Tech Challenge Fase 4

Natalia Alexandra Leite Trippo – RM363623

O Problema Proposto

A obesidade tem sido um problema principal de saúde pública, afetando indivíduos de diversas faixas etárias e em diversos contextos sociais, genéticos, ambientais e comportamentais, associadas à complicações clínicas.

Diante desse cenário, foi proposto um problema onde é preciso um auxílio aos médicos para prevenção de obesidade em indivíduos, oferecendo uma ferramenta preditiva que possa complementar avaliações clínicas que ajudem a realizar estratégias para prevenção e tratamentos mais eficazes.

Este relatório apresenta o desenvolvimento de um modelo de aprendizado de máquina, construído a partir da base de dados fornecida, ***obesity.csv***, com o propósito de prever a probabilidade de ocorrência de obesidade em pacientes de diferentes grupos.

Separando e Tratando os Dados

Na primeira etapa deste trabalho, foi realizada uma exploração da base de dados disponibilizada, acompanhada do dicionário de variáveis fornecido na proposta do problema. A base continha *gênero*, *idade*, *altura*, *peso*, *histórico familiar de excesso de peso*, *frequência de consumo de alimentos calóricos*, *ingestão de vegetais*, *número de refeições diárias*, *hábitos de alimentação entre refeições*, *tabagismo*, *consumo de água*, *monitoramento de calorias*, *prática de atividade física*, *tempo de uso de dispositivos tecnológicos*, *consumo de álcool*, *meio de transporte utilizado* e a variável alvo, *nível de obesidade*.

Foram escolhidos os seguintes dados:

- Idade
- Altura
- Peso
- Frequência de consumo de alimentos calóricos (FAVC)
- Ingestão de vegetais (FCVC)
- Número de refeições principais (NCP)
- Hábito de comer entre refeições (CAEC)
- Consumo de água (CH2O)
- Prática de atividade física (FAF)
- Consumo de álcool (CALC)

Esses dados são elementos-chave na avaliação de risco de obesidade pois refletem tanto características físicas quanto padrões de comportamento alimentar e estilo de vida, enquanto as demais, embora relevantes em análises mais amplas, foram consideradas de impacto secundário para o estudo, sendo descartadas da etapa de modelagem inicial, buscando garantir análise objetiva e focada em fatores de maior influência direta sobre o resultado final.

Na etapa de preparação dos dados, inicialmente foram removidos os valores nulos, limpando os dados para melhor análise. Em seguida, realizei a seleção das variáveis relevantes para a análise, por meio do código em Python que especifica as colunas de interesse, o que resultou em um subconjunto de dados contendo atributos como idade, altura, peso, gênero, frequência de consumo de alimentos calóricos, ingestão de vegetais, número de refeições principais, hábito de comer entre refeições, consumo de água, prática de atividade física e consumo de álcool. O comando utilizado para isso foi

```
selected_columns = ['Age', 'Height', 'Weight', 'Gender', 'FAVC', 'FCVC', 'NCP', 'CAEC', 'CH2O', 'FAF', 'CALC']
dados_selecionados = dados[selected_columns]
display(dados_selecionados.head())
```

Que teve como função restringir a base apenas às colunas selecionadas e exibir suas primeiras linhas para conferência.

Outro ajuste importante foi a tradução dos dados da tabela, no geral, originalmente em inglês, para o português, de modo a facilitar a interpretação por parte da equipe médica que fará uso do sistema. Além disso, os valores numéricos foram padronizados: idade, frequência de consumo de alimentos calóricos, prática de atividade física, ingestão de vegetais e número de refeições principais foram convertidos para números inteiros; altura foi mantida em metros; peso em quilogramas; e consumo de água em litros. As variáveis que eram respostas textuais foram mantidas. Esses ajustes também anularam a desconfiguração dos valores ao longo da base de dados.

Masculino	18	1,72	53	yes	2	3	Sometimes	2	0	Sometimes
Masculino	20	1,56	45	no	2	3	Sometimes	2	1	Sometimes
Feminino	25,196214	1,686306	104,572712	yes	3	3	Sometimes	1,152736	0,319156	Sometimes
Feminino	18,503343	1,683124	126,67378	yes	3	3	Sometimes	1,115967	1,541072	Sometimes
Feminino	26	1,622397	110,79263	yes	3	3	Sometimes	2,704507	0	Sometimes
Feminino	21,853826	1,755643	137,796884	yes	3	3	Sometimes	2,184707	1,978631	Sometimes
Feminino	21,90012	1,843419	165,057269	yes	3	3	Sometimes	2,406541	0,10032	Sometimes
Feminino	18,306615	1,7456	133,03441	yes	3	3	Sometimes	2,984323	1,586525	Sometimes
Feminino	26	1,630927	111,485516	yes	3	3	Sometimes	2,444125	0	Sometimes
Feminino	26	1,629191	104,826776	yes	3	3	Sometimes	2,654702	0	Sometimes
Feminino	21,849705	1,770612	133,963349	yes	3	3	Sometimes	2,825629	1,399183	Sometimes
Masculino	19,799054	1,743702	54,927529	yes	2	3,28926	Sometimes	2,847264	1,680844	Sometimes

Base de dados Obesity.csv antes do tratamento.

Para que tudo isso pudesse ser feito, o arquivo foi baixado em extensão ***.xlsx*** para que o Google Colab pudesse ler o arquivo de forma eficaz.

Valores de altura apareciam estar em metros e também em centímetros, e o peso em quilogramas e em gramas. Para corrigir isso, foram aplicadas conversões e arredondamentos, garantindo que todos os registros estivessem padronizados. Além disso, colunas como idade, frequência de atividade física e número de refeições principais foram convertidas para valores inteiros, eliminando possíveis erros de digitação, deixando os dados mais limpos.

```
selected_columns = ['Age', 'Height', 'Weight', 'Gender', 'FAVC', 'FCVC', 'NCP', 'CAEC', 'CH2O', 'FAF', 'CALC']
dados_selecionados = dados[selected_columns].copy()

# Mapeamentos de tradução para valores dentro das colunas
translation_maps = {
    'Gender': {'Female': 'Feminino', 'Male': 'Masculino'},
    'FAVC': {'yes': 'sim', 'no': 'não'},
    'CAEC': {'Sometimes': 'Às vezes', 'Frequently': 'Frequentemente', 'Always': 'Sempre', 'no': 'não'},
    'CALC': {'Sometimes': 'Às vezes', 'Frequently': 'Frequentemente', 'no': 'não'}
}

# Aplicar as traduções aos valores das colunas
for column, mapping in translation_maps.items():
    if column in dados_selecionados.columns:
        dados_selecionados.loc[:, column] = dados_selecionados[column].replace(mapping)

# Mapeamento de tradução para os nomes das colunas
column_name_translations = {
    'Age': 'Idade',
    'Height': 'Altura',
    'Weight': 'Peso',
    'Gender': 'Gênero',
    'FAVC': 'Frequência_Alimentar_Calorica',
    'FCVC': 'Consumo_Vegetais_Frutas',
    'NCP': 'Refeições_Principais',
    'CAEC': 'Comer_Entre_Refeições',
    'CH2O': 'Consumo_Agua',
    'FAF': 'Atividade_Fisica_Frequencia',
    'CALC': 'Consumo_Alcool'
}

# Aplicar as traduções aos nomes das colunas
dados_selecionados = dados_selecionados.rename(columns=column_name_translations)

# Converter colunas para inteiros
dados_selecionados['Idade'] = dados_selecionados['Idade'].astype(int)
dados_selecionados['Consumo_Vegetais_Frutas'] = dados_selecionados['Consumo_Vegetais_Frutas'].astype(int)
dados_selecionados['Refeições_Principais'] = dados_selecionados['Refeições_Principais'].astype(int)
dados_selecionados['Atividade_Fisica_Frequencia'] = dados_selecionados['Atividade_Fisica_Frequencia'].astype(int)
dados_selecionados['Consumo_Agua'] = dados_selecionados['Consumo_Agua'].astype(int)

# 'Altura' e 'Peso' já estão em formato float e representam metros e quilos, respectivamente.

print("Tipos de dados após as conversões:")
display(dados_selecionados.dtypes)
display(dados_selecionados.head())
```

Novamente o arquivo foi instalado em .xlsl (Obesity 5.0.xlsx) para realização da etapa de teste e treino com o modelo escolhido.

Teste, Treino e Resultados

O treino e teste foram feitos em um arquivo diferente para a preservação do progresso do trabalho sem correr riscos de danificar o mesmo caso ocorresse um erro.

O modelo escolhido foi o Random Forest, pois por se tratar de um problema de classificação supervisionada multiclasse, onde o objetivo é prever diferentes níveis de obesidade o Random Forest se mostrou adequado porque lida bem com variáveis numéricas e categóricas, reduz o risco de overfitting ao combinar diversas árvores de decisão e permite identificar a importância das variáveis na previsão. Além disso, é um modelo robusto e amplamente utilizado em problemas de saúde e comportamento, garantindo boa acurácia e interpretabilidade. O modelo foi treinado com os dados de treino, aprendendo padrões que relacionam características individuais (idade, hábitos alimentares, atividade física, etc.) com a categoria de obesidade.

O modelo Random Forest foi treinado e avaliado utilizando primeiro todo o conjunto de dados, incluindo indivíduos do gênero masculino e feminino para avaliação do modelo no geral. A acurácia obtida foi de **93,38%**, demonstrando um bom desempenho na classificação das categorias de obesidade.

Para aprofundar a análise, os dados foram segmentados por gênero e modelos independentes foram treinados para cada grupo. Essa abordagem permitiu verificar se havia diferenças relevantes no desempenho do algoritmo em populações distintas. Nos dados femininos, o modelo alcançou uma acurácia de **94,74%**, enquanto nos dados masculinos obteve **93,93%**.

```
Shape of X_feminino: (1043, 15)
Shape of y_feminino: (1043,)
Shape of X_train_feminino: (834, 15)
Shape of X_test_feminino: (209, 15)
Shape of y_train_feminino: (834,)
Shape of y_test_feminino: (209,)

Modelo RandomForestClassifier treinado com sucesso para dados femininos.

Acurácia do modelo (Feminino): 0.95
Assertividade total (em porcentagem - Feminino): 94.74%

Relatório de Classificação (Feminino):
                    precision    recall  f1-score   support

    Abaixo do Peso       1.00      1.00      1.00        35
  Obesidade Grau I       1.00      0.93      0.97        30
 Obesidade Grau II       1.00      1.00      1.00        16
Obesidade Grau III       1.00      1.00      1.00        51
       Peso Normal       0.89      0.86      0.88        37
         Sobrepeso       0.84      0.90      0.87        40

          accuracy                           0.95       209
         macro avg       0.95      0.95      0.95       209
      weighted avg       0.95      0.95      0.95       209
```

```
Shape of X_masculino: (1068, 15)
Shape of y_masculino: (1068,)
Shape of X_train_masculino: (854, 15)
Shape of X_test_masculino: (214, 15)
Shape of y_train_masculino: (854,)
Shape of y_test_masculino: (214,)

Modelo RandomForestClassifier treinado com sucesso para dados masculinos.

Acurácia do modelo (Masculino): 0.94
Assertividade total (em porcentagem - Masculino): 93.93%

Relatório de Classificação (Masculino):
                    precision    recall  f1-score   support

    Abaixo do Peso       1.00      0.97      0.98        29
  Obesidade Grau I       1.00      0.85      0.92        39
Obesidade Grau II        0.95      1.00      0.97        55
       Peso Normal       0.87      0.90      0.89        30
         Sobrepeso       0.91      0.95      0.93        61

          accuracy                           0.94       214
         macro avg       0.95      0.93      0.94       214
      weighted avg       0.94      0.94      0.94       214
```

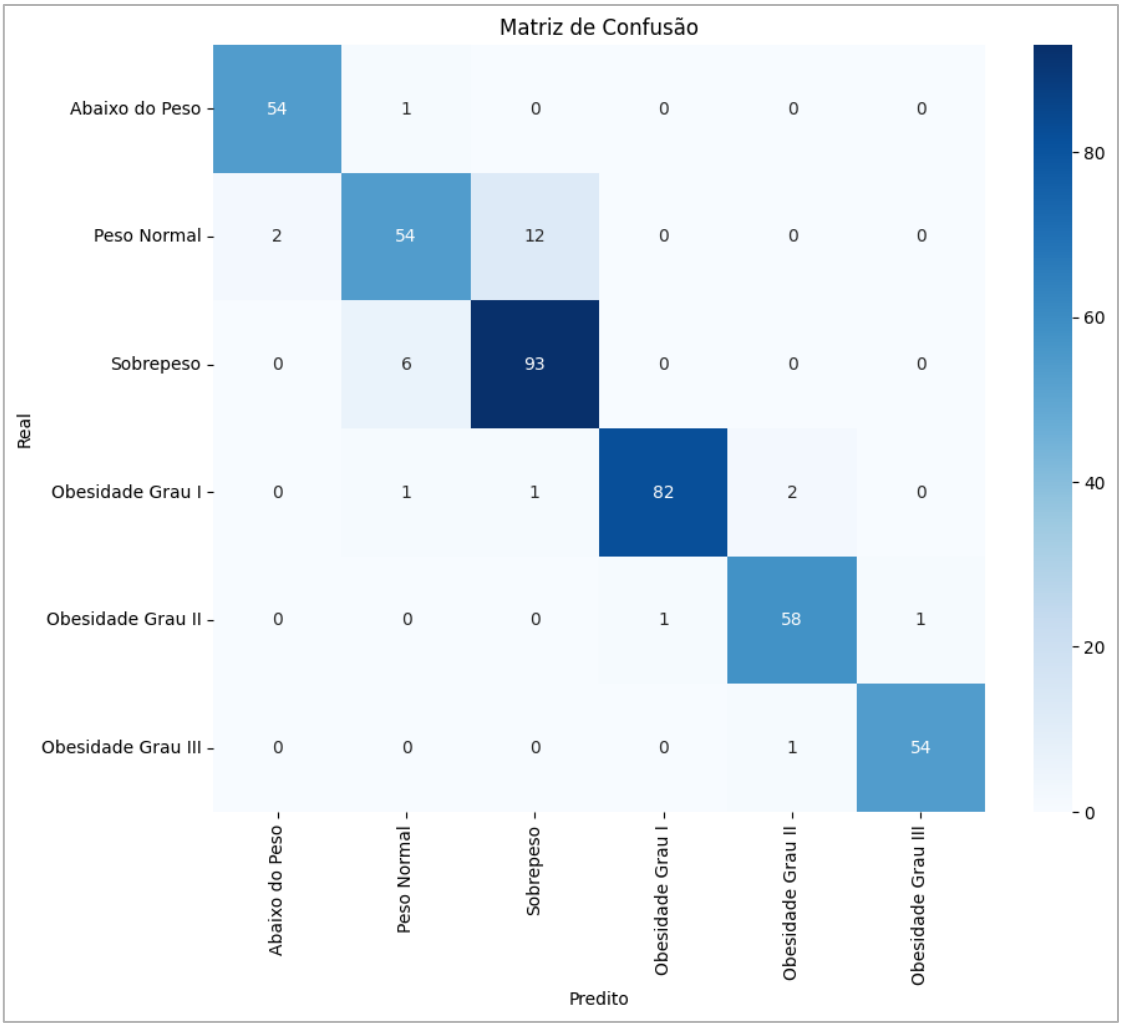
Esses resultados indicam que o modelo manteve alta assertividade em ambos os grupos, com uma ligeira superioridade no conjunto feminino.

A análise separada por gênero contribui para compreender melhor os padrões de obesidade em diferentes populações e reforça a robustez do modelo, que apresentou desempenho consistente em cenários distintos.

A Análise

Após o treinamento e avaliação dos modelos, foram gerados gráficos que permitiram visualizar os resultados de forma mais clara e intuitiva. Essas representações gráficas reforçaram os achados da pesquisa sobre obesidade, evidenciando padrões nos dados e diferenças entre grupos específicos.

A matriz de confusão, por exemplo, mostrou os acertos e erros do modelo em cada categoria de obesidade, permitindo identificar que a maior parte das classificações incorretas ocorreu em categorias adjacentes (como Sobrepeso e Obesidade Grau I).



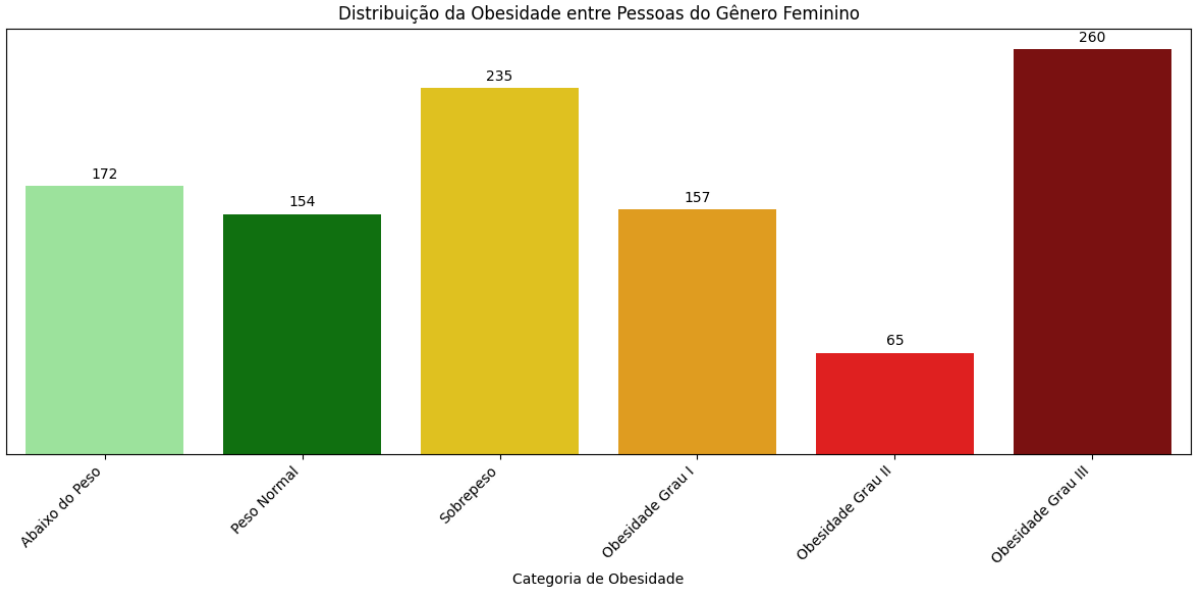
Ela revela que o modelo teve bom desempenho na classificação das categorias intermediárias, como sobrepeso (93 acertos) e obesidade grau I (82 acertos), mas apresentou confusões entre categorias vizinhas, como peso normal e sobrepeso, ou obesidade grau I e II. Esses erros são compreensíveis à luz dos dados analisados:

- **Distribuição de idade:** as faixas etárias médias entre categorias vizinhas são próximas, especialmente entre sobrepeso e obesidade grau I, o que pode dificultar a distinção pelo modelo.

- **Hábitos alimentares e de atividade física:** muitos indivíduos com sobrepeso e obesidade grau I compartilham padrões semelhantes de consumo de alimentos calóricos e baixa prática de exercícios, o que reduz a separação entre classes.

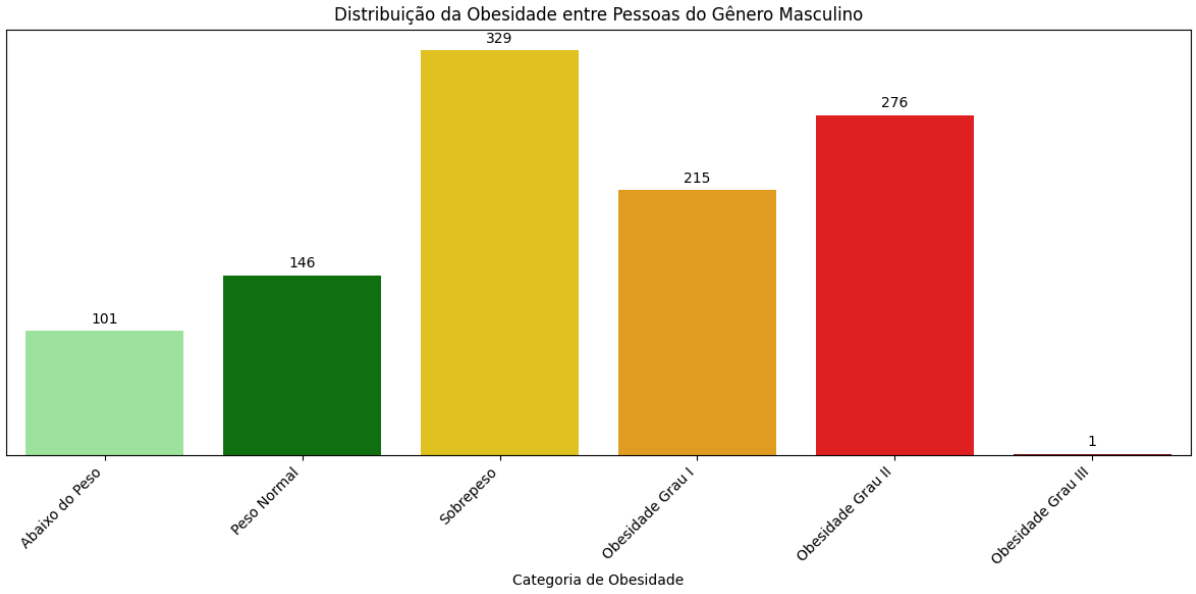
- **Diferenças entre gêneros:** como os dados mostram que mulheres concentram-se mais em obesidade grau III e homens em sobrepeso e grau II, o modelo pode ter aprendido padrões mais claros para esses grupos, mas menos precisos para extremos ou faixas menos representadas.

Essas relações indicam que os erros do modelo não são aleatórios, mas refletem sobreposições reais nos dados. Isso reforça a importância de considerar variáveis complementares — como estilo de vida, histórico familiar e consumo de vegetais — para melhorar a acurácia nas classificações mais difíceis.



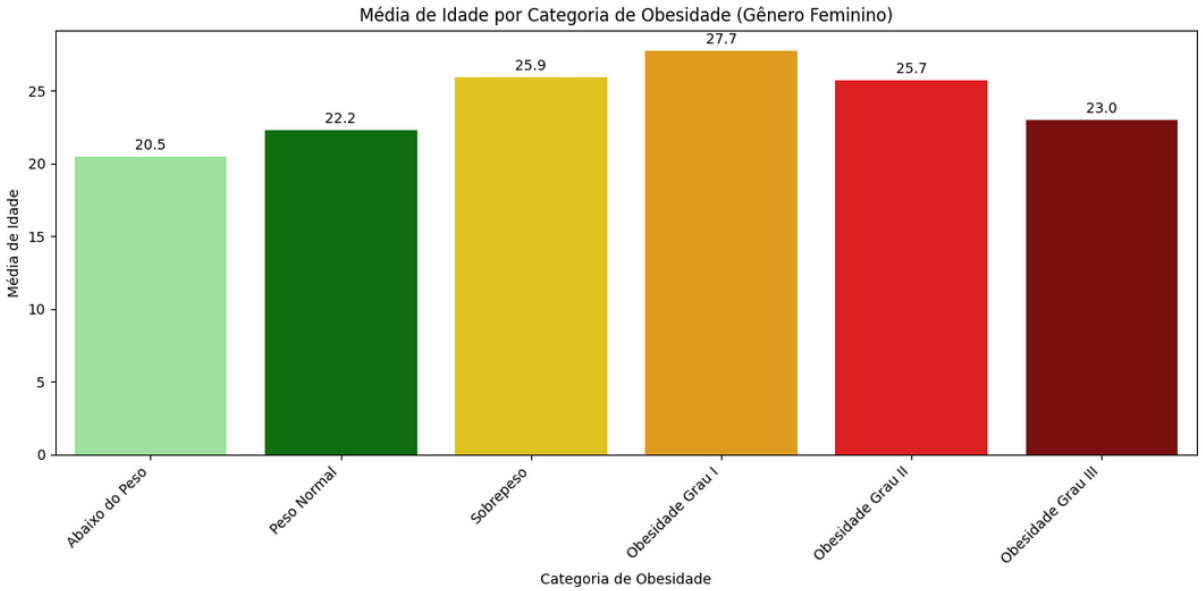
Entre mulheres, o maior número de casos está na categoria de **obesidade grau III**, o que contrasta com os dados masculinos, onde essa categoria é **quase inexistente**. Isso indica uma prevalência mais alta de obesidade severa entre mulheres, o que exige atenção especial em estratégias de prevenção e intervenção precoce.

Estudos epidemiológicos apontam que essa diferença pode estar relacionada a **fatores históricos e sociais**, como a maior exposição das mulheres a **mudanças no padrão alimentar**, especialmente em contextos de vulnerabilidade socioeconômica. Segundo pesquisa realizada com usuárias do sistema público de saúde no Brasil, há uma associação entre obesidade e características demográficas como baixa escolaridade, maternidade precoce e acesso limitado à informação nutricional. Além disso, fatores culturais e de gênero influenciam o comportamento alimentar e a prática de atividade física. Mulheres tendem a assumir responsabilidades domésticas e de cuidado que limitam o tempo disponível para autocuidado, incluindo alimentação saudável e exercícios regulares. Historicamente, **políticas de saúde pública voltadas para obesidade não consideraram as especificidades femininas**, o que contribui para a manutenção de desigualdades.



A maior parte dos homens está concentrada nas categorias de **sobrepeso** e **obesidade grau II**, indicando que o excesso de peso é predominante, mas os casos mais graves (**grau III**) são raros. Isso sugere que, embora o problema seja amplo, ainda há espaço para intervenções antes que evolua para quadros clínicos mais críticos.

Estudos populacionais indicam que homens tendem a apresentar maior prevalência de sobrepeso em relação às mulheres, mas menor incidência de obesidade severa. Essa diferença pode estar relacionada a fatores comportamentais, como **maior prática de atividade física entre homens**, mesmo em contextos de sobrepeso, e **menor procura por serviços de saúde**, o que **pode atrasar o diagnóstico e acompanhamento nutricional**. Estratégias como **campanhas voltadas ao público masculino, incentivo à prática regular de exercícios e ampliação do acesso à orientação nutricional** podem ser eficazes para evitar a progressão para obesidade grau III, que está associada a riscos elevados de doenças cardiovasculares, metabólicas e ortopédicas.

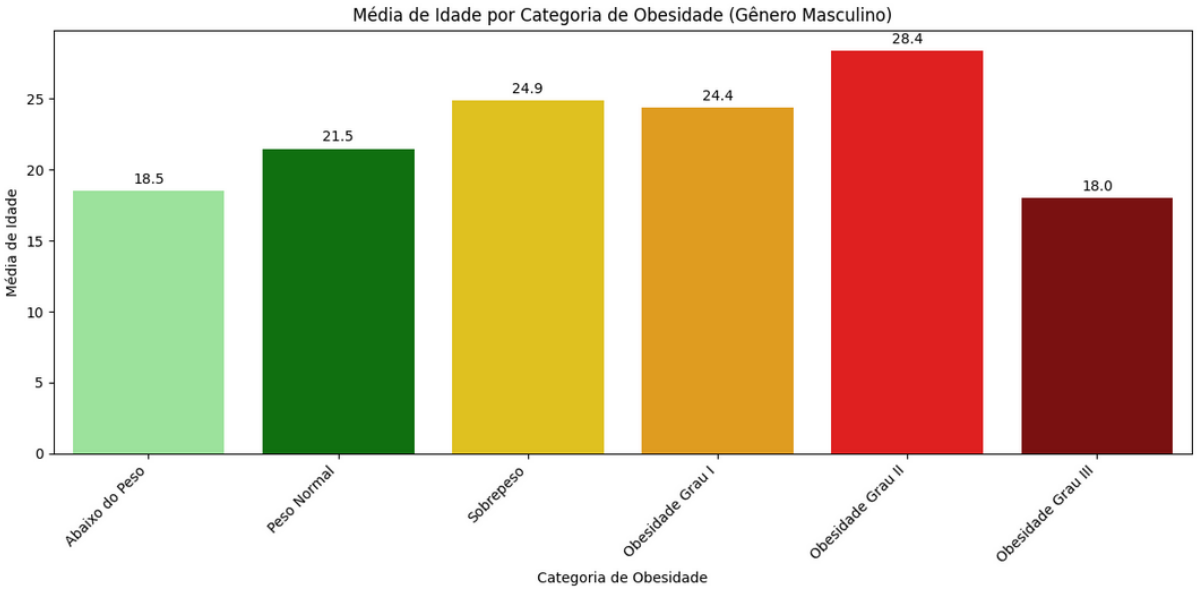


A distribuição da média de idade entre as categorias de obesidade **feminina** mostra uma **progressão até a obesidade grau I,** com pico de idade média em **27 anos**. A partir daí, observa-se uma leve queda: **mulheres com obesidade grau II têm média de 25 anos**, e aquelas com **obesidade grau III, 23 anos**.

Esse padrão sugere que os casos mais graves de obesidade estão se manifestando em **mulheres mais jovens**, o que pode indicar um **início precoce de hábitos alimentares inadequados, sedentarismo ou fatores metabólicos**. A presença de obesidade grau III em faixas etárias mais baixas é preocupante, pois **implica maior exposição prolongada aos riscos associados**, como diabetes tipo 2, hipertensão e doenças cardiovasculares.

Além disso, o aumento gradual da idade média nas categorias de peso normal, sobrepeso e obesidade grau I pode refletir mudanças de estilo de vida ao longo do tempo, como redução da atividade física, aumento da carga de trabalho e menor atenção à alimentação saudável.

Esse cenário reforça a importância de **ações preventivas voltadas para mulheres jovens, com foco em educação alimentar, incentivo à prática de exercícios e acompanhamento clínico precoce**, especialmente em contextos urbanos como São Paulo, onde o estilo de vida acelerado pode contribuir para o desenvolvimento da obesidade em idades cada vez mais baixas.



A distribuição da média de idade entre as categorias de obesidade masculina revela padrões onde os homens com obesidade **grau II** apresentam a **maior média de idade** (**28 anos**), enquanto os com obesidade **grau III** têm a **menor média** (**18 anos**). Esse contraste pode indicar que os casos mais graves de obesidade estão se manifestando em **indivíduos mais jovens**, o que é **preocupante do ponto de vista clínico e epidemiológico**.

Por outro lado, a progressão entre as categorias "peso normal", "sobrepeso" e "obesidade grau I" mostra um aumento gradual da média de idade, sugerindo que o **ganho de peso pode estar relacionado ao envelhecimento e à mudança de hábitos ao longo do tempo**.

Esse padrão reforça a importância de **intervenções precoces, especialmente entre jovens com obesidade grau III, que podem estar expostos a riscos metabólicos mais cedo e por mais tempo**.