Modelagem Preditiva Aplicada à Saúde e Bem Estar

Projeto de Machine Learning – AP1

Ciência de Dados e Inteligência Artificial – IBMEC

Lucca Lanzellotti – 202107395273
Bernardo Rodrigues Borges Loureiro – 202108075094

Abril de 2025

Resumo

Este trabalho apresenta a aplicação de técnicas de aprendizado supervisionado em um conjunto de dados real proveniente do Kaggle, relacionado a hábitos e características de usuários de academia. Foram implementados dois modelos de machine learning: regressão linear simples, para prever o gasto calórico com base na frequência semanal de treino, e regressão logística, para classificar o gênero a partir da altura. O processo envolveu análise exploratória dos dados, testes de normalidade, cálculo de correlação e avaliação de desempenho dos modelos com métricas estatísticas. Ao final, os modelos foram integrados a uma API REST funcional. Os resultados demonstram a viabilidade da modelagem preditiva em contextos relacionados à saúde e bem-estar, com possibilidade de aplicação prática em sistemas de recomendação e monitoramento físico.

Sumário

1 Apresentação do Dataset

Nome do Dataset: Gym Members Exercise Dataset
Origem: Kaggle - gym-members-exercise-dataset

Descrição:
O dataset contém **973 registros** com informações sobre idade, altura, IMC, frequência de treino, calorias queimadas, tipo de treino, ingestão de água e mais. Ele permite aplicar regressão linear (variável contínua: calorias queimadas) e regressão logística (variável binária: gênero).

Justificativa da Escolha:
É um conjunto de dados realista e bem estruturado, ideal para exercícios práticos de machine learning voltados à saúde, bem-estar e comportamento humano.

2 Pré-processamento e Análise Exploratória

Valores ausentes: foram verificados e não comprometeram a análise.

Renomeação de colunas: espaços e caracteres especiais foram padronizados.

Distribuição etária e composição por gênero: a população amostrada concentra-se entre 20 e 50 anos, com proporções equilibradas entre homens e mulheres.

Relação entre variáveis:

- Idade tende a se distribuir de forma simétrica.
- BMI (Índice de Massa Corporal) apresenta leve assimetria positiva.
- A frequência de treino varia entre 1 e 7 dias por semana, com média próxima a 4.

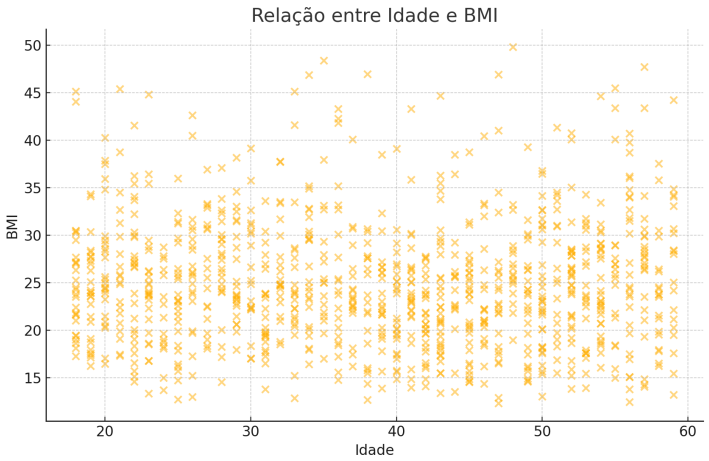

Figura 1: Relação entre Idade e BMI

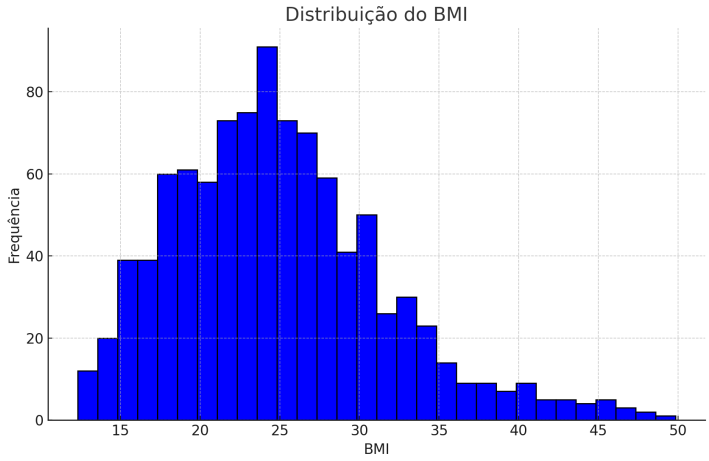

Figura 2: Distribuição da variável BMI

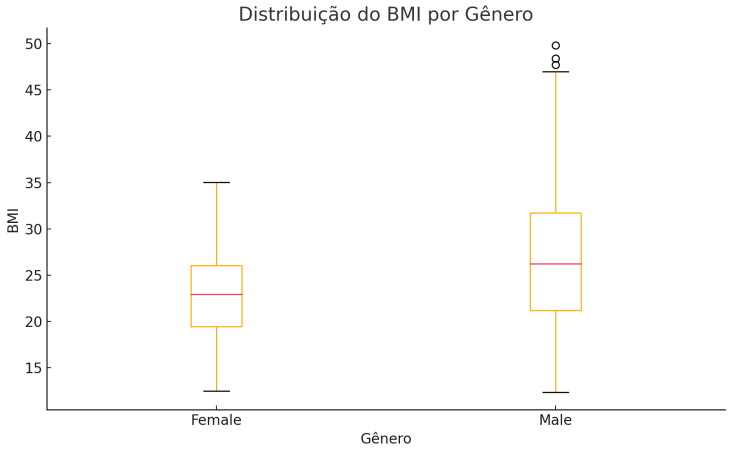

Figura 3: Boxplot do BMI por Gênero

3 Testes de Normalidade

Foi aplicado o teste de Shapiro-Wilk às variáveis `BMI` e `Workout Frequency`.

Resultados:

- `BMI`: p $< 1.44 \times 10^{-14}$ – não segue distribuição normal.
- `Workout Frequency`: p $< 2.2 \times 10^{-16}$ – também não segue distribuição normal.

Conclusão: Recomenda-se uso de métodos não paramétricos ou transformação dos dados.

4 Correlação

Correlação de Spearman foi utilizada entre idade e frequência de treino.

Resultado:

- Coeficiente: 0.0063 – praticamente nula.
- Valor-p >0.8 – sem significância estatística.

Conclusão: Não há evidência de associação monotônica entre idade e frequência semanal de treinos.

5 Regressão Linear Simples

Modelo: `Calories_Burned ~ Workout_Frequency`

- MAE: 175.34
- RMSE: 215.44
- R^2: 0.343

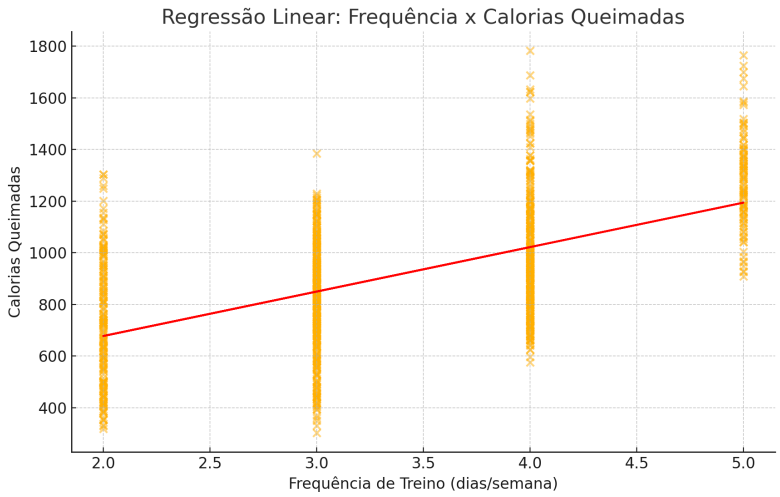

Figura 4: Regressão Linear: Frequência de Treino vs. Calorias Queimadas

Interpretação: 34,3% da variância das calorias queimadas é explicada pela frequência de treino.

Observações:

- O erro médio é relevante, mas o modelo capta uma tendência clara.
- Há espaço para melhorias com mais variáveis ou modelos multivariados.

6 Regressão Logística

Objetivo: prever o gênero com base na altura.

Métricas:

- Acurácia: 72,26%
- Kappa: 0,445

- Sensibilidade e especificidade: ≈73%
- p-valor de McNemar: 0.6567 – sem viés de classe.

Conclusão: O modelo apresenta desempenho razoável e previsões equilibradas, ainda que com espaço para aumento da acurácia com mais variáveis preditoras.

7 API REST

A API REST foi implementada com o pacote `plumber` em R e disponibiliza dois endpoints funcionais:

- **/predicao** – retorna a estimativa de calorias queimadas com base na frequência semanal de treino (`workout_frequency_days_week`), usando o modelo salvo.
- **/classificacao** – recebe a altura (variável `height_m`) e retorna a classe prevista (gênero binário) e a probabilidade associada, com base no modelo de regressão logística salvo como `modelo_regressao_logistica.rds`.

A documentação interativa da API pode ser acessada via Swagger:
`http://localhost:8000/__swagger__/`
Repositório: `https://github.com/bernalourodri/ap1_projeto_ML`

8 Conclusão

O trabalho cumpriu os objetivos propostos, aplicando técnicas fundamentais de machine learning supervisionado com base em regressão linear e logística. O uso de um dataset real permitiu interpretar relações típicas do ambiente fitness, como a influência da frequência de treino no gasto calórico e a diferenciação de gêneros pela altura.

As análises mostraram a importância do pré-processamento, testes de suposições estatísticas e avaliação com métricas robustas. A construção e publicação de uma API REST completou o ciclo de aprendizado, mostrando como um modelo pode ser usado em sistemas reais.