Exploratory Data Analysis on Obesity Risk Dataset for Cardiovascular Disease Prediction

Juan de Souza Holanda
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
juanholanda@alu.ufc.br

Matheus de Castro Vieira
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
matheusdcastro@alu.ufc.br

Diego Duarte de Lima
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
diegolimaufcalu.ufc.br

Mateus Andrade Maia
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
mateusmaia45@alu.ufc.br

***Abstract*—A Análise Exploratória de Dados é uma etapa fundamental para extrair informações úteis de um conjunto de dados. Por meio dela, é possível visualizar de forma intuitiva e resumida conhecimentos que estão nas ocultos nas camadas internas do conjunto de dados. No âmbito médico, identificar padrões em meio aos dados é de extrema importância para realizar diagnósticos precisos rapidamente, identificar qual tratamento adequado e determinar riscos à saúde. O dataset "Obesity or CVD Risk" foi obtido da plataforma online Kaggle. Uma metodologia foi adotada para estudar as 2111 amostras e 17 atributos, consistindo em usar a ferramentas de visualização de gráficos do Python, realizando análises monovaridas, bivariadas, PCA, condicionais e e incondicionais para obter representações em plots e tabelas. Os resultados do estudo são promissores e têm utilidade em prover insights para elaboração de políticas de saúde pública e prevenção de doenças cardiovasculares.**

***Index Terms*—Análise Exploratória de Dados, Pré-Processamento, obesidade, redução de dimensionaldade, PCA**

I. Introdução

A obesidade tornou-se um dos principais desafios de saúde pública nas últimas décadas, afetando milhões de pessoas em todo o mundo e contribuindo para o desenvolvimento de doenças crônicas como diabetes, hipertensão e problemas cardiovasculares. Na América Latina, esse cenário se mostra particularmente preocupante, com taxas crescentes de sobrepeso e obesidade em diferentes faixas etárias. Fatores como mudanças nos hábitos alimentares, aumento do consumo de alimentos processados e redução da atividade física têm sido apontados como elementos centrais nesse processo, tornando fundamental a compreensão dos padrões comportamentais associados ao ganho excessivo de peso.

A identificação precoce de fatores de risco relacionados à obesidade permite não apenas intervenções mais eficazes, mas também a formulação de políticas públicas direcionadas à promoção de estilos de vida mais saudáveis. Nesse contexto, a análise de dados se apresenta como uma ferramenta valiosa para revelar associações entre características demográficas, hábitos alimentares, nível de atividade física e a prevalência de diferentes graus de obesidade.

O presente trabalho tem como objetivo realizar uma análise exploratória de um conjunto de dados sobre estimativa de níveis de obesidade em indivíduos provenientes do México, Peru e Colômbia, com idades entre 14 e 61 anos. Os dados foram coletados por meio de uma plataforma web contendo um questionário anônimo, resultando em 2111 registros e 17 atributos que abrangem informações demográficas, hábitos alimentares e condições físicas. Entre as variáveis analisadas estão o consumo de alimentos altamente calóricos, frequência de ingestão de vegetais, número de refeições principais, consumo de água e álcool, além de indicadores de atividade física e uso de tecnologia. A partir da aplicação de métodos estatísticos descritivos e técnicas de visualização de dados, busca-se compreender a distribuição dos níveis de obesidade na população estudada e investigar quais fatores apresentam maior associação com as diferentes categorias de peso, contribuindo para futuras estratégias de prevenção e intervenção em saúde pública.

II. Métodos

A. Descrição do dataset

1) Origem e contexto: O conjunto de dados utilizado neste estudo corresponde à estimativa do nível de obesidade entre indivíduos. Os dados foram coletados por meio de uma pesquisa on-line na qual os participantes relataram anonimamente informações demográficas, dietéticas e de atividade física. O conjunto de dados processado possui um total de 2111 registros e 17 atributos, abrangendo indivíduos com idade entre 14 e 61 anos.

O conjunto de dados foi obtido a partir do repositório Kaggle [1], plataforma de aprendizagem e competição. O nome do dataset para este trabalho está especificado, na plataforma Kaggle, como: "Obesity or CVD risk (Classify/Regressor/Cluster)".

2) Estrutura e características: O conjunto de dados é composto por variáveis preditores agrupadas em três categorias principais: atributos demográficos, hábitos alimentares e condição física. A variável resposta (alvo) representa o nível de obesidade estimado para cada indivíduo.

Categoria	Variáveis	Descrição (resumo)
Demográfica	Gênero, Idade, Altura, Peso	Características pessoais básicas
Hábitos alimentares	FAVC, FCVC, NCP, CAEC, CH2O, CALC	Indicadores de hábitos de consumo de alimentos e bebidas
Condição física	SCC, FAF, TUE, MTRANS	Fatores relacionados ao estilo de vida e atividade
Alvo	NObesity	Rótulo categórico indicando o nível de obesidade (seis classes baseadas no IMC)

TABLE I
CATEGORIZAÇÃO DAS VARIÁVEIS DO ESTUDO.

As variáveis numéricas e categóricas do conjunto representam diferentes aspectos dos hábitos e condições dos participantes. As variáveis FAVC (Frequent consumption of high caloric food), FCVC (Frequency of consumption of vegetables), NCP (Number of main meals), CAEC (Consumption of food between meals), CH2O (Daily water consumption) e CALC (Consumption of alcohol) descrevem padrões alimentares relacionados à ingestão de alimentos calóricos, frequência de consumo de vegetais e quantidade de refeições diárias, assim como, avaliam comportamentos complementares à dieta, ingestão hídrica e consumo de bebidas alcoólicas. As variáveis SCC (Calories consumption monitoring), FAF (Physical activity frequency), TUE (Time using technology devices) e MTRANS (Transportation used) estão associadas à condição física e ao estilo de vida, refletindo o controle de calorias, a frequência de atividade física, o tempo de exposição a dispositivos eletrônicos e o principal meio de transporte utilizado. Por fim, a variável NObesity corresponde à classe alvo, categorizando os indivíduos em seis níveis de obesidade com base em faixas de IMC, variando de Underweight a Obesity III.

B. Procedimentos analíticos

1) Análise monovariada incondicional: Está analise tem por objetivo caracterizar, de forma isolada, a distribuição de cada variável do conjunto de dados, sem estratificação por *NObesity*. Inicialmente serão verificados a integridade e o tipo das variáveis, assegurando a ausência de valores faltantes e a consistência de codificação entre atributos demográficos (*Age*, *Height*, *Weight*) e de hábitos/estilo de vida (*FAVC*, *FCVC*, *NCP*, *CAEC*, *CH2O*, *CALC*, *SCC*, *FAF*, *TUE*, *MTRANS*). Em seguida serão calculadas estatísticas descritivas para cada variável, incluindo média, desvio padrão, mediana, quartis, amplitude interquartil, assimetria e curtose, além de valores mínimos e máximos, a fim de caracterizar tendência central, dispersão e forma das distribuições.

Para a inspeção gráfica, serão gerados histogramas com largura de classe definida por critério adequado à amostra e, quando pertinente, curvas de densidade suavizada para variáveis contínuas. Variáveis discretas ou ordinais, como *FAF*, *TUE* e *NCP*, serão representadas por gráficos de frequência. Boxplots complementares serão utilizados para evidenciar a presença de valores potencialmente atípicos, tomando como referência o critério de uma vez e meia a amplitude interquartil.

2) Análise monovariada condicional de classe: Está analise tem por finalidade descrever o comportamento de cada variável quando condicionada aos níveis de *NObesity*. O conjunto de dados será estratificado segundo as classes *Insufficient_Weight*, *Normal_Weight*, *Overweight_Level_I*, *Overweight_Level_II*, *Obesity_Type_I*, *Obesity_Type_II* e *Obesity_Type_III*, registrando-se, para cada estrato, o tamanho amostral e a proporção correspondente. Para cada par variável–classe serão obtidas estatísticas descritivas condicionais, contemplando média, desvio padrão, mediana, quartis, amplitude interquartil e taxa de valores potencialmente atípicos; no caso de variáveis ordinais, serão também reportadas distribuições de frequência por nível.

A comparação visual entre classes será realizada por meio de painéis de boxplots com a mesma ordem de categorias no eixo horizontal para todas as variáveis analisadas, favorecendo a leitura consistente entre atributos. Para variáveis discretas ou ordinais serão utilizados gráficos de barras condicionais, com padronização de rótulos, cores e escalas ao longo dos painéis.

3) Análise bivariada incondicional: Após estudar as variáveis isoladamente, é importante explorar as relações entre pares de preditores. A análise bivariada consiste em averiguar como dois atributos se realcionam entre si,independentemente da classe ou saída do modelo. Isso é essencial para:

- Entender informações internas do dataset;
- Planejar a seleção de ou redução de dimensionalidade (PCA).

Uma das principais métricas obtidas durante essa etapa é a correlação, que indica se existe relação linear entre duas variáveis. No entanto, para calculá-la, é necessário transformar os atributos categóricos (*Gender, FAVC, CAEC, CALC, SCC, MTRANS*) em valores numéricos. O valor da correlação entre os preditores no mapa de calor resultante desse processo deve ser analisado conforme a seguinte diretriz:

- Próximo de $0 \implies$ Sem relação linear significativa;
- 0 a $\pm 0.3 \implies$ Fraca correlação;
- ± 0.3 a $\pm 0.7 \implies$ Moderada;
- ± 0.7 a $\pm 1.0 \implies$ Forte.

4) Análise multivariada incondicional: A Análise de Componentes Principais (PCA) é uma ferramenta crucial para entender as relações entre vários preditores de maneira simultânea sem considerar a classificação de saída. Essa técnica permite a investigação de padrões, dependências, redundâncias e agrupamentos. Em especial, ela tem como objetivo reduzir a dimensionalidade dos dados por meio da transformação de um conjunto de variáveis correlacionadas em um conjunto

de componentes principais - uma combinação dos atributos originais -, que melhor explicam a variância dataset.

A implementação do PCA foi feita manualmente, usando a biblioteca NumPy e aplicando os seguintes passos:

C. Pré-Processamento

As variáveis categóricas ordinais (*CAEC, CALC*) foram mapeadas para valores inteiros (ex: 0,1,2,3 etc.), de forma a preservar sua ordem. Já as variáveis categóricas nominais (*Gender, MTRANS*) foram transformadas, usando a técnica one-hot enconding.

Então, o conjunto de dados resultante foi padronizado (centralizando com média 0 e escalonando com variância 1) para que todas a variáveis tivessem o mesmo peso, contribuindo de forma igual.

D. Matriz de Covariância

É uma matriz que expressa como os preditores variam conjuntamente, possibilitando descobrir direções em que há maior variabilidade dos dados.

E. Decomposição dos Autovalores e Autovetores

Essa é a etapa principal do PCA. Nela, é necessário decompor a matriz de covariância em:

- Autovetores, os quais indicam um uma direção no espaço dos dados.
- Autovalores, que representam a quantidade de variaância explicada nessa direção.

Assim, ao ordenar os autovalores decrescentemente, os os autovetores correspondentes determinam as novas dimensões, e os autovalores normalizados quantificam quanta variância cada componente explica.

F. Projeção

Concluída a etapa anterior, é necessário projetar os dados originais nos novos eixos, a fim de reduzir a dimensionalidade, mas mantendo a maior parte da variabilidade dos dados. Ao fim desta etapa, é possível gerar uma visualização transformada 2D - neste dataset, em específico - dos dados complexos.

III. Resultados e Discussão

A. Análise Monovariada

A análise monovariada incondicional mostrou algumas informações interessantes sobre o comportamento das variáveis, bem como alguns padrões e distorções dos dados.

Ao observar os dados tabulares mostrados e os plots gerados no Jupyter Notebook, é notável que:

- As variáveis *Age* e *NCP* apresentam muitos outliers na respresentação de boxplot. Isso ocore pois as frequências dessas variáveis estão muito concentradas em um ponto, apresentando grande assimetria. Assim, é imprescindível analisar com cuidado essas métricas, visto que, aqui, os outliers não representam necessariamente erros, mas valores em regiões pouco povoadas da distribuição. No caso de *Age*, isso representa que há uma maioria jovem e um pequeno subgrupo de idosos. Já em *NCP*, os outliers podem representar hábitos alimentares extremos, que caracterizam comportamentos alimentares relevantes à análise do risco de obesidade.

Variável	**Média (μ)**	**Desvio padrão (σ)**	**Assimetria (γ)**
Age	24.31	6.35	1.53
Height	1.70	0.09	-0.01
Weight	86.59	26.19	0.26
FCVC	2.42	0.53	-0.43
NCP	2.69	0.78	-1.11
CH2O	2.01	0.61	-0.10
FAF	1.01	0.85	0.50
TUE	0.66	0.61	0.62

TABLE II
Resumo da Análise Monovariada Incondicional

- Os atributos *Height* e *Weight* apresentam uma distribuição proxima à distribuição normal, seguindo um comportamento equilibrado. Apesar de *Weight* possuir uma assimetria maior que a de *Height*, ainda possui a maior parte dos valores em torno da média. Esse tipo de comportamento ocasiona menor necessidade de transformar, normalizar ou padronizar essas variáveis.

- As outras variáveis (*FCVC,CH2O,FAF* e *TUE*) possuem algo em comum: possuem distribuição com alguns picos e possuem assimetria moderada (exceto *CH2O*, que possui baixa assimetria), indicando que há mais casos com valores menores. Esse comportamento de picos ocorre devido a muitas dessas variáveis representarem quantidades discretas ou inteiras. Vale ressaltar que, a assimetrias, nesses casos, não representam outlier, mas possíveis comportamento raros, com impacto na análise clinica.

Já a análise monovariada condicional fornece alguns insights interessantes sobre como os preditores se comportam em função do nível de obesidade, dentre os quais:

- Observou-se que quatro variáveis apresentam menor assimetria na classe Overweight Level I, indicando distribuições mais equilibradas e homogêneas nessa faixa de obesidade. Esse comportamento sugere que essa classe atua como um ponto de transição entre indivíduos da amostra populacional, refletindo padrões médios de comportamento e medidas corporais. Por outro lado, variáveis como *Age, FAF, TUE* e *NCP* não seguem essa tendência, apresentando assimetrias mais acentuadas, possivelmente devido à maior variabilidade de fatores comportamentais e demográficos.

- As variáveis *Height,FCVC* e *CH2O* apresentam variação significativa na assimetria entre os níveis de obesidade, indicando que esses atributos possuem distribuição fortemente dependente da classe. Além disso, são variáveis relevantes para a análise, visto que essas mudanças na assimetria podem indicar a existência de subgrupos com

comportamentos distintos. Por exemplo, ,m *FCVC* (consumo de vegetais), há assimetria positiva em *Obesity Type I* e negativa em *Obesity Type II* pode indicar que dentro de cada grupo há padrões distintos de alimentação saudável, refletindo hábitos e estilos de vida diversos.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	19.783	2.670	2.759
Normal_Weight	21.739	5.097	3.135
Obesity_Type_I	25.885	7.756	1.024
Obesity_Type_II	28.234	4.868	0.821
Obesity_Type_III	23.496	2.764	-0.518
Overweight_Level_I	23.418	6.125	1.691
Overweight_Level_II	26.997	8.061	1.184

TABLE III
ANÁLISE CONDICIONAL DE AGE POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	1.691	0.100	-0.101
Normal_Weight	1.677	0.095	0.365
Obesity_Type_I	1.694	0.098	0.217
Obesity_Type_II	1.772	0.073	-0.318
Obesity_Type_III	1.688	0.065	0.439
Overweight_Level_I	1.688	0.096	0.043
Overweight_Level_II	1.704	0.089	-0.401

TABLE IV
ANÁLISE CONDICIONAL DE HEIGHT POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	49.906	6.011	0.188
Normal_Weight	62.155	9.296	0.303
Obesity_Type_I	92.870	11.486	0.349
Obesity_Type_II	115.305	8.024	-0.552
Obesity_Type_III	120.941	15.532	0.693
Overweight_Level_I	74.267	8.471	0.009
Overweight_Level_II	82.085	8.451	-0.226

TABLE V
ANÁLISE CONDICIONAL DE WEIGHT POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	2.481	0.585	-1.007
Normal_Weight	2.335	0.591	-0.254
Obesity_Type_I	2.186	0.432	0.233
Obesity_Type_II	2.391	0.490	-0.592
Obesity_Type_III	3.000	0.000	NaN
Overweight_Level_I	2.265	0.483	0.053
Overweight_Level_II	2.261	0.453	0.194

TABLE VI
ANÁLISE CONDICIONAL DE FCVC POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	2.914	0.901	-0.990
Normal_Weight	2.739	0.872	-1.117
Obesity_Type_I	2.432	0.789	-0.911
Obesity_Type_II	2.745	0.579	-1.536
Obesity_Type_III	3.000	0.000	NaN
Overweight_Level_I	2.504	0.963	-0.551
Overweight_Level_II	2.496	0.753	-0.869

TABLE VII
ANÁLISE CONDICIONAL DE NCP POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	1.871	0.602	0.120
Normal_Weight	1.850	0.638	0.139
Obesity_Type_I	2.112	0.625	-0.207
Obesity_Type_II	1.878	0.553	-0.126
Obesity_Type_III	2.208	0.604	-0.636
Overweight_Level_I	2.059	0.615	-0.092
Overweight_Level_II	2.025	0.554	0.052

TABLE VIII
ANÁLISE CONDICIONAL DE CH2O POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	1.250	0.857	-0.135
Normal_Weight	1.247	1.016	0.312
Obesity_Type_I	0.987	0.895	0.608
Obesity_Type_II	0.972	0.581	-0.060
Obesity_Type_III	0.665	0.733	0.537
Overweight_Level_I	1.057	0.852	0.646
Overweight_Level_II	0.958	0.825	0.678

TABLE IX
ANÁLISE CONDICIONAL DE FAF POR CATEGORIA DE NOBEYESDAD.

NObeyesdad	Média (μ)	Desvio Padrão (σ)	Assimetria (γ)
Insufficient_Weight	0.839	0.643	0.203
Normal_Weight	0.676	0.687	0.518
Obesity_Type_I	0.677	0.688	0.564
Obesity_Type_II	0.515	0.564	0.908
Obesity_Type_III	0.605	0.282	-0.455
Overweight_Level_I	0.613	0.678	0.792
Overweight_Level_II	0.697	0.588	0.403

TABLE X
ANÁLISE CONDICIONAL DE TUE POR CATEGORIA DE NOBEYESDAD.

B. Análise Bivariada

A análise bivariada revelou padrões significativos entre as variáveis de hábitos alimentares, fatores comportamentais e indicadores antropométricos, permitindo compreender de forma integrada os elementos que mais influenciam os níveis de obesidade. Embora várias combinações de variáveis apresentem distribuições dispersas, alguns eixos de associação se destacam pela consistência e relevância epidemiológica.

Inicialmente, verificou-se que as variáveis alimentares, como o número de refeições principais (NCP), o consumo de água (CH2O) e a frequência de ingestão de vegetais (FCVC), mantêm distribuições semelhantes entre todas as classes de obesidade. Esse padrão, visível no diagrama *FCVC* × *CH2O*, sugere que práticas alimentares equilibradas estão presentes de forma ampla na amostra, inclusive entre indivíduos com níveis mais elevados de obesidade. Dessa forma, fatores dietéticos isolados não se mostraram determinantes para a diferenciação entre classes, apontando para uma relativa homogeneidade nos hábitos alimentares autorrelatados.

Em contrapartida, as variáveis associadas ao comportamento físico — especialmente a frequência de atividade física (FAF)

e o tempo de uso de dispositivos eletrônicos (TUE) — apresentaram as relações mais expressivas com o peso corporal. Os gráficos *Weight* × *FAF* e *Weight* × *TUE* evidenciam uma clara tendência inversa: indivíduos com menor nível de atividade física e maior tempo de exposição a telas concentram-se nas classes *Obesity Type II* e *Obesity Type III*, enquanto aqueles com maior FAF e menor TUE distribuem-se predominantemente entre *Normal Weight* e *Overweight Level I*. Tal comportamento reforça o papel crítico do sedentarismo como fator de risco dominante, superando inclusive a influência dos hábitos alimentares na explicação do ganho de peso.

As combinações de variáveis mistas, como *FCVC* × *TUE*, reforçam essa interpretação. Mesmo entre participantes que relatam consumo adequado de vegetais e número regular de refeições, observam-se elevados tempos de uso de dispositivos eletrônicos, indicando que bons hábitos alimentares não neutralizam os efeitos adversos de um estilo de vida sedentário. Assim, o comportamento físico se destaca como o principal determinante observável da obesidade nas dimensões bivariadas.

Do ponto de vista antropométrico, a relação *Height* × *Weight* apresentou o padrão mais esperado: correlação linear positiva, com as classes de obesidade distribuídas ordenadamente ao longo do eixo do peso, validando a consistência da classificação por IMC utilizada no conjunto de dados. Ademais, a variável altura mostrou-se relativamente independente dos demais fatores, sugerindo que o aumento da massa corporal é predominantemente explicado por desequilíbrios entre consumo e gasto energético, e não por diferenças estruturais de porte físico.

Os resultados observados ao relacionar o peso com variáveis de estilo de vida (*Weight* × *FAF*, *Weight* × *FCVC* e *Weight* × *TUE*) consolidam uma visão integrada do fenômeno: embora os padrões alimentares sejam majoritariamente adequados, o déficit de atividade física e o comportamento sedentário sustentam os níveis mais altos de obesidade. Essa constatação reforça achados da literatura que destacam a inatividade como o principal fator de risco comportamental, sobretudo em contextos urbanos marcados por alta conectividade digital e baixo dispêndio energético diário.

Em síntese, as principais relações identificadas evidenciam que:

- Hábitos alimentares (CH2O, FCVC e NCP) são consistentes entre as classes e pouco discriminantes para a obesidade;
- Fatores comportamentais (FAF e TUE) apresentam forte correlação inversa com o peso corporal, sendo os principais marcadores visuais das classes de obesidade;
- As variáveis antropométricas (Height e Weight) mantêm coerência interna e confirmam o gradiente de massa corporal esperado;
- A combinação entre bons hábitos alimentares e baixo nível de atividade física caracteriza um perfil comum de obesidade comportamental, mais associado ao sedentarismo do que à dieta em si.

Esses achados reforçam a necessidade de estratégias de saúde pública voltadas não apenas à alimentação equilibrada, mas também à redução do tempo sedentário e ao incentivo à prática regular de atividades físicas como pilares centrais no enfrentamento da obesidade.

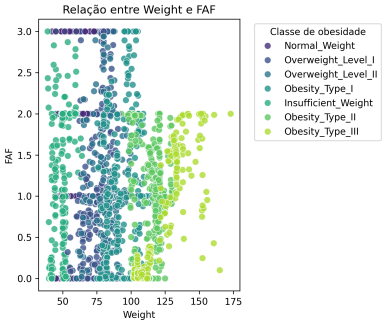

Fig. 1. Relação entre o peso corporal (Weight) e a frequência de atividade física (FAF).

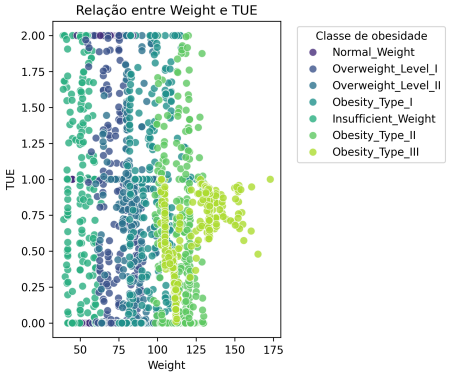

Fig. 2. Relação entre o peso corporal (Weight) e o tempo de uso de dispositivos eletrônicos (TUE).

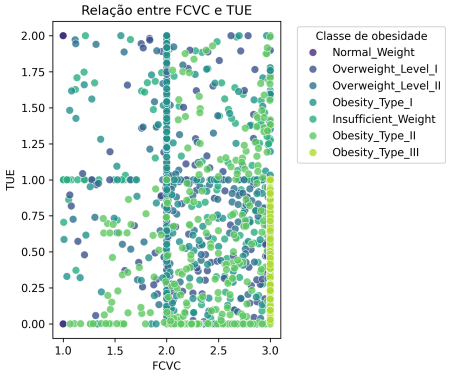

Fig. 3. Relação entre a frequência de consumo de vegetais (FCVC) e o tempo de uso de dispositivos eletrônicos (TUE).

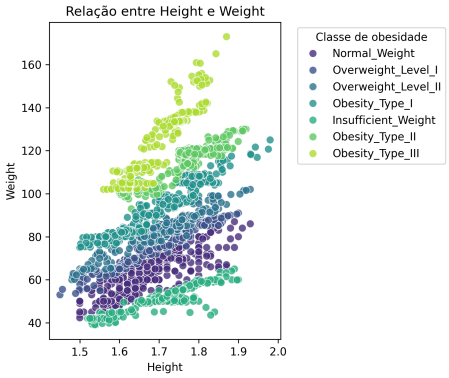

Fig. 4. Correlação entre altura (Height) e peso corporal (Weight) por nível de obesidade.

C. Análise Multivariada (PCA)

A projeção nos dois primeiros componentes principais (Fig. 5) evidencia um gradiente claro ao longo do eixo PC1, no qual as classes associadas a maior adiposidade tendem a deslocar-se para a direita, enquanto *Insufficient_Weight* e *Normal_Weight* concentram-se majoritariamente no semieixo esquerdo. Observa-se uma zona central de sobreposição entre *Overweight_Level_I*, *Overweight_Level_II* e *Obesity_Type_I*, compatível com a continuidade do espectro ponderal. Os contornos de densidade ajudam a distinguir padrões latentes: *Obesity_Type_III* forma um núcleo mais compacto no quadrante superior direito, enquanto *Obesity_Type_II* ocupa preferencialmente a porção direita inferior, sugerindo que, além do eixo de adiposidade capturado por PC1, há um segundo eixo (PC2) que separa subgrupos de obesidade segundo características de estilo de vida.

A leitura dos carregamentos indica que PC1 agrega sobretudo variáveis antropométricas e marcadores de ingestão calórica, funcionando como um eixo de adiposidade, ao passo que PC2 reúne variáveis associadas a atividade e sedentarismo (por exemplo, *FAF* e *TUE*), servindo como um eixo de balanço comportamental. Nesse plano, classes mais altas de obesidade aparecem em regiões de PC1 elevado e, a depender do padrão de PC2, distribuem-se entre quadrantes superior e inferior direitos, ao passo que classes com menor peso relativo aproximam-se da origem ou de valores negativos de PC1. O arranjo global confirma que uma combinação linear de poucos atributos é suficiente para sintetizar a estrutura do conjunto, ainda que a sobreposição entre classes adjacentes permaneça relevante e impeça separações lineares estritas.

Do ponto de vista prático, a PCA cumpre papel descritivo e comunicativo: os dois eixos latentes capturam dimensões epidemiologicamente plausíveis (adiposidade e atividade–sedentarismo), fornecendo uma visão compacta da heterogeneidade do conjunto e guiando etapas subsequentes de modelagem. Ao mesmo tempo, a presença de áreas de interpenetração entre classes sugere que tarefas de classificação exigirão técnicas supervisionadas e, possivelmente, variáveis adicionais ou transformações não lineares para ganhos discriminativos além do que se observa na projeção PC1–PC2.

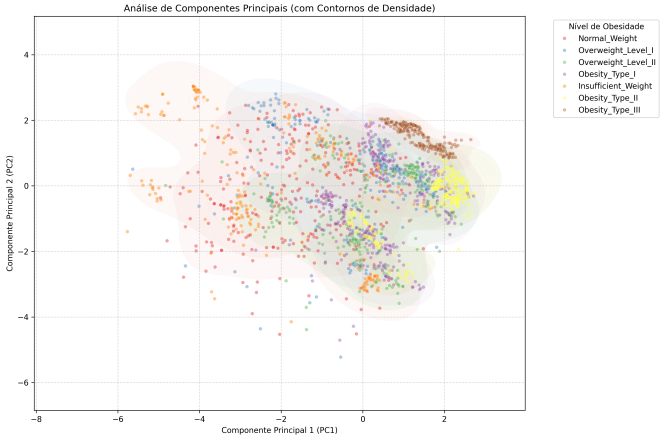

Fig. 5. Projeção nos dois primeiros componentes principais colorida por *NObesity*, com contornos de densidade por classe.

IV. Conclusão

A análise exploratória de dados demonstrou ser uma ferramenta essencial para a compreensão dos fatores associados à obesidade e aos riscos cardiovasculares. Através do estudo detalhado das 2.111 amostras do dataset "Obesity or CVD Risk", foi possível identificar relações relevantes entre hábitos alimentares, nível de atividade física e indicadores antropométricos. Os resultados evidenciaram que o sedentarismo e o uso prolongado de dispositivos eletrônicos apresentam forte correlação inversa com o peso corporal, constituindo os principais fatores comportamentais associados aos níveis mais elevados de obesidade.

Além disso, observou-se que, embora os padrões alimentares relatados se mantenham relativamente equilibrados entre as classes de peso, o déficit de atividade física se destaca como o principal determinante da obesidade comportamental. A análise multivariada (PCA) confirmou essa tendência ao revelar um gradiente de adiposidade ao longo do primeiro componente principal e um eixo secundário de atividade–sedentarismo, sintetizando de forma eficaz a estrutura do conjunto de dados.

De modo geral, a aplicação de técnicas de EDA proporcionou uma visão integrada e intuitiva dos fatores de risco, permitindo não apenas identificar variáveis críticas, mas também orientar estratégias de prevenção e políticas públicas de saúde voltadas à promoção de estilos de vida mais ativos e equilibrados.

References

[1] Kaggle, "Obesity or CVD Risk Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data

[2] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, New York, NY: Springer, 2013.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, NY: Springer, 2008.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in Python*, New York, NY: Springer, 2023.