Universidade do Rio de Janeiro
Tópicos Especiais em Matemática e Computação
Trabalho 2

Gustavo Dias de Oliveira

02/06/2024

1 Descrição dos Dados

1.1 Informações relevantes

O banco de dados selecionado foi ”Estimation of Obesity Levels Based On Eating Habits and Physical Condition”, encontrado no repositório UCI, esta base de dados apresenta dados para a estimativa dos níveis de obesidade em indivíduos dos países México, Peru e Colômbia, com base em seus hábitos alimentares e condição física. As amostras são rotuladas com a variável de classe NObesity (Nível de Obesidade), que permite a classificação dos dados utilizando os valores de Peso Insuficiente, Peso Normal, Sobrepeso Nível I, Sobrepeso Nível II, Obesidade Tipo I , Obesidade Tipo II e Obesidade Tipo III. 77% dos dados foram gerados de forma sintética utilizando a ferramenta Weka e o filtro SMOTE, 23% dos dados foram coletados diretamente dos usuários através de uma plataforma web.

1.2 Número de amostras

Os dados contêm 17 atributos e 2111 amostras.

1.3 Atributos

Descrição geral dos atributos presentes neste conjunto de dados:

1. **Gender:** Gênero do participante.
   Resposta: Masculino(Male), Feminino(Female).
2. **Age:** Idade do participante.
   Resposta: Numérica (anos).
3. **Height:** Altura do participante.
   Resposta: Numérica (metros).
4. **Weight:** Peso do participante.
   Resposta: Numérica (quilogramas).
5. **family_history_with_overweight:** Histórico familiar de sobrepeso.
   Resposta: Sim(yes), Não(no).

6. **FAVC:** Consumo de alimentos com alto teor de calorias.

   Resposta: Sim(yes), Não(no).

7. **FCVC:** Consumo de vegetais.

   Resposta: Numérica (1-Nunca, 2-Às vezes, 3-Sempre).

8. **NCP:** Número de refeições principais por dia.

   Resposta: Numérica (1-Uma, 2-Duas ,3-Três , 4-Mais de três).

9. **CAEC:** Consumo de comida entre as refeições.

   Resposta: Não(No), Às vezes(Sometimes), Frequentemente(Frequently), Sempre(Always).

10. **SMOKE:** Hábito de fumar.

    Resposta: Sim(yes), Não(no).

11. **CH2O:** Consumo de água diário.

    Resposta: Numérica (1-Menos de um litro, 2-Entre 1 e 2 litros, 3-mais de 2 litros).

12. **SCC:** Monitora o consumo de calorias.

    Resposta: Sim(yes), Não(no).

13. **FAF:** Frequência de prática de atividade física.

    Resposta: Numérica (0-não pratica, 1-um ou dois dias, 2-dois ou quatro dias, 3-quatro ou cinco dias).

14. **TUE:** Tempo gasto com dispositivos tecnológicos como celular, videogame, televisão, computador e outros.

    Resposta: Numérica (0-zero a duas horas, 1-três a cinco horas, 2-mais de cinco horas).

15. **CALC:** Consumo de álcool.

    Resposta: Não(No), Às vezes(Sometimes), Frequentemente(Frequently), Sempre(Always).

16. **MTRANS:** Meio de transporte usado pelo participante.

    Resposta: Transporte Publico(Public_Transportation), Automóvel(Automobile), Bicicleta(Bicycle), A pé(Walking), Moto(Motorbike).

17. **NObeyesdad:** Nível de obesidade do participante.

    Resposta: Insuficientemente Pesado(Insufficient_Weight), Normal(Normal_Weight), Excesso de Peso Nivel I(Overweight_Level_I), Excesso de Peso Nivel II(Overweight_Level_II), Obesidade Tipo I(Obesity_Type_I), Obesidade Tipo II(Obesity_Type_II), Obesidade Tipo III(Obesity_Type_III).

1.4 Classes

A classe 'NObeyesdad' é considerada o atributo alvo neste conjunto de dados, sendo fundamental para a análise, pois representa o nível de obesidade dos participantes. É a variável que estamos interessados em prever ou estimar com base nas outras características disponíveis, como gênero, idade, hábitos alimentares e estilo de vida.

1.5 Balanceamento

Para avaliar a balanceamento da classe, comparamos o número de amostras em cada categoria, se o número de amostras em cada categoria for parcialmente igual, então a classe é considerada balanceada.

- Obesity_Type_III: 324 amostras
- Obesity_Type_II: 297 amostras
- Obesity_Type_I: 351 amostras
- Overweight_Level_II: 290 amostras
- Overweight_Level_I: 290 amostras
- Normal_Weight: 287 amostras
- Insufficient_Weight: 272 amostras

Neste caso, não há uma grande disparidade entre o número de amostras em cada categoria. Embora haja algumas diferenças, a variação não é significativa. Portanto, pode-se dizer que a classe "NObeyesdad"está relativamente balanceada.

1.6 Valores Nulos ou Faltantes

Utilizando o comando 'dados.isnull().sum()', que retorna a contagem de valores nulos em cada coluna dos dados, foi observado que os atributos não possuiam nenhum valor nulo ou faltante.

1.7 Valores Duplicados

Usando o comando 'dados[dados.duplicated() == True]', que retorna a tabela com todas as linhas duplicadas encontradas nos dados, pudemos analisar a tabela com todos os valores duplicados, e vimos que eles são existentes.

1.8 Outliers (Boxplot)

Um boxplot é um gráfico que representa a distribuição de um conjunto de dados. Ele é composto por uma caixa que mostra a mediana e os quartis dos dados, com linhas que se estendem a partir da caixa para indicar a variabilidade dos dados. Pontos fora dessa linha são considerados outliers, ou seja, valores incomuns. O boxplot é uma ferramenta eficaz para identificar padrões e discrepâncias em conjuntos de dados.

Na Figura 1 abaixo, conseguimos observar que Age(Idade), Height(Altura), Weight(Peso) e NCP(Numero de refeições principais por dia) possuem outliers.

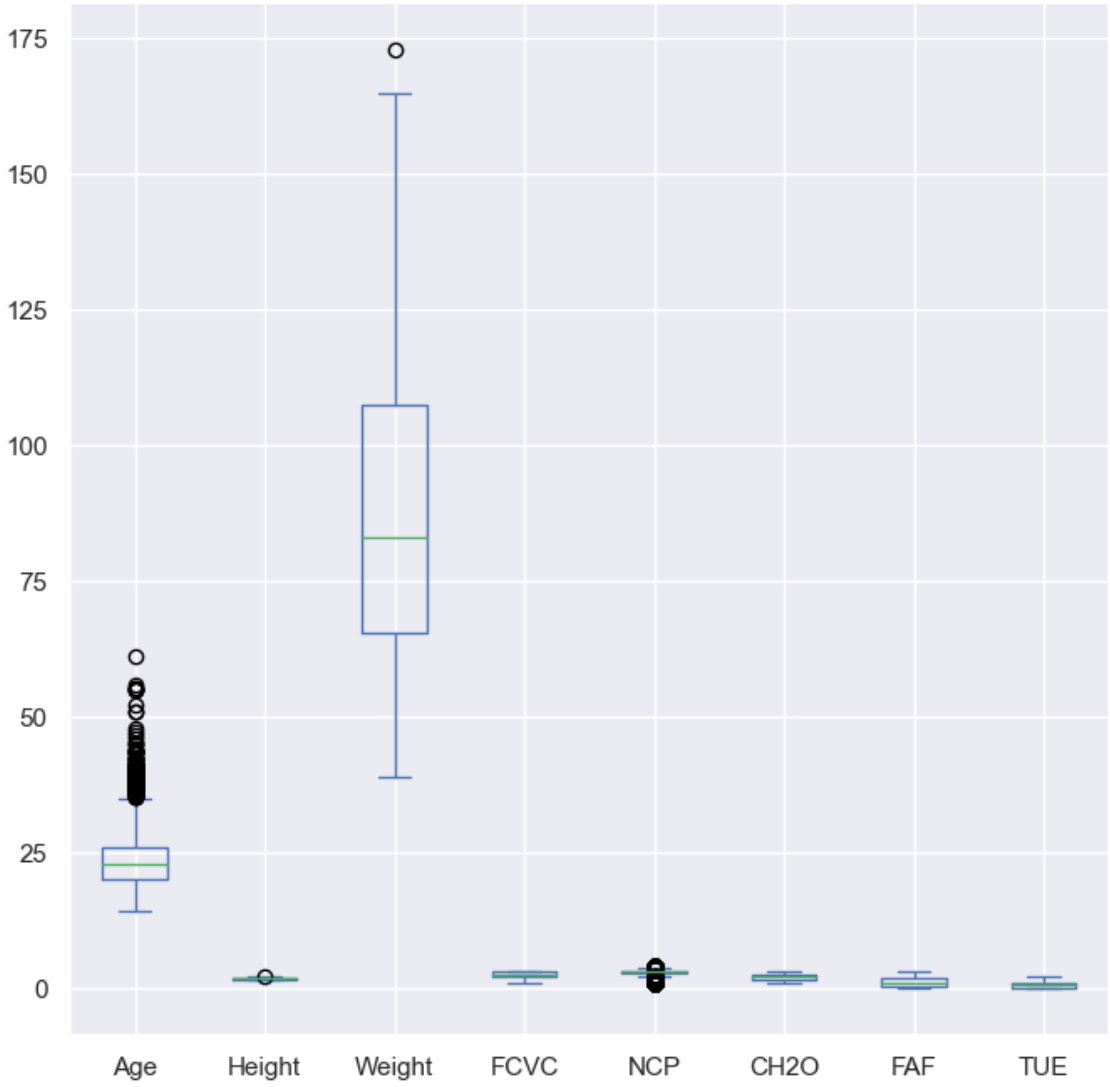

Figura 1: Boxplot dos dados numéricos

2 Análise dos Dados

Será apresentada a média, o desvio padrão, a variância, a matriz de correlação, o valor mínimo e máximo de cada atributo do nosso banco de dados.

2.1 Média

Na tabela 1 abaixo, será apresentados os atributos numéricos e suas determinadas médias.

Tabela 1: Valores Médios dos Atributos

Atributo	**Média**
Idade	24.31
Altura	1.70
Peso	86.59
FCVC	2.42
NCP	2.69
CH2O	2.01
FAF	1.01
TUE	0.66

2.2 Desvio Padrão

O desvio padrão é uma medida estatística que indica o quanto os valores de um conjunto de dados estão dispersos em relação à média.
Na tabela 2 abaixo, será apresentados os atributos numéricos e seus determinados desvios padrão.

Tabela 2: Desvios Padrão dos Atributos

Atributo	**Desvio Padrão**
Idade	6.35
Altura	0.09
Peso	26.19
FCVC	0.53
NCP	0.78
CH2O	0.61
FAF	0.85
TUE	0.61

2.3 Variância

A variância é uma medida estatística que descreve a dispersão dos valores de um conjunto de dados em relação à média
Na tabela 3 abaixo, será apresentados os atributos numéricos e suas determinadas variância.

Tabela 3: Variâncias dos Atributos

Atributo	**Variância**
Idade	40.27
Altura	0.01
Peso	685.98
FCVC	0.29
NCP	0.61
CH2O	0.38
FAF	0.72
TUE	0.37

2.4 Matriz de correlação

A matriz de correlação é uma tabela que mostra como cada variável em um conjunto de dados está relacionada às outras. Ela usa coeficientes de correlação para representar a força e a direção dessas relações. Valores próximos de 1 indicam uma forte correlação positiva, valores próximos de -1 indicam uma forte correlação negativa, e valores próximos de 0 indicam pouca ou nenhuma correlação.

Na Figura 2 abaixo, será apresentado a matriz de correlação dos dados numéricos, nela observamos que as variáveis que possuem maior correlação são Height(Altura) e Weight(Peso) com correlação de 0.46, as outras variáveis possuem correlação de 0.3 para menos, acredito que não tenha grande relevância.

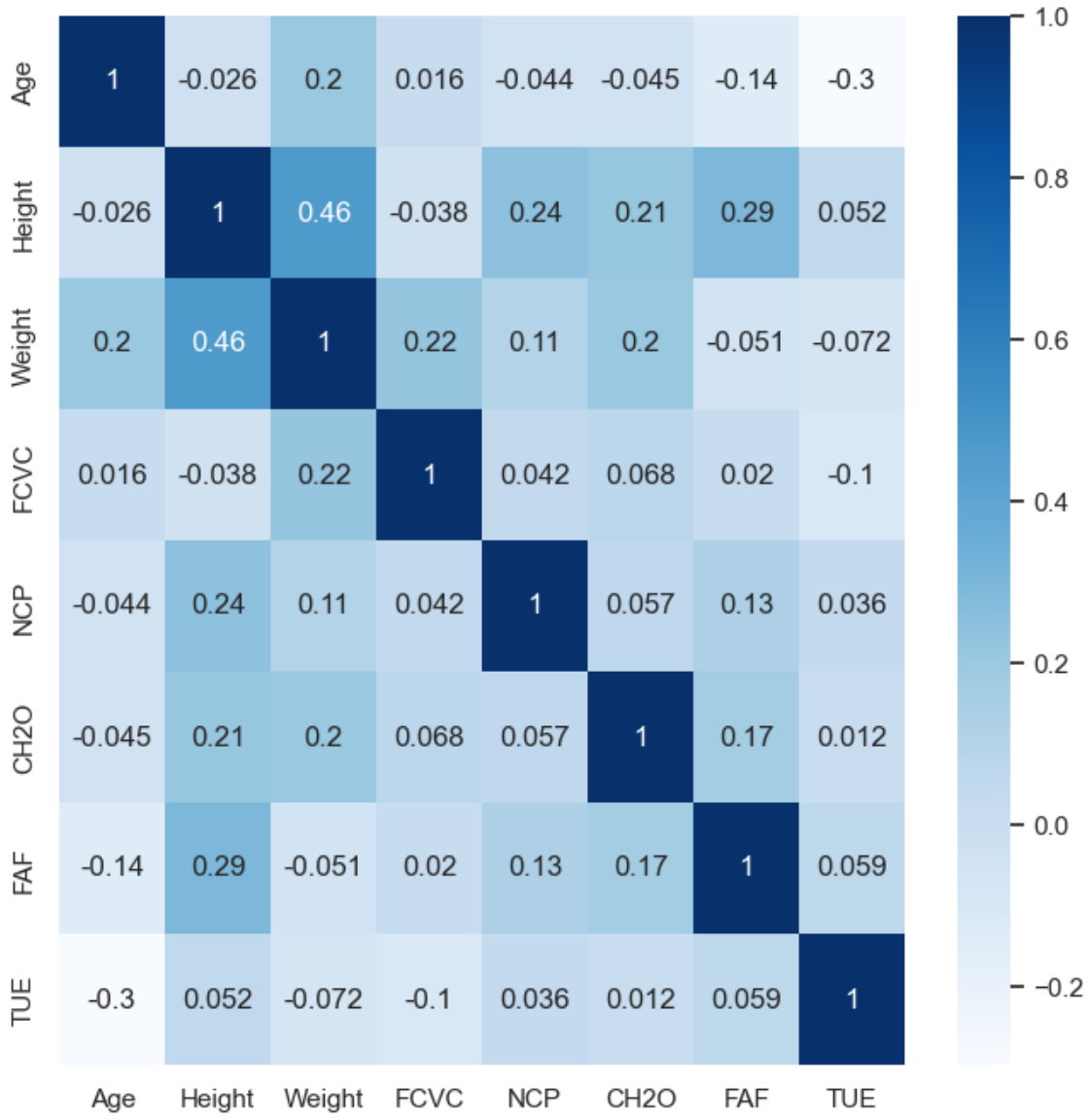

Figura 2: Matriz de correlação

2.5 Valor mínimo

Na tabela 4 abaixo, será apresentados os atributos numéricos e seus determinados valores mínimos.

Tabela 4: Valores Mínimos dos Atributos

Atributo	**Valor Mínimo**
Idade	14
Altura	1.45
Peso	39
FCVC	1
NCP	1
CH2O	1
FAF	0
TUE	0

2.6 Valor Máximo

Na tabela 5 abaixo, será apresentados os atributos numéricos e seus determinados valores maximos.

Tabela 5: Valores Máximos dos Atributos

Atributo	**Valor Máximo**
Idade	61
Altura	1.98
Peso	173
FCVC	3
NCP	4
CH2O	3
FAF	3
TUE	2

3 Metodologia

O pré-processamento dos dados, que consistiu em remover os valores duplicados, remover os outliers dos atibutos [Age(Idade), Height(Altura), Weight(Peso), NCP(Numero de refeições principais por dia)] e transformar as variáveis qualitativas[Gender(Gênero), family_history_with_overweight (Histórico familiar de sobrepeso), FAVC(Consumo de alimentos calóricos com frequência), CAEC (Consumo de alimentos entre refeições), SMOKE(Fumante), SCC(Monitora o consumo de calorias), CALC(Consumo de alcool), MTRANS(Meio de transporte), NObeyesdad(Nível de obesidade)] para quantitativas.

Após o término do pré-processamento foi dado inicio de fato as metodologias.

3.1 Grid Search

Grid Search é uma técnica de otimização utilizada para encontrar os melhores hiperparâmetros para um modelo de machine learning.

A ideia principal é realizar uma busca exaustiva através de um espaço pré-definido de hiperparâmetros, avaliando a performance do modelo para cada combinação possível desses hiperparâmetros. Isso é especialmente útil em problemas de machine learning, onde a escolha dos hiperparâmetros corretos pode ter um impacto significativo na performance do modelo.

3.2 Leave-One-Out (LOO)

Leave-One-Out (LOO) é uma técnica de validação cruzada usada para avaliar a performance de um modelo de machine learning. A ideia básica é treinar o modelo em todos os dados, exceto um único ponto de dados, e então usar o ponto deixado de fora para avaliar o desempenho do modelo. Esse processo é repetido para cada ponto de dados no conjunto de dados.

3.3 K-Vizinhos Mais Próximos (KNN)

O algoritmo KNN (K-Nearest Neighbors) é um método de aprendizado supervisionado utilizado tanto para classificação quanto para regressão em problemas de machine learning. Sua abordagem é simples e intuitiva: Treinamento, Previsão, Seleção dos Vizinhos, Classificação ou Regressão. O parâmetro ”k”representa o número de vizinhos a serem considerados e é crucial para a eficácia do algoritmo. Um valor baixo de ”k”pode levar a uma variância alta e sensibilidade a ruído, enquanto um valor alto de ”k”pode suavizar fronteiras de decisão e mascarar padrões locais.

4 Experimentos Computacionais

Foi realizado um processo de seleção de hiperparâmetros para o classificador K-Nearest Neighbors usando Grid Search combinado com validação Leave-One-Out(LOO). Além disso, foi rodado este processo várias vezes (30 iterações) para obter uma boa estimativa do desempenho do modelo.

4.1 K-Vizinhos Mais Próximos

No K-Vizinhos Mais Próximos, foram utilizadas algumas variações de parâmetros, foi variado o numero de vizinhos mais proximos com os valores de [1,2,3,4,5,6,10] e tambem o peso, utilizando o peso com valores [uniforme ou distância], como na tabela 6 abaixo.

Tabela 6: Variações de parâmetros no K-Vizinhos Mais Próximos

Modelo	**Parâmetro**	**Variações**
K-Vizinhos Mais Próximos	Número de Vizinhos	1, 2, 3, 4, 5, 6, 10
	Peso	Uniforme, Distância

4.2 Validação Cruzada Leave-One-Out (LOO)

Para realizar a avaliação da classificação, foi utilizada a validação cruzada Leave-One-Out (LOO). Em minha opinião, o LOO não é um bom método de validação cruzada, pois utiliza praticamente todos os dados para treino e um único ponto para teste em cada iteração. Como resultado, a classificação pode sofrer de alta variabilidade nas previsões e baixa estabilidade, onde o modelo se ajusta muito bem ao conjunto de treinamento, mas não generaliza bem para novos dados. Além disso, o custo computacional elevado torna o LOO ineficiente para datasets grandes.

Após as 30 iterações, obtivemos o seguinte resultado na tabela 7:

Tabela 7: Resultados da classificação utilizando KNN e validação Leave-One-Out

Run	Classificador	Acurácia	F1	Recall
0	**KNN**	1.0	1.0	1.0
1	KNN	1.0	1.0	1.0
2	KNN	1.0	1.0	1.0
3	KNN	1.0	1.0	1.0
4	KNN	1.0	1.0	1.0
5	KNN	1.0	1.0	1.0
6	KNN	1.0	1.0	1.0
7	KNN	1.0	1.0	1.0
8	KNN	1.0	1.0	1.0
9	KNN	1.0	1.0	1.0
10	KNN	1.0	1.0	1.0
11	KNN	1.0	1.0	1.0
12	KNN	1.0	1.0	1.0
13	KNN	1.0	1.0	1.0
14	KNN	1.0	1.0	1.0
15	KNN	1.0	1.0	1.0
16	KNN	1.0	1.0	1.0
17	KNN	1.0	1.0	1.0
18	KNN	1.0	1.0	1.0
19	KNN	1.0	1.0	1.0
20	KNN	1.0	1.0	1.0
21	KNN	1.0	1.0	1.0
22	KNN	1.0	1.0	1.0
23	KNN	1.0	1.0	1.0
24	KNN	1.0	1.0	1.0
25	KNN	1.0	1.0	1.0
26	KNN	1.0	1.0	1.0
27	KNN	1.0	1.0	1.0
28	KNN	1.0	1.0	1.0
29	KNN	1.0	1.0	1.0

Como o modelo pegou praticamente todos os dados para treino, o único dado que foi utilizado para teste teve uma classificação ótima, porém, em novos conjuntos de dados, provavelmente nosso modelo não se sairia tão bem assim.

Como resultado, a média do resultado após 30 iteraçoes ficou como na tabela 8 a seguir:

Tabela 8: Média dos resultados da classificação utilizando KNN e LOO

Classificador	Média Acurácia	Média F1	Média Recall
KNN	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

4.3 Resultado final do Grid Search

Ao final de todas as iterações, percebemos que todas elas obteram o mesmo resultado para o dado de teste, onde todas elas tiveram o seguinte resultado da tabela 9:

Tabela 9: Melhores Parâmetros e Métricas de Desempenho

Melhores Parâmetros	Acurácia	F1	Recall
{'n_neighbors': 1, 'weights': 'uniform'}	1.0	1.0	1.0

5 Conclusão

Neste trabalho, foi realizado um processo de seleção de hiperparâmetros para o classificador K-Nearest Neighbors (KNN) utilizando Grid Search combinado com validação Leave-One-Out (LOO). Além disso, o processo foi executado 30 vezes para obter uma estimativa robusta do desempenho do modelo.

No KNN, foram exploradas diversas combinações de parâmetros, variando o número de vizinhos mais próximos (valores de [1, 2, 3, 4, 5, 6, 10]) e o tipo de peso (uniforme ou distância).

A validação cruzada Leave-One-Out foi escolhida para a avaliação da classificação. Embora o LOO utilize praticamente todos os dados para treino e apenas um único ponto para teste em cada iteração, resultando em uma boa classificação para o ponto de teste, ele possui desvantagens significativas. O LOO pode levar a alta variabilidade nas previsões e baixa estabilidade, com o modelo ajustando-se muito bem ao conjunto de treinamento, mas não generalizando adequadamente para novos dados. Além disso, o custo computacional elevado do LOO torna-o ineficiente para datasets grandes.

Os resultados das 30 iterações mostraram que o classificador KNN com os parâmetros (n_neighbors': 1, 'weights': 'uniform') obteve uma acurácia, F1 score e recall perfeitos (1.0) em todos os testes. Esse resultado consistente indica que o modelo se ajustou perfeitamente aos dados fornecidos para teste em cada iteração.

No entanto, essa alta performance nos dados de teste pode ser enganosa. O uso do LOO, que treina o modelo com quase todos os dados e testa em apenas um ponto, pode levar a uma sobreajuste (overfitting). Portanto, apesar dos resultados ideais observados, o modelo pode não ter um desempenho tão bom em conjuntos de dados novos e não vistos.

Em resumo, embora o processo de seleção de hiperparâmetros tenha identificado uma configuração que maximiza o desempenho nos testes realizados, a metodologia LOO pode não ser a melhor escolha para avaliar a generalização do modelo, especialmente em contextos de dados grandes e diversos. O próximo passo seria validar o modelo com outras técnicas de validação cruzada e com novos conjuntos de dados para obter uma avaliação mais realista e robusta de seu desempenho.

Referências

- https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition
- https://doi.org/10.1016/j.dib.2019.104344