Uso de Algoritmos de Aprendizado de Máquina para Previsão de Distúrbios do Sono Baseado em Biométricas Amplamente Disponíveis

André Souza Santos (123456), Gabrielly Castilho Guimarães (805757), Miguel Antônio de Oliveira (772180), Natália Bachiega Magalhães (769846)

Departamento de Computação – Universidade Federal de São Carlos (UFSCar) São Carlos – SP – Brasil

{andresouza,gabriellycastilho,miguel.oliveira,nataliabachiega}@estudante.ufscar.br

Abstract. *The following work aims to perform a brief analysis of the Sleep Health and Lifestyle dataset for understanding the subject and furthermore, to develop a model capable of predicting which characteristics, among those available in the mentioned set, relate to the predisposition to sleep disorders in an individual, more precisely insomnia and apnea. The results of this study may contribute to a better understanding of sleep and provide an effective means for detecting these disorders.*

Resumo. *O trabalho a seguir tem como objetivo realizar uma breve análise do conjunto de dados "Sleep Health and Lifestyle" para a compreensão do assunto e para além disso, desenvolver um modelo capaz de prever quais características, dentre as disponíveis no conjunto aqui citado, se relacionam à predisposição à distúrbios do sono em um indivíduo, mais precisamente insônia e apneia. Os resultados deste estudo podem contribuir para um melhor entendimento do sono e fornecer um meio eficaz para a detecção destes distúrbios.*

1 Introdução

A qualidade do sono é um fator fundamental para a saúde humana. Dormir mal se relaciona fortemente com um amplo rol de doenças e problemas de saúde. Como parte dos esforços para compreender os fatores que influenciam a qualidade do sono, este artigo desenvolve modelos preditivos, utilizando técnicas do Aprendizado de Máquina, para classificar dados que possam indicar a presença de distúrbios do sono em pacientes. Tais dados podem ser obtidos a partir de aferências simples do próprio usuário, em conjunto com dados coletados a partir de tecnologias vestíveis, como relógios ou pulseiras inteligentes. Os resultados obtidos neste artigo podem colaborar com uma melhor compreensão do sono, bem como a diminuição do subdiagnóstico dos distúrbios aqui tratados.

2 Objetivos

Este trabalho objetiva o desenvolvimento de um modelo baseado em classificadores de aprendizado supervisionado para classificar se determinado conjunto de métricas sobre o sono pode ou não indicar a existência de distúrbios, a saber: apneia do sono e insônia.

3 Materiais e Métodos

O conjunto de dados utilizado para o desenvolvimento deste trabalho foi obtido na plataforma Kaggle, um repositório online que disponibiliza dados para uso em estudos de aprendizado de máquina. A fim de alcançar os objetivos aqui propostos, implementou-se classificadores baseados nos algoritmos de classificação kNN (*K*-nearest neighbors), algoritmo de árvore de decisão e Naive Bayes Gaussiano. Para esta finalidade, utilizamos os pacotes Pandas e Sklearn com a linguagem Python.

4 Análise e Pré-Processamento de Dados

4.1 Pré-Processamento dos Dados

Os dados obtidos passaram pelo processo de pré-processamento para garantir a qualidade e a confiabilidade antes da modelagem. Esse processo envolveu a verificação de dados faltantes, a remoção de outliers e a normalização das características para que todas estivessem na mesma escala.

Para uma primeira análise, foi utilizado a função `info()` provida pelo Pandas para observar a quantidade de classes e atributos e o tipo de cada um:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           374 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

Figura 1 – Resultado do comando info()

Observa-se a presença de atributos numéricos, dos quais discretos e contínuos, e outros do tipo *object*, nominais e ordinais. Disso, elenca-se a necessidade de tratamento e normalização. Além disso, constata-se a presença de 13 atributos sobre 374 amostras.

Para garantir a ausência de dados faltantes, contamos as instâncias que possuem dados nulos em algum de seus atributos, constatando por fim a devida inexistência de qualquer dado nessa categoria. Com isso, tratamentos acerca desta questão não são necessários.

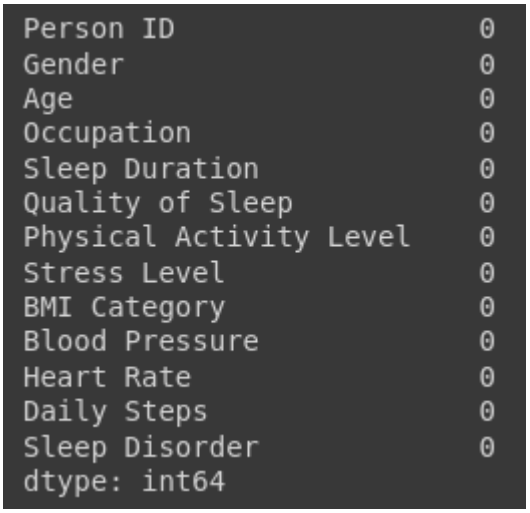


Figura 2 – Contagem de instâncias com dados faltantes.

Visando transformar os dados em entradas aceitáveis para os algoritmos, foram realizadas algumas alterações nos dados, dispostas abaixo.

- **Identificação da Pessoa:** O identificador foi omitido, dado que ele não apresenta informações úteis para inferência da classe.
- **Gênero:** O conjunto apresentou dois gêneros: "masculino" e "feminino". Mapeamos esses para 0 e 1 respectivamente, dado que o algoritmo somente aceita valores numéricos.
- **Idade:** Mapeamos as idades usando um dimensionamento *min-max*, para manter uma escala padronizada e evitar destaques excessivos dados para o atributo.
- **Ocupação:** A ocupação também foi omitida, pois não foi constatada influência significativa na qualidade de sono do sujeito. Além disso, seria um problema mapeá-la para um valor numérico sem introduzir vieses nos resultados.
- **Duração do Sono:** A duração do sono, inicialmente entre 0 e 24 horas, foi redimensionada por um fator de 24.
- **Qualidade do Sono:** Fizemos um dimensionamento *min-max*, baseado nos extremos presentes no conjunto.
- **Nível de atividade física:** Fizemos um dimensionamento *min-max*, baseado nos extremos presentes no conjunto.
- **Nível de Estresse:** Fizemos um dimensionamento *min-max*, baseado nos extremos presentes no conjunto.
- **Índice de Massa Corporal:** O conjunto divide o Índice de Massa Corporal (IMC) em categorias, "*Normal Weight*", "*Overweight*" e "*Obese*". Mapeamos cada

categoria usando o valor médio dos intervalos apresentados em (Gadzik, 2006). Os valores mapeados são listados na Tabela 1.

Classe	Mínimo em (Gadzik, 2006)	Máximo em (Gadzik, 2006)	Média Mapeada
Normal	0,74	0,99	0,865
Sobrepeso	1,00	1,19	1,095
Obeso	1,20	1,39	1,295

Tabela 1 – Mapeamento de categorias Índice de Massa Corporal

- **Pressão Arterial:** O conjunto apresenta as medidas de pressão como uma *string*, por exemplo: "125/80". Os números são a pressão sistólica e diastólica, respectivamente, em milímetros de mercúrio (mmHg). Fizemos um dimensionamento *min-max* usando mínimos e máximos teóricos, já que isso deveria aproximar o dimensionamento correto para conjuntos muito grandes. Definimos o mínimo da pressão sistólica (resp. diastólica) como 120 mmHg (resp. 80 mmHg) e o máximo como 180 mmHg (resp. 110 mmHg).
- **Frequência Cardíaca:** Também fizemos dimensionamento *min-max* com uma aproximação de patamares teóricos. Definimos a frequência máxima usando a fórmula $220 - \text{idade}$ BPM (Robergs, et al., 2002) e a mínima como 60 bpm (Spodick, 1993).
- **Contagem de passos diários:** Fizemos um dimensionamento *min-max*, baseado nos extremos presentes no conjunto.

4.2 Análise da Correlação

Após a etapa de pré-processamento, foi decidido pela análise de correlação entre os atributos presentes. Para tal finalidade, foi gerado uma matriz de correlação com o *dataset* normalizado.

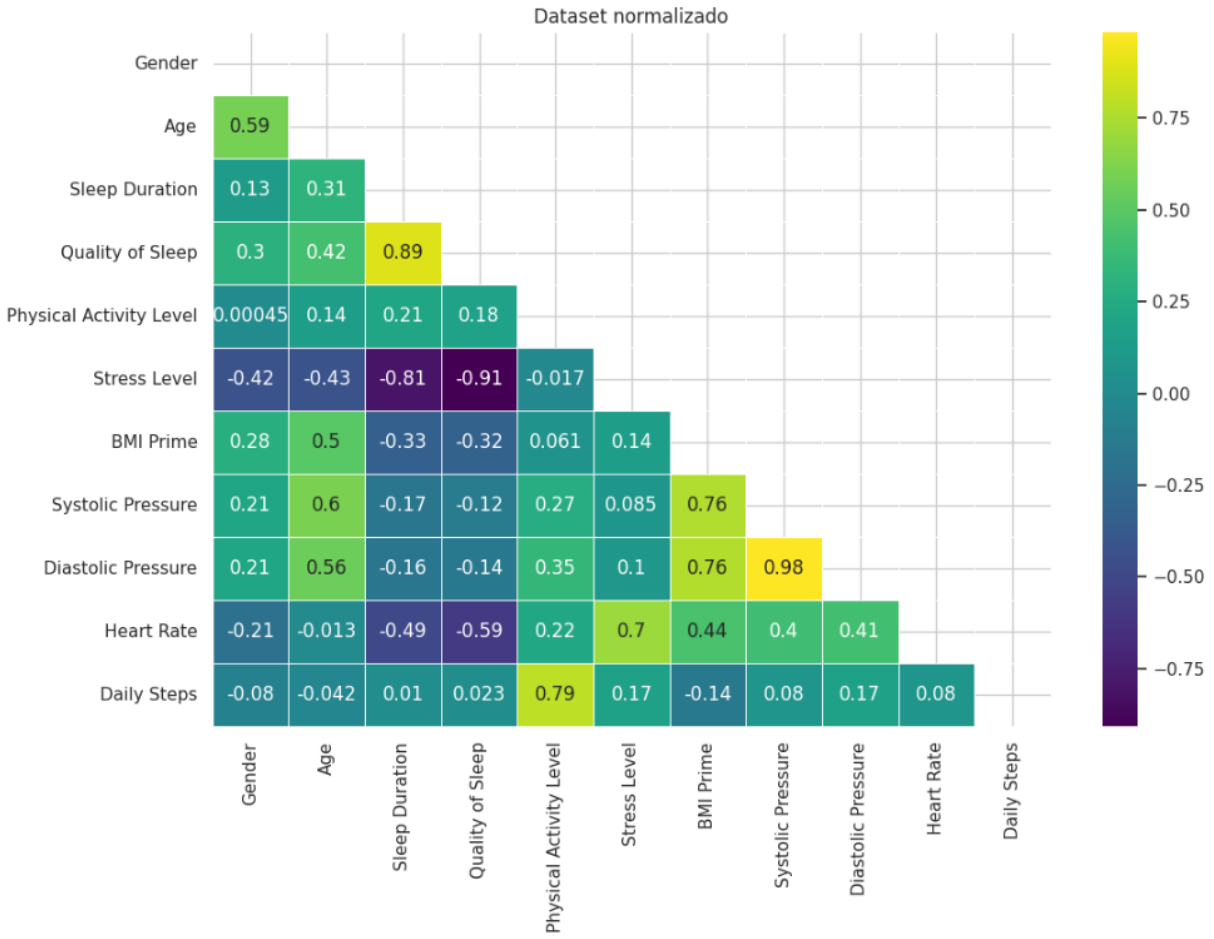


Figura 3 – *Heat Map* de correlação

É possível observar a partir da figura acima que alguns atributos possuem baixa correlação, como *"Daily Steps"*, *"Age"*, *"Physical Activity Level"*, *"Diastolic Pressure"* e "*Gender*".

Para resolver o problema de baixa correlação explicitada acima, aplicou-se a técnica de seleção de atributos e, para o possível viés de seleção, foi utilizada a técnica de amostragem estratificada e o teste será feito com os classificadores com cada dataset/amostra.

4.3 Amostragem estratificada

Para gerar a amostra 100 objetos aleatórios foram selecionados, porém ainda mantendo a proporção original do dataset:

```
None           58
Insomnia       21
Sleep Apnea    21
Name: Sleep Disorder, dtype: int64
```

Observa-se que a correlação da amostra é um pouco maior que a do *dataset*.

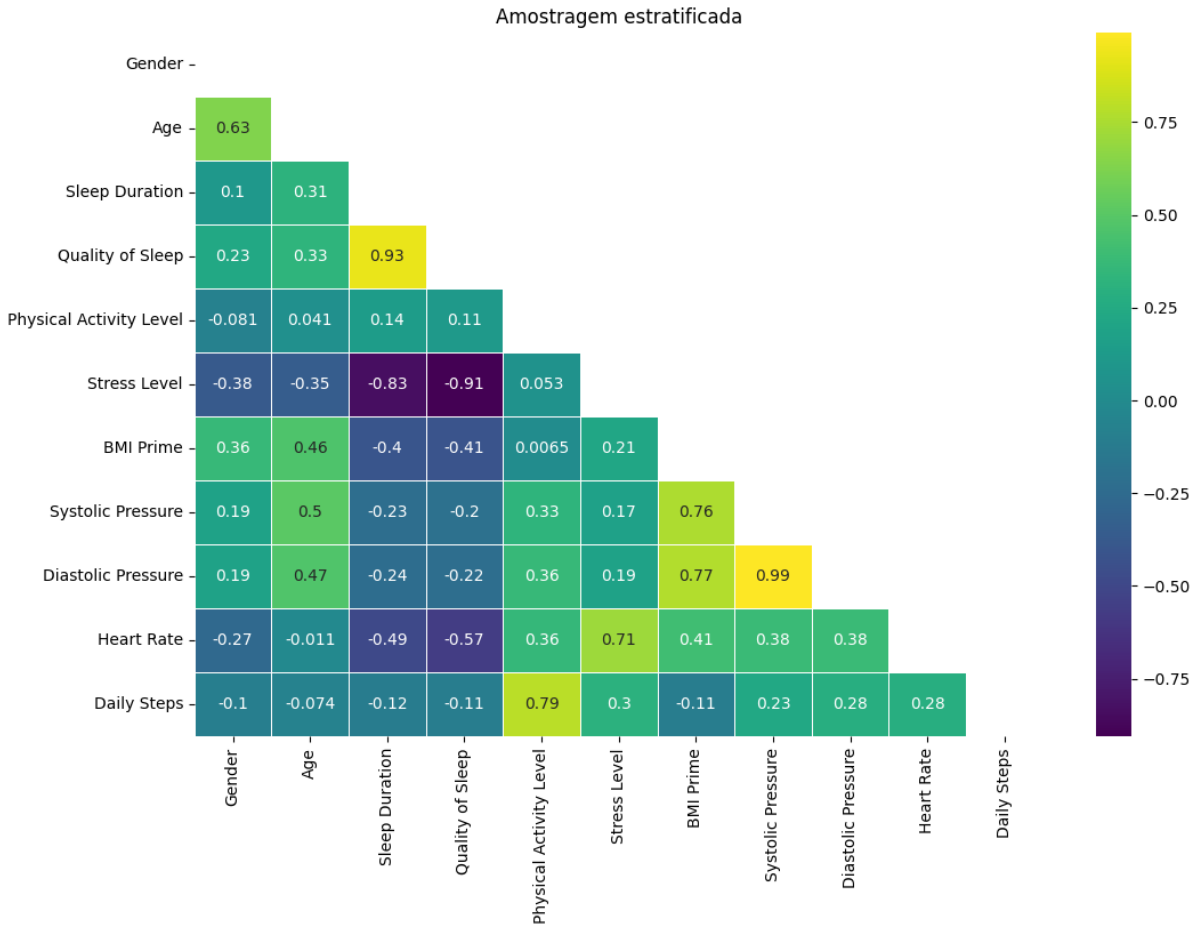


Figura 4 – *Heat Map* da correlação da amostra estratificada.

4.4 Seleção de atributos

Na etapa de seleção de atributos foram excluídos aqueles atributos com menor correlação, são eles: *"Daily Steps"*, *"Age"*, *"Physical Activity Level"*, *"Diastolic Pressure"* e *"Gender"*. Com isso gerou-se um dataset que conta com 7 atributos em vez das 11 anteriores.

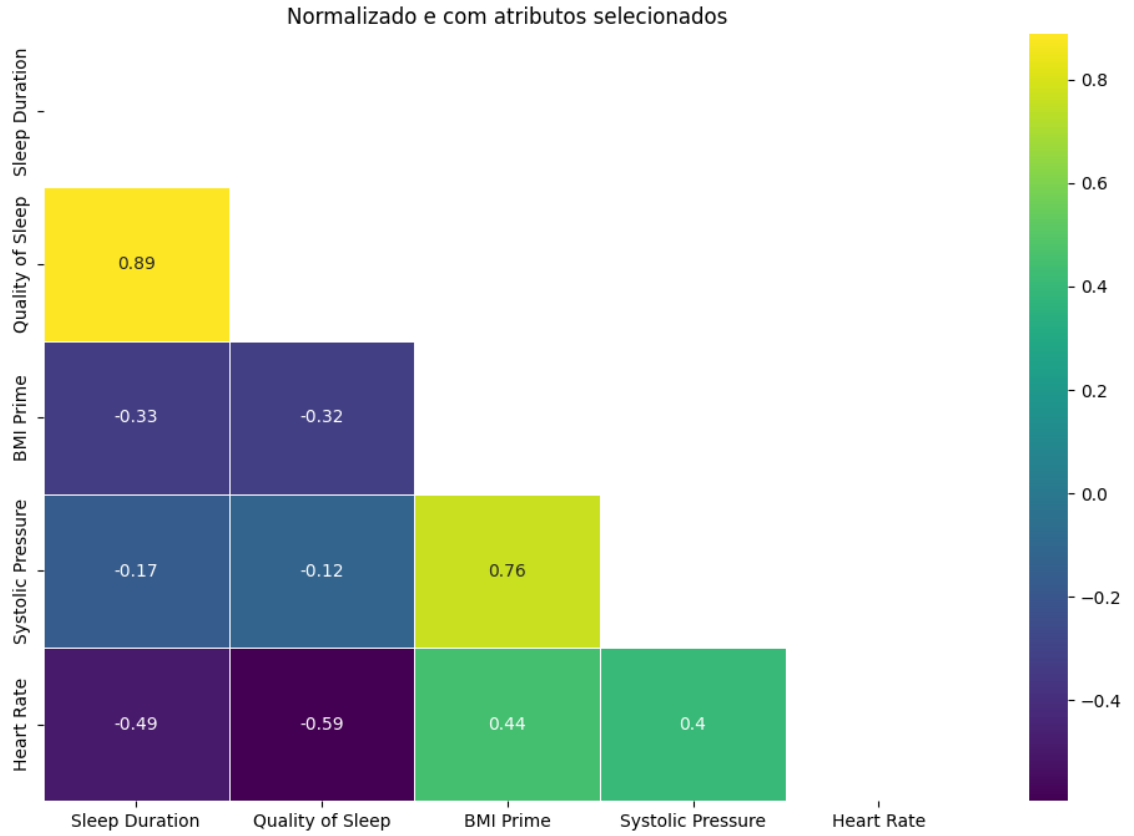


Figura 5 – *Heat Map* após a seleção de atributos.

5 Seleção e Implementação dos Modelos de Aprendizado de Máquina

Para a análise dos dados, foram considerados os modelos de aprendizado de máquina baseados nos seguintes algoritmos: kNN (k-nearest neighbors), Árvores de Decisão e Naive Bayes Gaussiano. Cada modelo foi treinado usando uma parte do conjunto de dados (conjunto de treinamento) e validado com outra parte (conjunto de validação). A seleção do modelo foi realizada com base na capacidade de generalização, que foi medida através da acurácia no conjunto de validação.

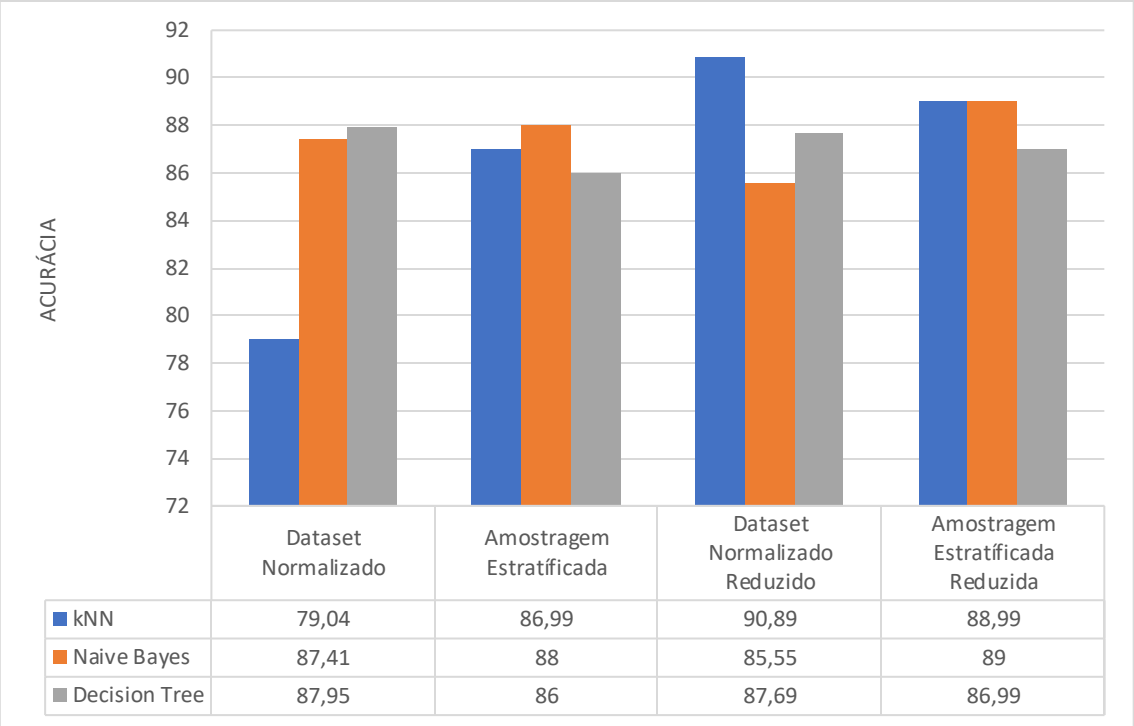
5.1 Validação Cruzada

Para garantir que nossos modelos foram efetivamente capazes de generalizar a partir dos dados, utilizamos a validação cruzada de *k-fold*. Esta técnica envolve a divisão do conjunto de dados em k subconjuntos e a realização de treinamento e validação em diferentes combinações desses subconjuntos.

Os próximos passos do estudo envolveram a interpretação dos resultados e as discussões pertinentes, que são apresentados nas seções subsequentes.

6 Resultados

Após a realização de todos os passos aqui descritos, obteve-se os resultados dispostos na Figura 4.



	Dataset Normalizado	Amostragem Estratificada	Dataset Normalizado Reduzido	Amostragem Estratificada Reduzida
kNN	79,04	86,99	90,89	88,99
Naive Bayes	87,41	88	85,55	89
Decision Tree	87,95	86	87,69	86,99

Figura 6 – Resultados obtidos.

7 Conclusão

O trabalho objetivou desenvolver um modelo baseado em classificadores de aprendizado supervisionado para distinguir conjuntos de métricas que poderiam indicar a existência de tais distúrbios do sono.

A aplicação de algoritmos de classificação permitiu a identificação de padrões e relações entre os dados, contribuindo para uma compreensão mais aprofundada dos fatores que podem indicar uma predisposição à insônia e à apneia. Estes resultados podem somar esforços a intervenções no âmbito da saúde, uma vez que facilitam a detecção precoce de distúrbios, e consequentemente, a implementação de estratégias de intervenção mais eficazes.

Em conclusão, este trabalho reafirma a importância de estudos interdisciplinares que combinam o campo da saúde com técnicas avançadas de ciência de dados. O desenvolvimento de tais modelos preditivos, baseados em aprendizado de máquina, mostra-se como um instrumento poderoso para a compreensão e detecção de distúrbios do sono, com potencial para influenciar positivamente a qualidade de vida de muitos indivíduos.

8 Bibliografia

Gadzik, James. 2006. How much should I weigh? - Quetelet's equation, upper weight limits, and BMI prime. *Connecticut medicine.* 2006, Vol. 70, pp. 81-88.

Robergs, Robert A e Landwehr, Roberto. 2002. The surprising history of the HR_{max} = 220 - age equation. *J Exerc Physiol.* 2002, pp. 1,10.

Spodick, David H. 1993. Survey of selected cardiologists for an operational definition of normal sinus heart rate. *The American journal of cardiology.* 1993, pp. 487-488.