# Estudo de caso sobre a previsão do tempo de sono baseado nos estilos de vida: Uma abordagem comparativa dos modelos árvores de decisão e rede neural

*Lia Nicole C. da Costa[1] Nayra G. S. Monteiro [2] José Ribeiro[3]*

[1] Ciência da Computação – Instituto Federal do Pará (IFPA)
CEP 67125-000 – Ananindeua – PA – Brasil

{lianicolechaves@gmail.com, nayrabccifpa@gmail.com, jose.souza.filho@gmail.com}

***Resumo.*** *Este artigo investiga a relação entre a qualidade do sono e seus impactos na saúde, comparando modelos de aprendizado de máquina para prever variáveis associadas ao sono. Foram analisados modelos de classificação baseados em árvores de decisão e redes neurais. O conjunto de dados utilizado foi composto por informações sobre hábitos de sono, características demográficas e indicadores de saúde, de modo que permitiu que houvesse uma análise abrangente dos fatores que influenciam as horas de sono. As métricas de desempenho foram utilizadas para comparar os modelos. Os resultados indicam que as redes neurais apresentam maior precisão na previsão da qualidade do sono em comparação com o modelo baseado em árvore de decisão, sugerindo que a complexidade dos padrões no sono humano pode ser melhor capturada por estruturas mais profundas de aprendizado. Dessa forma, a aplicação de técnicas avançadas de aprendizado de máquina pode contribuir para um melhor entendimento dos fatores que afetam a qualidade do sono, auxiliando na formulação de estratégias para melhorar a saúde.*

## 1. Introdução

O sono tem um papel fundamental na vida de um ser humano, de modo que afeta tanto na saúde física quanto mental. De acordo com uma pesquisa feita pelo site Science Direct sobre a qualidade do sono na população geral brasileira em que obtida uma amostra da pontuação média do PSQI dos participantes da pesquisa, cerca de 65% dos brasileiros foram classificados como dormidores ruins. No Brasil, muitas pessoas sofrem com problemas de insônia, e são geralmente ocasionados pelas escolhas dos estilos de vida diário dos indivíduos.

Escolhas como o uso excessivo de dispositivos eletrônicos, em que a exposição à luz de smartphones e computadores antes de dormir pode inibir a produção de melatonina prejudicando o sono. Outros fatores como alimentação inadequada com dietas ricas em cafeína podem dificultar o relaxamento noturno e a falta de atividade física durante o dia reduz a necessidade de recuperação durante o sono, o que ocasiona a dificuldade em um descanso profundo. Além disso, uma rotina com muito estresse e sem repousos durante o dia pode causar a ausência do sono durante a noite. Outro fator que dificulta longas horas de sono é a questão das extensas jornadas de trabalho.

Dessa forma, a má qualidade do sono, está atrelada ao estilo de vida das pessoas, em que é um problema com grande relevância por conta dos impactos profundos na saúde individual e coletiva da população. Em um cenário onde o uso excessivo de dispositivos eletrônicos, jornadas de trabalho extensas, diversas pessoas enfrentam problemas como noites mal dormidas, o que contribui para uma má qualidade de vida e o aumento de problemas como ansiedade e depressão. Além disso, a privação de sono afeta a cognição e o desenvolvimento do ser humano, gerando uma péssima qualidade de vida.

A falta de sono pode se estender para além dos efeitos da saúde, pode estar atrelado também com ocorrências de acidente de trânsito, erros do ambiente corporativo, é crucial que reconhecermos a importância do sono como um pilar essencial da saúde pública e bem estar geral para uma sociedade coesa e produtiva.

Esta pesquisa visa resolver um tipo de problema de classificação.

Classificação é um tipo de problema de aprendizado supervisionado em que o principal objetivo é prever a categoria ou classe de uma nova observação com base nos dados. O modelo de classificação aprende a partir de dados de treinamento e depois pode classificar novos dados.

O modelo de árvore de decisão é um modelo preditivo baseado em uma estrutura hierárquica de decisões. O uso da árvore de decisão permite uma análise clara e intuitiva do problema, é muito utilizado para medir o desempenho de um modelo mais sofisticado.

O modelo de redes neurais são mais complexos, porque são eficazes em capturar padrões complexos e não lineares nos dados. O uso de redes neurais demonstra como um modelo mais sofisticado pode lidar com a complexidade de dados que não seriam bem representados em uma árvore de decisão.

A comparação entre os dois modelos evidencia a questão do equilíbrio entre um modelo básico com resultados rápidos e interpretáveis(árvore de decisão) e um modelo mais robusto, que é capaz de lidar com dados complexos, com maior custo computacional( rede neural) . Esta análise busca entender os impactos das escolhas de estilo de vida com reflexos na duração e qualidade do sono. Desta forma, será possível identificar cenários que possam ser prejudiciais a um determinado grupo, avaliar o desempenho

dos modelos para ter previsões precisas e informar hábitos que ajudem os indivíduos a ajustar os hábitos para melhorar a qualidade do sono.

O objetivo desta pesquisa é realizar uma análise comparativa simplificada entre dois modelos de aprendizado de máquina de diferentes níveis de complexidade, um baseado em árvore de decisão e o outro em rede neural, aplicados ao dataset previsão do tempo de sono, visando avaliar e comparar o desempenho dos modelos com base em medidas clássicas de performance com a finalidade de identificar qual o melhor modelo para predição do sono ideal baseado nas atividades diárias do indivíduo.

## 2. Metodologia

O dataset Sleeptime Prediction Dataset (Previsão do tempo de sono), é um conjunto de dados encontrado no site da Kaggle em que pode ser usado para prever a duração do sono com base no estilo de vida das pessoas. Este dataset representa um problema de classificação.

O dataset de previsão do sono possui 7 atributos e 2.000 instâncias.

WorkoutTime é tipo numérico em que representa em horas o tempo de exercício físico diário.

ReadingTime é o tempo de leitura em horas que o usuário gasta lendo. É um tipo numérico.

PhoneTime: Tempo de uso do telefone diariamente. É um tipo numérico.

WorkHours: Horas de trabalho trabalhadas por dia. É um tipo numérico.

CaffeineIntake: Ingestão de cafeína medido em mg por dia. É um tipo numérico.

RelaxationTime:Tempo de relaxamento dedicado ao descanso diário. É um tipo numérico.

SleepTime: Duração do sono em horas. É um tipo numérico.

### 2.1 Tratamento dos outliers

Os outliers são pontos de dados que se distanciam de forma significativa da tendência central dos dados. São pontos que possuem valores muito altos ou muito baixos em relação ao resto do conjunto de dados. Dessa forma, a retirada dos outliers é importante porque pode prejudicar nas análises estatísticas e nos modelos de aprendizado. Ao

realizar a implementação do dataset nas estimativas de localização, com cálculos para as variáveis, foi identificado a presença de outliers no atributo objetivo (sleepTime).

Atributo: SleepTime Outliers: [15.94, 8.71, 17.77, 0.74, 17.87, 15.57, 15.69, 0.54, 19.81, 13.8, 9.19, 1.1, 10.97, 11.92, 16.63, 12.33, 19.19, 18.35, 17.4, 1.28, 0.15, 11.52, 13.16, 17.58, 13.18, 8.58, 15.1, 9.43, 18.7, 14.95, 10.6, 9.41, 19.76, 17.11, 12.15, 13.89, 15.09, 16.66, 8.08, 16.87, 12.54, 17.78, 0.38, 15.49, 8.92, 19.2, 0.92, 11.92, 13.72, 18.67, 19.51, 8.85, 16.73, 17.66, 15.06, 16.81, 9.62, 11.62, 18.26, 0.16, 12.98, 17.3, 10.41, 8.41]

Dessa forma, a presença dos outliers indica que possui uma maior variabilidade nos padrões do sono. Desse modo, foi necessário tratar os outliers em dados e realizar a normalização dos dados de modo que reduza a variabilidade entre as variáveis.

Foi criada uma função para tratar outliers, em que dentro dessa função é calculado o 1° quartil(valor abaixo do qual 25% dos dados se encontram) e o 3° quartil que é o valor abaixo de 75% dos dados se encontram. O intervalo(IQR) é uma diferença entre o 3° e o 1° quartil, usado para identificar a presença de outliers. Após isso, são definidos os limites inferiores e superiores, em que qualquer valor fora dessas métricas( 1.5) são considerados outliers.

**2. 2 Normalização de dados**

A normalização de dados é uma técnica de pré- processamento realizada para ajustar a escala dos valores numéricos em um conjunto de dados. No caso do dataset de previsão do sono não foi necessário realizar a normalização, pois os dados numéricos possuem valores muito próximos.

O dataset a princípio era um problema de regressão, pois o atributo objetivo possuía valores contínuos e numéricos. Foi transformado para um problema de classificação, no qual, o atributo sleeptime foi alterado de numérico para categórico.

A análise exploratória dos dados do dataset faz parte do processo de análise de dados, pois podemos compreender o comportamento do conjunto de dados antes de aplicar os

modelos de árvores e redes neurais. Os cálculos estatísticos  e os gráficos ajudam a identificar padrões, outliers e a atual distribuição dos dados.

### 3.3 Histogramas

Os gráficos de histogramas são gráficos que representam a frequência de uma determinada variável, de modo que se possa compreender a distribuição de dados e identificar os padrões.

O histograma ajuda a visualizar  a distribuição das frequências e a compreender a formas como os dados estão distribuídos.

### 3.4 Boxplots

São gráficos de visualização estatística usados para mostrar a distribuição de um conjunto de dados, eles são importantes para demonstrar a dispersão dos dados, presença de outliers e a simetria dos dados.

### 3.5 Árvore de decisão ID3

O objetivo desse modelo é dividir os dados em subconjuntos baseados nos valores dos atributos das instâncias, o ID3 é uma variação da árvore de decisão que utiliza a entropia para determinar qual atributo melhor divide os dados em cada nó.

Foi utilizado a entropia para que a árvore de decisão procure a maior redução na entropia a cada divisão, este modelo foi treinado no conjunto de dados de treinamento e depois fez previsões no conjunto de teste juntamente com as métricas de desempenho do modelo.

### 3.6 Rede Neural MLP

O MLP é composto por uma camada de entrada até a camada de saída, cada camada de neurônios aplica uma função matemática sobre as entradas e transmite o resultado para os neurônios da próxima camada. O modelo utiliza 3 camadas ocultas com 40, 35 e 46 neurônios, esses números influenciam na capacidade de aprendizado do modelo, a função de ativação utilizada é a logística(sigmoid) que transforma a saída do neurônio

em um valor entre 0 e 1.  O código divide os dados em treinamento e teste e preserva a distribuição das classes por meio da estratificação. Logo após, é feita a avaliação do modelo por meio de métricas de desempenho e é gerada a matriz de confusão para visualizar os erros de classificação.

### 3.7 Bibliotecas utilizadas

Numpy: forneceu suporte para matrizes e arrays, foi utilizada para manipulações de dados, cálculos matemáticos e operações numéricas.

Pandas: Essa biblioteca é ideal para manipulação de dados tabulares, além de que ajudou no carregamento, limpeza, transformação e visualização dos dados estruturados.

Seaborn: Uma biblioteca de visualização de dados baseada no matplotlib, foi usada para criar gráficos estatísticos.

Matplotlib: essa biblioteca é capaz de criar gráficos de linhas, histogramas e barras, específica para visualização de dados.

Math: Foi usada para realizar cálculos matemáticos.

Sklearn.datasets: Usada para carregar e usar datasets direto da biblioteca.

Sklearn.preprocessing.StandardScaler: Usada para transformar variáveis numéricas para que tenham determinada escala.

Sklearn.tree.DecisionTreeClassifier: Usada para construir e treinar uma árvore de decisão.

Sklearn.model_selection.train_test_split: Foi utilizada para dividir os dados em partes de treinamento e teste.

Sklearn.metrics.accuracy_score: Usada para avaliar o desempenho dos modelos de classificação.

Sklearn.metrics.classification_report: Usada para avaliar o desempenho de um modelo de classificação, com as métricas de avaliação.

Sklearn.neural_network.MLPClassifier: Foi utilizada para construir e treinar uma rede neural com múltiplas camadas.

Sklearn.metrics.confunsion_matrix: Foi utilizada para visualizar e avaliar o desempenho do modelo, além de criar a matriz de confusão.

A divisão do dataset em treino e teste utilizou 70% para treinamento do modelo e 30% para avaliar o desempenho do teste. Logo após, os modelos são treinados e avaliados usando a acurácia que irá medir a porcentagem de previsões corretas feitas pelos modelos.

Além disso, a matriz de confusão exibe os números de previsões corretas, oferecendo uma visão detalhada do desempenho do modelo.

A acurácia é uma métrica que calcula a proporção de previsões corretas em relação ao total de previsões feitas. Na implementação da árvore de decisão a acurácia foi equivalente a 50,54% enquanto para a rede neural foi de 94%,22. O que a rede neural obteve um desempenho muito superior à árvore.

A precisão é uma métrica que avalia a qualidade das previsões positivas, no qual para cada classe, é calculado como o número de previsões corretas de uma classe em relação ao total de previsões feitas. A precisão para a classe “muito baixo” na árvore de decisão foi de 44%, enquanto que na rede neural foi de 87%, diante desses dados a rede neural teve um desempenho melhor para prever as classes.

O recall mede a capacidade do modelo de identificar corretamente todas as instâncias de uma classe específica. O recall da árvore de decisão para a classe “muito baixo” foi de 57% enquanto que na rede neural foi de 85%. Isso indica que a árvore de decisão teve dificuldade de identificar as classes menores, como “muito baixo”.

O F1-score é uma média ponderada entre a precisão e o recall, sendo útil quando há um desbalanceamento nas classes. O F1-score médio da árvore foi de 50% enquanto que na rede neural foi de 94%, em que indica que a rede neural teve um desempenho muito melhor na precisão e recall.

Disponível as implementações no collab: Projeto_LiaNicoleCosta

## 4. Resultados e Discussões

A análise dos atributos “WorkoutTime”, “ReadingTime”, “PhoneTime”, “WorkHours”, “CaffeineIntake”, “RelaxationTime” não possui muita variabilidade nos cálculos da média, desvios etc, o que indica que foram distribuídos de maneira relativamente uniforme. Não foram encontrados outliers para esses atributos, o que significa que eles são mais consistentes.

O atributo sleeptime apresenta uma média de sono de 4.88 horas enquanto a mediana é de 4.6 horas, em que retrata que os indivíduos dormem em média 4 a 5 horas diárias. Essa variável “sleeptime” precisou ser transformada para uma variável categórica com 4 categorias: muito baixo, baixo, normal e alta.

Foram identificados outliers com valores de sono altos ou baixos em relação à média. Esses dados precisavam ser tratados.

O gráfico abaixo(figura 1) demonstra que o atributo sleeptime possui valores muito baixos e muito altos, demonstrando a presença dos outliers.

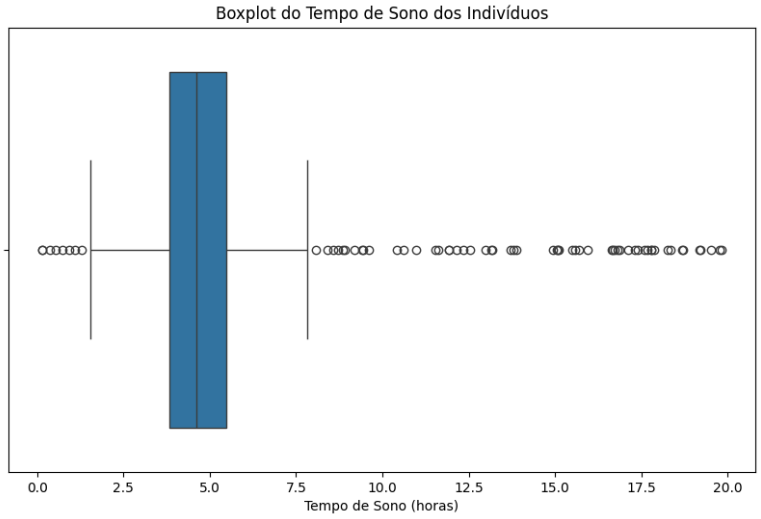


Figura 1

Após categorizar o atributo “sleeptime”, obteve os seguintes resultados na figura 2:

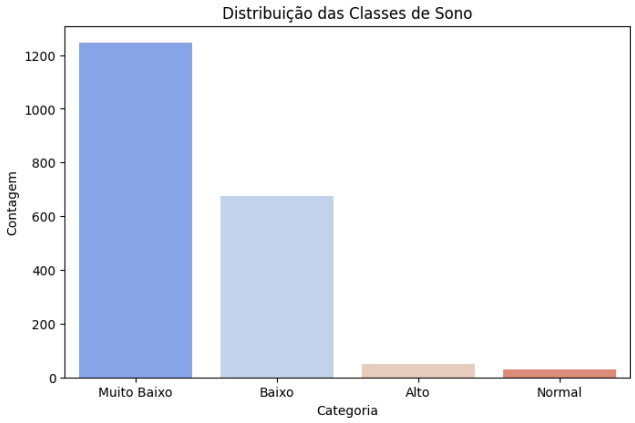


Figura 2

Possui muitas pessoas com o tempo de sono muito baixo, conforme o gráfico.

As correlações realizadas com o atributo sleeptime, indicam que:

- Correlação positiva 0.2957, quanto mais tempo de exercício, maior o tempo de sono.
- Correlação negativa -0.5311, Quanto mais tempo no celular, menor o tempo de sono.
- Correlação negativa -0.5059, quanto mais tempo trabalhando, menos tempo dormindo.
- Correlação negativa -01530, quanto maior o consumo de cafeína, menos horas de sono a pessoa possui.
- Correlação positiva 01998, quanto mais tempo de relaxamento, mais horas de sono o indivíduo possui.

Na árvore de decisão(figura 3), os dados foram divididos em 70% de treino e 30% de teste. O modelo teve um desempenho de 50.54%, o que representa que o modelo acerta

aproximadamente metade das previsões.

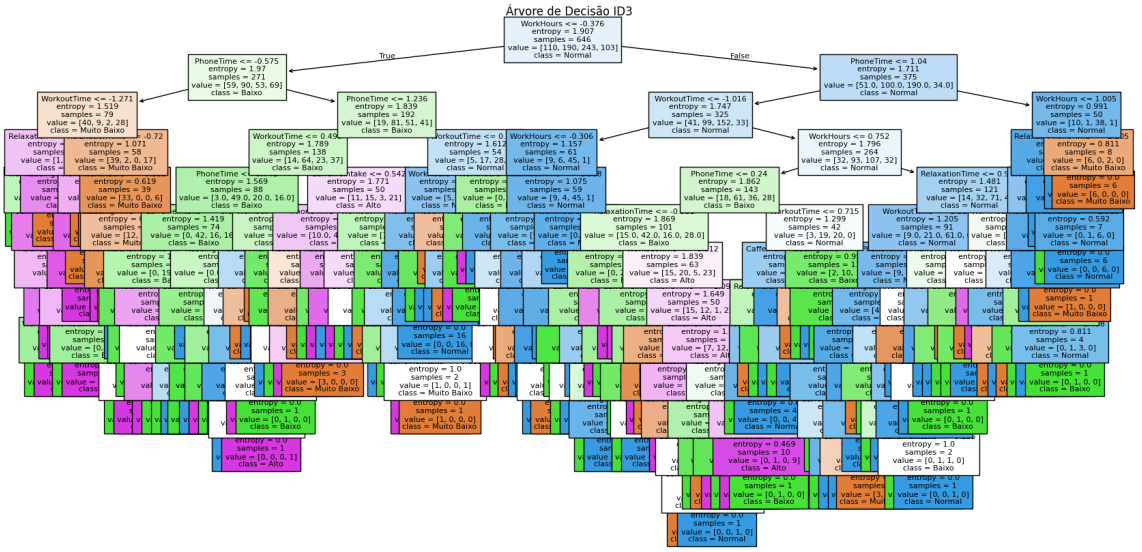


Figura 3

A precisão indica a precisão correta das classes, a classe normal tem o melhor desempenho, em que é mais fácil de prever. A classe “alto" tem o pior desempenho.

Na rede neural, os dados foram divididos em 70% para treino e 30% para teste, a rede neural possui 3 camadas ocultas de tamanhos (40, 35, 46).

A acurácia do modelo foi equivalente a 94.22%, o que significa que o modelo teve um desempenho muito bom acertando 94% das previsões. A categoria normal teve o melhor desempenho com o f1-score de 0.97 além disso, a classe “muito baixo” teve o menor desempenho. Conforme o gráfico abaixo (figura 4), a classe muito baixo possui maior quantidade e valores na célula, todos os valores na diagonal indicam os acertos do modelo.

- 104 amostras na classe “muito baixo”
- 77 amostras na classe “baixa”
- 40 amostras na classe “normal
- 40 amostras na classe 40

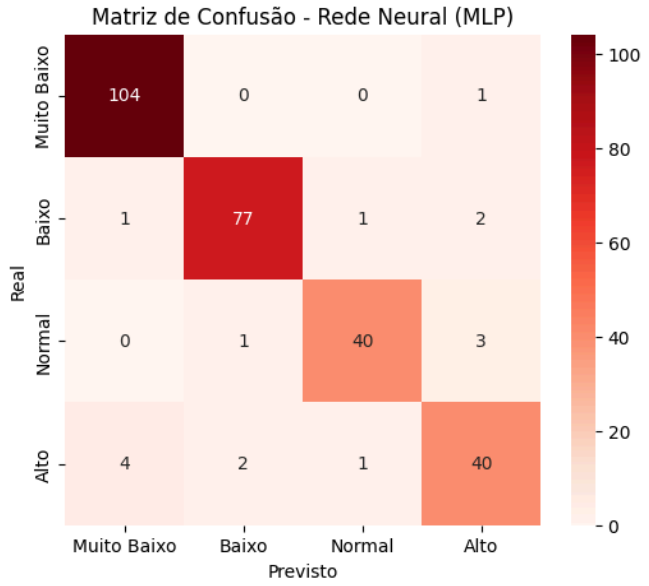


Figura 4

De acordo com o gráfico abaixo(figura 5), é demonstrado a comparação das acurácias do modelo da árvore de decisão e da rede neural. É visualizado a diferença do valores da rede neural(acurácia de 94.22%) e da árvore de decisão (acurácia de 50.54%), em que no problema do dataset o modelo da rede neural conseguiu captar melhor os padrões para prever as categorias do tempo de sono.

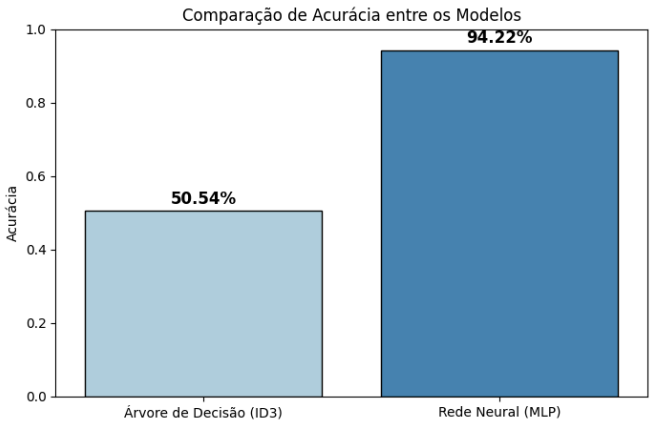


Figura 5

## 5 . Conclusão

Os dados da pesquisa analisaram a influência do estilo de vida baseado na previsão do tempo de sono utilizando modelos de como árvore de decisão e redes neurais. Os resultados indicam que variáveis como tempo de exercício, uso de celular, carga horária

de trabalho, consumo de cafeína e tempo de relaxamento apresentam correlações significativas com a duração do sono. A rede neural demonstrou um desempenho superior à árvore de decisão, alcançando uma acurácia de 94%, enquanto a árvore de decisão teve apenas 50,54%. A análise dos outliers evidenciou uma grande variabilidade no tempo de sono, justificando a necessidade de tratamento dos dados para melhorar a qualidade das previsões. Entre as limitações, houve a restrição do conjunto de dados utilizado, que pode não representar toda a diversidade da população, e a necessidade de maior balanceamento das classes para melhorar a precisão do modelo.

Para trabalhos futuros:

- Foi identificado o uso de técnicas de balanceamento de classes para melhorar a capacidade preditiva dos modelos.

O estudo da análise do sono contribui para a compreensão dos fatores que afetam o sono e destaca a importância do uso de modelos avançados para prever padrões de sono, fornecendo subsídios para futuras pesquisas e possíveis aplicações na área da saúde coletiva.

**6. Referências**


1. DRAGER, L.F; PACHITo, D. V; MORIHISA, R; CARVALHO,P. Qualidade do sono na população geral brasileira: um estudo transversal. Sleep Medicine: X, v. 4. Disponível em: https://www.sciencedirect.com/science/article/pii/S2667343622000014 Acesso em: 16 jan. 2025
2. Alura. Estatística para ciência de dados. 2023. Disponível em: https://www.alura.com.br/artigos/estatistica-ciencia-de-dados?srsltid=AfmBOopuqsDK2ovd8fNjn4kOeMwW1PT2amCwOKUB-iGuM0A8QV1iZ0JV Acesso em: 20 jan. 2025
3. DEVMEDIA. Classificação e análise de dados usando árvores de decisão. 2023. Disponível em: https://www.devmedia.com.br/classificacao-e-analise-de-dados-usando-arvore-de-decisao/7066 Acesso em: 25 jan. 2025
4. BRASIL ESCOLA. Medidas de dispersão: variância e desvio padrão. 2023. Disponível em: https://brasilescola.uol.com.br/matematica/medidas-dispersao-variancia-desvio-padrao.htm Acesso em: 26 jan. 2025.
5. DATACAMP. Multilayer Perceptrons in Machine Learning. DataCamp, 2023. Disponível em: https://www.datacamp.com/pt/tutorial/multilayer-perceptrons-in-machine-learning. Acesso em: 26 jan. 2025.
6. REICHERT JR, Igor. Boxplot e histograma: como interpretar. Medium, 2021. Disponível em: https://medium.com/@ingoreichertjr/boxplot-e-histograma-como-interpretar-e044b2bb2e2d. Acesso em: 29 jan.. 2025.