Universidade do Rio de Janeiro
Tópicos Especiais em Matemática e Computação
Trabalho 1

Gustavo Dias de Oliveira

05/05/2024

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.

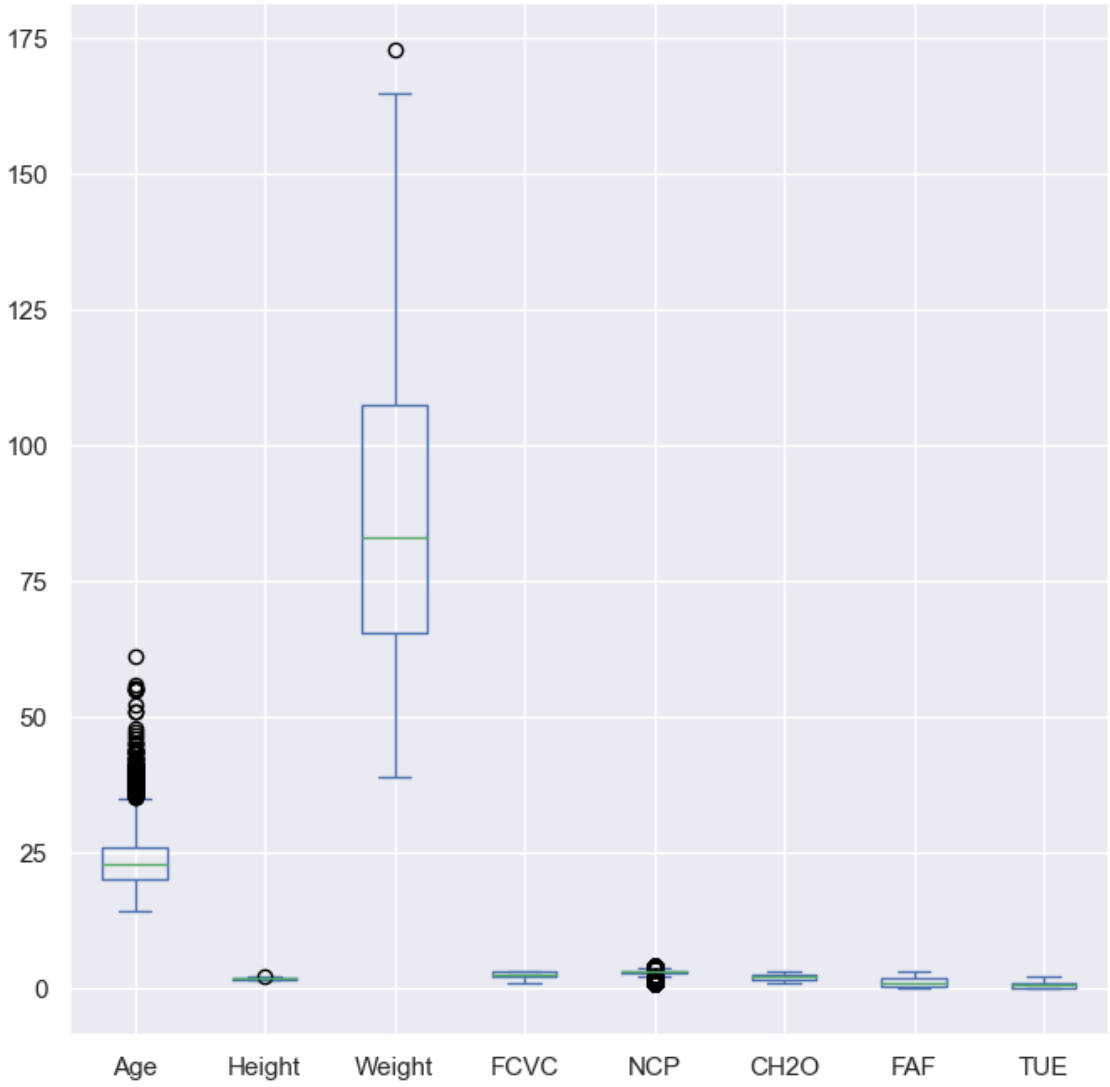

Figura 1: Boxplot dos dados numéricos

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 apresentados o gráfico da matriz de correlação dos dados numéricos.

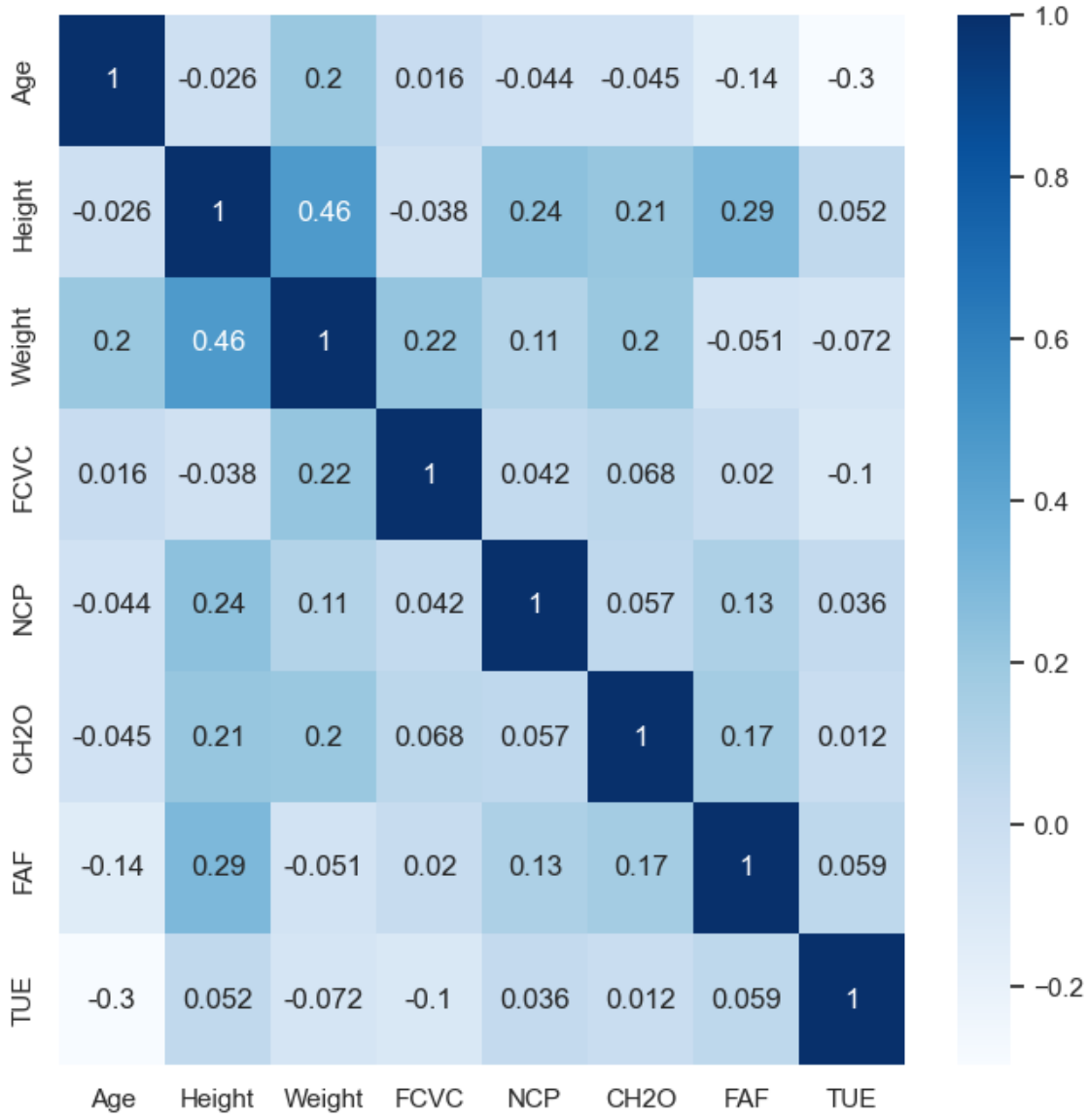

Figura 2: Matriz de correlação

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 Análise dos componentes principais (PCA)

O PCA (Principal Component Analysis) é uma técnica de redução de dimensionalidade frequentemente usada em análise de dados e aprendizado de máquina. A sua principal finalidade é simplificar conjuntos de dados complexos, preservando ao máximo a sua estrutura de variância.

Após aplicar no conjunto de dados com o número de componentes igual a três, obtivemos a seguinte Tabela 6:

Característica	1ª Componente	2ª Componente	3ª Componente
Gender	0.0019	0.0024	0.2666
Age	0.0433	0.9962	0.0431
Height	0.0015	-0.0022	0.0581
Weight	0.9966	-0.0458	0.0449
Family History with Overweight	0.0064	0.0076	-0.0295
FAVC	0.0034	0.0016	-0.0254
FCVC	0.0049	-0.0048	-0.0789
NCP	-0.0011	-0.0075	0.0161
CAEC	-0.0046	-0.0038	0.0439
Smoke	0.0002	0.0015	0.0108
CH2O	0.0047	-0.0103	0.1227
SCC	-0.0017	-0.0030	0.0016
FAF	-0.0025	-0.0294	0.6784
TUE	-0.0025	-0.0269	0.0644
Calc	0.0043	-0.0003	-0.0374
MTRANS	-0.0004	0.0491	0.0856
NObeyesdad	0.0688	0.0345	-0.6507

Tabela 6: Resultados do PCA

3.2 Método de Seleção de Carcterísticas (K-Best)

O método SelectKBest é uma técnica de seleção de características, utilizada para selecionar as K melhores características(nesse caso, k=3, entao ele seleciona 3 características) de um conjunto de dados com base em uma determinada função de pontuação. Essa função de pontuação é geralmente uma medida estatística que avalia a relação entre cada característica e a variável de saída.

O objetivo foi reduzir a dimensionalidade dos dados, mantendo apenas as características mais importantes para o modelo de aprendizado de máquina. Pois isso pode melhorar o desempenho do modelo, reduzir o tempo de treinamento e evitar overfitting.

No código, 'X' representa as características do conjunto de dados, ou seja, as informações que são usadas para prever alguma coisa e 'y' é a variável de destino, também conhecida como a classe. Após aplicar o K-Best ele nos deu que as melhores atributos são ['Gender', 'Weight', 'family_history_with_overweight'], e criamos um novo conjunto de dados apenas com essas variáveis, esse novo conjunto foi nomeado "dados_new".

3.3 Método de Agrupamento (K-Means)

O Método de Agrupamento K-Means é um algoritmo de aprendizado não supervisionado usado para agrupar dados sem rótulos em grupos distintos, chamados de "clusters".

Neste trabalho, foi a aplicado o metodo K-Means com variação dos clusters(número de agrupamento de dados) e n_init(número de inicializações diferentes).

3.4 Críterio de Validação utilizado (DaviesBouldin)

Critério de Validação de Davies-Bouldin é uma métrica de avaliação interna usada para avaliar a qualidade dos clusters produzidos por algoritmos de agrupamento, como o K-Means. Ele é utilizado para determinar quão bem definidos e separados os clusters estão dentro de um conjunto de dados.

Quanto menor o valor do Critério de Validação de Davies-Bouldin, melhor a separação entre os clusters. Um valor mais baixo indica que os clusters são mais compactos e bem definidos, com boa separação entre eles.

3.5 Variação dos parâmetros que foram testados (Parameter Grid)

Parameter Grid, é uma técnica usada em aprendizado de máquina para encontrar os melhores parâmetros para um modelo preditivo.

Muitas vezes você tem vários parâmetros que podem ser ajustados para otimizar o desempenho do modelo. Em vez de tentar manualmente diferentes combinações de parâmetros, o Parameter Grid permite que você defina uma grade de valores para cada parâmetro que deseja ajustar. Em seguida, o algoritmo testa todas as combinações possíveis desses parâmetros e retorna a combinação que produz o melhor desempenho de acordo com uma métrica específica

Foi utilizado o parameter grid nos dois conjuntos de dados, o conjunto normal(sem alteraçoes pós pré-processamento) e o conjunto reduzido(o conjunto obtido atravez do K_best), juntamente com o K-Means e o Davies-Bouldin, foi definindo como numero de clusters 7 valores diferentes e como numero de inicialização 4 valores.

4 Experimentos Computacionais

4.1 Características que mais influênciam as componentes principais

Após aplicar o PCA o atributo que tiver como resultado o maior valor em módulo em determinada componente, é a que vai ter maior influência nesse determinado componente.
Logo, segundo a Tabela 6 já mencinada antes, podemos concluir que temos como caracteristicas mais influentes no PCA:

Componente 1:

- Característica 4: Weight

Componente 2:

- Característica 2: Age

Componente 3:

- Característica 13: FAF

4.2 Parâmetros do K-Means

Para testar diferentes valores para o K-Means e encontrar o valor ótimo foi usado o Parameter Grid, utilizamos 7 valores para o numero de cluster conforme a tabela 7 abaixo, e em 4 valores o numero de inicialização conforme a tabela 8 abaixo. E como metodo de avaliação tivemos o DaviesBouldin.

Tabela 7: Número de Clusters Testados para o K-Means

Número de Clusters
2
3
4
5
6
7
8

Tabela 8: Número de Inicializações Testadas para o K-Means

Número de Inicializações
5
10
15
20

O Parameter Grid foi utilizado tanto no Conjunto de dados completo(com todas as variáveis) quanto no reduzido(reduzido com o K-Best) onde obtivemos o seguintes resultados:

Conjunto de Dados completo
Os parâmetros ótimos encontrados:

- n_clusters: 2
- n_init: 5

O melhor valor do critério de validação:

- DaviesBouldin: 0.5710340360481823

Com o seguinte gráfico na Figura 3 abaixo:

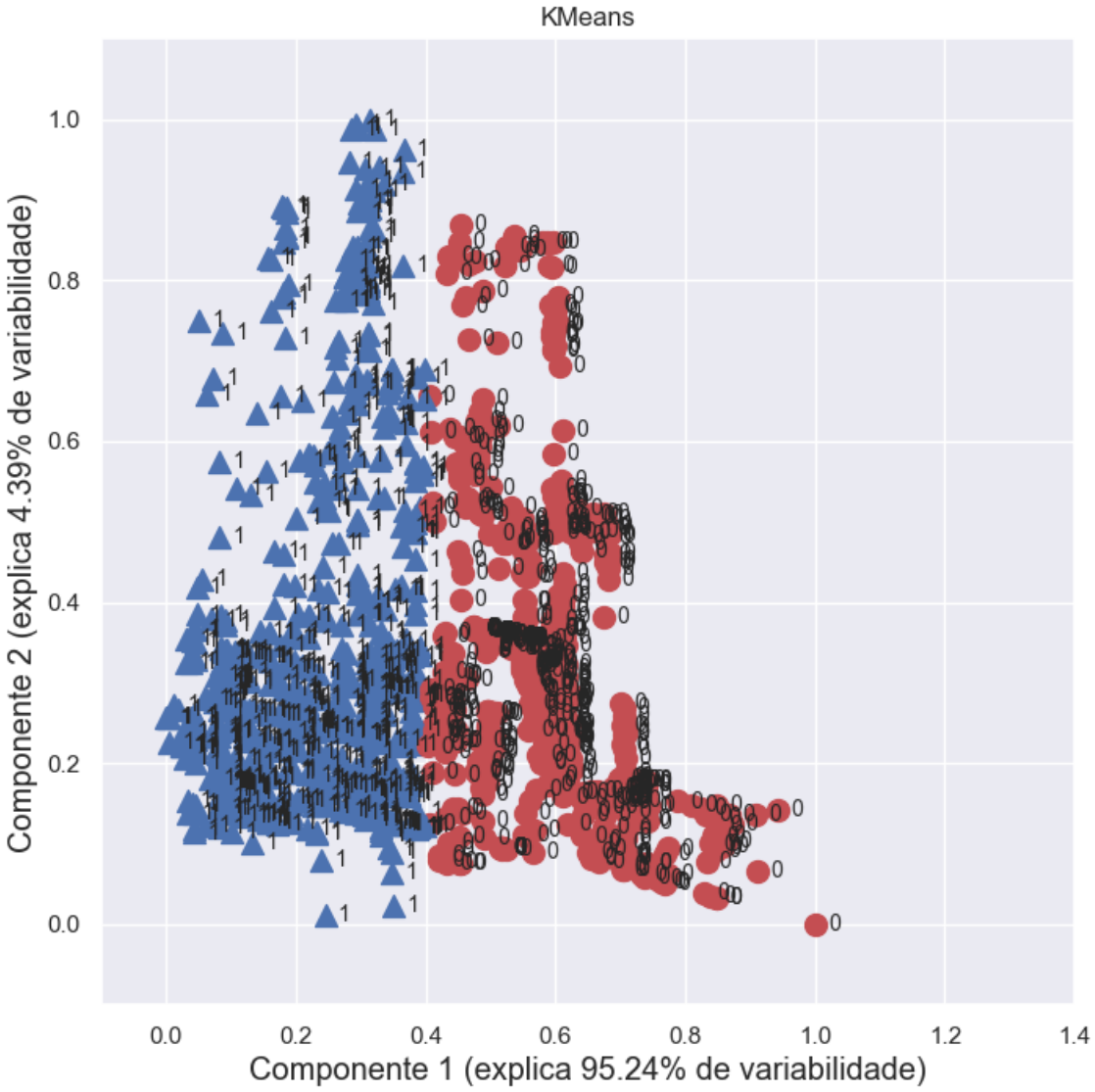

Figura 3: Gráfico resultado K-Means Normal

Conjunto de Dados Reduzidos (K-Best)
Os parâmetros ótimos encontrados:

- n_clusters: 7
- n_init: 10

O melhor valor do critério de validação:

- DaviesBouldin: 0.47520105266791113

Com o seguinte gráfico na Figura 4 abaixo:

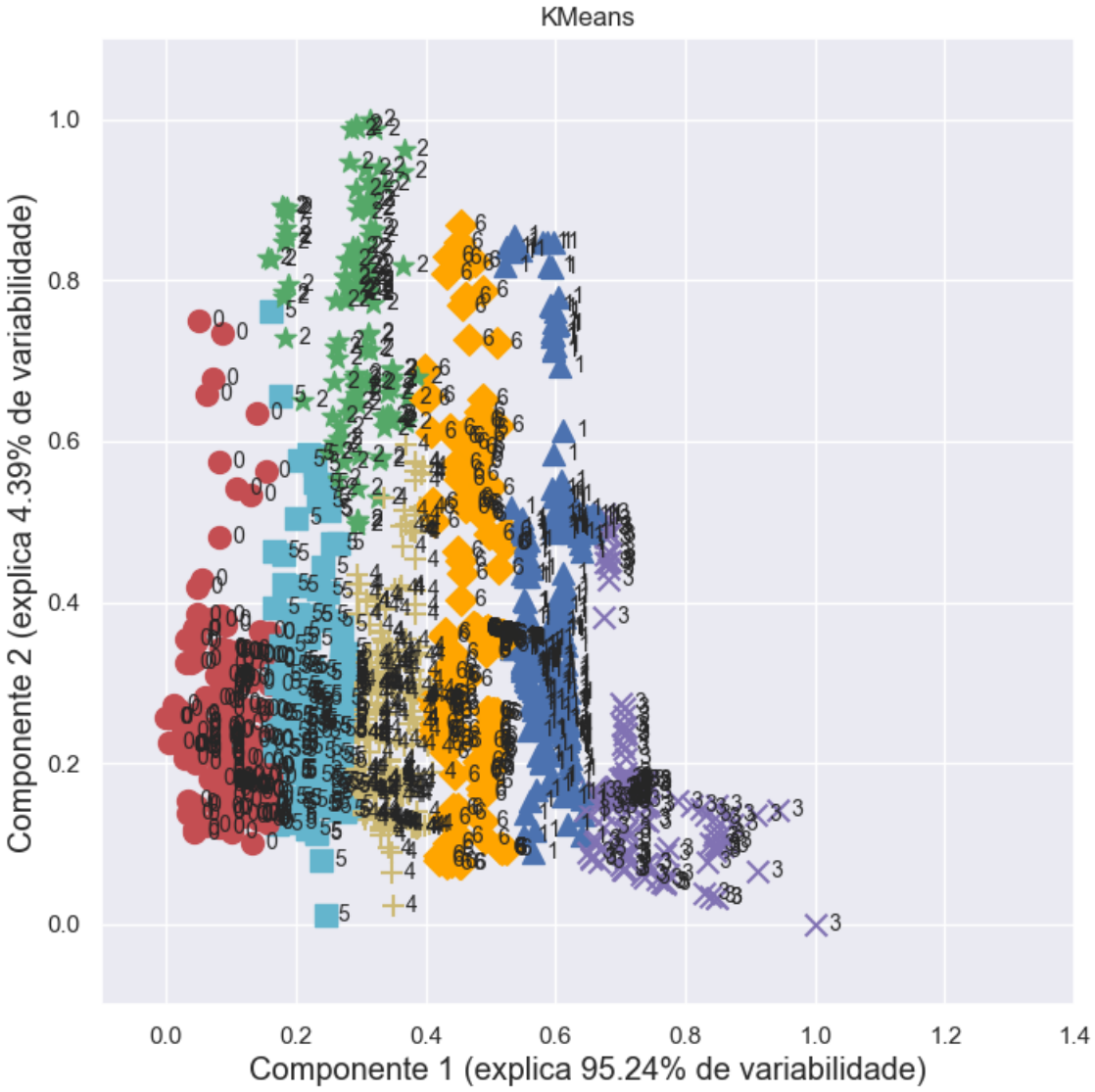

Figura 4: Gráfico resultado K-Means Reduzido

5 Conclusão

O trabalho apresentou uma boa análise do conjunto de dados sobre os níveis de obesidade com base em hábitos alimentares e condição física em indivíduos do México, Peru e Colômbia. Começando pela descrição dos dados, foram detalhadas as informações relevantes, o número de amostras, os atributos e as classes, além de verificar o balanceamento, a presença de valores nulos, duplicados e outliers.

Em seguida, foram realizadas análises estatísticas, como média, desvio padrão, variância e matriz de correlação, proporcionando conhecer melhor sobre as características do conjunto de dados. Posteriormente, foram aplicadas metodologias como Análise de Componentes Principais (PCA), Seleção de Características (K-Best) e Agrupamento (K-Means), com o objetivo de simplificar os dados, selecionar as melhores características e identificar padrões de agrupamento, além disso, a métrica de validação Davies-Bouldin foi utilizada para avaliar a qualidade dos clusters formados.

Como o resultado do Davies-Bouldin do conjunto de dados reduzido foi menor do que o conjunto de dados por inteiro, conseguimos afirmar que a metodologia utilizada, além de reduzir os dados, conseguiu melhorar a performace do programa.

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