ajustado é 0.4419. Dessa vez não há a rejeição da hipótese nula (reject=False), indicando que não há uma diferença significativa no tempo de execução entre os modelos.

A diferença média no tempo de execução entre Random Forest e Logistic Regression é de -1446.5238 segundos. O intervalo de confiança está entre -1639.9742 e -1253.0734 segundos, com o p-valor ajustado é significativamente baixo (0.001), logo rejeição da hipótese nula indica que há uma diferença significativa no tempo de execução entre os modelos.

5 CONCLUSÕES E TRABALHOS FUTUROS

Os três modelos apresentaram uma desenvoltura notável na resolução do problema de identificação de fraudes de cartão de crédito. Mas é possível concluir que o modelo de Floresta Aleatória é estatisticamente preferido devido ao desempenho superior em todas as métricas consideradas.Já a Árvore de Decisão é uma escolha intermediária, superando a Regressão Logística em todas as métricas, mas ficando abaixo da Random Forest.

No entanto, em relação ao tempo o Random Forest apresentou um tempo de execução significativamente maior em comparação com a Decision Tree, mas significativamente menor em comparação com a Logistic Regression.

Sendo assim, se o tempo de execução é uma consideração crítica, a Decision Tree pode ser uma escolha mais eficiente em comparação com RandomForest e Logistic Regression. Porém o Random Forest pode ser uma opção aceitável se o ganho em desempenho justificar o custo adicional de tempo de execução.

No geral, O modelo Random Forest destaca-se como uma escolha forte, oferecendo desempenho superior nas métricas de classificação, apesar de um maior tempo de execução. A Decision Tree pode ser uma escolha intermediária, oferecendo bom desempenho com menor tempo de execução. A Logistic Regression pode ser a escolha preferida apenas se a eficiência computacional for a prioridade máxima, pois teve o menor tempo de execução, mas o desempenho nas métricas de classificação foi inferior

REFERÊNCIAS

BC. 2023. Disponível em: <https://www.bcb.gov.br/detalhenoticia/687/noticia>.

BORGNE, Y.-A. L.; SIBLINI, W.; LEBICHOT, B.; BONTEMPI, G. **Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook**. Université Libre de Bruxelles, 2022. Disponível em: <https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook>.

DADOS. 2023. Disponível em: <https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023>.

VALOR Investe. 2020. Disponível em: <https://valorinveste.globo.com/produtos/servicos-financeiros/noticia/2020/02/12/brasil-e-2o-pais-da-america-latina-com-mais-fraudes-no-cartao-em-compras-online.ghtml>.