Introdução

Atualmente, mediante ao avanço tecnológico e à modernização da sociedade como um todo, os humanos têm enfrentado desafios de adaptação de seu estilo de vida; muitos desses desafios acabam por afetar a saúde humana, principalmente no que se refere à saúde do sono. Este trabalho de relatório tem como motivação mostrar essa relação entre a saúde do sono e o estilo de vida por meio de uma análise estatística de uma base de dados específica. A motivação deste trabalho reside na importância de entender como fatores influenciam a qualidade do sono. Além disso, serão apresentados dois modelos de Machine Learning que têm como propósito classificar e prever uma variável específica da base de dados com base em outras variáveis, oferecendo insights valiosos para a área da saúde.

Fundamentos Teóricos e Metodológicos

A análise realizada neste estudo baseou-se no conjunto de dados "Sleep Health and Lifestyle Dataset", composto por 13 variáveis e 374 observações. A análise foi conduzida por meio da linguagem de programação Python, utilizando técnicas de análise exploratória de dados para extrair insights iniciais. Posteriormente, foram usados os modelos de Árvore de Decisão e KNN para classificar a variável “Sleep Disorder”. A seguir, pode se ver as variáveis presentes na base de dados:

- **Person ID (ID pessoal, categórica)**: Um identificador pessoal e único;
- **Gender (Gênero, categórica):** O gênero da pessoa (Masculino/Feminino);
- **Age (Idade, numérica):** A idade da pessoa;
- **Occupation (Profissão, categórica):** A profissão da pessoa;
- **Sleep Duration (Duração do sono, numérica):** Número de horas dormidas por dia;
- **Quality of Sleep (Qualidade do sono, numérica):** Uma escala de 0-10 que mede a qualidade do sono;
- **Physical Activity Level (Nível de atividade física, numérica):** O número de minutos de atividade física por dia;
- **Stress Level (Nível de estresse, numérica):** Uma escala de 0-10 que mede o nível de estresse;

- **BMI Category (Categoria do IMC, categórica):** O Índice de Massa Corporal;
- **Blood Pressure (Pressão arterial, categórica):** Medição da pressão arterial do indivíduo em sistólica ou diastólica;
- **Heart Rate (Frequência cardíaca, numérica):** Frequência cardíaca da pessoa em repouso medida em batimentos por minuto;
- **Daily Steps (Passos diários, numérica):** Passos que a pessoa dar por dia;
- **Sleep Disorder (Distúrbio do sono, categórica):** Presença ou Ausência de algum distúrbio do sono.

Aplicação

Primeiro, na Análise Exploratória, serão apresentadas as informações estatísticas das variáveis numéricas na tabela abaixo:

Tabela 1: informações estatísticas das variáveis numéricas

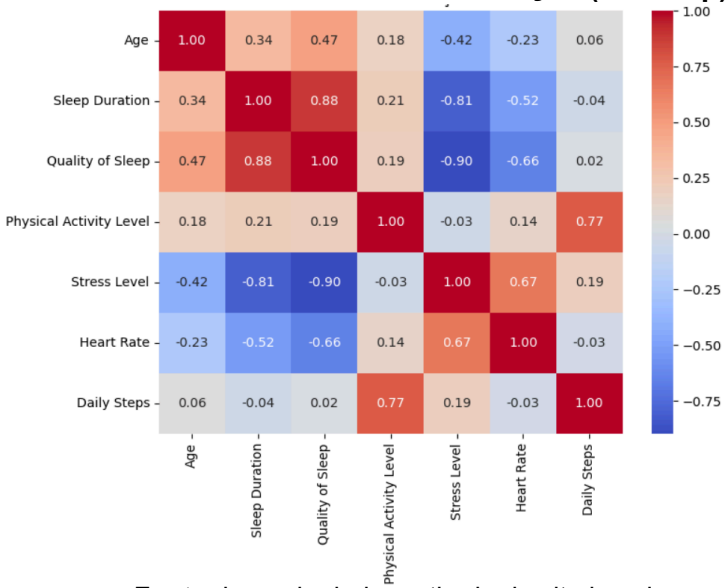
\\\\\\\\\\\\\\\\ \\\\\\\\\\\\\\\\ \\\\\\\\\\\\\\\\	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps
Média	42.18	7.13	7.31	59.17	5.38	70.16	6816.84
DP	8.67	0.79	1.19	20.83	1.77	4.13	1617.91
Min	27	5.8	4	30	3	65	3000
Q1	35.25	6.4	6	45	4	68	5600
Q2	43	7.2	7	60	5	70	7000
Q3	50	7.8	8	75	7	72	8000
Max	59	8.5	9	90	8	86	10000

Fonte: base de dados retirada do site kaggle

Observando a tabela 1, nota-se que as pessoas fazem, em média, cerca de 59 minutos de atividade física por dia, um número um pouco alto considerando o valor máximo obtido de 90 minutos por dia. Percebe-se, também, que, numa escala de 0-10, houve uma média de 7,31 na qualidade do sono, um valor bem alto considerando o máximo, que é 9. Além disso, o quartil 2 (mediana) do nível de estresse foi igual a 5, ou seja, metade das pessoas tem o nível de estresse menor que 5 e a outra metade tem o nível de estresse maior que 5.

Vamos observar a matriz de correlação das variáveis numéricas:

Gráfico 1: matriz de correlação (heatmap)

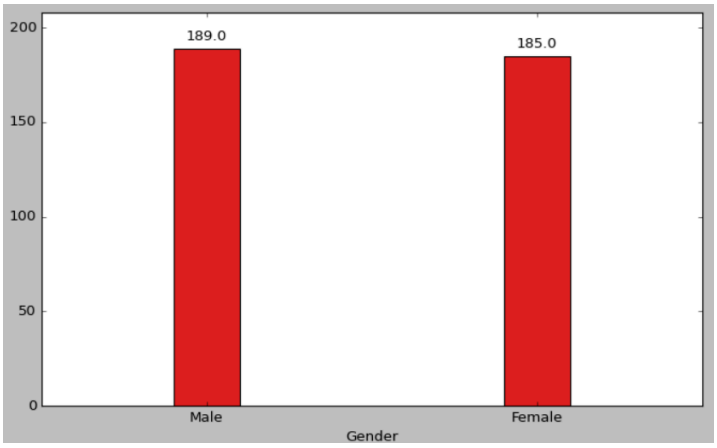


Fonte: base de dados retirada do site kaggle

Tem-se que a maior correlação positiva é entre a 'qualidade do sono' e a 'duração do sono', com o valor de 0,88, que é bem alto. Em contraste a isso, tem-se que a maior correlação negativa é entre a 'qualidade do sono' e o 'nível de estresse', com valor de -0,9, que é bem alto também. A diagonal principal foi desconsiderada, obviamente.

Vamos ver alguns gráficos de barras de algumas variáveis categóricas:

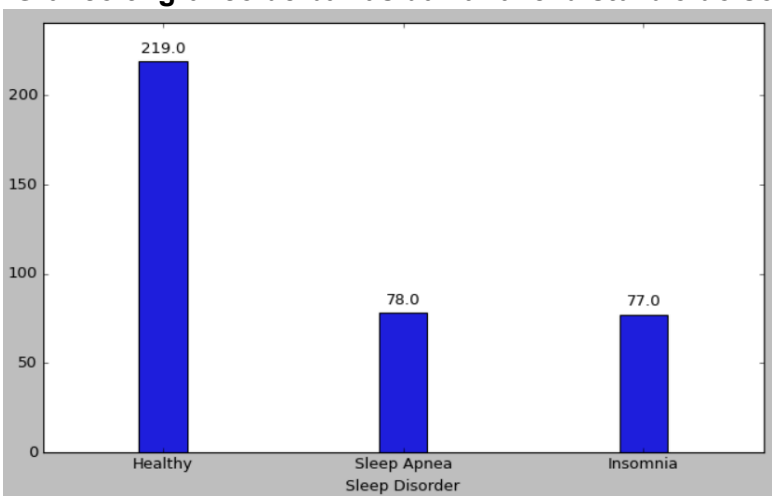
Gráfico 2: gráfico de barras da variável gênero



Analisando o gráfico 2, percebe-se que há só 4 homens a mais que mulheres, totalizando 189 homens e 185 mulheres das 374 pessoas estudadas.

Fonte: base de dados retirada do site kaggle

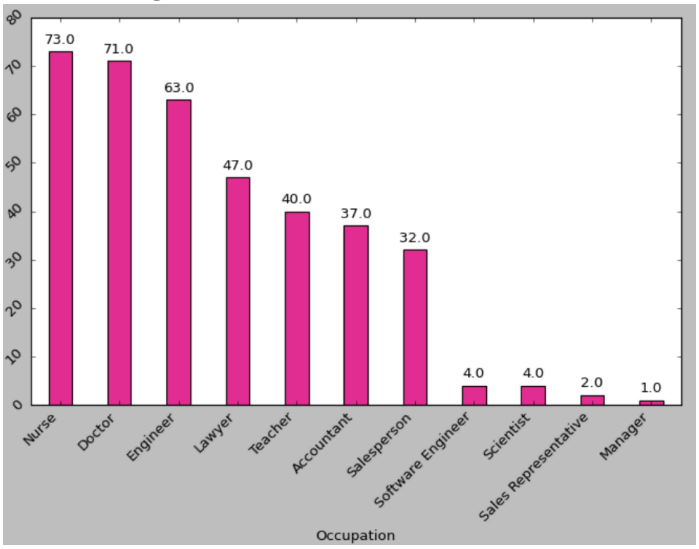
Gráfico 3: gráfico de barras da variável distúrbio do sono



Das 374 pessoas, apenas 155 sofrem com algum distúrbio, sendo 78 de apneia do sono e 77 de insônia.

Fonte: base de dados retirada do site kaggle

Gráfico 4: gráfico de barras da variável profissão



Observe que as 3 profissões com o maior número de profissionais são enfermeiro, doutor e engenheiro com 73, 71 e 63 pessoas. Em contraste disso, as profissões com o menor número de profissionais são cientista, representante de vendas e gerente com 4, 2 e 1 pessoas respectivamente.

Fonte: base de dados retirada do site kaggle

Agora, será vista a aplicação dos modelos de Machine Learning para classificar a variável “Sleep Disorder” (para ambos os modelos, serão apresentadas as métricas, o gráfico de curva de aprendizado e a matriz de confusão):

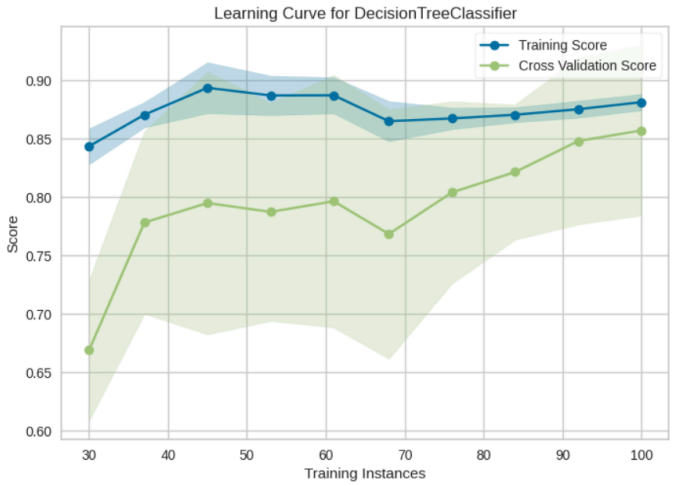
Árvore de Decisão:

Tabela 2: Métricas da A. D.

Acurácia	AUC	Sensibilidade	Precisão	F1	Kappa	MCC
0.8568	0.8788	0.8568	0.8920	0.8560	0.7861	0.8025

Fonte: base de dados retirada do site kaggle

Gráfico 5: curva de aprendizado A. D



Fonte: base de dados retirada do site kaggle

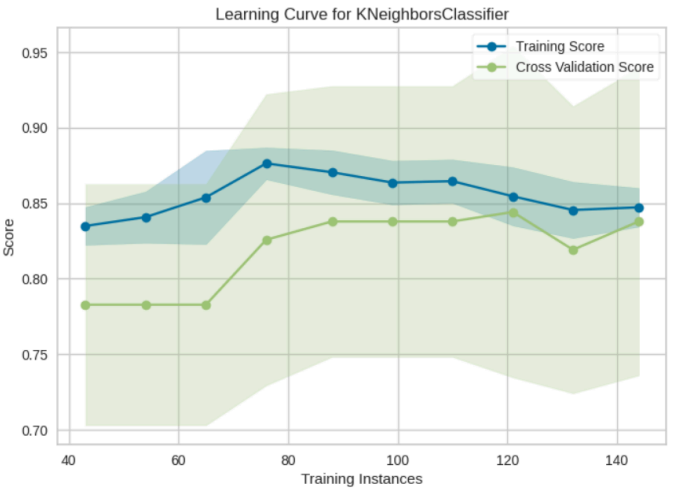
KNN:

Tabela 3: Métricas do KNN

Acurácia	AUC	Sensibilidade	Precisão	F1	Kappa	MCC
0.8379	0.9016	0.8379	0.8602	0.8328	0.7553	0.7672

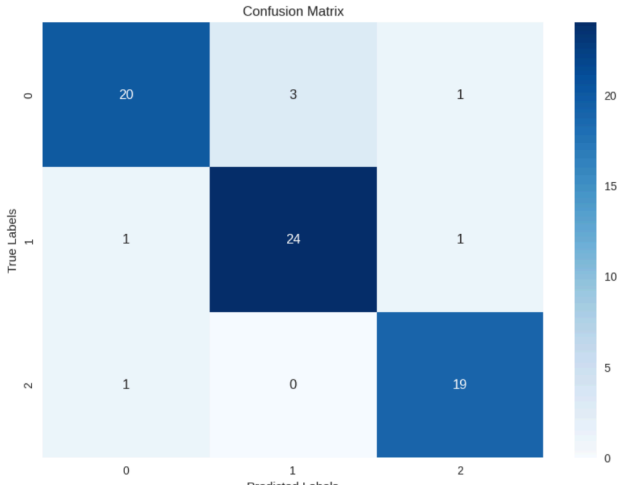
Fonte: base de dados retirada do site kaggle

Gráfico 6: curva de aprendizado KNN



Fonte: base de dados retirada do site kaggle

Gráfico 7: matriz de confusão da A. D.



Accuracy: 0.9

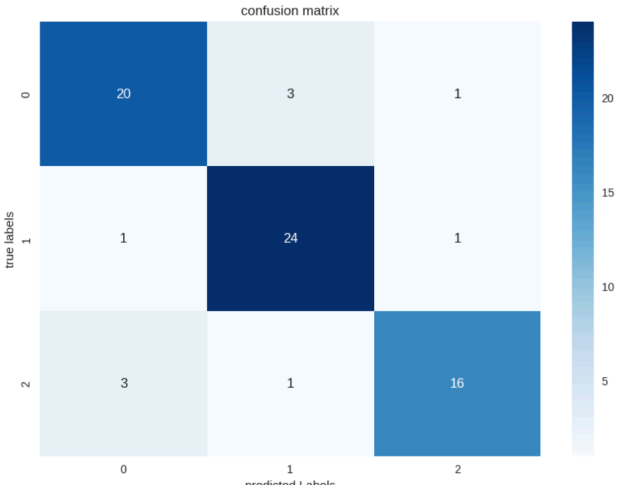
Kappa: 0.8492307692307692

Classification Report:

	precision	recall	f1-score	support
Insomnia	0.91	0.83	0.87	24
Healthy	0.89	0.92	0.91	26
Sleep Apnea	0.90	0.95	0.93	20
accuracy			0.90	70
macro avg	0.90	0.90	0.90	70
weighted avg	0.90	0.90	0.90	70

Fonte: base de dados retirada do site kaggle

Gráfico 8: matriz de confusão do KNN



Accuracy: 0.8571428571428571

Kappa: 0.7836835599505563

Classification Report:

	precision	recall	f1-score	support
Insomnia	0.83	0.83	0.83	24
Healthy	0.86	0.92	0.89	26
Sleep Apnea	0.89	0.80	0.84	20
accuracy			0.86	70
macro avg	0.86	0.85	0.85	70
weighted avg	0.86	0.86	0.86	70

Fonte: base de dados retirada do site kaggle

Analisando, primeiramente, as métricas de ambos os modelos, percebe-se que não houve tanta diferença entre os dois. A acurácia, a sensibilidade e a precisão da Árvore de Decisão foram, respectivamente, de 0.8568, 0.8568 e 0.8920 e a acurácia, a sensibilidade e a precisão do KNN foram, respectivamente, de 0.8379, 0.8379 e 0.8602. Nesse quesito, a Árvore mostrou-se ser o melhor modelo dos dois por uma ínfima diferença de valores nas métricas. Além disso, ao analisar os gráficos de curva de aprendizado de ambos os modelos, nota-se que a diferença entre os valores associados ao treinamento (training score) e ao teste (cross validation score) na Árvore de Decisão foi um pouco maior que no KNN, o que mostra que o KNN, de acordo com os gráficos, teve maior eficiência nesse quesito. E, por fim, ao analisar os resultados da matriz de confusão de ambos os modelos, a Árvore de Decisão mostrou-se mais promissora do que o KNN, percebe-se que os

valores (0 - Sleep Apnea; 1 - Insomnia; 2 - Nenhum (Healthy)) foram iguais em ambas as matrizes para a apneia do sono e para a insônia, porém a Árvore acertou mais valores para os resultados ‘2 - saudáveis’, sendo mais precisa que o KNN.

Conclusão

Com as informações apresentadas anteriormente, o relatório visou mostrar a relação entre o estilo de vida e a qualidade da saúde do sono por meio de uma análise exploratória de dados, onde foram investigadas diversas variáveis como nível de atividade física, qualidade do sono, nível de estresse, duração do sono, distúrbio do sono, etc. Esta análise trouxe uma visão geral sobre algumas influências nessa relação de estilo de vida e a qualidade da saúde do sono, com destaque na matriz de correlação, onde isso é mais visivelmente identificável. Além disso, foram apresentados dois modelos de Machine Learning com o objetivo de classificar a variável ‘Sleep Disorder’, Distúrbio do Sono em português, tais modelos foram a Árvore de Decisão e o KNN, onde a Árvore apresentou métricas e valores mais assertivos que o KNN, sendo a mesma mais preferível de escolha entre ambos se o objetivo é ter um modelo com o melhor desempenho imediato no conjunto de dados.

Contribuições da equipe

Nomes e contribuições:

- **Pedro Lucas de Oliveira Marques:** Fez o relatório e escreveu os programas da Análise Exploratória de dados (33,33% de contribuição);
- **Jorge Antônio Silva Santos**: Escreveu os programas dos modelos de Machine Learning (33,33% de contribuição);
- **George Willians da Silva:** Fez o slide de apresentação (33,33% de contribuição).

Referências:

https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
(Link da base de dados usada no estudo).