Análise de Modelos de Aprendizado de Máquina Supervisiondo na Resolução do Desafio Machine Learning From Disaster

Hugo V. Bianquini[1], **Savio Suhett**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

Abstract. *This paper presents a method for solving the "Machine Learning From Disaster"challenge and an analysis of the performance of the machine learning models employed in the solution. The approach includes a comprehensive data analysis and preprocessing stage. The models investigated were Logistic Regression, Decision Trees, Random Forest, and Extra Trees. The different performance results provide valuable insights into the advantages and limitations of each model in the context of the proposed challenge.*

Resumo. *Este artigo apresenta um método para resolver o desafio "Machine Learning From Disaster"e uma análise de desempenho dos modelos de aprendizado de máquina empregados na resolução. A abordagem inclui uma etapa abrangente de análise e pré-processamento dos dados. Foram investigados os modelos de Regressão Logística, Árvores de Decisão, Random Forest e Extra Trees. Os diferentes resultados de desempenho fornecem insights valiosos sobre as vantagens e limitações de cada modelo no contexto do desafio proposto.*

1. Introdução

O avanço exponencial das técnicas de aprendizado de máquina tem permitido a aplicação dessas tecnologias em desafios complexos. As máquinas não apenas executam tarefas manuais, mas também desempenham funções cognitivas, abordando problemas que tradicionalmente requerem inteligência humana. [Ludermir 2021]. Em um cenário complexo como este, é de grande importância que os profissionais de tecnologia compreendam os algoritmos e técnicas utilizadas por modelos de aprendizado de máquina, sendo capazes de selecionar os melhores modelos a partir da interpretação do contexto.

O aprendizado de máquina supervisionado, por exemplo, é amplamente utilizado em escala global devido à sua eficácia em resolver uma vasta gama de problemas de classificação e regressão [Baranauskas and Monard 2000]. Um caso frequentemente utilizado no meio acadêmico trata-se do desafio "Machine Learning From Disaster", cuja tarefa é prever a sobrevivência de indivíduos em cenários de desastre (naufrágio do Titanic) utilizando dados históricos. Existem diversos métodos para a resolução do desafio, fato que poder ser utilizado para melhor compreensão de modelos de aprendizado supervisionado, o que constitui a principal motivação deste experimento.

A descrição, documentação e regras do desafio estão presentes na plataforma Kaggle, que é essencial para a área de aprendizado de máquina, pois oferece uma variedade de competições, conjuntos de dados, e recursos que permitem a pesquisadores e profissionais aprimorarem suas habilidades. [Banachewicz and Massaron 2022]

O presente trabalho visa propôr um método eficaz para resolver o desafio em questão, englobando uma etapa inicial abrangente de análise e pré-processamento dos dados, seguida pelo treinamento dos modelos e por fim a análise de desempenho das diferentes técnicas aplicadas. Os modelos de aprendizado de máquina investigados neste estudo incluem Regressão Logística, Árvores de Decisão, Random Forest e Extra Trees.

2. Trabalhos Relacionados

O desafio "Machine Learning From Disaster"é amplamente utilizado no meio acadêmico como uma ferramenta para expandir a experiência prática com aprendizado de máquina. Este desafio, que envolve a previsão de sobrevivência dos passageiros do Titanic com base em dados históricos, serve como um excelente ponto de partida para estudantes e pesquisadores aplicarem e testarem diferentes algoritmos de aprendizado supervisionado. Devido à sua popularidade e relevância, há uma abundância de experimentos e análises disponíveis que examinam diversas abordagens e técnicas, desde modelos mais simples com baixo custo computacional, até técnicas mais complexas com maior custo de processamento.

No estudo "Prevendo os Sobreviventes do Titanic Kaggle, Machine Learning From Disaster" [Farag and Hassan 2018], foram utilizadas técnicas de aprendizado de máquina, especificamente Árvores de Decisão e Naïve Bayes, para analisar e prever a sobrevivência dos passageiros do Titanic. O artigo destaca a importância da seleção de características e do pré-processamento dos dados para melhorar a precisão dos modelos preditivos. Os resultados mostraram que o algoritmo de Árvores de Decisão alcançou uma precisão de 90,01% na previsão da sobrevivência dos passageiros, enquanto o Naïve Bayes Gaussiano obteve uma precisão de 92,52%. Este trabalho ilustra a eficácia das técnicas de aprendizado de máquina supervisionado na resolução de problemas de classificação e fornece uma base comparativa valiosa para estudos subsequentes sobre predição de sobrevivência em cenários de desastre.

Em outro trabalho, entitulado "Analisando o desastre do Titanic usando algoritmos de aprendizado de máquina" [Singh et al. 2017], os autores buscaram determinar a correlação entre fatores como idade, sexo, classe do passageiro, tarifa, entre outros, e a chance de sobrevivência dos passageiros do Titanic. No artigo, foram implementados vários algoritmos de aprendizado de máquina, incluindo Regressão Logística, Naive Bayes, Árvore de Decisão e Random Forest, para prever a sobrevivência dos passageiros. Em particular, este trabalho de pesquisa compara os algoritmos com base na porcentagem de acurácia em um conjunto de dados de teste.

Na pesquisa "Análise exploratória de dados e aprendizado de máquina no conjunto de dados de desastres do Titanic" [Singh et al. 2020], o objetivo foi realizar uma análise exploratória de dados para compreender os efeitos dos parâmetros chave na sobrevivência de uma pessoa, caso ela estivesse a bordo do Titanic. A previsão de sobrevivência foi realizada aplicando vários algoritmos, como Regressão Logística, K-vizinhos mais próximos, Máquinas de Vetores de Suporte e Árvore de Decisão. Ao final, as acurácias dos algoritmos, com base nas características fornecidas, foram comparadas em formato tabular.

3. Implementação do Método

Nesta seção, serão detalhadas as diversas etapas envolvidas na implementação do método proposto. O início consiste na descrição dos materiais e tecnologias utilizados, proporcionando uma base sólida para a replicação do estudo. Em seguida, será abordada a análise exploratória do conjunto de dados e posteriormente o pré-processamento dos dados. Por fim, será tratado o processo de treinamento dos modelos de aprendizado de máquina.

3.1. Material Utilizado

Os conjuntos de dados de treino e teste utilizados neste estudo foram obtidos através do Kaggle, especificamente da página do desafio "Machine Learning From Disaster". Esses conjuntos de dados fornecem informações detalhadas sobre os passageiros do Titanic, permitindo a análise e a construção de modelos preditivos.

A implementação foi realizada utilizando a linguagem de programação Python, devido à sua versatilidade e ampla adoção na comunidade de ciência de dados. A IDE escolhida para o desenvolvimento foi o Visual Studio Code, equipada com a extensão para suporte à programação em Python e Jupyter, facilitando a escrita e execução do código de maneira interativa e organizada.

Diversos módulos e bibliotecas de Python foram empregados para a análise e modelagem dos dados, incluindo pandas para manipulação de dados, numpy para operações numéricas, matplotlib e seaborn para visualização de dados, e, principalmente, scikit-learn para a construção e avaliação dos modelos de aprendizado de máquina. O uso de Jupyter Notebooks foi fundamental para a organização do código e a documentação do processo, permitindo uma abordagem incremental e transparente no desenvolvimento do projeto.

3.2. Análise Exploratória dos Dados

A análise exploratória de dados (EDA) é uma etapa crucial no processo de ciência de dados e aprendizado de máquina, pois permite obter uma compreensão profunda do conjunto de dados, incluindo sua estrutura, distribuição e relações entre variáveis. A EDA facilita a detecção de anomalias e outliers, que podem distorcer os resultados da análise, e ajuda a avaliar a qualidade dos dados, revelando problemas como valores ausentes e inconsistências [Lopes et al. 2019].

Na prática, após a importação dos módulos descritos na seção "Material Utilizado", utilizou-se a biblioteca pandas para carregar os dados de treinamento e teste em dataframes, que são estruturas bidimensionais que organizam os dados em linhas e colunas, facilitando a manipulação e a análise. Com os dataframes carregados, foi possível descrever os atributos presentes no conjunto de dados, bem como examinar seu conteúdo de maneira eficiente. Uma maneira alternativa para se obter conhecimento genérico da base de dados, neste primeiro instante, é acessar a própria documentação do desafio no Kaggle, a qual descreverá cada um dos atributos, assim como seus possíveis valores ou tipos de conteúdo.

Ao todo, o conjunto de dados de treinamento possui 891 linhas, enquanto o conjunto de teste possui 418 linhas. Ambos possuem 12 atributos, são eles: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. Essas informações foram adquiridas utilizando o pandar e estão levemente distintas das

informações encontradas na documentação do Kaggle, como é possível notar através da tabela abaixo.

Tabela 1. Tabela de atributos na documentação do Kaggle

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Entre os atributos, destaca-se que ”Survived” é a variável de saída, indicando se o passageiro sobreviveu ou não. O atributo ”Pclass” representa a classe do passageiro, dividida em primeira classe, segunda classe e terceira classe. Os atributos ”SibSp” e ”Parch” descrevem as relações familiares a bordo, onde ”SibSp” representa o número de irmãos/cônjuges a bordo e ”Parch” o número de pais/filhos a bordo. A variável ”Fare” corresponde ao valor da tarifa da passagem comprada pelo passageiro.

Após uma interpretação básica e generalista, o próximo passo é avaliar a distribuição dos dados no conjunto de treinamento. Para isso, foram utilizados os módulos Seaborn e Matplotlib, que são fundamentais porque fornecem ferramentas robustas e versáteis para criar visualizações de alta qualidade. Os gráficos gerados a partir dessas ferramentas ajudam a transformar dados brutos em insights visuais claros, facilitando a identificação de padrões, tendências e outliers.

Iniciando com a avaliação dos atributos categóricos, os mais relevantes foram as variáveis ”Survived”, ”Pclass” e ”Sex”, que serão abordadas na continuação.

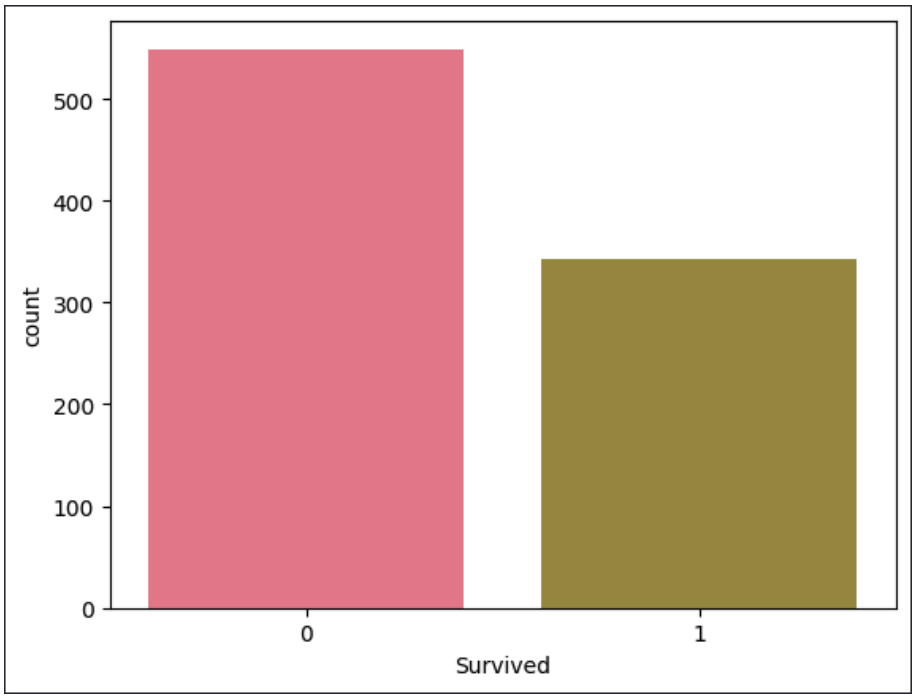

Figura 1. Distribuição de passageiros que sobreviveram ou não

Com base na figura acima, é possível notar que a maior parte dos passageiros não sobreviveu ao naufrágio. Ao todo, 342 passageiros sobreviveram, enquanto 549 não conseguiram sobreviver. Essa observação pode impactar no treinamento do modelo, já que trata-se da classe de saída e esta não está perfeitamente balanceada, em uma proporção de aproximadamente de 38.4% de sobreviventes e 61.6% de não sobreviventes.

O desbalanceamento da classe de saída afeta os modelos de aprendizado de máquina por várias razões, incluindo a tendência para a classe majoritária e aprendizado desigual. Técnicas mais avançadas de balanceamento, como ajustes de pesos ou conjuntos de dados sintéticos, poderiam ser utilizadas para garantir melhor distribuição das classes, o que seria uma das responsabilidades da etapa de pré-processamento [Batista et al. 2003]. Neste caso, como o desbalanceaento não é extremo, a decisão tomada foi de manter a distribuição original.

A análise prosseguiu com o atributo Pclass, que indica a distribuição dos passageiros de acordo com suas classes de viagem. Essa variável também serve como um indicador do status socioeconômico, visto que a primeira classe (valor mais alto) representava a elite do navio, enquanto a terceira classe (valor mais baixo) acomodava os passageiros com menor poder aquisitivo.

Compreender a divisão das classes de viagem dos passageiros no conjunto de treinamento é crucial, pois essa informação pode apresentar forte correlação com a variável de saída (sobrevivência dos passageiros). A distribuição não apenas influencia diretamente as previsões do modelo, mas também pode refletir profundamente nas características dos passageiros em diferentes classes socioeconômicas.

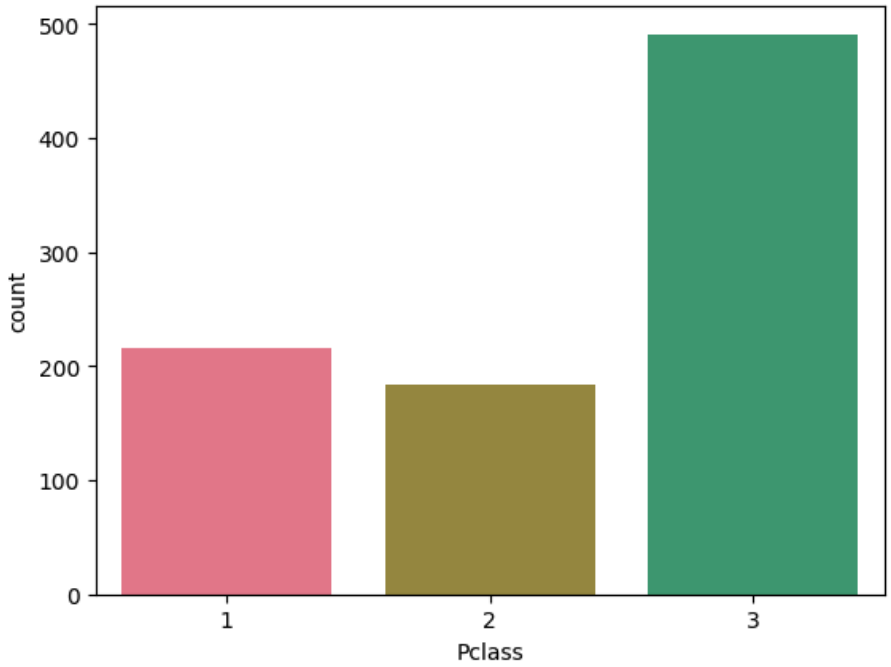

Figura 2. Distribuição das classes de viagem dos passageiros

Como é possível observar a partir da Figura 2, o número de passageiro de primeira e segunda classe são muito próximos, enquanto a quantidade de passageiros na terceira classe é mais que o dobro. Isso pode ser explicado socioeconomicamente, já que muitos não possuiam poder aquisitivo suficiente para comprar tickets de classes superiores.

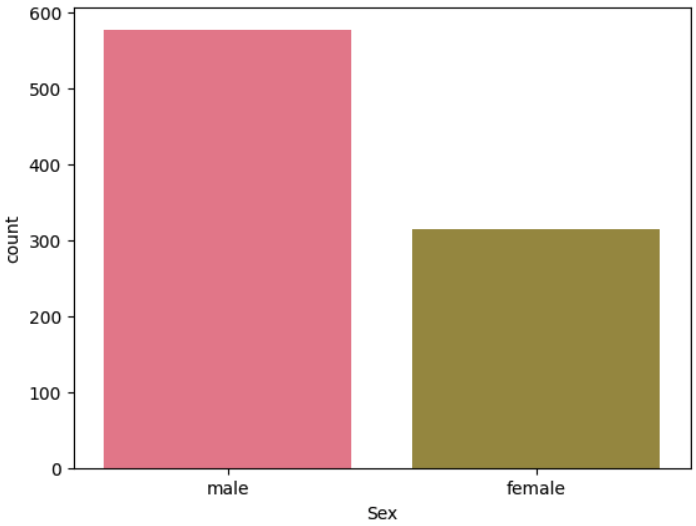

Figura 3. Distribuição do sexo dos passageiros

A figura 3 ilustra como a maior quantidade de passageiros é do sexo masculino, possuindo uma quantidade de 577 passageiros, enquanto 314 passageiras são do sexo feminino. Uma proporção de aproximadamente 65% de homens e 35% de mulheres.

Seguindo adiante, o próximo passo consiste na mesma análise de distribuição,

porém desta vez com atributos numéricos. Serão destacados os atributos ”Age” e ”Fare”.

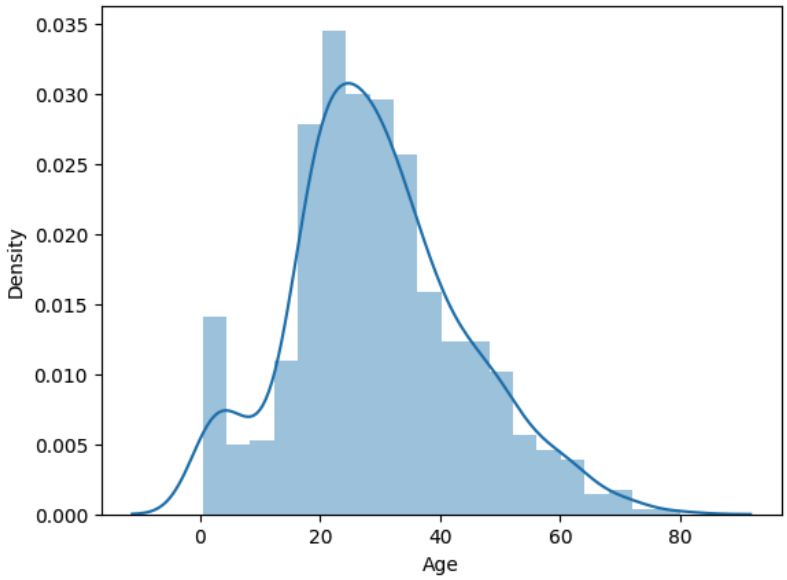

Figura 4. Distribuição de idade dos passageiros

A composição demográfica a bordo do Titanic mostra uma variedade de faixas etárias, refletindo a diversidade de passageiros embarcados. O gráfico ilustra que a maior densidade de idades é de passageiros entre 20 e 40 aos, sugerindo uma população predomunantemente jogem e em idade produtiva.

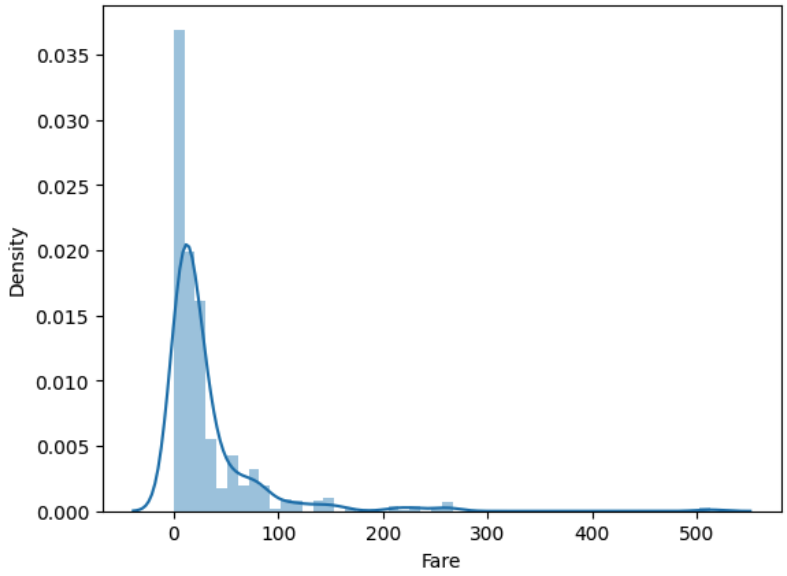

Figura 5. Distribuição da tarifa de passagem dos passageiros

O valor da passagem dos passageiros possui variações consideráveis e está potencialmente correlacionado às classes de viagem dos mesmos. As variações desde 0 até 500 sugerem que este atributo está passível de pré-processamento, objetivando transformar os valores para obter uma curva mais uniforme.

Após a análise individual dos principais atributos, é possível identificar algumas correlações intuitivas entre eles. A próxima etapa consiste em criar gráficos que relacionem dois ou mais desses atributos, permitindo uma compreensão mais profunda das interações e padrões presentes nos dados. Isso pode fornecer insights valiosos para as etapas subsequentes da análise. Os atributos mais relevantes para essas correlações são: ”Survived”, ”Pclass” e ”Fare”.

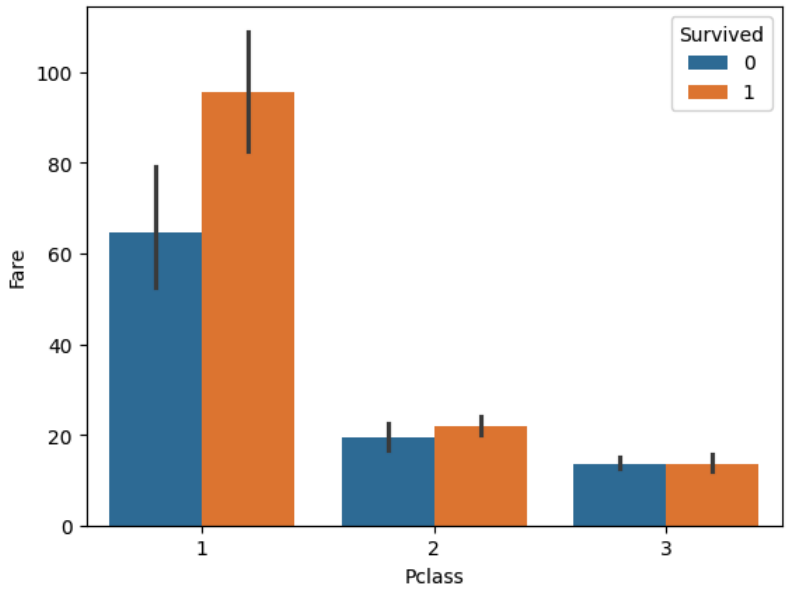

Figura 6. Relação entre classe de viagem, valor da passagem e sobreviventes

A partir do gráfico apresentado acima, é possível extrair duas informações importantes. Primeiramente, conforme esperado, os maiores valores de passagem são predominantemente associados à primeira classe de viagem. Em segundo lugar, observa-se que a maioria dos passageiros da primeira classe sobreviveu, enquanto as taxas de sobrevivência para as segunda e terceira classes estão muito mais equilibradas, bem dividida entre sobreviventes e não sobreviventes. O resultado pode ser justificado devido a melhores acomodações e acesso prioritário aos botes salva-vidas para a primeira classe.

3.3. Pré-processamento de Dados

Nesta seção, será abordado o pré-processamento dos dados, uma etapa essencial para garantir a qualidade e a eficácia dos modelos de aprendizado de máquina. O pré-processamento envolve a limpeza e transformação dos dados brutos, tratando valores ausentes, normalizando variáveis e codificando atributos categóricos [Batista et al. 2003]. Essas ações são fundamentais para preparar o conjunto de dados do desafio ”Machine Learning From Disaster” para o treinamento de modelos de aprendizado.

O primeiro passo do método adotado é combinar os conjuntos de dados de treinamento e teste, para garantir uma maior consistência no pré-processamento e reduzir chances de overfitting por tratar exclusivamente valores presentes nos dados de treinamento. Em seguinda, o segundo passo é constituído por uma busca por valores nulos ou ausentes.

Ao listar a quantidade de valores nulos ou ausentes em cada uma das colunas, observou-se a presença de 418 valores ausentes no atributo ”Survived”, 263 valores ausentes no atributo ”Age”, 1014 valores ausentes no atributo ”Cabin”, 2 valores nulos para o atributo ”Embarked” e 1 valor nulo para o atributo ”Fare”. Lembrando que ambos os conjuntos de dados de treinamento e teste foram combinados. Cada um dos atributos deve receber um tratamento de pré-processamento de forma que não prejudique o treinamento dos modelos.

Para o atributo ”Cabin”, a quantidade de valores nulos é tão grande que torna inviável a aplicação de qualquer técnica de preenchimento para esses registros. Diante disso, a decisão tomada foi simplesmente excluir essa coluna do conjunto de dados, considerando que o preenchimento não seria prático nem confiável.

Para os atributos numéricos ”Age” e ”Fare”, que são essenciais para o treinamento do modelo e possuem a maioria de seus valores preenchidos corretamente, optou-se pela utilização de técnicas de preenchimento de valores ausentes. A técnica escolhida para o método proposto é o preenchimento com base na média dos valores da coluna. Assim, a média foi calculada individualmente para cada atributo e, em seguida, utilizada para preencher os valores ausentes correspondentes. Dessa forma, mantém-se a integridade dos dados e assegura-se que todos os registros contenham informações completas para o treinamento do modelo.

Para atributos categóricos, como o atributo ”Embarked”, utilizou-se a técnica de substituição de valores nulos pela moda. Dessa forma, ao aplicar a moda da coluna, identificou-se a classe mais frequente nesse atributo, a qual foi utilizada para preencher os registros ausentes. Esse método garante que os valores imputados sejam representativos da maioria dos casos, mantendo a consistência do conjunto de dados.

Continuando na etapa de pré-processamento, conforme observado na análise exploratória, o atributo ”Fare”apresenta uma distribuição com alta variância, o que o torna um candidato ideal para transformação. Com o objetivo de estabilizar a variância dos valores e mitigar a influência de outliers, optou-se por aplicar uma transformação logarítmica nos dados do atributo ”Fare”. Essa transformação foi realizada utilizando o módulo numpy, o qual aplicou o logaritmo natural a cada valor da coluna, somando 1 para evitar problemas com valores zero ou negativos, conforme prática comum. O resultado desta operação poderá ser observado na Figura 7, que apresenta uma variância muito mais estável em relação à Figura 5, que continha o conteúdo original da variável ”Fare”.

Além de remover colunas com muitos valores ausentes, é essencial também excluir aquelas que são consideradas desnecessárias para o treinamento do modelo. Os atributos ”Name”, ”Ticket”e ”PassengerId” são exemplos claros disso, pois não contribuem diretamente para a predição de sobrevivência dos passageiros no contexto do desafio do Titanic. Portanto, essas colunas foram removidas do conjunto de dados para simplificar e focar nas variáveis que são mais relevantes para a modelagem preditiva.

O próximo passo é transformar oa atributos categóricos restantes em atributos numéricos. Realizar essa transformação é essencial para preparar dados para análise e modelagem em aprendizado de máquina. Essa prática permite que algoritmos processem e interpretem informações qualitativas de forma eficaz, atendendo aos requisitos de muitos modelos que só aceitam dados numéricos. Os atributos transformados foram ”Sex” e

”Embarked”. A ferramenta utilizada foi o LabelEncoder do módulo scikit-learn.

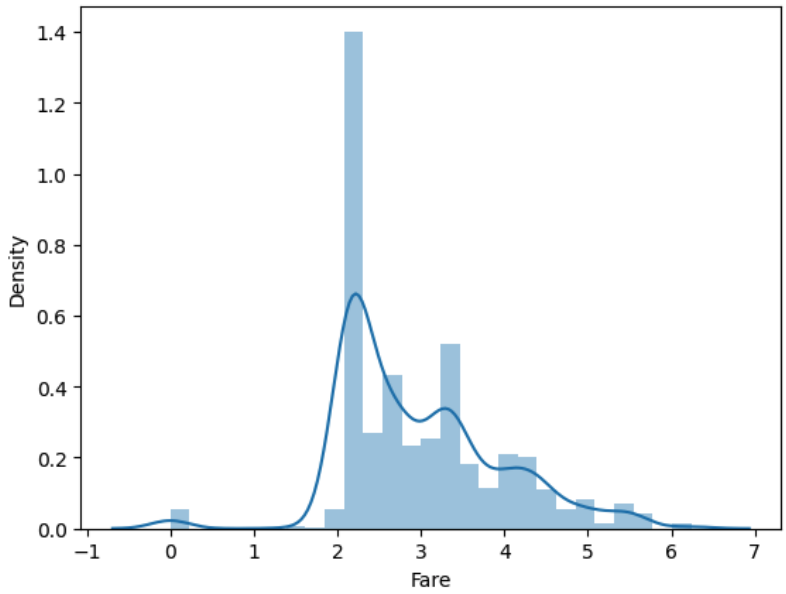

Figura 7. Valor da passagem após pré-processamento

3.4. Treinamento dos Modelos de Aprendizado de Máquina

Após a etapa de pré-processmaento, é necessário dividir os conjunto de dados, que foi combinado entre treinamento e teste anteriormente, fazendo com que retornem ao seu número de registros originais.

Antes de se aprofundar no processo de treinamento e análise de desempenho, é necessário retomar conceitos como a validação cruzada. É uma técnica fundamental em aprendizado de máquina e estatística para avaliar o desempenho de um modelo em dados não vistos. O objetivo é estimar a capacidade de generalização do modelo para novos dados que não foram usados no treinamento [Kuhn 2014].

Com este conceito em mente, foi desenvolvida uma função genérica para facilitar o treinamento e validação de modelos de classificação. Essencialmente, esta função recebe um modelo de classificação como parâmetro, utiliza a função train_test_split do módulo scikit-learn para dividir o conjunto de dados em atributos e rótulos, treina o modelo com os dados de treinamento resultantes, calcula sua precisão no conjunto de teste e realiza validação cruzada para uma avaliação mais abrangente do desempenho. Esse procedimento é amplamente adotado para comparar e avaliar modelos de aprendizado de máquina em problemas de classificação, oferecendo uma medida do desempenho geral do modelo.

A função de classificação e validação foi invocada quatro vezes, utilizando modelos de aprendizado supervisionado importados do scikit-learn. Os modelos treinados e validados foram, na sequência: Regressão Logística (LogisticRegression), Árvore de Decisão (DecisionTreeClassifier), Random Forest (RandomForestClassifier) e Extra Trees (ExtraTreesClassifier).

4. Análise de Desempenho

Nesta seção serão abordados os resultados e métricas de desempenho dos modelos utilizados, incluindo acurácia e o score da validação cruzada.

A pontuação da validação cruzada foi calculada como a média das pontuações obtidas em cada partição (fold) utilizada pelo modelo. Essa avaliação foi realizada utilizando a função cross_val_score do scikit-learn, que automatiza o processo de dividir os dados, treinar o modelo em múltiplos conjuntos de treino-teste e calcular as métricas de desempenho para cada fold individualmente.

Modelo	Acurácia	CV Score
LogisticRegression	80,72%	78,34%
DecisionTreeClassifier	71,75%	77,44%
RandomForestClassifier	80,27%	80,81%
ExtraTreesClassifier	79,37%	78,90%

Figura 8. Métricas de Desempenho dos Modelos

A figura apresenta as métricas de acurácia e score da validação cruzada (CV Score) para os modelos analisados. Observa-se que, de maneira geral, o desempenho foi satisfatório, com a menor acurácia registrada de 71,25% e o menor CV Score de 77,44%, ambos obtidos pelo modelo de Árvore de Decisão, que, segundo essas métricas, teve o desempenho mais baixo entre os modelos avaliados.

Para determinar o melhor modelo, foram ponderadas as métricas de acurácia e CV Score, atribuindo um peso ligeiramente maior ao CV Score devido à sua capacidade de avaliar melhor a generalização e evitar overfitting do modelo. Com base nessa análise, concluiu-se que o Random Forest se destacou como o melhor modelo entre os utilizados na resolução do desafio.

5. Conclusão

Em suma, este estudo explorou diversas técnicas de aprendizado de máquina supervisionado para resolver o desafio ”Machine Learning From Disaster” utilizando o conjunto de dados do Titanic. A análise abordou desde a preparação inicial dos dados até a avaliação detalhada do desempenho dos modelos. Observou-se que modelos como Regressão Logística, Árvore de Decisão, Random Forest e Extra Trees oferecem diferentes níveis de precisão e capacidade de generalização. Através da validação cruzada, foi possível estimar com maior confiança o desempenho dos modelos em novos conjuntos de dados. Com base nos resultados obtidos, o Random Forest emergiu como o modelo mais robusto, destacando-se pela sua capacidade de equilibrar acurácia e generalização.

Referências

Banachewicz, K. and Massaron, L. (2022). *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd.

Baranauskas, J. A. and Monard, M. C. (2000). Reviewing some machine learning concepts and methods.

Batista, G. E. d. A. P. et al. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. PhD thesis, Universidade de São Paulo.

Farag, N. and Hassan, G. (2018). Predicting the survivors of the titanic kaggle, machine learning from disaster. In *Proceedings of the 7th international conference on software and information engineering*, pages 32–37.

Kuhn, M. (2014). Futility analysis in the cross-validation of machine learning models. *arXiv preprint arXiv:1405.6974*.

Lopes, G. R., Almeida, A. W. S., Delbem, A., and Toledo, C. F. M. (2019). Introdução à análise exploratória de dados com python. *Minicursos ERCAS ENUCMPI*, 2019:160–176.

Ludermir, T. B. (2021). Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, 35:85–94.

Singh, A., Saraswat, S., and Faujdar, N. (2017). Analyzing titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 406–411. IEEE.

Singh, K., Nagpal, R., and Sehgal, R. (2020). Exploratory data analysis and machine learning on titanic disaster dataset. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 320–326. IEEE.