**Efeitos do Sedentarismo na Saúde Pública**

**Integrantes:**

**Bruno Zovaro Nascimento - 10424880**

**Douglas Novaes Dias – 10423666**

**Milan Mirco Moraes Mazur – 10363757**

**Paulo Cesar Masson Junior - 10416023**

**Universidade Presbiteriana Mackenzie**

**São Paulo**

**2025**

**SUMÁRIO**

1 Introdução (p. 3)

2 Referencial teórico (p. 5)

3 Metodologia (p. 8)

4 Resultados (p.22)

5 Conclusão (p.23)

6 Referências bibliográficas (p.24)

1. **Introdução**

1.1 Contexto do trabalho

O estudo "Estimativa dos Níveis de Obesidade por Meio de uma Análise Preditiva Baseada em Atividade Física e Alimentação" surge da preocupação com o aumento da obesidade em várias partes do mundo. A obesidade está ligada a problemas de saúde como doenças cardíacas, diabetes tipo 2 e outras condições graves. Com uma vida cada vez mais sedentária e o consumo frequente de alimentos pouco saudáveis, é fundamental encontrar formas mais precisas de identificar e classificar os níveis de obesidade.

1.2 Motivação

Devido ao aumento da obesidade no mundo identificamos a oportunidade de auxiliar as pessoas a terem uma vida mais saudável através de recomendações baseadas em seus históricos.

1.3 Justificativa

A pesquisa se justifica pela necessidade de analisar a obesidade de maneira mais ampla, levando em conta não apenas medidas físicas, mas também os hábitos alimentares e de exercício das pessoas. Os métodos tradicionais podem não captar toda a complexidade do problema, o que pode levar a avaliações imprecisas ou à falta de entendimento sobre os fatores que aumentam a obesidade.

* 1. Objetivo geral e objetivos específicos da pesquisa

O objetivo geral está alinhado com o Objetivo de Desenvolvimento Sustentável (ODS): Desenvolver um sistema de recomendação para proporcionar saúde e bem-estar.

Objetivos específicos:

Desenvolver um modelo que utilize aprendizado de máquina para prever e classificar os níveis de obesidade com base em hábitos de alimentação e atividade física;

Ajudar na criação de estratégias mais eficazes para prevenir e tratar a obesidade, fornecidas aos profissionais de saúde ferramentas mais precisas para adaptar recomendações e tratamentos;

Contribuir para a saúde pública, ajudando a reduzir os impactos da obesidade na sociedade e incentivando um estilo de vida mais saudável por meio de ações bem fundamentadas

1. **Referencial teórico**

Há estudos científicos recentes que buscaram percorrer o mesmo caminho, de clusterização para perquirir padrões e hábitos em crianças da rede escolar relacionando-os com a ocorrência de obesidade realizado por Universidade Paranaenses[[1]](#footnote-1), cujo cerne do estudo foi investigar se existem padrões (ou agrupamentos) de comportamento relacionados ao estilo de vida entre estudantes, levando em conta hábitos como alimentação (consumo de frutas, vegetais, alimentos açucarados e refrigerantes), sono, atividade física e tempo em comportamento sedentário. Após identificar esses padrões, o estudo também analisou se eles estavam associados ao excesso de peso corporal entre os participantes.

Trata-se de uma pesquisa de natureza transversal (realizada em um único momento no tempo), com base em dados coletados de aproximadamente 17 mil estudantes com idades entre 4 e 20 anos, todos participantes do Projeto Paraná Saudável. Os dados foram obtidos por meio de um questionário estruturado, que avaliou diversos comportamentos do estilo de vida. O excesso de peso foi determinado com base no índice de massa corporal (IMC) dos estudantes.

Para descobrir os padrões de comportamento, foi utilizada a técnica de análise de cluster, que agrupa indivíduos com características semelhantes. Em seguida, usou-se análise de covariância e regressão logística para avaliar a relação entre esses grupos e a presença de excesso de peso.

Foram encontrados cinco grupos distintos de comportamentos em ambos os sexos. Entre os principais resultados, os estudantes que pertenciam ao grupo com menor consumo de frutas e vegetais e maior tempo de comportamento sedentário apresentaram o dobro de chance de ter excesso de peso em comparação ao grupo considerado referência. Além disso, os estudantes agrupados por alto consumo de alimentos açucarados e refrigerantes e com menor tempo de sono também apresentaram maior probabilidade de estarem com excesso de peso (69% a mais entre as moças e 73% entre os rapazes).

Em conclusão, a combinação de hábitos saudáveis — como alta ingestão de frutas e vegetais, baixo consumo de alimentos açucarados e refrigerantes, maior tempo de sono e menos tempo sedentário — foi associada à manutenção de um peso corporal mais saudável entre os estudantes avaliados.

Desse estudo extrai-se uma forma de comparação entre os clusters que pode ser aproveitada em trabalhos científicos congêneres, que consiste em comparar os hábitos encontrados nos clusters identificados por meio dos algoritmos de agrupamento, ganhando especial relevância as métricas utilizadas nesse processo específico.

Não obstante, pode-se formular novas métricas a depender das peculiaridades da base de dados, uma vez que os atributos, achados e classificações podem permitir novas conclusões sobre aspectos inexplorados em pesquisas anteriores.

Pode-se mencionar também outro estudo[[2]](#footnote-2), publicado na Revista Científica da Unifenas em setembro de 2024, intitulado “Estimativas de Níveis de Obesidade Utilizando Machine Learning”, que explora a aplicação de algoritmos de aprendizado de máquina na previsão dos níveis de obesidade em indivíduos da Colômbia, México e Peru. A pesquisa destaca a importância de fatores demográficos e comportamentais — como idade, gênero, IMC, hábitos alimentares e prática de atividades físicas — na determinação do estado de peso dos indivíduos. A partir da coleta e análise de dados estruturados, o objetivo central foi identificar padrões preditivos que auxiliem em estratégias de intervenção e prevenção da obesidade, condição que vem crescendo globalmente.

O trabalho empregou diversas técnicas de machine learning, tais como Regressão Logística, Árvores de Decisão, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) e Redes Neurais Artificiais (ANN), aplicadas sobre um conjunto de dados com mais de dois mil registros. Para assegurar a representatividade dos dados, parte do conjunto foi gerado sinteticamente utilizando SMOTE, enquanto o restante foi obtido por meio de uma plataforma web. O processo metodológico envolveu etapas como análise exploratória dos dados, pré-processamento, divisão entre treino e teste, e ajuste de hiper parâmetros via Grid Search.

Os resultados obtidos revelaram que os modelos Random Forest e Redes Neurais Artificiais se destacaram quanto à capacidade preditiva, superando os demais algoritmos em termos de acurácia e F1-score. A combinação de dados demográficos e comportamentais proporcionou uma abordagem mais robusta na modelagem preditiva. O estudo também enfatizou a importância da escolha apropriada dos algoritmos e da preparação dos dados, incluindo a normalização e a transformação de variáveis categóricas.

Apesar dos avanços e resultados promissores, a pesquisa identificou limitações relacionadas à generalização dos modelos para outras populações, especialmente devido ao uso de dados sintéticos e às características culturais específicas dos países estudados. Ainda assim, os modelos preditivos desenvolvidos apresentam potencial aplicação em ambientes clínicos, podendo ser utilizados por profissionais da saúde como ferramentas para identificar grupos de risco e direcionar ações preventivas personalizadas.

Por fim, o estudo reforça a relevância das técnicas de aprendizado de máquina como suporte para decisões em saúde pública. Os autores sugerem, para trabalhos futuros, a inclusão de novas variáveis (como fatores genéticos e socioeconômicos), a expansão da base de dados com amostras reais e a adoção de técnicas de aprendizado profundo. A pesquisa não apenas contribui para a área da Ciência de Dados aplicada à saúde, como também demonstra o poder transformador dos modelos preditivos na formulação de políticas públicas mais eficazes no combate à obesidade.

1. **Metodologia**

Obtida a base de dados, iniciou-se a preparação, incluindo a análise exploratória, que possui grande importância no trabalho de um cientista de dados

Em nosso treinamento não supervisionado buscamos agrupar utilizando a linguagem Python e, especialmente, o algoritmo DBSCAN[[3]](#footnote-3) para realizar o agrupamento, sendo um componente do conhecido pacote scikit-learn. Em tradução livre, trata-se de “clusterização espacial baseada em densidade de aplicações com ruído. Encontra amostras de núcleo de alta densidade e expande clusters a partir delas. Bom para dados que contêm clusters de densidade similar”.

Além das etapas de preparação e pré-processamento da base, foi realizada uma análise da qualidade dos agrupamentos formados com o algoritmo DBSCAN. Como se trata de uma técnica de aprendizado não supervisionado, não há um rótulo definido que permita avaliar diretamente a acurácia dos grupos gerados. Por isso, foram empregadas métricas internas, como o índice de silhueta, que mede o grau de coesão e separação dos clusters. Valores próximos de 1 indicam que os dados estão bem agrupados, enquanto valores negativos apontam sobreposição entre os

grupos.  
  
A escolha do DBSCAN — Density-Based Spatial Clustering of Applications with Noise — se deu por sua capacidade de identificar clusters de formatos variados e por lidar de forma eficiente com ruídos e outliers, que são comuns em dados comportamentais. Os parâmetros eps (distância máxima para considerar pontos vizinhos) e min\_samples (número mínimo de pontos para formar um grupo) foram ajustados de maneira empírica, com o objetivo de reduzir a quantidade de dados classificados como ruído e garantir maior coesão dentro dos clusters.

Com os agrupamentos definidos, conduziu-se uma análise descritiva dos perfis médios de cada cluster, buscando destacar características específicas, como hábitos alimentares, frequência de atividade física e práticas relacionadas ao cuidado com a saúde. A identificação dos clusters com predominância de indivíduos com peso corporal normal permitiu a construção de perfis considerados saudáveis, servindo de referência para comparação com outros grupos.

Por fim, abre-se a possibilidade de utilizar os clusters como rótulos para o desenvolvimento de modelos supervisionados. Isso permitiria prever o grupo de pertencimento de novos indivíduos com base em seus hábitos, o que se mostra especialmente útil para fins de recomendação personalizada. Podem ser considerados algoritmos como Random Forest, KNN e redes neurais, com a aplicação de validação cruzada como técnica para assegurar maior generalização dos modelos e evitar overfitting.

As etapas estão descritas abaixo em detalhes, passo a passo, no tópico “Base de dados – exploração e preparação”.

3.1 Exploração, preparação da base de dados e treinamentos

A base de dados foi baixada do seguinte local:

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

As características iniciais da base de dados foram obtidas por meio de exploração utilizando a ferramenta Spyder do pacote Anaconda.

A base traz uma legenda para o que consta dela conforme quadro a seguir:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Role** | **Type** | **Demographic** | **Description** |
| Gender | Feature | Categorical | Gender |  |
| Age | Feature | Continuous | Age |  |
| Height | Feature | Continuous |  |  |
| Weight | Feature | Continuous |  |  |
| family\_history\_with\_overweight | Feature | Binary |  | Has a family member suffered or suffers from overweight? |
| FAVC | Feature | Binary |  | Do you eat high caloric food frequently? |
| FCVC | Feature | Integer |  | Do you usually eat vegetables in your meals? |
| NCP | Feature | Continuous |  | How many main meals do you have daily? |
| CAEC | Feature | Categorical |  | Do you eat any food between meals? |
| SMOKE | Feature | Binary |  | Do you smoke? |
| CH2O | Feature | Continuous |  | How much water do you drink daily? |
| SCC | Feature | Binary |  | Do you monitor the calories you eat daily? |
| FAF | Feature | Continuous |  | How often do you have physical activity? |
| TUE | Feature | Integer |  | How much time do you use technological devices such as cell phone, videogames, television, computer and others? |
| CALC | Feature | Categorical |  | How often do you drink alcohol? |
| MTRANS | Feature | Categorical |  | Which transportation do you usually use? |
| NObeyesdad | Target | Categorical |  | Obesity level |

Utilizando a linguagem Python e por meio dos seguintes comandos foram obtidas informações gerais, mas importantes, sobre a base, com as saídas respectivas, abaixo comentados:

***Neste ponto inicial, obteve-se o número total de instâncias e de atributos, respectivamente:***

print(df\_obesidade.shape)

(2111, 17)

***Após, imprimiu-se as 10 primeiras linhas com seus valores:***

Texto

O conteúdo gerado por IA pode estar incorreto.

Teclado de computador

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

***Na sequência, obteve-se as características dos dados, mais precisamente os tipos de dados em cada atributo:***

Texto

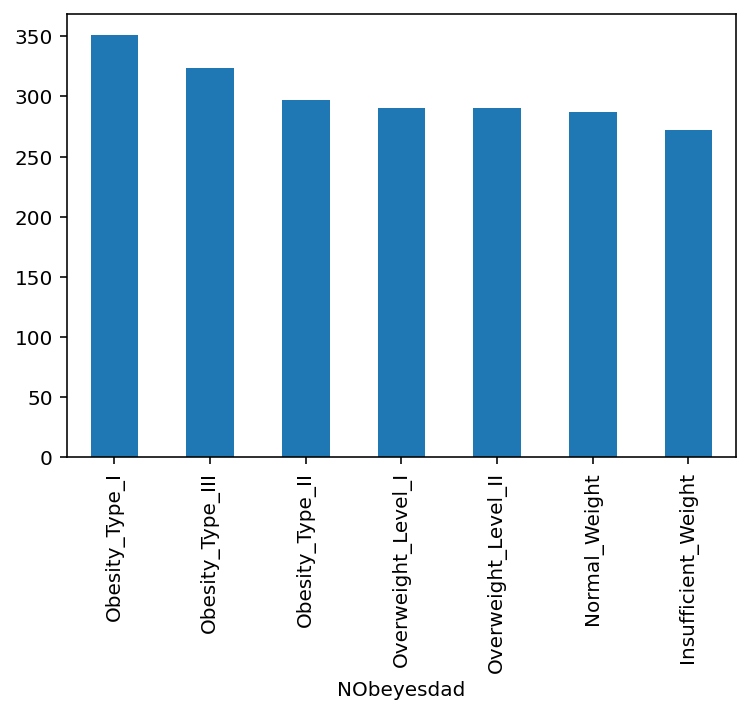
O conteúdo gerado por IA pode estar incorreto.

***Em seguida, buscou-se eventuais valores ausentes:***

Texto

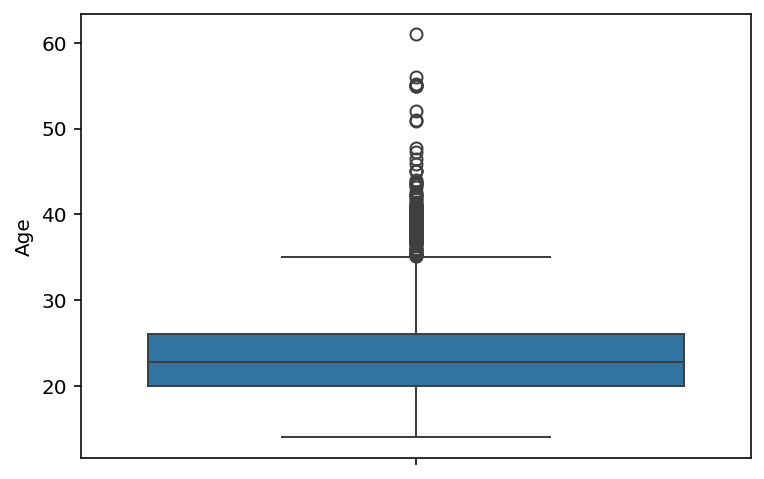
O conteúdo gerado por IA pode estar incorreto.

***Pode-se obter gráficos exploratórios:***



No Gráfico de Barras acima temos a quantidade de cada classificação acerca da obesidade presente na base de dados. Podemos constatar que há mais pessoas com “Obesidade Tipo I”, seguidas pelas com “Obesidade Tipo III”.

***Pode-se obter também um box plot para termos uma visão geral sobre as idades:***



Extrai-se ainda que:

- a linha dentro da caixa representa a mediana, que parece estar em torno dos 23-24 anos. Isso indica que metade das pessoas tem idade abaixo desse valor e metade acima.

- Os quartis (Caixa do box plot)

Q1 (Primeiro quartil - 25%): Aproximadamente 21 anos.

Q3 (Terceiro quartil - 75%): Aproximadamente 30 anos.

O IQR (Intervalo Interquartil = Q3 - Q1): A maior parte dos dados está entre 21 e 30 anos.

- Os bigodes (Whiskers) estendem-se até um limite razoável dos dados antes de considerar algo um outlier. Aqui, o limite superior parece estar próximo dos 40 anos.

- Os outliers (círculos acima do bigode superior) são os pontos acima do bigode representam valores atípicos.

Há várias idades acima de 40 anos, chegando aos 60 anos. Isso sugere que há um número pequeno de pessoas mais velhas no conjunto de dados.

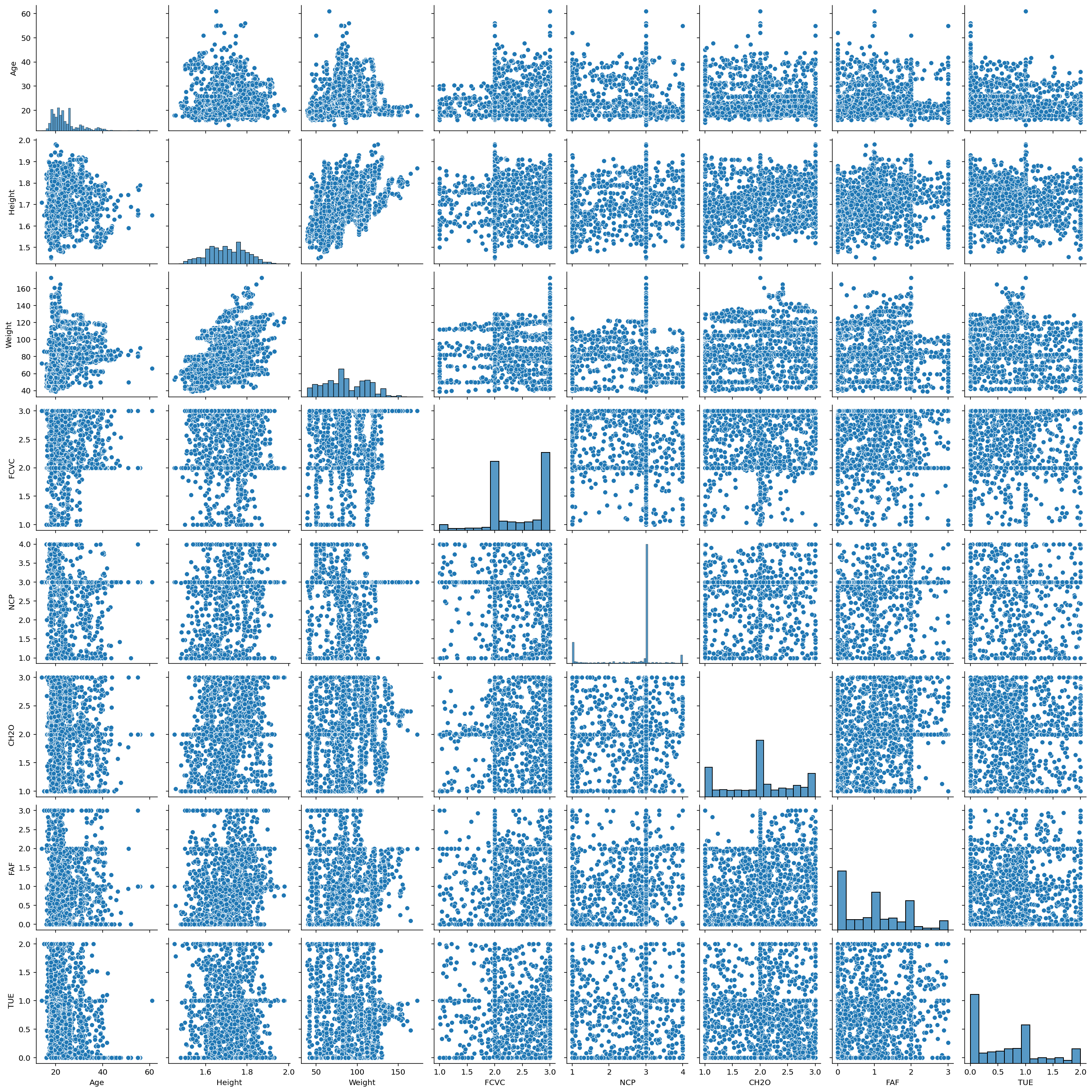
***Calculados os quartis do atributo “Age” chega-se em um resultado preciso:***

|  |  |
| --- | --- |
| Quartil | Média de idade |
| 0.25 | 19.947192 |
| 0.50 | 22.777890 |
| 0.75 | 26.000000 |

***Outra forma de apresentação pode ser assim:***

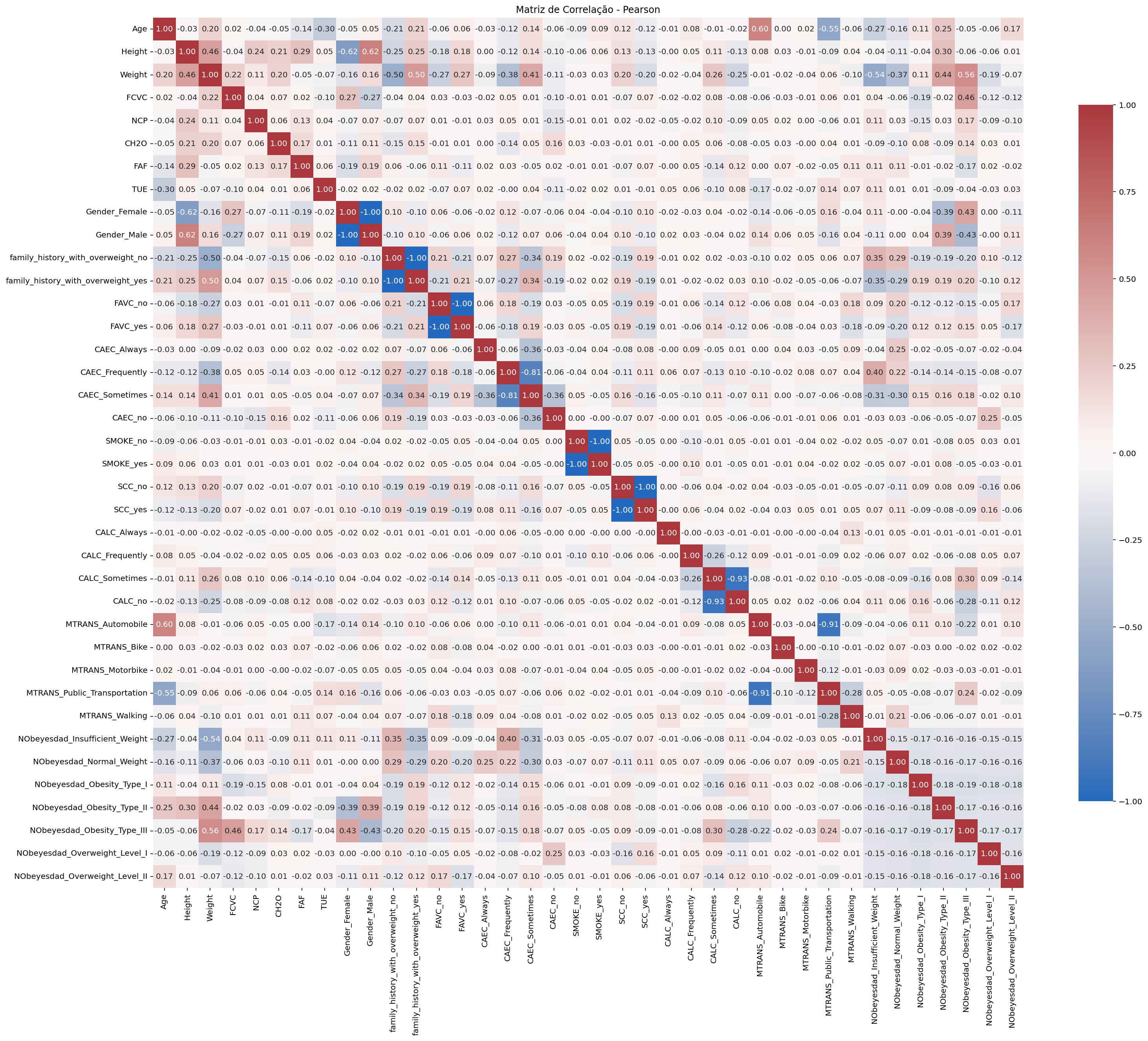


***O gráfico de dispersão é muito útil para, de antemão, verificarmos as relações entre os atributos:***



***Após, foi realizado o one hot encode, transformando os valores categóricos em colunas separadas com o nome das variáveis e com 0 (False) ou 1 (True).***

***Em sequência, produziu-se uma matriz de correlação:***



***Após o encode o DataFrame já tem o seguinte formato, com mais colunas:***

***Em seguida normalizou-se o data frame:***

Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

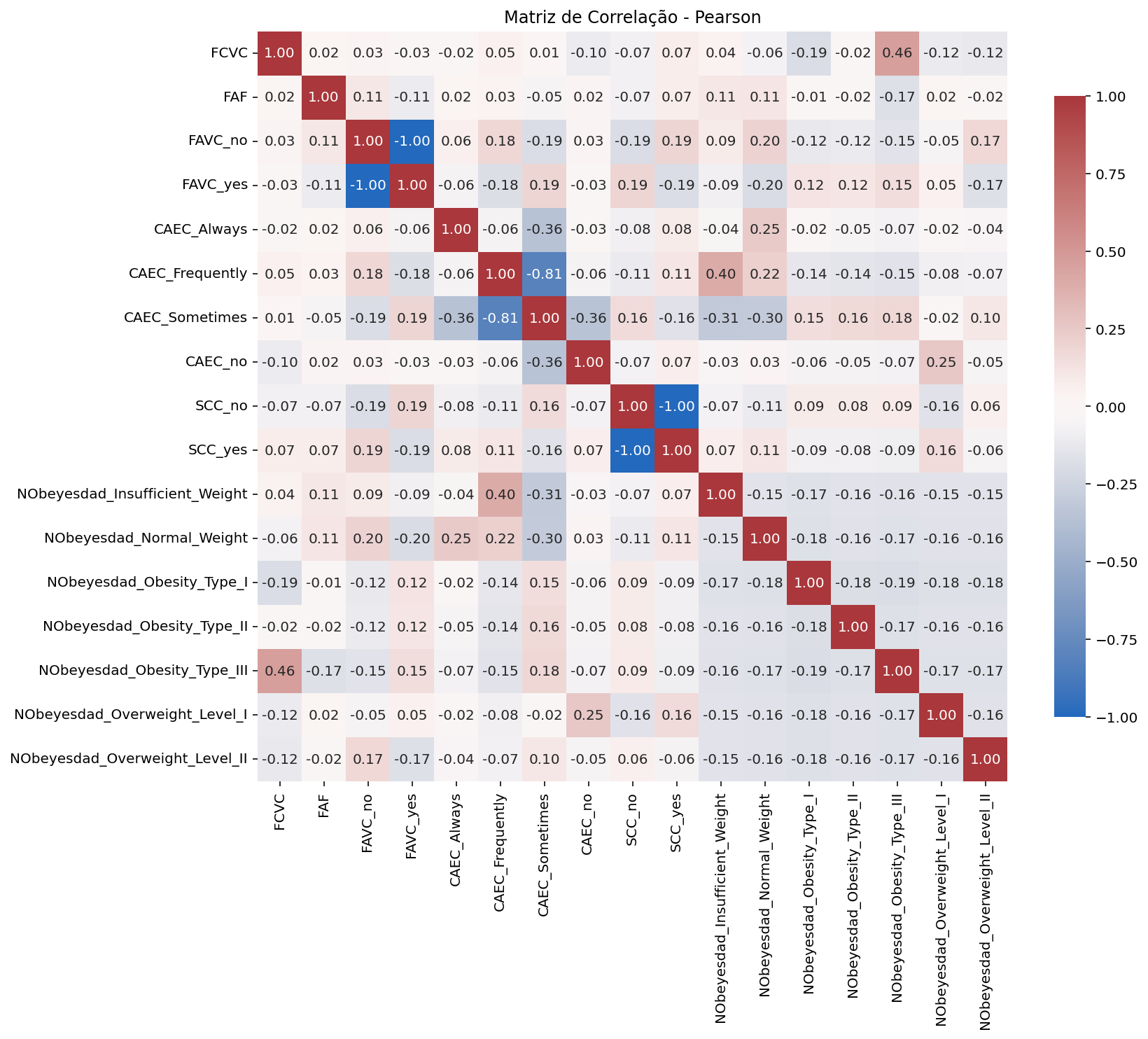
***E foi realizada a primeira tentativa de treinamento, por meio de DBSCAN, que acabou resultando em muitos clusters e ainda trouxe a faixa de pessoas com peso considerado normal para junto dos outliers pois ficaram alocadas no grupo -1.***



Tela de computador com texto preto sobre fundo branco

O conteúdo gerado por IA pode estar incorreto.

***O que levou a novos tratamentos da base para reduzir a dimensionalidade no sentido de manter somente os atributos (colunas) mais relevantes para a obtenção de grupos.***



***Desta forma, com a dimensionalidade reduzida foi refeita a tentativa de treinamento não supervisionado com o DBSCAN***

***E com os resultados foram verificados quais clusters têm mais pessoas com peso normal:***

Clusters com pessoas de peso normal:

[ 0 -1 2 5]

***E foram exibidas as características das pessoas nestes clusters específicos (pessoas com peso normal):***

Tabela

O conteúdo gerado por IA pode estar incorreto.

***Analisados os dados, extraiu-se que:***

Os clusters que têm 100% das pessoas com peso normal são: Cluster 0, Cluster 2, Cluster 5.

É possível, então, analisar as características médias dessas pessoas e compará-las com os outros clusters, como diferenciais das pessoas com peso normal identificamos: Menor consumo de alimentos altamente calóricos (FAVC),  
100% das pessoas nesses clusters responderam que não consomem frequentemente alimentos altamente calóricos (FAVC\_no = 1.0). Isso é diferente de muitos outros clusters, onde há um percentual significativo de pessoas que consomem esses alimentos;

Menos consumo frequente de comida fora de casa (CAEC), Cluster 0, 2 e 5 têm CAEC\_Sometimes = 1.0, ou seja, essas pessoas comem fora de casa às vezes, mas não frequentemente. Em comparação, alguns outros clusters têm um consumo frequente ou sempre de comida fora de casa;

Atividade física (FAF) um pouco maior. Esses clusters têm valores de FAF (Frequência de atividade física) entre 0.31 e 0.50, o que é um pouco mais alto do que a média de muitos outros clusters.  
SCC (Cuidado consciente com a saúde), no Cluster 0 e 2, 100% das pessoas responderam SCC\_no = 1.0, indicando que não fazem um acompanhamento de saúde tão frequente enquanto o Cluster 5, SCC\_yes = 1.0, ou seja, as pessoas desse grupo se preocupam mais com a saúde.

Em resumo, pessoas com peso normal, nesses clusters, estão se diferenciando pelos seguintes hábitos:

Evitam alimentos altamente calóricos, comem fora de casa apenas ocasionalmente,

praticam um pouco mais de atividade física, têm uma variação no acompanhamento de saúde (algumas cuidam mais, outras não tanto)

A partir daqui, e uma vez rotulados os clusters, abre-se a possibilidade de:

1) Identificar a qual cluster pertencem indivíduos e comparar hábitos com outros clusters e, eventualmente, sugerir mudanças para alcançar um perfil desejado (por exemplo, peso normal).

2) Viabiliza-se a utilização de aprendizado de máquina supervisionado (como, por exemplo, Random Forest ou Redes Neurais) para prever o cluster de novos usuários com base em seus hábitos.

1. **Resultados**

Os resultados obtidos a partir da aplicação do algoritmo de clusterização DBSCAN permitiram identificar agrupamentos distintos de indivíduos com base em seus hábitos alimentares, estilo de vida e fatores relacionados à saúde. Dentre os vários clusters formados, três deles — os Clusters 0, 2 e 5 — chamaram a atenção por serem compostos exclusivamente por pessoas com peso corporal normal.

A análise dos hábitos predominantes nesses clusters revelou padrões de comportamento considerados mais saudáveis, tais como:

Baixo consumo de alimentos altamente calóricos (FAVC);

Alimentação entre as refeições de forma ocasional (CAEC às vezes);

Frequência ligeiramente maior de atividade física (FAF);

Variação na preocupação com o controle calórico (SCC), sendo que em um dos clusters a maioria das pessoas realiza esse controle.

Esses resultados demonstram que, mesmo dentro de uma população com diversidade de comportamentos, é possível identificar grupos homogêneos cujas características podem servir como referência para intervenções em saúde pública. A partir desses perfis, é possível desenvolver estratégias de recomendação personalizadas, visando aproximar indivíduos com hábitos menos saudáveis de padrões de comportamento associados ao peso normal.

Além disso, a estrutura dos clusters poderá ser utilizada futuramente como base para modelos supervisionados, permitindo prever o grupo ao qual um novo indivíduo pertence, com base em seus dados comportamentais.

1. **Conclusão**

Este projeto demonstrou como a aplicação de técnicas de **ciência de dados**, especialmente algoritmos de aprendizado não supervisionado como o **DBSCAN**, pode contribuir significativamente para o entendimento de padrões de comportamento relacionados à obesidade. Ao analisar um conjunto de dados contendo informações sobre alimentação, atividade física, hidratação, consumo de álcool, entre outros fatores, conseguimos identificar **clusters com perfis saudáveis**, especialmente aqueles compostos apenas por indivíduos com peso normal.

Os resultados obtidos reforçam a importância de **hábitos equilibrados** para a manutenção da saúde corporal e apontam caminhos para ações educativas e preventivas, tanto em nível individual quanto coletivo. A possibilidade de utilizar esses agrupamentos como rótulos para futuras aplicações supervisionadas ampliam ainda mais o potencial prático deste trabalho, permitindo a criação de **sistemas de recomendação personalizados**, voltados à promoção da saúde e bem-estar.

Concluímos, portanto, que a análise preditiva baseada em dados comportamentais é uma ferramenta promissora na luta contra o sedentarismo e a obesidade, podendo ser aplicada em contextos clínicos, escolares e institucionais como suporte a **políticas públicas de saúde**.

1. **Referências bibliográficas e sites consultados**

CASTRO, L. N.; FERRARI, D. G. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

SILVA, L. A; PERES, S. M; BOSCARIOLI, C. Introdução à mineração de dados: com aplicações em R. Rio de Janeiro: LTC, 2016.

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html> – Acesso em 29/05/2025

<https://periodicos.ufsc.br/index.php/rbcdh/article/view/82910> – Acesso em 29/05/2025

<https://revistas.unifenas.br/index.php/revistaunifenas/article/view/1064/412> – Acesso em 13/04/2025

1. https://periodicos.ufsc.br/index.php/rbcdh/article/view/82910 [↑](#footnote-ref-1)
2. https://revistas.unifenas.br/index.php/revistaunifenas/article/view/1064/412 [↑](#footnote-ref-2)
3. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html [↑](#footnote-ref-3)