**UNIVERSIDADE PRESBITERIANA MACKENZIE**

**CURSO DE TECNOLOGIA EM CIÊNCIA DE DADOS**

**ANDERSON APARECIDO DA SILVA ALVES**

**ANDRÉIA DOMINGOS DOS SANTOS**

**GERSON SOARES RODRIGUES**

**SAMUEL BONFIM DA SILVA**

PREDIÇÃO MULTICLASSE DE RISCO À OBESIDADE

**SÃO PAULO**

**2024**

**ANDERSON APARECIDO DA SILVA ALVES**

**ANDRÉIA DOMINGOS DOS SANTOS**

**GERSON SOARES RODRIGUES**

**SAMUEL BONFIM DA SILVA**

**PREDIÇÃO MULTICLASSE DE RISCO À OBESIDADE**  
Análise de dados fornecidos pela empresa Kaggle

Trabalho acadêmico para conclusão da disciplina de Projeto Aplicado I do Curso de Tecnologia em Ciência de Dados pela Universidade Presbiteriana Mackenzie.

Professor: Felipe Albino dos Santos

**SÃO PAULO**

**2024**

A001

Silva, Anderson Aparecido da Silva.

Dos Santos, Andréia Domingos.  
Rodrigues, Gerson Soares.

Da Silva, Samuel Bonfim.

Predição Multiclasse de Risco à Saúde / Anderson Aparecido da Silva Alves; Andréia Domingos dos Santos; Gerson Soares Rodrigues; Samuel Bonfim da Silva - 2024. 23 f.: il.; 30 cm.

Trabalho Acadêmico (Projeto Aplicado I) – Universidade Presbiteriana Mackenzie,

Presbiteriana Mackenzie, São Paulo, 2024.

Bibliografia: f. 8.

1. Ciência. 2. Tecnologia. 3. Sociedade.

CDD01

A001 Rodrigues, Gerson Soares.

Relação entre Ciência, Tecnologia e Sociedade / Gerson Soares Rodrigues - 2023. 6 f.: il.; 30 cm.

Trabalho Acadêmico (Ciência, Tecnologia e Sociedade) – Universidade Presbiteriana Mackenzie,

Presbiteriana Mackenzie, São Paulo, 2022.

Bibliografia: f. 6.

1. Ciência. 2. Tecnologia. 3. Sociedade.

CDD01

SUMÁRIO

[1. INTRODUÇÃO 1](#_Toc166678481)

[2. GLOSSÁRIO 1](#_Toc166678482)

[3. PREMISSAS E APRESENTAÇÃO DA EMPRESA 2](#_Toc166678483)

[4. OBJETIVO E METAS 3](#_Toc166678484)

[5. ESCOLHA DA TEMÁTICA 4](#_Toc166678485)

[6. COMUNICAÇÃO 4](#_Toc166678486)

[7. PROPOSTA PRODUTO ANALÍTICO E ANÁLISE EXPLORATÓRIA DE DADOS 4](#_Toc166678487)

[7.1 PENSAMENTO COMPUTACIONAL EM CONTEXTOS ORGANIZACIONAIS 4](#_Toc166678488)

[8. SCRIPT DA ANÁLISE EXPLORATÓRIA DOS DADOS DO PROJETO 4](#_Toc166678489)

[8.1 IDENTIFICAÇÃO DE OUTLIERS 14](#_Toc166678490)

[9. OBJETIVOS OBTIDOS POR MEIO DAS ANÁLISES ANTERIORES 19](#_Toc166678491)

[9.1 CONVERSÃO DAS VARIÁVEIS CATEGÓRICAS 'Gender' E 'family\_history\_with\_overweight' EM NUMÉRICAS 19](#_Toc166678492)

[9.2 CONVERSÃO VARIÁVEL CATEGÓRICA DE IMC EM NUMÉRICA 20](#_Toc166678493)

[10. CONCLUSÃO 22](#_Toc166678494)

[11. REFERÊNCIA BIBLIOGRÁFICA 23](#_Toc166678495)

Parte superior do formulário

1. INTRODUÇÃO

A obesidade é uma condição global que afeta predominantemente milhões de adolescentes e adultos, sendo ocasionada por doenças pré-existentes, fatores hereditários e também pelo aumento da ingestão de alimentos ricos em calorias e pela diminuição da atividade física. Desde 1980, a prevalência da obesidade tem aumentado significativamente, tornando-se um problema global de saúde pública. Além de ser um risco bem estabelecido a uma série de problemas de saúde, buscamos saber se existem algumas predisposições, tanto no sentido de estilo de vida quanto de fatores hereditários. Para combater esse problema, diversas ferramentas e soluções foram desenvolvidas, incluindo a mineração de dados, que permite identificar padrões e prever a obesidade com alta precisão.

Neste projeto, exploraremos a relação entre obesidade e suas predisposições, utilizando dados fornecidos pela National Library of Medicine – NIH através da plataforma Kaggle que utilizamos para a realização de nossa análise inicial.

1. GLOSSÁRIO

Foram disponibilizados pela plataforma Kaggle alguns dados os quais foram coletados (em formato CSV) que serão armazenados no GitHub do projeto e analisados nas próximas etapas;

Os dados contêm valores numéricos e categóricos, que serão utilizados para análise baseada em algoritmos de classificação, predição, segmentação e associação. Os dados estão disponíveis em formato CSV. Os conjuntos de dados que estão presentes nas colunas que serão utilizadas na análise do projeto são os seguintes:

* id - Identificação
* Gender - Gênero
* Age - Idade
* Height - Altura
* Weight - Peso
* family\_history\_with\_overweight - histórico familiar com sobrepeso
* FAVC - Frequência de consumo de alimentos calóricos
* FCVC - Frequência de consumo de Vegetais
* NCP - Número de refeições principais
* CAEC - Consumo de alimentos durantes das refeições
* SMOKE - Fumante ou não
* CH2O - Consumo diário de água
* SCC - Monitoramento de consumo calórico
* FAF - Frequência de Atividades Físicas
* TUE - Tempo Utilizado com Dispositivos Eletrônicos
* CALC - Consumo de Álcool
* MTRANS - Meio de transporte utilizado
* NObeyesdad - Nível de obesidade (Abaixo, Normal, Sobrepeso, Obesidade Grau I, Obesidade Grau II, Obesidade Grau IIII

Com o desenvolvimento do Script, nos deparamos com algumas informações categóricas que estão inseridas no conjunto de dados presentes nas colunas, as quais serão descritas abaixo:

* Female – Feminino
* Male – Masculino
* Yes – Sim
* No – Não
* Sometimes – Ocasionalmente
* Frequently – Frequentemente
* Always – Sempre
* Public Transportation – Transporte Público
* Walking – Caminhando
* Automobile – Automóvel
* Motorbike – Moto
* Bike – Bicicleta
* Insufficient Weight – Peso Abaixo do Normal
* Normal Weight – Peso Normal
* Overweight Level I – Sobrepeso Nível I
* Overweight Level II – Sobrepeso Nível II
* Obesity Type I – Obesidade Tipo I
* Obesity Type II – Obesidade Tipo II
* Obesity Type III – Obesidade Tipo III

1. PREMISSAS E APRESENTAÇÃO DA EMPRESA

A empresa Kaggle é conhecida por suas competições entre o público de Data Science, com uma comunidade diversificada chegando a mais de 4 milhões de membros. Ela também é responsável por ajudar a comunidade, lançando alguns desafios com o intuito de praticar as habilidades sobre a análise de dados. Visando isso, aceitamos o desafio proposto pela Kaggle, da Tabular Playground Series, especificamente na quarta temporada, episódio 2, onde o nosso desafio será analisar dados fornecidos pela empresa sobre a “Multi-Class Prediction of Obesity Risk”. Tentamos fazer uma análise nos baseando nos padrões brasileiros através do website dados.gov.br, no entanto, como não obtivemos dados concisos referentes ao tema abordado, utilizamos o website do NIH (National Library of Medicine) em nos foram fornecidos dados para estimativa dos níveis de obesidade com base em hábitos alimentares, fatores hereditários, estilo de vida e condição física em indivíduos situados na Colômbia, Peru e México e por serem países latino-americanos, assim como o Brasil, optamos por usá-los como nossa base de dados reais. O objetivo é utilizar os vários fatores fornecidos, para prever o risco de obesidade em indivíduos, os quais podem estar relacionados ao seu estilo de vida e fatores hereditários para implementarmos tais métodos à realidade da população brasileira através da atenção primária.

1. OBJETIVO E METAS

O projeto tem como objetivo destacar a importância da prevenção de doenças crônicas, relacionadas à obesidade, e a necessidade de implementar programas direcionados a essas condições. Além disso, colaborar com a redução de hospitalizações e tratamentos emergenciais e promover estilos de vida saudáveis. Essas medidas têm como ênfase a atenção primária e preventiva, contribuindo assim para a redução dos custos associados à busca por especialistas e diminuição da progressão de doenças que exijam tratamentos de alta complexidade. É essencial frisar que os bancos de dados do DataSUS não contêm estudos abrangentes sobre a prevenção da obesidade e das doenças relacionadas a ela. Portanto, este projeto visa prevenir a obesidade e as doenças decorrentes dessa condição, para melhora da qualidade de vida da população brasileira e, consequentemente, reduzir os gastos públicos relacionados ao tratamento de doenças crônicas.

Através dos dados obtidos, na plataforma Kaggle, adotamos uma abordagem multifatorial para prever o risco de obesidade. Para entendermos e anteciparmos os fatores que levem ao aparecimento e progressão da obesidade. Deste modo, teremos como metas, ao realizar a análise exploratória sobre os dados, adotar intervenções preventivas e terapêuticas mais assertivas. Essas intervenções abrangem a identificação precoce de padrões de risco e a customização de estratégias de intervenção, considerando os dados fornecidos pela plataforma.

1. ESCOLHA DA TEMÁTICA

Após uma pesquisa de artigos publicados na plataforma Kaggle, foi consenso de todos os participantes a escolha do tema a ser utilizado visando a implementação deste mesmo tipo de análise para os padrões da realidade brasileira.

1. COMUNICAÇÃO

Foram feitas reuniões periódicas entre os participantes do grupo para a organização do assunto abordado da temática escolhida e a delegação de tarefas para cada membro do grupo.

1. PROPOSTA PRODUTO ANALÍTICO E ANÁLISE EXPLORATÓRIA DE DADOS
   1. Pensamento computacional em contextos organizacionais

*Definição do Problema:* Investigar a relação entre obesidade e suas predisposições relacionadas à estilo de vida e fatores hereditários. Diante do aumento global da obesidade como um desafio de saúde pública.

*Coleta de Dados*: Utilização de dados disponibilizados pela Kaggle através do desafio da Playground Series, temporada 4, episódio 2, focado na Previsão Multiclasse do Risco de Obesidade e realizar com dados reais de países latino-americanos obtidos através da National Library of Medicine.

*Análise*: Adoção de uma abordagem multifatorial para analisar os dados, visando prever o risco de obesidade e sua relação com o estilo de vida de cada indivíduo da amostra além de seus fatores hereditários.

*Desenvolvimento do Modelo:* Criação de modelos preditivos baseados nos dados explorados para identificar padrões de risco e antecipar o surgimento da obesidade.

*Intervenção:* Proposição de intervenções preventivas e terapêuticas personalizadas, fundamentadas na análise dos dados.

*Impacto:* Fornecer insights para influenciar políticas de saúde pública, práticas clínicas e promover a saúde, enfrentando o desafio da obesidade e doenças relacionadas, além de a contribuição à redução de custos com especialista e tratamentos para doenças de alta complexidade.

1. Script da Análise Exploratória dos Dados do Projeto

Iniciamos nosso projeto utilizando como compilador dos dados o Google Colaboratory, comumente chamado de Google Colab. O script foi escrito na linguagem Python e nele foi feito a importação de algumas bibliotecas, que foram utilizadas para o melhoramento da análise descritiva dos dados, que poderão ser visualizadas a seguir:

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder

from sklearn.metrics import accuracy\_score

from google.colab import files

Após realizarmos a importação das bibliotecas, destacadas acima, foi feita a importação do arquivo em formado CSV do Dataset, presente na plataforma Kaggle, conforme ilustrado abaixo:

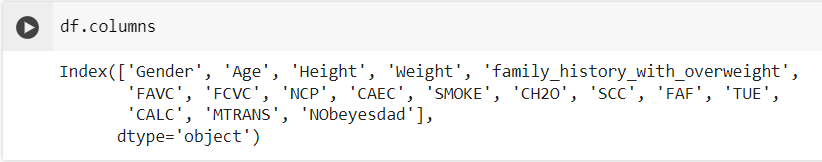
df = pd.read\_csv('/content/ObesityDataSet.csv')

Obs.: Para compilar a programação, a leitura do arquivo em formado csv (‘.csv‘) dependerá da descrição completa de todo o caminho até onde o usuário o inseriu.

Logo após a leitura da planilha contendo o Dataset, foi feita a apresentação bruta dos dados:

Inicialmente, através do comando *columns*, fizemos a classificação dos dados presentes em cada coluna do dataframe, conforme ilustrado na Figura 1.

df.columns

  
Figura 1: Classificação dos dados presentes em cada coluna do dataframe

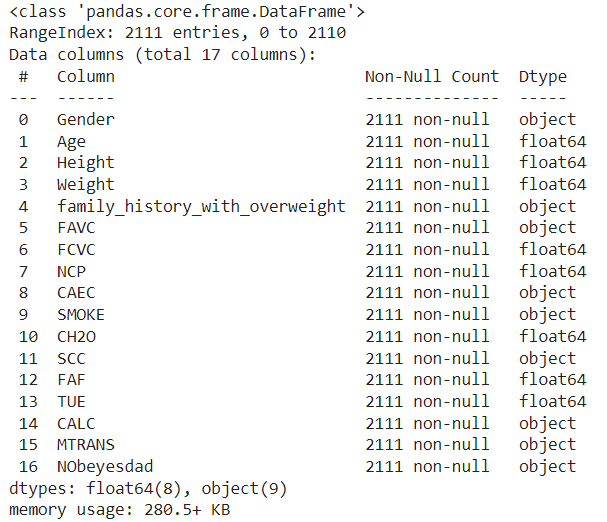
Feita a classificação, utilizamos o comando *head* para verificarmos os primeiros valores de saída de cada coluna, conforme ilustrado na Figura 2:

df.head()

  
Figura 2: Visualização dos primeiros valores presentes em cada coluna.

Através do comando *info*, analisamos as colunas e identificamos quais variáveis são categóricas, numéricas e também se há a presença de valores nulos, conforme ilustrado na Figura 3:

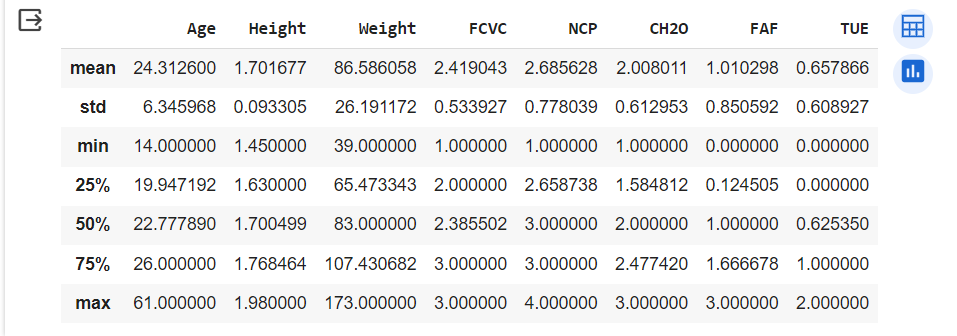
df.info()

  
Figura 3: Leitura das colunas presentes no Dataset para verificação das variáveis de valores categóricos, numéricos e valores nulos.

Através do comando *describe*, muito utilizado na biblioteca pandas para a análise de dados, nos fornecendo estatísticas descritivas como: valores de mediana, desvio padrão, mínimo, máximo e quartis, presentes somente nas colunas que possuem valores numéricos, conforme podemos observar na Figura 4.

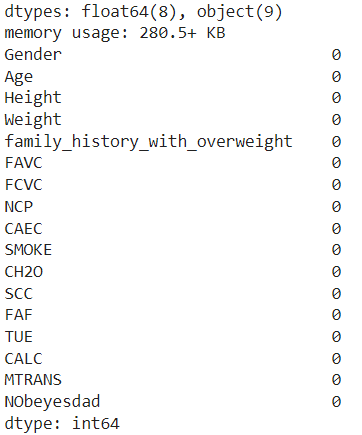
Obs.: No comando, retiramos o COUNT, devido se referenciar as linhas com exceção das colunas com variáveis categóricas.

df.describe().drop('count')

  
Figura 4: Valores de mediana, desvio padrão, mínimo, máximo e quartis, presentes somente nas colunas que possuem valores numéricos no dataframe.

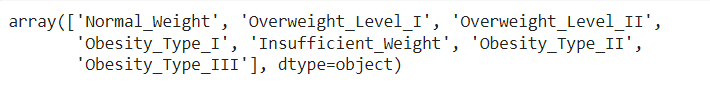
Por fim, para o teste final de valores nulos, utilizamos o comando *isnull().sum()*, da biblioteca pandas que verifica de valores ausente em um dataframe, para a leitura das colunas, e com isso, analisamos que não houve a presença de valores nulos, conforme ilustrado na Figura 5.

df.isnull().sum()# passou no teste, não há valores nulos

  
Figura 5: Leitura das colunas para a verificação de valores nulos.

Como análise inicial, relacionamos os dados à saúde de indivíduos, com uma coluna final, a qual mostra uma variável categórica representando o nível de obesidade. A seguir, através da Figura 6, será ilustrada a saída da linha de código utilizada para o processo:

df.NObeyesdad.unique()#podemos ver o nome de cada categoria

  
Figura 6: Saída da linha de comando, destacando de que forma foram classificados os níveis de Índice de Massa Corpórea (IMC) dos indivíduos presentes na amostra.

Através da função, que será demostrada abaixo, a qual irá especificar um gráfico de barras que mostre a distribuição das categorias em uma coluna categórica do dataframe, em que serão exibidas as porcentagens de ocorrência de cada categoria. Nela serão ajustados o tamanho da figura, definição de cores para as barras, adição de rótulos com porcentagens às barras, ajuste das margens do gráfico e rótulo do eixo y como "Porcentagem".

def analise\_coluna\_cat(df, coluna):

    plt.figure(figsize=(8, 6))

    ac = (df[coluna].value\_counts(normalize = True) \* 100 ) .plot(kind = 'bar', color = ['#20b2aa',

        'orange',

        '#b8860b',

        '#bacdff',

        '#ffbaf1',

        'yellow',

        '#fe7272']

    )

    ac.bar\_label(ac.containers[0] ,fmt='%.2f%%', label\_type = 'edge')

    ac.margins(y = 0.1)

    plt.ylabel('Porcentagem');

Caso o usuário deseje fazer as análises das porcentagens através dos gráficos com plot em pizza, deverá utiliza o código a seguir:

def analise\_coluna\_cat(df, coluna):

    plt.figure(figsize=(10, 8))

    valores = df[coluna].value\_counts(normalize=True)

    cores = ['#20b2aa', 'orange', '#b8860b', '#bacdff', '#ffbaf1', 'yellow', '#fe7272']

    explode = tuple([0.1] + [0.03] \* (len(valores) - 1))

    plt.pie(valores, labels=None, autopct=lambda p: '{:.0f}%'.format(p), pctdistance=1.1, labeldistance=1.1, colors=cores, explode=explode)

    plt.title(f'Distribuição de {coluna}')

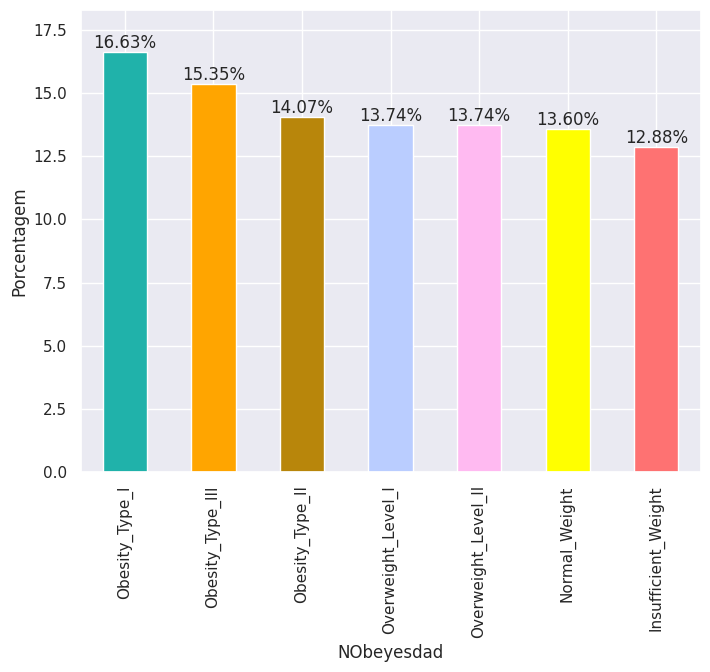
    plt.legend(valores.index, loc="center left", bbox\_to\_anchor=(1, 0, 0.5, 1))

    plt.show()

Após feitas as especificações acima, serão analisadas, de forma individual, a porcentagem de cada variável categórica. No caso os gráficos escolhidos para serem plotados nas análises seguintes será o plot em colunas.

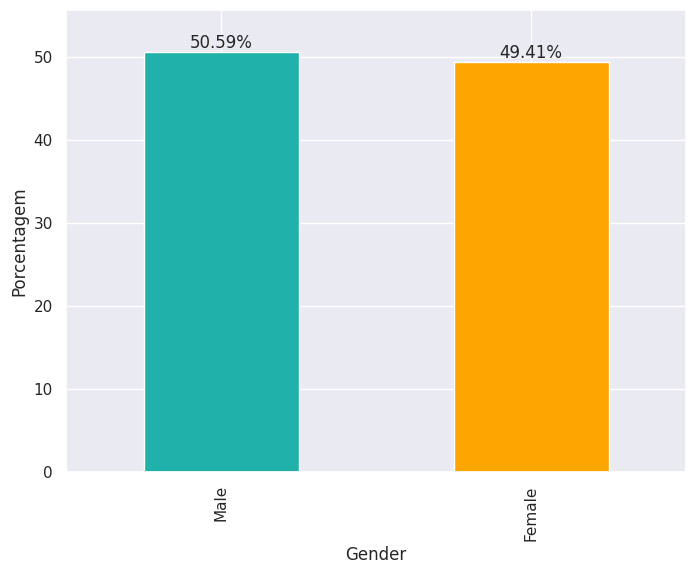
Na Figura 7, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra conforme seus níveis de IMCs, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'NObeyesdad')

  
Figura 7: Gráfico de porcentagem de indivíduos presentes em cada nível de IMC.

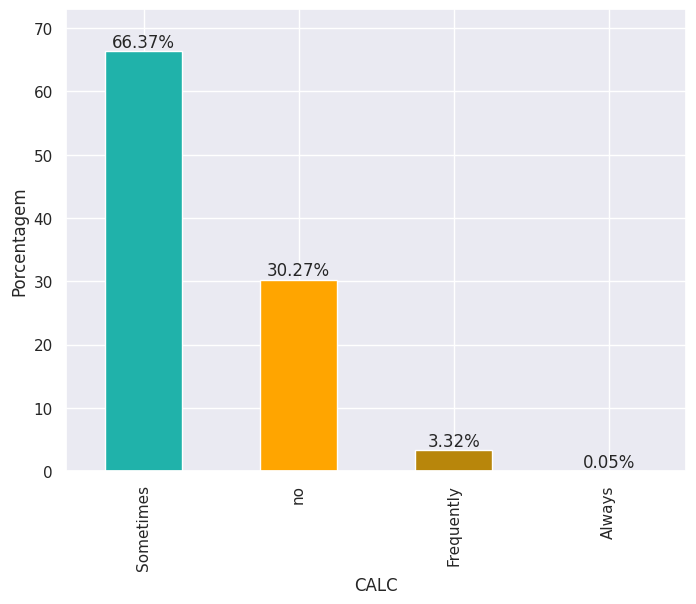
Na Figura 8, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra conforme estes são separados por gêneros, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'Gender')

  
Figura 8: Gráfico de porcentagem de indivíduos conforme o gênero.

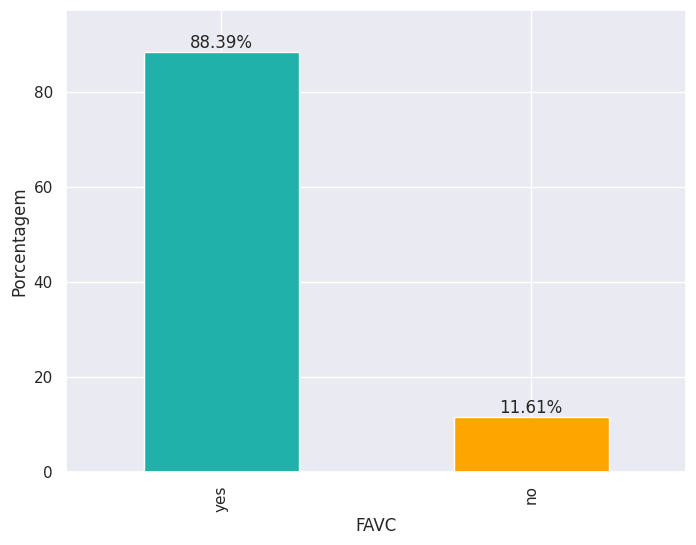
Na Figura 9, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com o consumo de bebidas alcoólicas, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'CALC')

  
Figura 9: Gráfico de porcentagem de indivíduos classificados pelo consumo de álcool.

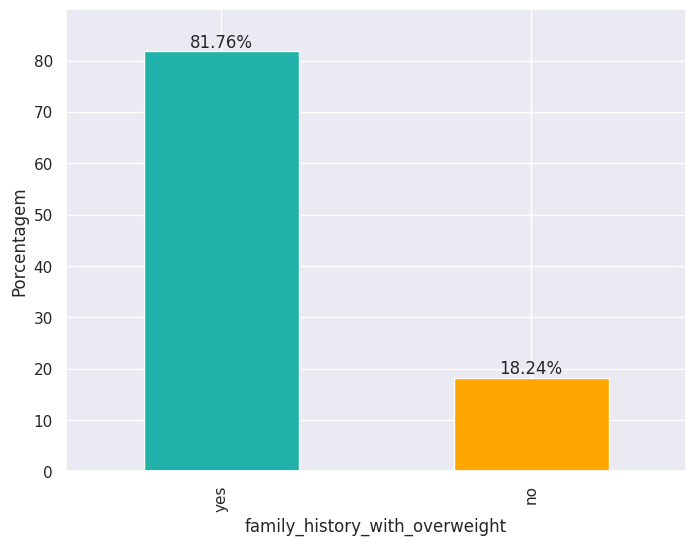
Na Figura 10, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com o consumo de alimentos com alto valor calórico, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'FAVC')

  
Figura 10: Gráfico de porcentagem de indivíduos que consomem alimentos de alto valor calórico.

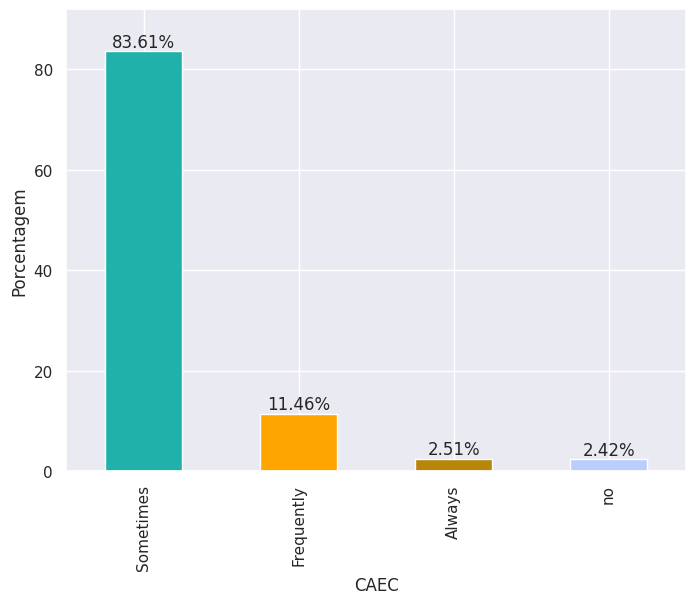
Na Figura 11, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com histórico familiar de sobrepeso, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'family\_history\_with\_overweight')

  
Figura 11: Gráfico de porcentagem de indivíduos que possuem histórico familiar de sobrepeso.

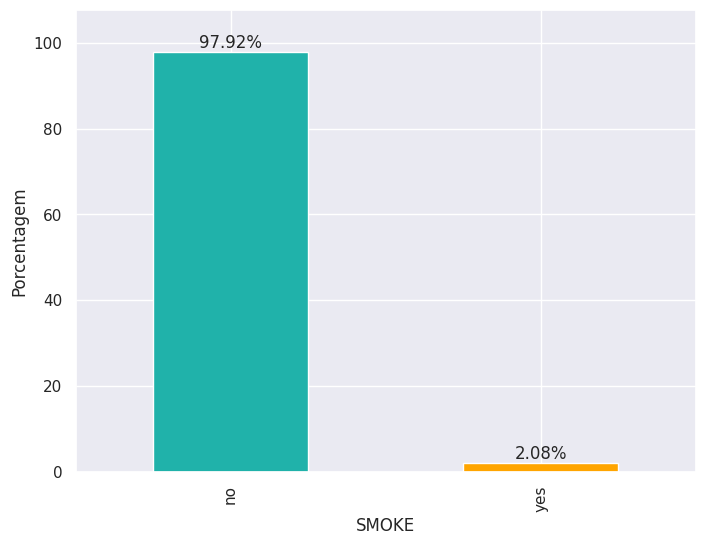
Na Figura 12, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com consumo de alimentos entre as refeições, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'CAEC')

  
Figura 12: Gráfico de porcentagem de indivíduos que consomem alimentos entre as refeições.

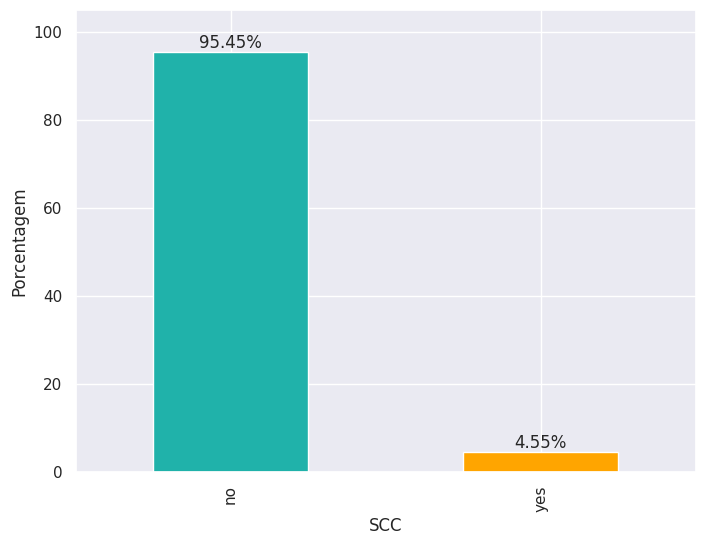
Na Figura 13, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com consumo cigarro, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'SMOKE')

  
Figura 13: Gráfico de porcentagem de indivíduos que são fumantes e não-fumantes.

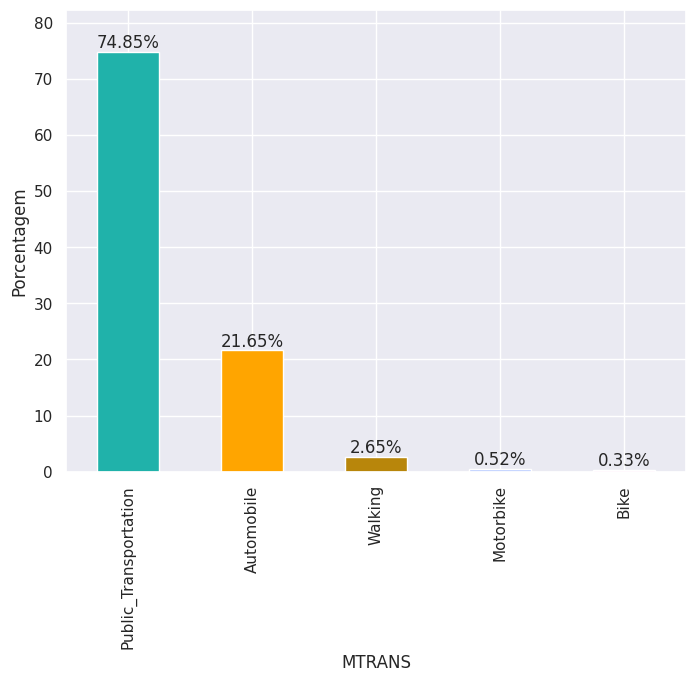
Na Figura 14, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados de acordo com monitoramento do consumo calórico, ou seja, se estes conferem os valores calóricos para refeição feita, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'SCC')

  
Figura 14: Gráfico de porcentagem de indivíduos que fazem monitoramento de consumo calórico.

E, conforme Figura 15, será ilustrado o gráfico com as porcentagens de indivíduos presentes na amostra classificados pelos meios de transporte utilizados para locomoção, valores estes obtidos através do código abaixo:

analise\_coluna\_cat(df, 'MTRANS')

  
Figura 15: Gráfico de porcentagem de indivíduos classificados pelos meios de transporte utilizados para locomoção.

* 1. Identificação de Outliers

Como teste dos outliers, decidimos plotar o gráfico de caixa (BoxPlot), ilustrado na Figura 16, em que poderemos visualizar a distribuição da idade entre os gêneros e seus respectivos outliers, através do código descrito abaixo:

plt.figure(figsize=(10,5))

plt.subplot(1,1,1) # (1 equivale ao padrão, o 1 tamanho ocupado nos quadrantes, 1 a posição em que ele está localizado)

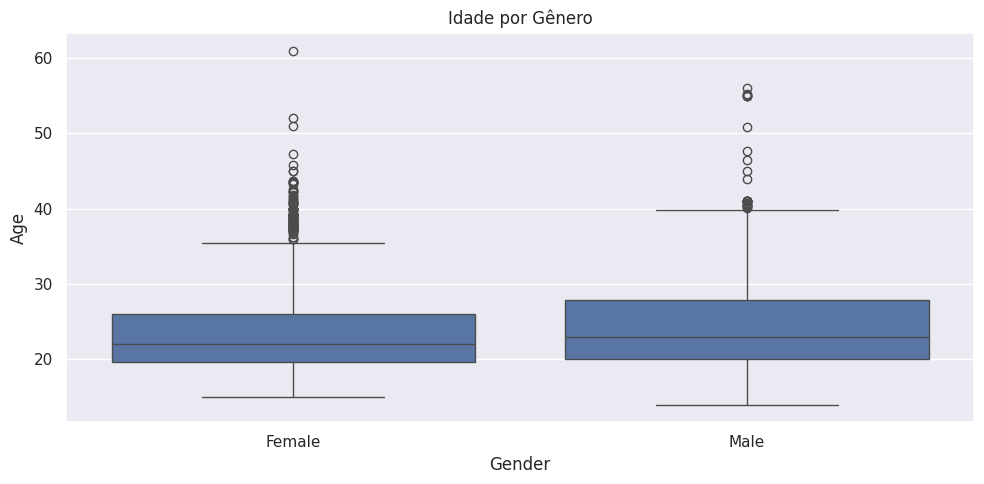
sns.boxplot( data=df, x='Gender', y= 'Age')

plt.title('Idade por Gênero')

plt.tight\_layout()

# Mostrar os gráficos

plt.show()

  
Figura 16: Gráfico de caixa (Boxplot) com a distribuição da idade entre os gêneros e seus respectivos outliers.

Analisando o Gráfico acima podemos notar que na amostra tanto a idade das mulheres quanto as dos homens se concentram entre 20 e 30 anos, sendo que nos homens a mediana é um pouco maior e os outliers presentes se concentram mais entre 35 e 45 anos nas mulheres e próximo dos 40 anos nos homens. Entretanto, mesmo com essa pequena variação a mediana na coluna gênero condiz com os valores presentes na Figura 4.

Na Figura 17, iremos ilustrar o gráfico em Boxplot contendo os Outliers referentes aos pesos e a altura conforme o gênero dos indivíduos junto de seus respectivos outliers, através do código descrito abaixo:

plt.figure(figsize=(16,5))

plt.subplot(1,2,1) # (1 equivale ao padrão, o 3 tamanho ocupado nos quadrantes, 1 a posição em que ele está localizado)

sns.boxplot( data=df, x='Gender', y= 'Weight')

plt.title('Peso por Gênero')

plt.subplot(1, 2, 2)

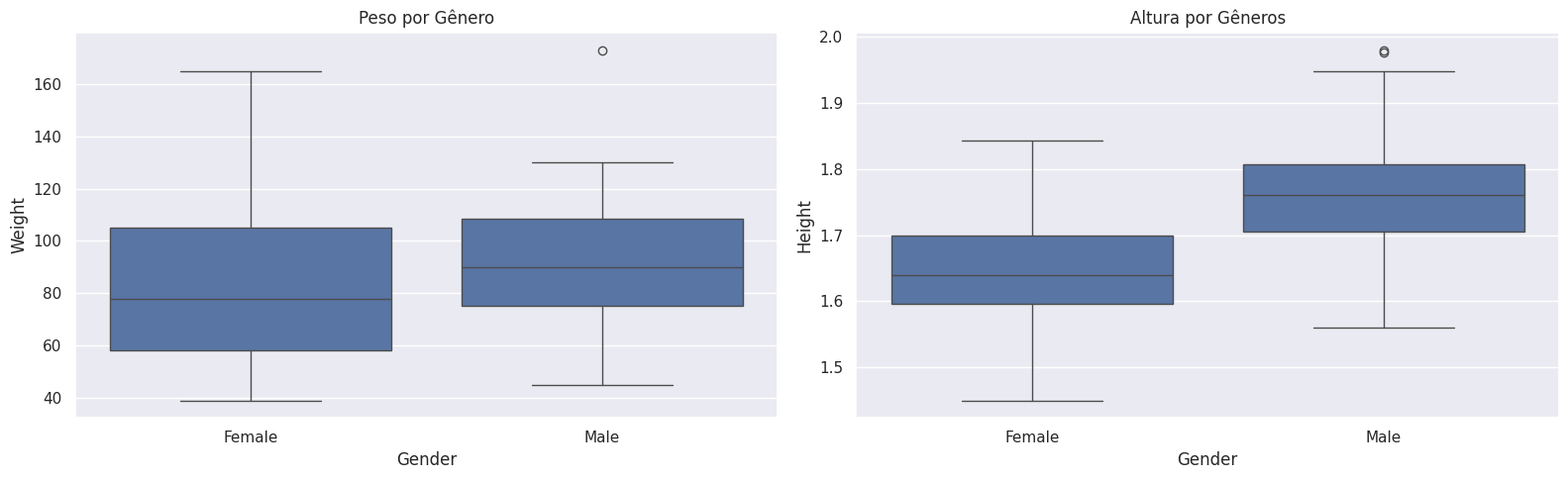
sns.boxplot(data=df, x='Gender', y='Height')

plt.title('Altura por Gêneros')

plt.tight\_layout()

# Mostrar os gráficos

plt.show()

  
Figura 17: Gráfico de caixa (Boxplot) com a distribuição do peso e altura entre os gêneros e seus respectivos outliers.

Analisando o Gráfico acima, na relação Peso por Gênero, podemos notar na amostra que há uma maior variação interquartis no peso das mulheres em relação ao dos homens, sendo a mediana das mulheres com valor próximo de 80Kg e a dos homens com valor próximo de 90Kg. Apesar de existir essa variação interquartis mais elevada entre as mulheres, estas possuem um menor valor de mediana e os valores próximos ao terceiro quartil estão próximos em ambos os gêneros. No entanto, quando analisamos a relação Altura por Gênero, podemos notar uma diferença significativa entre as medianas e os valores interquartis, sendo que as mulheres possuem como medianas alturas próximas à 1,65m e os homens com mediana próxima à 1,75m e tanto na mediana quantos nos valores interquartis, os homens possuem números mais elevados. Apesar, estas diferenças notamos ao analisar a mediana das alturas na Figura 4, o valor presente lá é um valor intermediário tanto nas às alturas quantos nos pesos entre os gêneros.

Na Figura 18, será ilustrado o gráfico em Boxplot contendo os outliers referentes aos históricos familiares de sobrepeso por Peso e Consumo de Alimentos de alta caloria por Peso junto de seus respectivos outliers, através do código descrito abaixo:

plt.figure(figsize=(16,5))

plt.subplot(1, 2, 1)

sns.boxplot(data=df,x='family\_history\_with\_overweight',y='Weight')

plt.title('Peso por Histórico Familiar Sobrepeso')

plt.subplot(1, 2, 2)

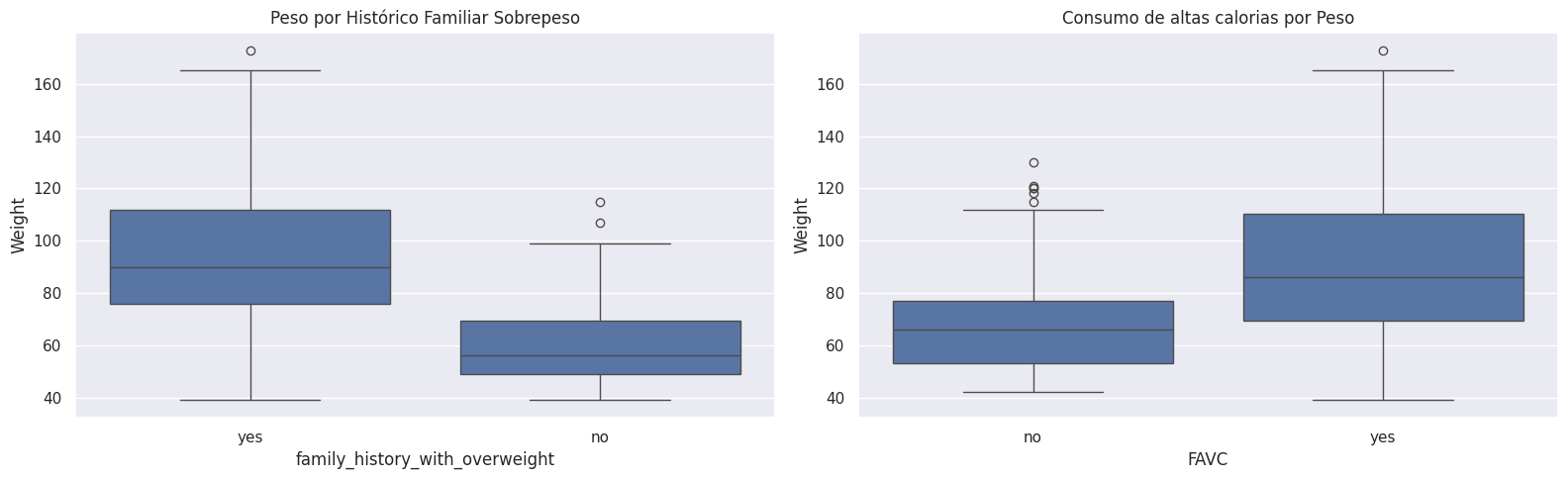
sns.boxplot(data=df, x='FAVC', y='Weight')

plt.title('Consumo de altas calorias por Peso')

plt.tight\_layout()

# Mostrar os gráficos

plt.show()

  
Figura 18: Gráfico de caixa (Boxplot) contendo os outliers referentes aos históricos familiares de sobrepeso por Peso e Consumo de Alimentos de alta caloria por Peso.

Analisando o Gráfico acima, na relação Peso por Histórico Familiar com Sobrepeso, podemos notar na amostra que há um maior ganho de peso nos indivíduos que possuam histórico na família de sobrepeso, com um valor de mediana próximo de 90Kg, e isto pode sugerir que estes tenham a tendência a se tornarem obesos quando comparados àqueles sem o histórico familiar que possuem valor de mediana próximo à 60Kg.

No caso dos indivíduos que consomem alimentos com altos valores calóricos, notamos a relação no aumento de peso, com valores interquartis de aproximadamente 70 a 110Kg e uma mediana próxima à 90Kg, quando comparados aos que não consomem estes tipos de alimentos que possuem valores interquartis de aproximadamente 50 a 80 Kg e uma mediana próxima à 70Kg.

Com base nas informações contidas nos gráficos anteriores, plotamos os gráficos das Figuras 19, 20 e 21 notamos que estas possuem informações de grande relevância em nosso objeto de estudo.

Na Figura 19 o gráfico através da identificação de quais faixas etárias os indivíduos estão conforme seus níveis de IMC, ou seja, em quais idades se concentram a maior parte desta população nos diferentes níveis, informações estas obtidas no através do código abaixo:

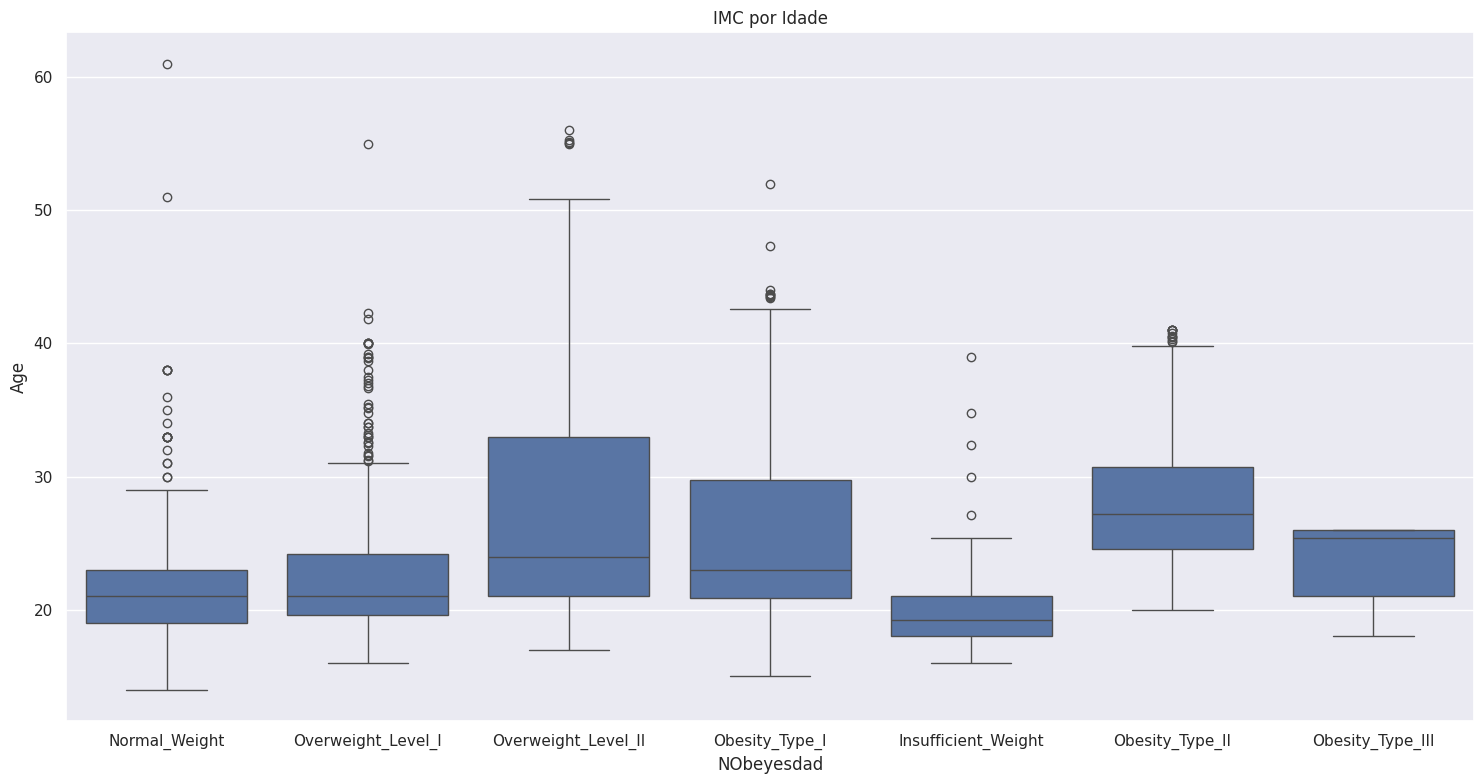
plt.figure(figsize=(15,8))

plt.subplot(1, 1, 1 )

sns.boxplot(data=df, x='NObeyesdad', y='Age',)

plt.title('IMC por Idade')

plt.tight\_layout()

  
Figura 19: Gráfico com a identificação de quais faixas etárias os indivíduos estão conforme seus níveis de IMC.

Analisando o Gráfico acima, na relação IMC por Idade, podemos notar na amostra que no grupo onde há indivíduos com *peso abaixo do normal*, estão os indivíduos que possuem idades com valores interquartis entre 15 a 22 anos e mediana abaixo dos 20 anos. Sugerindo assim que este grupo tem em sua maioria adolescentes e jovens adultos. Os valores de outliers presentes neste grupo estão bem dispersos entre as idades de 25 a 40 anos.

O grupo que possui indivíduos com *peso normal* notamos que nas idades, os valores interquartis se encontram entre 16 a 25 anos e mediana acima dos 20 anos. Os valores de outliers presentes neste grupo estão bem dispersos entre 30 a 40 anos.

No grupo que possui indivíduos com *sobrepeso nível I* notamos que nas idades, os valores interquartis se encontram entre 20 a 25 anos e mediana acima dos 20 anos, valores estes semelhantes aos encontrados no grupo de peso normal. Os valores de outliers presentes neste grupo, sendo este o grupo com maior concentração de outliers, estão bem dispersos entre valores acima de 30 a aproximadamente 40 anos.

Os indivíduos pertencentes ao grupo com *sobrepeso nível II* notamos que em suas idades, os valores interquartis se encontram entre 21 a 35 anos, este grupo possui o maior intervalo interquartil da amostra, e mediana em torno dos 25 anos. Não há uma presença de outliers neste grupo.

Seguindo ao grupo que possui indivíduos com *obesidade tipo I* é notado que nas idades, os valores interquartis se encontram entre 21 a 30 anos e mediana acima dos 20 anos, próxima aos 23 anos, valores estes um pouco abaixo aos encontrados no grupo de sobrepeso nível. Os valores de outliers presentes neste grupo se dispersam entre 43 a 53 anos, sendo sua maior concentração um pouco acima dos 43 anos.

Em sequência, seguimos ao grupo que possui indivíduos com *obesidade tipo II*, podendo ser observado que nas idades, os valores interquartis se encontram entre 25 a 31 anos e mediana acima dos 25 anos, próxima aos 27 anos. Os valores de outliers presentes neste grupo concentram por volta dos 40 anos.

Por fim, seguimos ao grupo que possui indivíduos com *obesidade tipo III*, podendo ser observado que nas idades, os valores interquartis se encontram entre 21 a 26 anos e mediana por volta dos 25 anos. Não foi notada a presença de outliers neste grupo.

Na Figura 20, continuando a análise do gráfico da Figura 19, foi plotado um gráfico de dispersão para determinar a distribuição dos indivíduos na amostra de acordo com suas idades e seus respectivos grupos nos níveis de IMC. Isso permitiu a identificação das faixas etárias em que a maioria da população está concentrada para os diferentes níveis de IMC, assim como os valores atípicos (outliers). Assim, foi possível relacionar com mais clareza os dados do gráfico abaixo com os apresentados anteriormente.

  
Figura 20: Gráfico de dispersão da distribuição de indivíduos em cada grupo na coluna de IMC de acordo com suas idades.

Na Figura 21, será ilustrado o gráfico com a relação de ganho de peso nos indivíduos que monitoram as quantidades de calorias consumidas durantes as refeições com seus ganhos ou não de peso, informações estas obtidas no através do código abaixo:

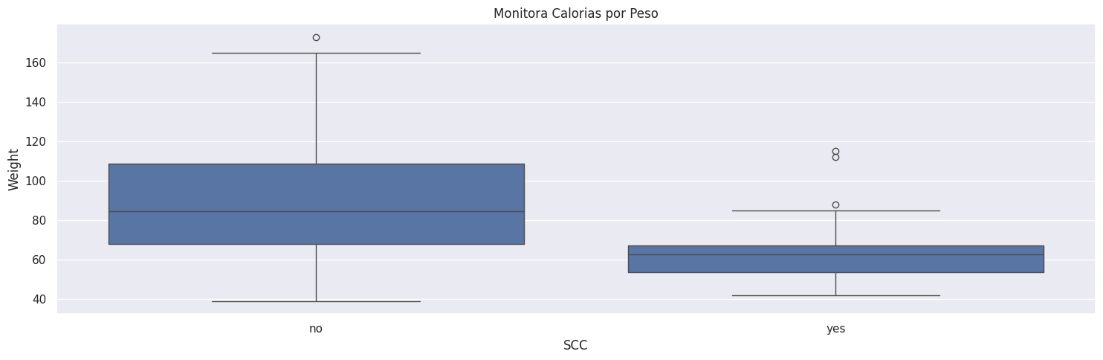
plt.figure(figsize=(15,5))

plt.subplot(1, 1, 1 )

sns.boxplot(data=df, x='SCC', y='Weight',)

plt.title('Monitora Calorias por Peso')

plt.tight\_layout()

  
Figura 21: Gráfico com a relação de ganho de peso nos indivíduos que monitoram as quantidades de calorias consumidas durantes as refeições com o peso.

Analisando o Gráfico acima, notamos que na relação Monitoramento da quantidade de calorias consumidas por refeições por Peso, podemos notar na amostra que os indivíduos que fazem o controle de suas calorias consumidas possuem a tendência de manterem seu peso em níveis mais estáveis e com baixa variação, como podemos notar na relação interquartil que possuem valores entre 58 a 70Kg e sua mediana em torno dos 65Kg. Os valores de outliers de nestes indivíduos se dispersam entre 90 a 115Kg.

No caso dos indivíduos que não contam suas calorias durante as refeições, podemos notar que há uma tendência de aumento no ganho de peso quando comparados ao outro grupo. Seus valores interquartis possuem uma maior variação, encontrando-se entre 70 a 110Kg com mediana de aproximadamente 85Kg. O valor de outlier nestes indivíduos se dispersa acima dos 160Kg.

1. Objetivos OBTIDOS POR MEIO DAS ANÁLISES ANTERIORES

Através dos dados obtidos através das informações anteriormente coletadas, decidimos buscar o aprimoramento desses resultados optamos por realizar as seguintes análises:

* 1. Conversão das variáveis categóricas 'Gender' e 'family\_history\_with\_overweight' em numéricas

le = LabelEncoder()

df['Gender'] = le.fit\_transform(df['Gender'])

df['family\_history\_with\_overweight'] = le.fit\_transform(df['family\_history\_with\_overweight'])

df['FAVC'] = le.fit\_transform(df['FAVC'])

Divisão os dados em conjunto de treinamento e conjunto de teste

X = df.drop('NObeyesdad', axis=1)  # Features

y = df['NObeyesdad']  # Target variable

Conversão das variáveis categóricas em variáveis dummy

X = pd.get\_dummies(X, columns=['Gender', 'family\_history\_with\_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS'])

Padronização dos dados

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

Inicializar e treinar o modelo de regressão logística com um número maior de iterações

model = LogisticRegression(max\_iter=1000)

model.fit(X\_train\_scaled, y\_train)

Fazer previsões no conjunto de teste

y\_pred = model.predict(X\_test\_scaled)

Avaliar a precisão do modelo

accuracy = accuracy\_score(y\_test, y\_pred)

print("Acurácia do modelo de regressão logística:", accuracy)

Obs.: Convertendo as variáveis categorias em numéricas, foi obtido que a acurácia do modelo de regressão logística foi de 0.8652, ou seja, aproximadamente 86%.

Gráfico plotado em colunas de todas as variáveis numéricas originais

sns.set( font\_scale=1.0, rc={'figure.figsize':(15,15)} )

eixo = df.hist(bins=20, color= 'blue')

* 1. Conversão variável categórica de IMC em numérica

A quantidade de valores numéricos varia de acordo com a quantidade de valores categóricos presentes em cada coluna.

df['NObeyesdad\_code'] = pd.factorize(df['NObeyesdad'])[0]

df['NObeyesdad\_code'].unique()

Identificação dos nomes dado a cada categoria convertida de IMC no dataset

df.NObeyesdad.unique()

As variáveis categóricas presentes em cada colunas e com seus respectivos valores categóricos que estão presentes no dataframe são os seguintes

df.select\_dtypes(include='object').describe().drop('count')

A função seguinte receberá um Dataframe e o nome de uma coluna categórica. Esta irá criar uma coluna no Dataframe com os códigos numéricos correspondentes às categorias da coluna especificada

def encode\_categorical\_variable(df, column\_name):

    # Cria uma nova coluna com códigos numéricos para as categorias

    df[column\_name + '\_code'] = pd.factorize(df[column\_name])[0]

    # Retorna os códigos únicos atribuídos

    return df[column\_name + '\_code'].unique()

Modelo de predição a partir dos dados coletados anteriormente do DataFrame

# Definir as features (X) e a variável alvo (y)

X = df.drop(['NObeyesdad'], axis=1)  # Todas as colunas exceto 'NObeyesdad'

y = df['NObeyesdad']

# Dividir os dados em conjuntos de treinamento e teste

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Pré-processamento das features

numeric\_features = X.select\_dtypes(include=['int64', 'float64']).columns

categorical\_features = X.select\_dtypes(include=['object']).columns

numeric\_transformer = StandardScaler()

categorical\_transformer = OneHotEncoder()

preprocessor = ColumnTransformer(

    transformers=[

        ('num', numeric\_transformer, numeric\_features),

        ('cat', categorical\_transformer, categorical\_features)

    ])

# Criando o pipeline com pré-processamento e modelo

pipeline = Pipeline([

    ('preprocessor', preprocessor),

    ('model', LogisticRegression(max\_iter=1000))

])

# Treinar o modelo dentro do pipeline nos dados de treinamento

pipeline.fit(X\_train, y\_train)

# Fazer previsões no conjunto de teste

y\_pred = pipeline.predict(X\_test)

# Avaliar o modelo

accuracy = accuracy\_score(y\_test, y\_pred)

print(f'Acurácia do modelo: {accuracy:.2f}')

print(classification\_report(y\_test, y\_pred))

Em resumo, o processo envolveu a conversão de variáveis categóricas em formatos numéricos compreensíveis para modelos de aprendizado de máquina, seguido pela divisão dos dados em conjuntos de treinamento e teste, o pré-processamento dos dados (incluindo padronização e criação de variáveis dummy) e a aplicação de um modelo de regressão logística para previsão da variável alvo (NObeyesdad). O modelo alcançou uma alta acurácia de 86% para 98% no conjunto de teste, indicando um bom desempenho na previsão da classe de obesidade com base nos dados fornecidos.

1. Conclusão

A partir da análise das amostras de dados relacionadas aos estilos de vida, histórico familiar e consumo de alimentos calóricos, foi possível identificar padrões significativos associados ao aumento de peso e à obesidade. Esses padrões revelam a importância do histórico familiar na predisposição ao sobrepeso, o impacto do consumo de alimentos calóricos na condição de peso e a influência do monitoramento da ingestão calórica na manutenção de um peso estável.

Para realizar um estudo semelhante aos padrões brasileiros, seria fundamental a coleta de dados abrangentes relacionados ao histórico familiar de sobrepeso, consumo de alimentos calóricos, índice de massa corporal por idade e monitoramento da ingestão calórica. Além disso, a análise exploratória destes dados permitiria identificar padrões específicos da população brasileira, auxiliando na elaboração de intervenções preventivas e terapêuticas mais assertivas.

Dessa forma, a aplicação desses métodos à realidade brasileira poderia fornecer insights valiosos para a implementação de programas direcionados à prevenção da obesidade e promoção de estilos de vida saudáveis, contribuindo para a redução de gastos públicos relacionados ao tratamento de doenças crônicas.

A análise detalhada desses dados poderia proporcionar uma visão abrangente da situação da obesidade no Brasil, oferecendo embasamento para a formulação de políticas de saúde pública e práticas clínicas direcionadas ao combate da obesidade e suas doenças associadas, resultando em uma melhoria significativa na qualidade de vida da população.

1. Referência Bibliográfica

* **Material de Apoio das disciplinas de: Introdução a Ciência de Dados, Pensamento Computacional, Análise Exploratória de Dados e Analytics.**
* [**https://www.kaggle.com/competitions/playground-series-s4e2**](https://www.kaggle.com/competitions/playground-series-s4e2) **(acessado em 23/02/2024)**
* **https://github.com/grupos4g4/PROJAPLIC1 (criado em 04/03/2024)**
* [**https://pubmed.ncbi.nlm.nih.gov/31467953/**](https://pubmed.ncbi.nlm.nih.gov/31467953/) **(acessado em 16/03/2024)**
* [**https://thescipub.com/pdf/jcssp.2019.67.77.pdf**](https://thescipub.com/pdf/jcssp.2019.67.77.pdf) **(acessado em 19/04/2024)**