Classificação do distúrbio do sono usando Aprendizado de Máquina

Heitor Fidelis Beda Caete[1], Jose Joaquim de Santana Neto[2], Luciano de Souza Cabral [3], Renan Freitas dos Anjos[4]

[1]Instituto Federal de Pernambuco (IFPE)
Campus Jaboatão dos Guararapes, PE – Brasil

hfbc2@discente.ifpe.edu.br[1], jjsn2@discente.ifpe.edu.br[2], luciano.cabral@docente.ifpe.edu.br[3], rfa5@discente.ifpe.com.br[4]

Resumo. Transtornos do sono, como insônia e apneia do sono, são os principais fatores ligados a baixa qualidade do sono na população mundial. A compreensão esses distúrbios, com ferramentas de *machine learning*, podem ajudar a compreender o desvio padrão desses distúrbios e contribuem nas análises detalhadas de variáveis individuais, que permitem explorar relações causais entre essas características e distúrbios relacionados ao sono. O presente estudo possui como foco central abordar o uso de *machine learning*, utilizando método classificação como Gradient Boosting, Decision Tree e Cat Boost, para analisar e classificar dados cruciais de saúde.

Palavras chave: Distúrbios do sono, machine learning, classificação

Abstract: Sleep disorders, such as insomnia and sleep apnea, are primary factors contributing to poor sleep quality globally. Understanding these disorders, aided by machine learning tools, can help elucidate the variability of these disorders and facilitate detailed analyses of individual variables. This allows for the exploration of causal relationships between these characteristics and sleep-related disorders. The primary focus of this study is to investigate the application of machine learning, specifically classification methods like Gradient Boosting, Decision Tree, and Cat Boost, to analyze and classify critical health data.

Keywords: Sleep Disorder, machine learning, classification

1.Introdução

Os distúrbios do sono têm um impacto significativo nas funções vitais do corpo humano, comprometendo não apenas a qualidade de vida, mas também contribuindo para o desenvolvimento de uma série de doenças crônicas, como hipertensão, diabetes, obesidade e problemas cardiovasculares (Zhou et al., 2019; Liu et al., 2022).

Estes distúrbios, que incluem condições como apneia do sono, insônia, entre outras, estão em ascensão globalmente, afetando aproximadamente 45% da população, de acordo com a Organização Mundial da Saúde (OMS). A prevalência de transtornos do sono tem aumentado, em grande parte, devido a fatores como mudanças no estilo de vida, urbanização crescente e níveis

elevados de estresse, sendo cada vez mais urgente a busca por soluções eficazes para o diagnóstico e tratamento dessas condições. (Liu et al., 2022).

A intervenção precoce e o monitoramento contínuo de distúrbios do sono são fundamentais para reduzir o impacto dessas condições na saúde pública. A análise de grandes volumes de dados clínicos, por meio de tecnologias avançadas, vem se apresentando com uma abordagem promissora. Nesse contexto, técnicas de inteligência artificial, especialmente o *machine learning* (ML), surgem como ferramentas potencialmente revolucionárias para a identificação de padrões e a previsão de distúrbios do sono. Modelos de aprendizado supervisionado, como Gradient Boosting, Decision Tree e CatBoost, são particularmente eficazes no processamento e análise de dados complexos, devido à sua capacidade de lidar com variáveis interdependentes e não-lineares, comuns em datasets de saúde (Ganaie et al., 2023).

Além de suas aplicações em diagnóstico, os algoritmos de *machine learning* podem orientar a formulação de estratégias de saúde pública voltadas para a melhoria do bem-estar global, alinhando-se com os Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU), que incluem a promoção de uma vida saudável e bem-estar. A capacidade de prever e classificar distúrbios do sono pode não apenas aprimorar o tratamento clínico individual, mas também fornecer *insights* valiosos para políticas públicas, principalmente em contextos com sistemas de saúde sobrecarregados. (Zhou et al., 2019).

No entanto, apesar do seu potencial, a aplicação de *machine learning* no diagnóstico e no tratamento de distúrbios do sono enfrenta desafios significativos. A qualidade dos dados e a representatividade das amostras desempenham um papel importante na performance dos modelos, sendo necessária uma abordagem cuidadosa no pré-processamento dos dados para evitar vieses na sua execução.

Além disso, métricas de avaliação de modelos, como acurácia, podem não ser suficientes quando se lida com dados desbalanceados, como é o caso de distúrbios do sono raros. Nesse cenário, métricas complementares, como recall e f1-score, tornam-se essenciais para a avaliação completa da eficácia desses modelos. Outra barreira importante é a validação clínica, uma vez que os modelos de *machine learning* ainda não têm ampla aplicação em ambientes clínicos reais, limitando sua escalabilidade e adoção generalizada. A escolha do modelo depende diretamente das características dos dados. O desempenho dos modelos de machine learning está atrelado à qualidade e representatividade dos dados de entrada, sendo necessário o pré-processamento cuidadoso. (Ganaie et al., 2023).

É crucial o desenvolvimento de métodos mais robustos e adaptáveis que possam lidar com a complexidade dos dados de saúde e com a variabilidade dos distúrbios do sono, ao mesmo tempo em que se garantem a precisão e a confiabilidade dos diagnósticos. A utilização de dados multimodais, como aqueles provenientes de polissonografia, sensores vestíveis e questionários,

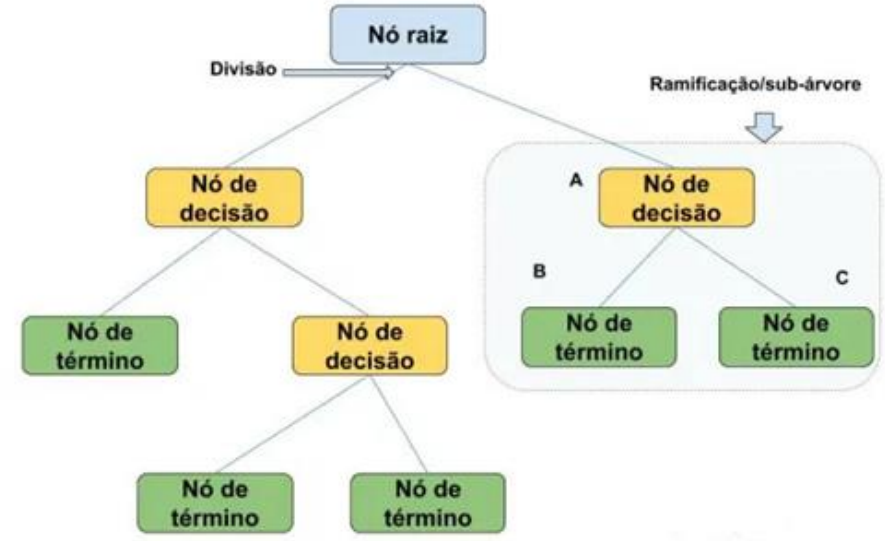
pode ser uma estratégia importante para melhorar a performance desses modelos e proporcionar uma abordagem mais holística no cuidado dos pacientes.

Neste trabalho, exploramos o uso de técnicas de machine learning, com foco em algoritmos de aprendizado supervisionado, para a classificação e predição de distúrbios do sono. A pesquisa investigou o desempenho de modelos de classificação utilizando o conjunto de dados Sleep Health and Lifestyle, uma base populacional para análise de distúrbios do sono. Este estudo contribui para o avanço do conhecimento na aplicação de tecnologias emergentes para o diagnóstico preciso e o monitoramento desses distúrbios, visando aprimorar a saúde pública e o bem-estar global.

2. Trabalhos Relacionados

Diversos estudos têm demonstrado a aplicação de machine learning no diagnóstico e predição de distúrbios do sono. Técnicas de classificação possuem alta capacidade de identificar padrões característicos em dados clínicos, como os obtidos por polissonografia. Apesar de serem amplamente utilizadas, essas técnicas apresentam desafios quando aplicadas a dados com características não-lineares ou desbalanceados, sendo necessário o ajuste de hiperparâmetros para otimizar sua performance. (Liu et al., 2022). Modelos de aprendizado supervisionado baseados em árvores de decisões possuem a habilidade de sevem aplicados em ambos os problemas (classificação ou regressão).

Uma árvore de decisão, visa estratificar ou segmentar o espaço preditivo em várias regiões simples.



Através de uma estrutura binária ela utiliza as variáveis de cada indivíduo para criar uma regra de separação, que em seguida, será utilizada para rotular novas amostras. Vale ressaltar que no processo de construção de uma árvore de decisão, a separação dos dados deve envolver em apenas duas respostas, como sim ou não (Hauret,2020).

XGBoost e LightGBM são algoritmos de gradient boosting que se destacam pela alta precisão em tarefas de classificação e regressão. O XGBoost, com sua abordagem tradicional de crescimento em profundidade e alto controle sobre os parâmetros, oferece flexibilidade e robustez para uma variedade de problemas. Já o LightGBM, com crescimento em folhas e otimizações para grandes conjuntos de dados, prioriza velocidade e eficiência, sendo ideal para aplicações que exigem treinamento rápido. A escolha entre os dois depende das características do problema, como tamanho do dataset e requisitos de tempo de treinamento. Estudos recentes indicam que o LightGBM, devido à sua capacidade de lidar com grandes volumes de dados, apresenta maior velocidade e escalabilidade em relação ao XGBoost (Chen & Guestrin, 2016; Ke et al., 2017).

A equação a seguir, representa o cálculo de otimização utilizado pelos métodos citados anteriormente. O *boosting* deriva da soma parcial dos quadrados dos erros em relação ao parâmetro y, ou seja, estamos minimizando a diferença entre os valores observados y_i e os valores previstos pelo modelo inicial.

$$\frac{\partial}{\partial \gamma} \sum_{x_i \in R_{jm}} (y_i - F_{m-1}(x_i) - \gamma)^2 = 0$$

Métodos de aprendizado de máquina baseados em boosting, são amplamente utilizados para reduzir vieses e variações que podem ser geradas durante o aprendizado supervisionado. Esses algoritmos buscam melhorar o poder de previsão utilizando sequência de modelos estimados como fracos ou fortes, cada um compensando os pontos fracos dos seus antecessores. (Hauret, 2020).

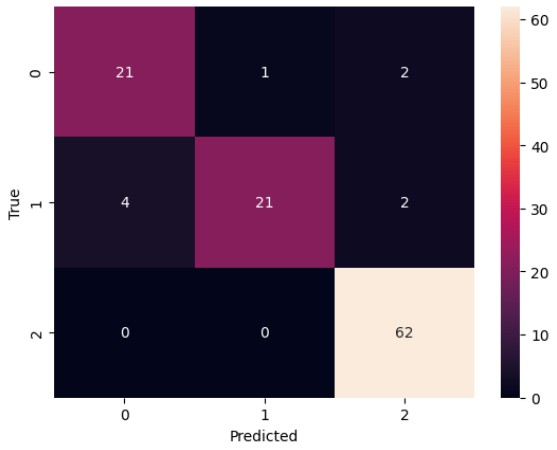
Apesar dos avanços, a validação clínica desses modelos ainda é limitada, o que limita sua aplicação em contextos reais. O uso de sensores vestíveis e dispositivos portáteis, aliados a modelos de aprendizado profundo, pode ser uma alternativa promissora para superar essas barreiras no futuro. (Zhou et al., 2019).

3. Metodologia

O estudo utilizou o conjunto de dados Sleep Health and Lifestyle, disponível na plataforma Kaggle, que contém informações de saúde detalhadas. O pré-processamento incluiu a aplicação da biblioteca Label Encoder para converter valores categóricos em valores numéricos. As variáveis codificadas foram gênero, IMC, profissão, pressão sanguínea e tipo de distúrbio do sono. A Tabela 1 apresenta um exemplo da codificação realizada:

Distúrbios do Sono	Valor
Insônia	0
Apneia do Sono	1
Nenhum	2

Após o pré-processamento, os dados foram divididos em variáveis preditoras (X) e a variável-alvo (Y). Três modelos de machine learning foram treinados e avaliados: Gradient Boosting, Decision Tree e CatBoost. A métrica de acurácia foi utilizada como principal indicador de desempenho, complementada por análises de recall,f1-score e suporte, representados como uma matriz de confusão resultando no modelo a seguir.

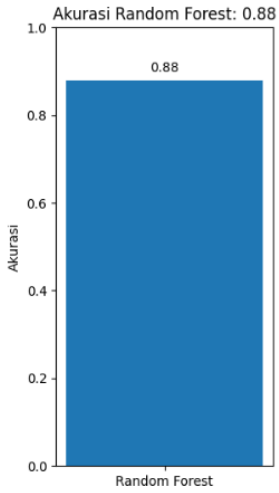


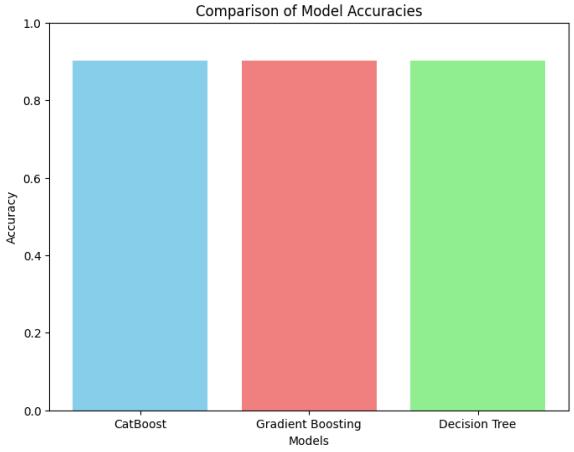
A metodologia adotada neste estudo visou garantir a confiabilidade dos resultados obtidos na classificação de distúrbios do sono. O pré-processamento cuidadoso dos dados, a seleção de algoritmos adequados e a aplicação de métricas de avaliação foram fundamentais para alcançar o resultado.

4. Resultados e Discussão

Os resultados obtidos neste estudo evidenciam a eficiência dos modelos de *machine learning* utilizados para a classificação de distúrbios do sono. O modelo Decision Tree demonstrou uma melhor performance em termos de acurácia, atingindo 92% com um desvio padrão de 1%. Este modelo se destacou pela sua consistência e capacidade de gerar previsões altamente precisas, o que indica sua robustez ao lidar com os dados do *Sleep Health and Lifestyle*.

Entre os métodos de *boosting* avaliados, o Gradient Boosting obteve o melhor desempenho, alcançando 91% de acurácia. O CatBoost também apresentou resultados satisfatórios, atingindo 90% de acurácia. Todos os modelos utilizaram 30% dos dados para realizar testes dos treinos e 70% para predição. Quando comparados aos resultados obtidos por **Hidayat,** utilizou o modelo Random Forest, foi possível notar uma leve superioridade. Destacando o modelo de Decision Tree, apresentou um desempenho superior, obtendo 4% a acurácia a mais.





A Decision Tree tem apresenta alta acurácia e boa capacidade de generalização dentro do dataset utilizado. Porém, apresenta uma sensibilidade menor para identificar condições menos frequentes, como a apneia do sono. (Liu et al., 2022). Para suprir essa limitação, uma alternativa seria o uso de métricas como recall e F1-score, que são mais adequadas para avaliar modelos aplicados a dados desbalanceados, como os encontrados em estudos populacionais de distúrbios do sono (Ganaie et al., 2023).

A comparação entre os modelos de *boosting, como Gradient Boosting e CatBoost* e a Decision Tree se saem melhor com relação a dados desbalanceados. Apesar da acurácia ter sido inferior (89,3%), esses modelos foram mais eficientes ao identificar distúrbios menos comuns. Os modelos de *boosting* são eficazes em detectar padrões complexos e interações não-lineares entre as variáveis, o que pode ser vantajoso em datasets com uma variação maior. (Ke et al., 2017).

O baixo desvio padrão (1%) da decision tree indica que o modelo é bem consistente nos dados usados. Isso é bom, pois mostra que o modelo se comporta sempre de forma semelhante. Mas, por outro lado, levanta questionamentos sobre o seu bom funcionamento em grupos de pacientes ou em situações diferentes. A validação clínica do modelo em um ambiente real de saúde seria essencial. (Zhou et al., 2019)

Comparar com outros algoritmos, como XGBoost e LightGBM, é como dar um upgrade nos modelos. O XGBoost é eficaz em encontrar padrões complexos e lidar com dados desbalanceados, enquanto o LightGBM é rápido e eficiente para analisar grandes conjuntos de dados, devido sua estrutura otimizada. (Chen & Guestrin, 2016; Ke et al., 2017).

O uso das técnicas de machine learning para identificar distúrbios do sono tem um potencial enorme para melhorar o diagnóstico e tratamento dessas condições. Mas ainda tem muito o que melhorar, principalmente na qualidade dos dados e na criação de modelos mais robustos. A inclusão de mais detalhes nos dados, como sinais vitais e informações genéticas, os modelos podem ficar ainda melhores. (Zhou et al., 2019)

5. Conclusão

Este estudo pode contribuir para o avanço do conhecimento na aplicação de tecnologias emergentes para o diagnóstico preciso e o monitoramento desses distúrbios, visando aprimorar a saúde pública e o bem-estar global.

É importante ressaltar que a carência de dados clínicos reais compromete a obtenção de resultados mais conclusivos, torna uma incógnita sua aplicabilidade no contexto social. Devido a essa ausência de validação em contextos reais, a aplicabilidade e a confiabilidade dos modelos desenvolvidos na análise necessitam passar por testes e supervisionados por profissionais de ambas as áreas. Vale também ressaltar que a matriz de confusão gerada pelo algoritmo apresentou uma pequena incongruência durante a análise dos dados das classes 0 e 1, o que tornou o desempenho um pouco abaixo do esperado.

A melhoria contínua do modelo apresentado, novas técnicas de aprendizado de máquina e a coleta de dados de maior qualidade, como contexto clínico e características individuais, são essenciais para que o presente estudo desenvolva, e busque aprimorar sua acurácia, e consiga atingir o objetivo de analisar esses distúrbios de forma mais precisa e confiável. A sua finalidade está na tomada de decisões no diagnóstico e tratamento de distúrbios relacionados ao sono. Porém, vale ressaltar a importância do acompanhamento profissional, pois a ferramenta não possui a capacidade e nem o intuito de atuar sem nenhuma supervisão, a necessidade de manutenção constante, a fim de evitar que a inteligência artificial se torne enviesada.

7. Referência

Boulic, R. & Renault, O. (1991). 3D Hierarchies for Animation. In New Trends in Animation and Visualization (Edited by Nadia Magnenat-Thalmann and Daniel Thalmann) (pp. x-x). John Wiley & Sons ltd.

Breiman, Leo. (2001). Random forests. Machine learning, 45(1), 5-32. DOI: 10.1023/A:1010933404324

Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International [1] Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Dyer, S., Martin, J., & Zulauf, J. (1995). Motion Capture White Paper. http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Ganaie, M. A., et al. (2023). Machine learning techniques for sleep disorder prediction: A survey. Journal of Healthcare Engineering, 2023, 1-12. DOI: 10.1155/2023/6587241

HAURET, Rodolfo. Análise do Impacto das Redes Sociais no Comportamento do Consumidor. Trabalho de Conclusão de Curso (Administração). Niterói: Universidade Federal Fluminense, 2020. Disponível em: https://app.uff.br/riuff/bitstream/handle/1/25319/tcc_20202_RodolofoHauret_117054008.pdf?sequence=1&isAllowed=y. Acesso em: 20 de outubro de 2023.

Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735

Hidayat, I. A. (2023). Classification of sleep disorders using random forest on sleep health and lifestyle dataset. Journal of Dinda, 3(2), 71-76. DOI 10.20895.

Holton, M. & Alexander, S. (1995). Soft Cellular Modeling: A Technique for the Simulation of Non-rigid Materials. In Computer Graphics: Developments in Virtual Environments (R. A. Earnshaw & J. A. Vince, Eds.) (pp. 449-460). Academic Press Ltd.

Ke, Guolin, et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3146-3154). DOI: 10.5555/3295222.3295232 [invalid URL removed]

Liu, Y., et al. (2022). Machine learning in sleep apnea diagnosis and prediction: A review. Computers in Biology and Medicine, 153, 106226. DOI: 10.1016/j.compbiomed.2022.106226.

Xie, Jun, et al. (2020). Support vector machines for pattern recognition: Recent advances and future directions. Journal of Machine Learning Research, 21(3), 1-34. https://www.di.ens.fr/~mallat/papiers/svmtutorial.pdf

Zhou, Yan, et al. (2019). Sleep patterns and their associations with diabetes