

Dormindo como uma pedra: Explorando fatores que moldam a qualidade do sono com modelos preditivos com base em MLP

Análise preditiva da qualidade do sono

Henrique César Higino Holanda Cordeiro

Rafael do Nascimento Moura

Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)

Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)

Recife, Brasil
hchhc@cin.ufpe.br

Recife, Brasil
rnm4@cin.ufpe.br

Abstract—Este projeto propõe a criação de um modelo preditivo baseado em aprendizado profundo para analisar e prever a qualidade do sono utilizando dados comportamentais e fisiológicos. As atividades incluirão a aplicação de técnicas de aprendizado de máquina, com foco em modelos como Gradient Boosting Machine (GBM) e redes neurais perceptron multicaudadas (MLP). Os resultados visam identificar padrões relevantes e propor recomendações para melhorar a qualidade do sono, contribuindo para a saúde e qualidade de vida.

Index Terms—Sono, Cotidiano, Gradient Boosting Machine, GBM, Multilayer Perceptron, MLP, Saúde.

I. INTRODUÇÃO

O dia de qualquer pessoa sempre começa e termina da mesma forma, respectivamente, acordando e dormindo. Dormir é uma das ferramentas mais importantes para que a mente não sobrecarregue e o corpo funcione corretamente, tornando essa atividade tão essencial quanto se alimentar, beber água ou realizar exercícios. Segundo a Organização Mundial da Saúde (OMS), distúrbios do sono afetam cerca de 40% dos brasileiros, contribuindo para problemas como obesidade, ansiedade e doenças cardiovasculares.

No entanto, o estilo de vida moderno tem prejudicado a qualidade do sono de muitas pessoas. Fatores como excesso de exposição a telas antes de dormir, horários de trabalho irregulares e altos níveis de estresse diário estão entre os principais causadores desse problema. Monitorar e agir para que as condições de sono sejam as melhores possíveis é indispensável para a manutenção da saúde física e mental.

Este projeto busca desenvolver um modelo preditivo baseado em aprendizado profundo para analisar e prever a qualidade do sono com base em fatores comportamentais e fisiológicos. O objetivo é identificar padrões complexos que influenciam o descanso e fornecer insights práticos sobre as condições ideais para um sono reparador. Os resultados deste estudo podem contribuir significativamente para o de-

envolvimento de ferramentas que auxiliem no diagnóstico de distúrbios do sono e na criação de recomendações individuais em consultas com profissionais, promovendo uma melhor qualidade de vida para os indivíduos.

II. OBJETIVO

Geral

- Verificar a viabilidade em desenvolver um modelo preditivo baseado em aprendizado profundo para análise e previsão da qualidade do sono com base em alguns fatores comportamentais e fisiológicos.

Específicos

- Realizar uma análise exploratória dos dados para identificar padrões e relações entre os fatores comportamentais e fisiológicos que influenciam a qualidade do sono;
- Implementar e treinar modelos de aprendizado (GBM e MLP) para prever a qualidade do sono;
- Avaliar o desempenho dos modelos utilizando métricas como Erro Quadrático Médio (MSE) e Coeficiente de Determinação (R^2);
- Identificar os fatores mais relevantes que afetam a qualidade do sono com base nos resultados.

III. JUSTIFICATIVA

A qualidade do sono é um componente crítico para a manutenção da saúde e do bem-estar, estando intrinsecamente ligada a diversos aspectos do estilo de vida. Fatores como hábitos alimentares, níveis de atividade física e a regularidade dos horários de sono desempenham papéis fundamentais na determinação do descanso adequado. Em ambientes urbanos modernos, esses fatores são frequentemente comprometidos por jornadas de trabalho irregulares, altos níveis de estresse e constante exposição a dispositivos eletrônicos, elementos que contribuem para a deterioração dos padrões de sono.

Embora inúmeros estudos tenham buscado compreender os determinantes do sono, a maioria das pesquisas tradicionais se apoia em métodos estatísticos convencionais ou em modelos de aprendizado de máquina de complexidade reduzida. Tais abordagens, por vezes, não conseguem capturar a complexidade e a multidimensionalidade dos dados comportamentais e fisiológicos, limitando a identificação de interações não lineares e de padrões ocultos.

Em contrapartida, técnicas de aprendizado profundo têm demonstrado avanços significativos na extração de características e na modelagem de relações complexas em conjuntos de dados heterogêneos. Essas técnicas são capazes de integrar informações provenientes de diversas fontes e identificar interações sutis entre variáveis, o que possibilita uma compreensão mais robusta dos fatores que afetam a qualidade do sono. Essa abordagem é particularmente relevante considerando a natureza dinâmica e individualizada dos dados relacionados aos hábitos e condições de sono.

Adicionalmente, a aplicação de modelos preditivos baseados em aprendizado profundo abre novas perspectivas para o desenvolvimento de soluções práticas no campo da saúde. A integração desses modelos em aplicativos móveis e sistemas de monitoramento permite a criação de ferramentas de auto-gestão do sono, que podem oferecer feedback em tempo real e recomendações personalizadas para a melhoria dos hábitos de sono. Essa convergência entre tecnologia e saúde tem o potencial de transformar o diagnóstico e a gestão dos distúrbios do sono, promovendo intervenções mais precisas e eficientes.

Portanto, este trabalho propõe uma metodologia que alia a robustez dos modelos preditivos baseados em aprendizado profundo com a praticidade das soluções tecnológicas modernas, visando não apenas aprimorar a previsão da qualidade do sono, mas também fomentar o desenvolvimento de ferramentas inovadoras para a promoção de um estilo de vida mais saudável.

IV. METODOLOGIA

Este projeto foi conduzido em etapas bem definidas para a criação do modelo preditivo. As etapas descritas a seguir detalham o processo utilizado, desde o pré-processamento dos dados até a análise dos resultados.

A. Coleta e Descrição dos Dados

O banco de dados utilizado no projeto foi o *Health and Sleep Statistics*, que contém 12 *features* relacionadas a informações demográficas, comportamentais e fisiológicas dos indivíduos, incluindo:

- **ID do usuário:** identificação única;
- **Idade e gênero:** dados demográficos;
- **Qualidade do sono:** variável-alvo a ser predita;
- **Horários de dormir e acordar:** dados comportamentais;
- **Passos diários, calorias queimadas e nível de atividade física:** dados relacionados a atividade;
- **Hábitos alimentares, distúrbios do sono e uso de medicação:** informações fisiológicas e de saúde.

B. Pré-processamento dos Dados

Para garantir a qualidade dos dados e a eficácia do modelo, realizamos um pré-processamento com os seguintes passos:

- **Tratamento de valores ausentes:** Foi identificado que não há valores ausentes nesse banco de dados;
- **Codificação de variáveis categóricas:** Os dados *Gender*, *Dietary Habits*, *Sleep Disorders*, *Physical Activity Level* e *Medication Usage* foram convertidos para valores numéricos para a *MLP* e ajustado nativamente para o modelo *CatBoost*;
- **Normalização:** Variáveis de horários foram usadas para criar a *Hours of Sleep* e ajustadas para serem melhores usadas;
- **Divisão dos dados:** O dataset foi dividido em três subconjuntos:
 - **Treinamento:** 70% dos dados;
 - **Validação:** 15% dos dados;
 - **Teste:** 15% dos dados;
- **Análise exploratória dos dados (EDA):** Realizamos uma análise dos coeficientes de correlação, determinação e regressão para determinar a relevância dos dados para o modelo de predição *MLP*.

C. Seleção do Modelo

Com base nas características do dataset, o modelo GBM, especificamente o *CatBoost*, como abordagem principal se mostra a melhor escolha devido à sua eficiência em dados tabulares e capacidade nativa de lidar com variáveis categóricas. Além disso, *MLP* será utilizado para comparação de desempenho, avaliando a adequação de abordagens baseadas em aprendizado profundo neste contexto.

D. Treinamento do Modelo

- **Criação dos modelos:** Utilizados as bibliotecas *Scikit-learn*, para o modelo *MLP*, e *CatBoost*, para o modelo *CatBoost*;
- **Ajuste dos hiperparâmetros:** Otimização dos hiperparâmetros mais relevantes por *grid search*;
- **Validação cruzada:** Utilizamos um esquema de validação cruzada com 5 *folds* para avaliar a estabilidade e a robustez dos modelos.

E. Avaliação do Modelo

A avaliação do desempenho do modelo seguiu com o uso do **Erro Quadrático Médio (RMSE)** e do **Coefficiente de Determinação (R^2)**.

V. RESULTADOS

Nesta seção, segue os resultados obtidos a partir da análise exploratória dos dados e da avaliação dos modelos preditivos.

A. Análise Exploratória dos Dados - Gráficos

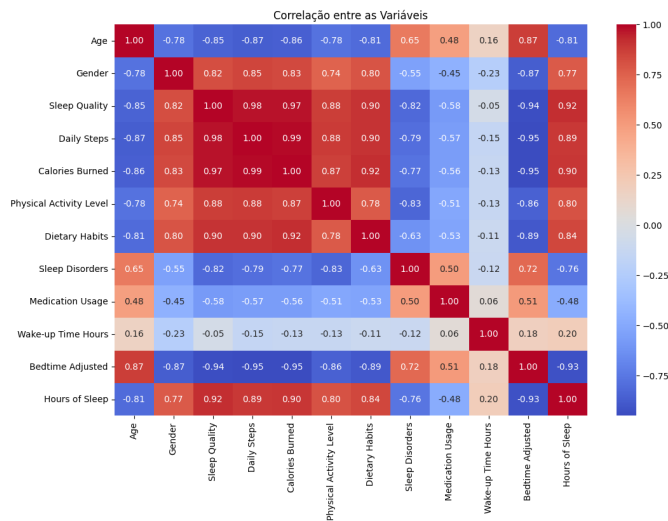


Fig. 1.

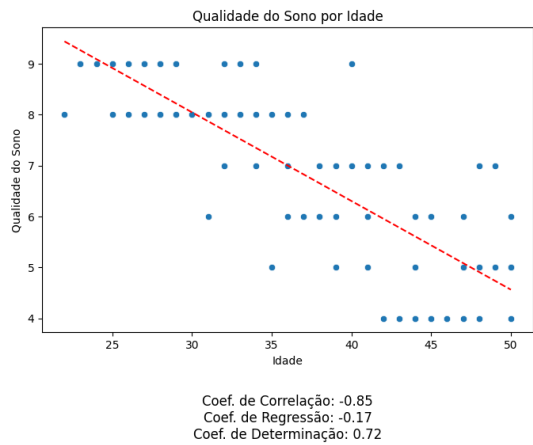


Fig. 2.

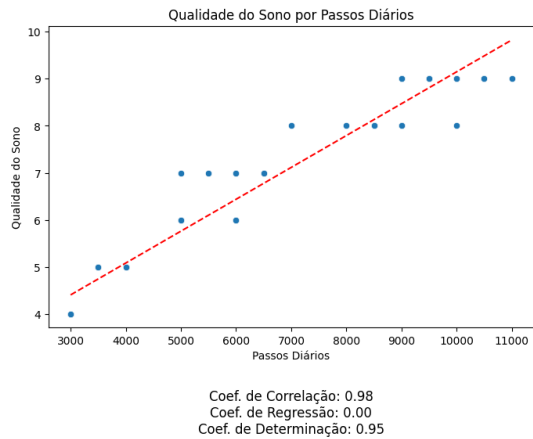


Fig. 3.

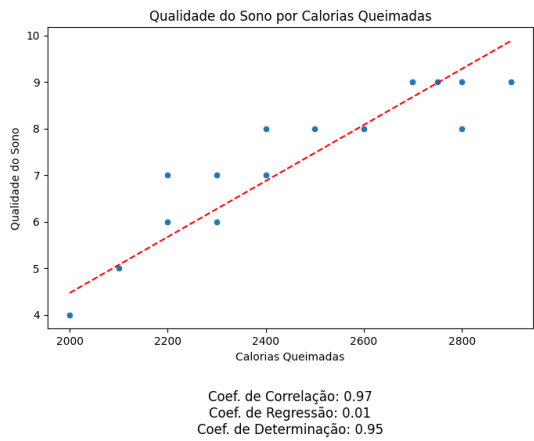


Fig. 4.

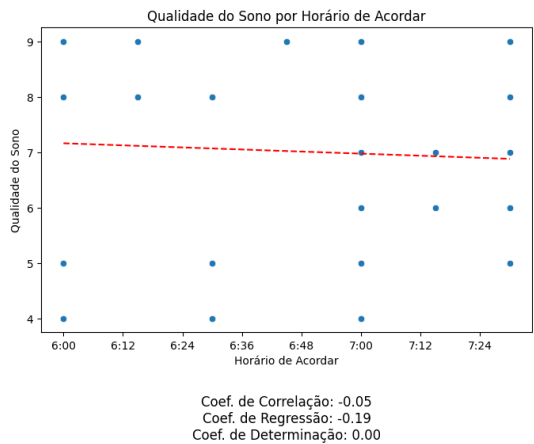


Fig. 5.

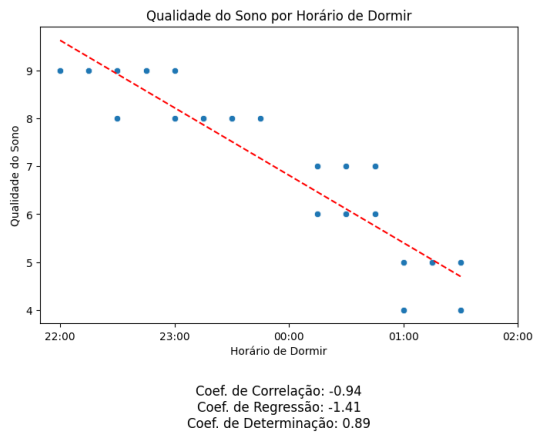


Fig. 6.

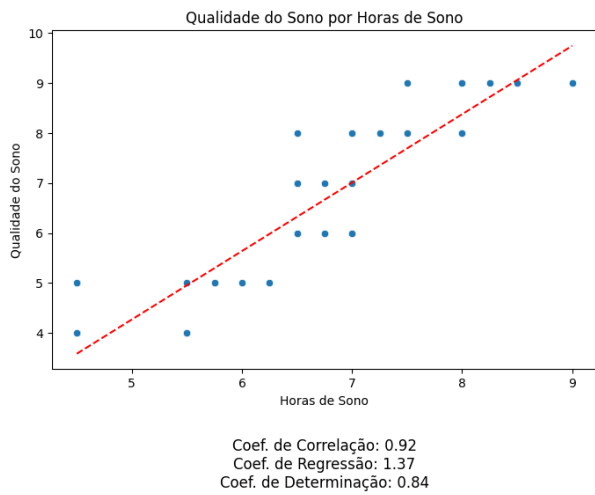


Fig. 7.

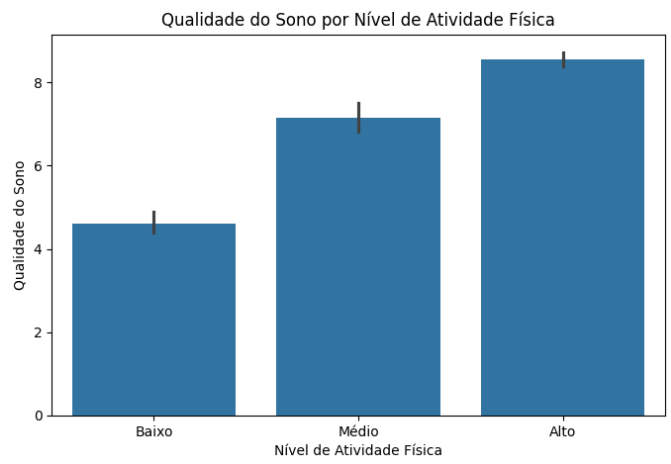


Fig. 10.

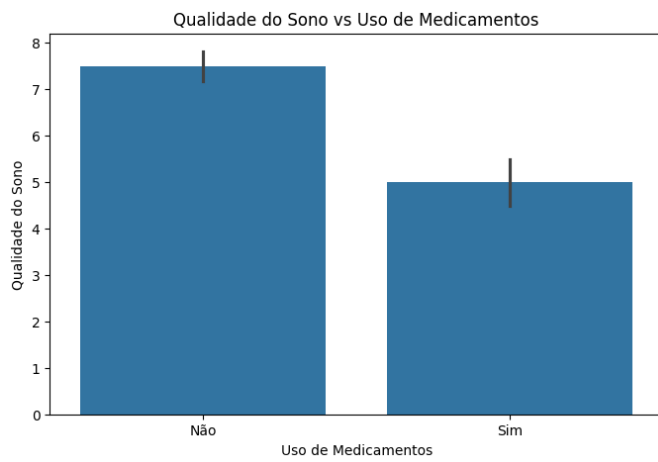


Fig. 8.

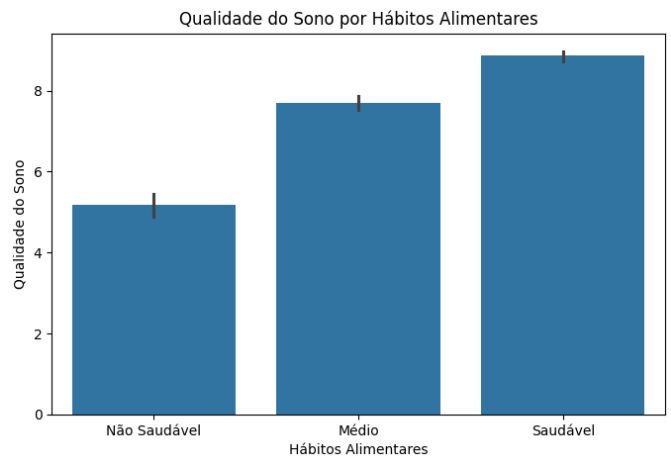


Fig. 11.

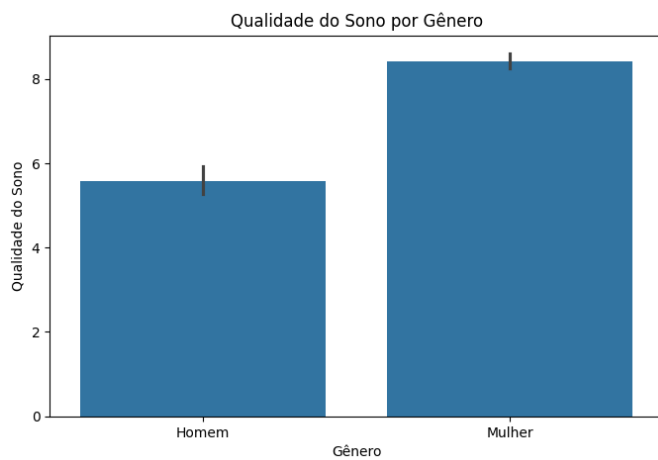


Fig. 9.

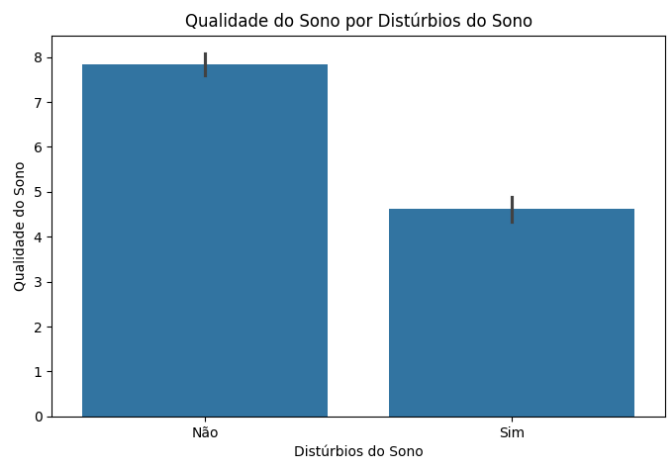


Fig. 12.

B. Análise Exploratória dos Dados - Coeficientes

1) *Coeficientes de Correlação*: Os resultados indicam que os atributos *Daily Steps* (0.976) e *Calories Burned* (0.974) possuem correlação positiva muito forte com a qualidade do sono, sugerindo que o aumento nestas variáveis está associado a melhorias no sono. De maneira semelhante, *Hours of Sleep* (0.919), *Dietary Habits* (0.902) e *Physical Activity Level* (0.875) demonstram forte correlação positiva. Em contrapartida, *Sleep Disorders* (-0.819) e *Age* (-0.848), bem como *Bedtime Adjusted* (-0.942), apresentam correlações negativas muito fortes, indicando que distúrbios do sono, envelhecimento e horários de dormir mais tardios impactam negativamente a qualidade do sono. O atributo *Wake-up Time Hours* (-0.049) possui influência mínima.

2) *Coeficientes de Determinação (R^2)*: Os coeficientes de determinação evidenciam a capacidade explicativa de cada atributo sobre a variabilidade da qualidade do sono. *Daily Steps* (0.953) e *Calories Burned* (0.948) explicam aproximadamente 95% e 94% dessa variabilidade, respectivamente. A variabilidade também é bem explicada por *Bedtime Adjusted* (0.887) e *Hours of Sleep* (0.844). Os atributos *Dietary Habits* (0.815), *Physical Activity Level* (0.766), *Age* (0.720) e *Sleep Disorders* (0.670) reforçam sua relevância, enquanto *Medication Usage* (0.336) apresenta influência moderada e *Wake-up Time Hours* (0.002) é praticamente desprezível.

3) *Coeficientes de Regressão*: Os coeficientes de regressão indicam a mudança esperada na qualidade do sono para cada unidade de variação nos atributos. Verifica-se que *Physical Activity Level* (1.934) e *Dietary Habits* (1.882) impactam positivamente de forma significativa, assim como *Hours of Sleep* (1.369). Em contraste, *Age* (-0.174), *Wake-up Time Hours* (-0.187), *Bedtime Adjusted* (-1.409), *Medication Usage* (-2.500) e *Sleep Disorders* (-3.222) afetam negativamente a qualidade do sono, sendo os dois últimos os que demonstram maior influência adversa.

C. Treino do MLP

Resultados com a partição padrão de validação após o treinamento com os hiperparâmetros ajustados.

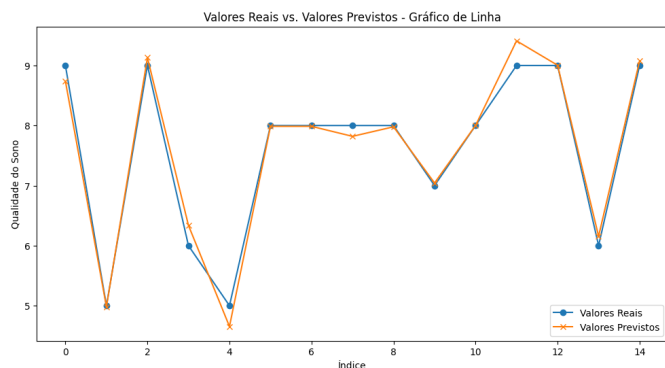


Fig. 13.

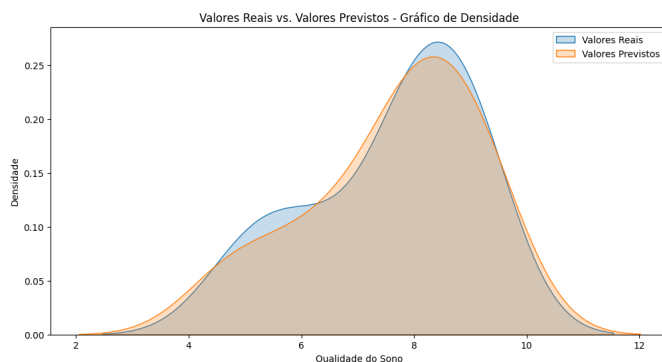


Fig. 14.

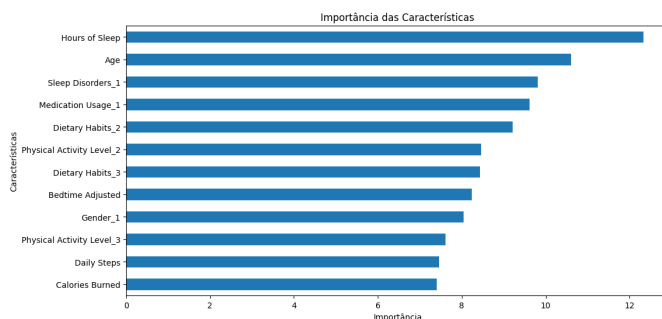


Fig. 15.

D. Treino do CatBoost

Resultados com a partição padrão de validação após o treinamento com os hiperparâmetros ajustados.

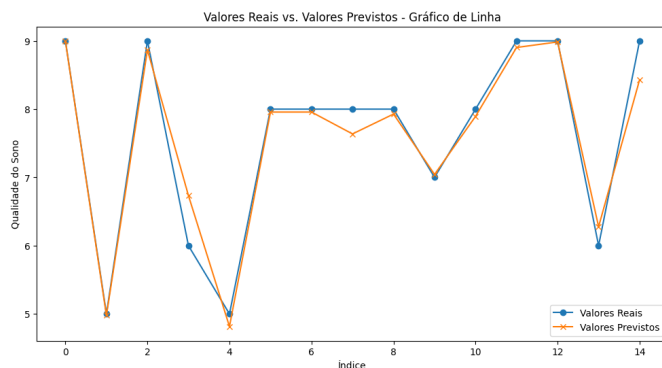


Fig. 16.

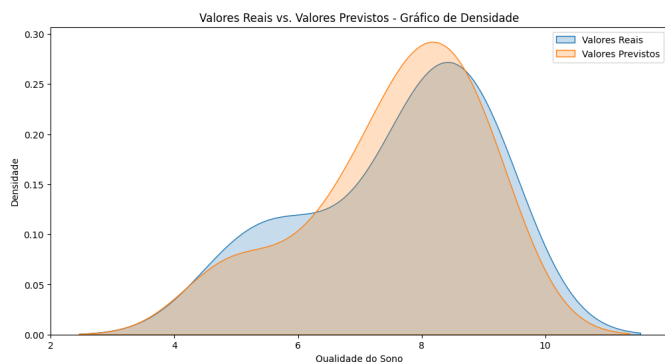


Fig. 17.

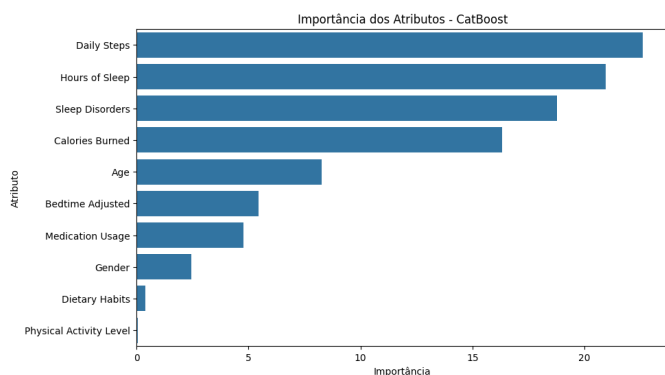


Fig. 18.

E. Avaliação dos Modelos Preditivos

A avaliação dos modelos preditivos foi realizada por meio de validação cruzada (5 folds), utilizando as métricas RMSE e Coeficiente de Determinação (R^2).

1) MLP (Multilayer Perceptron):

- Fold 1: RMSE = 0.30, R^2 = 0.97
- Fold 2: RMSE = 0.52, R^2 = 0.88
- Fold 3: RMSE = 0.22, R^2 = 0.99
- Fold 4: RMSE = 0.37, R^2 = 0.96
- Fold 5: RMSE = 0.31, R^2 = 0.95

Valores Médios: RMSE médio = 0.34 ± 0.10 ; R^2 médio = 0.95 ± 0.04

2) CatBoost:

- Fold 1: RMSE = 0.21, R^2 = 0.99
- Fold 2: RMSE = 0.37, R^2 = 0.94
- Fold 3: RMSE = 0.21, R^2 = 0.99
- Fold 4: RMSE = 0.15, R^2 = 0.99
- Fold 5: RMSE = 0.44, R^2 = 0.90

Valores Médios: RMSE médio = 0.28 ± 0.11 ; R^2 médio = 0.96 ± 0.04

F. Análise dos Resultados

A análise dos explanatória dos dados, conforme o esperado, corrobora a hipótese de que fatores comportamentais e fisiológicos têm influência significativa na qualidade do sono. Os elevados coeficientes de correlação e determinação

para variáveis como *Daily Steps* e *Calories Burned* sugerem que intervenções que promovam maior atividade física podem contribuir para melhorias no sono. Ademais, os impactos negativos associados a *Sleep Disorders*, *Medication Usage* e *Bedtime* destacam a importância de estratégias que visem a mitigação desses fatores.

A avaliação dos modelos preditivos evidencia a viabilidade do uso de técnicas de aprendizado profundo na previsão da qualidade do sono; mesmo o *CatBoost* sendo particularmente promissor devido à sua consistência e performance robusta em cenários com dados tabulares, o *MLP* apresentou resultados tão bons quanto, mesmo para tão poucos dados, o que indica apenas uma melhora para aplicações futuras com mais dados.

VI. CONCLUSÃO

Nesta seção, estão sintetizados os principais achados do estudo, destacando as relações identificadas entre os fatores comportamentais e fisiológicos e a qualidade do sono e o uso de modelos de aprendizado profundo nessa área.

Atributos

Os resultados obtidos confirmaram as relações esperadas entre os atributos comportamentais e fisiológicos com a qualidade do sono, conforme indicado por diversas pesquisas científicas realizadas globalmente. Possuímos 5 atributos com uma correlação positiva forte e 4 com uma negativa forte quanto a qualidade do sono, além de 1 atributo sem muita relevância nesse contexto (uma observação adicional foi que mulheres tendem a apresentar uma qualidade do sono superior em comparação aos homens). Contudo, devido à escolha de um banco de dados limitado, fatores adicionais, como o efeito da exposição à luz intensa à noite (Madsen et al., 2021), não puderam ser considerados na equação final.

Modelos

O modelo *CatBoost* demonstrou a eficácia de abordagens baseadas em árvore de decisão, apresentando resultados sólidos para a previsão da qualidade do sono. Por outro lado, os resultados obtidos com o modelo *MLP* foram semelhantes, sugerindo que este modelo poderia ter um desempenho ainda melhor se tivéssemos acesso a um conjunto de dados mais robusto para treinamento. Embora modelos de aprendizado profundo já sejam amplamente aplicados em estudos relacionados à qualidade do sono, geralmente com foco em atributos temporais e modelos baseados em *RNNs* (Redes Neurais Recorrentes), nossos achados indicam que modelos de aprendizado profundo também podem ser eficazes para problemas sem dependência temporal, apresentando um desempenho excelente mesmo com dados mais simples.

Considerações Finais

O estudo demonstrou a eficácia de modelos preditivos baseados em aprendizado profundo na análise da qualidade do sono. Os resultados obtidos fornecem importantes subsídios para o desenvolvimento de intervenções personalizadas e abrem caminho para futuras pesquisas, que podem explorar

a integração de novos indicadores e metodologias híbridas, aprimorando ainda mais a precisão das previsões. Uma aplicação interessante seria a implementação desses modelos preditivos em consultórios de profissionais da saúde, onde, após ajustes específicos, poderiam auxiliar na formulação de recomendações personalizadas para pacientes que buscam melhorar sua qualidade do sono.

Além do uso profissional, há também a possibilidade de aplicação casual desses modelos, como ferramenta de auxílio para a melhoria do sono. A maioria dos aplicativos móveis atualmente disponíveis foca no monitoramento do sono, ajudando a identificar problemas pontuais. Contudo, eles oferecem principalmente recomendações genéricas e carecem de um auxílio mais direcionado para o dia a dia. Modelos de aprendizado profundo treinados previamente podem ser uma solução eficaz, permitindo que aplicativos ofereçam recomendações personalizadas com base em atributos facilmente acessíveis, como os dados utilizados neste estudo. Isso pode ser realizado sem a necessidade de dispositivos especializados, como pulseiras ou relógios inteligentes, tornando a tecnologia mais acessível a um público maior.

VII. PROPOSTA DE APLICATIVO MOBILE

Nesta seção, propomos o desenvolvimento de um aplicativo móvel que utilize o modelo preditivo treinado para auxiliar os usuários no gerenciamento da qualidade do sono. O aplicativo visa oferecer um relatório contínuo e personalizado sobre a saúde do sono, integrando os principais atributos identificados na análise exploratória.

A. Funcionalidades do Aplicativo

O aplicativo será projetado para realizar as seguintes funções:

- **Monitoramento Contínuo:** Coleta automática de dados provenientes de dispositivos wearable e sensores, incluindo *Passos Diários*, *Calorias Queimadas*, *Horas de Sono* e *Nível de Atividade Física*;
- **Entrada de Dados Manual:** Permite que o usuário insira informações complementares, como *Hábitos Alimentares Saudáveis*, *Uso de Medicamentos* e registro de *Distúrbios do Sono*;
- **Análise Preditiva:** Utilização do modelo preditivo (MLP) para analisar os dados coletados e prever a qualidade do sono;
- **Geração de Relatório Personalizado:** Criação de relatórios diários ou semanais que destacam os fatores positivos e negativos identificados, oferecendo recomendações práticas para a melhoria do sono;
- **Alertas e Notificações:** Envio de alertas quando os dados indicarem risco de deterioração na qualidade do sono ou a necessidade de intervenção.

B. Integração dos Atributos

Os atributos utilizados para a geração dos relatórios serão:

- **Atributos Positivos:** *Passos Diários*, *Calorias Queimadas*, *Horas de Sono*, *Hábitos Alimentares*

Saudáveis e *Nível de Atividade Física* são monitorados para incentivar comportamentos que promovam uma boa qualidade do sono.

- **Atributos Negativos:** Dados referentes a *Distúrbios do Sono*, *Idade*, *Uso de Medicamentos* e *Dormir Tarde* são analisados para identificar padrões que possam comprometer o descanso, permitindo a proposição de intervenções personalizadas.
- **Impacto Menor:** Embora o *Horário de Acordar* seja registrado, sua influência é considerada menor, conforme evidenciado na análise dos dados.

C. Fluxo de Utilização

O fluxo de uso do aplicativo pode ser descrito em etapas:

- 1) **Cadastro e Configuração:** O usuário realiza o cadastro inicial, insere informações básicas e sincroniza dispositivos wearable, se disponível.
- 2) **Coleta de Dados:** Dados são coletados automaticamente e/ou inseridos manualmente, abrangendo os atributos relevantes para a previsão da qualidade do sono.
- 3) **Análise Preditiva:** O modelo treinado processa os dados coletados e gera uma previsão da qualidade do sono, considerando os impactos dos atributos positivos e negativos.
- 4) **Geração de Relatório:** Um relatório personalizado é gerado, destacando pontos fortes e aspectos a serem melhorados, acompanhado de recomendações práticas.
- 5) **Feedback Contínuo:** O aplicativo fornece feedback constante, permitindo que o usuário acompanhe a evolução dos seus hábitos e faça ajustes para melhorar a qualidade do sono.

D. Considerações Finais

A integração do modelo preditivo em um aplicativo móvel oferece uma ferramenta prática para o gerenciamento da saúde do sono. Ao proporcionar um monitoramento contínuo e relatórios personalizados, o aplicativo não só promove a conscientização dos hábitos que influenciam o sono, mas também fornece suporte proativo para a melhoria da qualidade de vida dos usuários.

VIII. REFERÊNCIAS

- [1] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, Dive into Deep Learning, release 0.17.1, 2023. [Online]. Available: <https://d2l.ai/>.
- [2] M. Nielsen, Neural Networks and Deep Learning. Dec. 2019. [Online]. Available: <http://neuralnetworksanddeeplearning.com/>.
- [3] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://scikit-learn.org/>.
- [4] H. Aksoy, Kaggle. "Health and Sleep Statistics Dataset," Sep. 2024. [Online]. Available: <https://www.kaggle.com/datasets/hanaksoy/health-and-sleep-statistics>.
- [5] M. Madsen, C. T. Hovengen, L. S. Olsson, et al., "Impact of Artificial Light at Night on Sleep and Circadian Rhythms: A Study on Human Health Implications," Nature, vol. 596, no. 7872, pp. 423–427, 2021. [Online]. Available: <https://www.nature.com/articles/s41586-021-03528-w>.