

Análise de Peso Corporal através de Hábitos Alimentares e Condições Físicas

1. Introdução

Este projeto propõe o desenvolvimento de modelos de Machine Learning para classificar níveis de obesidade com base em informações sobre hábitos alimentares e condições físicas.

Os dados utilizados neste estudo são provenientes do dataset “**ObesityDataSet**”, disponível no Kaggle. Ele inclui informações de indivíduos de países como México, Peru e Colômbia, com um total de **2111 registros e 17 atributos**. A variável alvo, **NObesity (nível_obesidade)**, permite classificar os registros em categorias:

- Abaixo do Peso
- Peso Normal
- Excesso de Peso Nível I
- Excesso de Peso Nível II
- Obesidade Tipo I
- Obesidade Tipo II
- Obesidade Tipo III

Os dados presentes nesse dataset foram balanceados utilizando a técnica **SMOTE (Synthetic Minority Oversampling Technique)**, garantindo um maior equilíbrio entre as classes e melhorando a representatividade das categorias menos frequentes.

Disponível em: <https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data>

Este projeto propõe a implementação de um **comitê de avaliação de modelos de Inteligência Artificial**, composto pelos seguintes modelos:

1. **Regressão Logística**
2. **Árvore de Decisão**
3. **Rede Neural Artificial (RNA)**
4. **Comitê de Avaliação**

Os modelos serão treinados individualmente e comparados para identificar aquele com maior eficácia na classificação de níveis de obesidade.

2. Objetivo

Desenvolver um sistema de avaliação que:

1. Treine e avalie modelos de IA em um conjunto de dados contendo informações relacionadas a obesidade.

2. Compare o desempenho dos modelos utilizando métricas apropriadas (e.g., acurácia, precisão, recall, curva AUC ROC, F1-score).
3. Identifique o modelo mais eficiente para classificação dos níveis de obesidade.

3. Escopo do Projeto

3.1. Escopo Funcional

- **Input:** Dataset “**ObesityDataSet**”, os dados estão estruturados conforme tabela abaixo.

Questões	Respostas Possíveis
Qual é o seu gênero?	Feminino
	Masculino
Qual é a sua idade?	Valor numérico
Qual é a sua altura?	Valor numérico em metros
Qual é o seu peso?	Valor numérico em quilogramas
Algum membro da família sofreu ou sofre de excesso de peso?	Sim
	Não
Você come alimentos altamente calóricos com frequência?	Sim
	Não
Você costuma comer vegetais nas suas refeições?	Nunca
	Às vezes
	Sempre
Quantas refeições principais você faz diariamente?	Entre 1 e 2
	Três
	Mais de três
Você come alguma coisa entre as refeições?	Não
	Às vezes
	Frequentemente
	Sempre
Você fuma?	Sim
	Não
Quanta água você bebe diariamente?	Menos de um litro
	Entre 1 e 2 L
	Mais de 2 L
Você monitora as calorias que ingere diariamente?	Sim
	Não
Com que frequência você pratica atividade física?	Não realizo
	1 ou 2 dias
	2 ou 4 dias
	4 ou 5 dias
Quanto tempo você usa dispositivos tecnológicos como celular, videogame, televisão, computador e outros?	0–2 horas
	3–5 horas
	Mais de 5 horas
Com que frequência você bebe álcool?	Eu não bebo
	Às vezes
	Frequentemente
	Sempre
Qual meio de transporte você costuma usar?	Automóvel
	Motocicleta
	Bicicleta
	Transporte público
	Andando

- **Processamento:**
 - Pré-processamento dos dados (limpeza, normalização, tratamento de variáveis categóricas).
 - Normalização da variável target “nível_obesidade” utilizando SMOTE.
 - Treinamento individual de três modelos:
 - Regressão Logística.
 - Árvore de Decisão.
 - Rede Neural Artificial.
 - Comitê.
 - Avaliação dos modelos com base em métricas de classificação.
 - Armazenamentos dos modelos em extensão “.pkl” para avaliação do comitê posteriormente
- **Output:** Modelo com melhor desempenho e relatório com análise comparativa.

3.2. Escopo Não Funcional

- O sistema deve ser desenvolvido em Python, utilizando bibliotecas como **Pandas, Numpy, Seaborn, Matplotlib, Scikit-learn, Joblib**
- Garantir reprodutibilidade dos experimentos utilizando métodos de seeds em buscas de amostragem aleatória.
- Eficiência Computacional: O sistema deve otimizar o uso de recursos computacionais, garantindo execução eficiente em hardware comum, como notebooks com capacidade média de processamento.

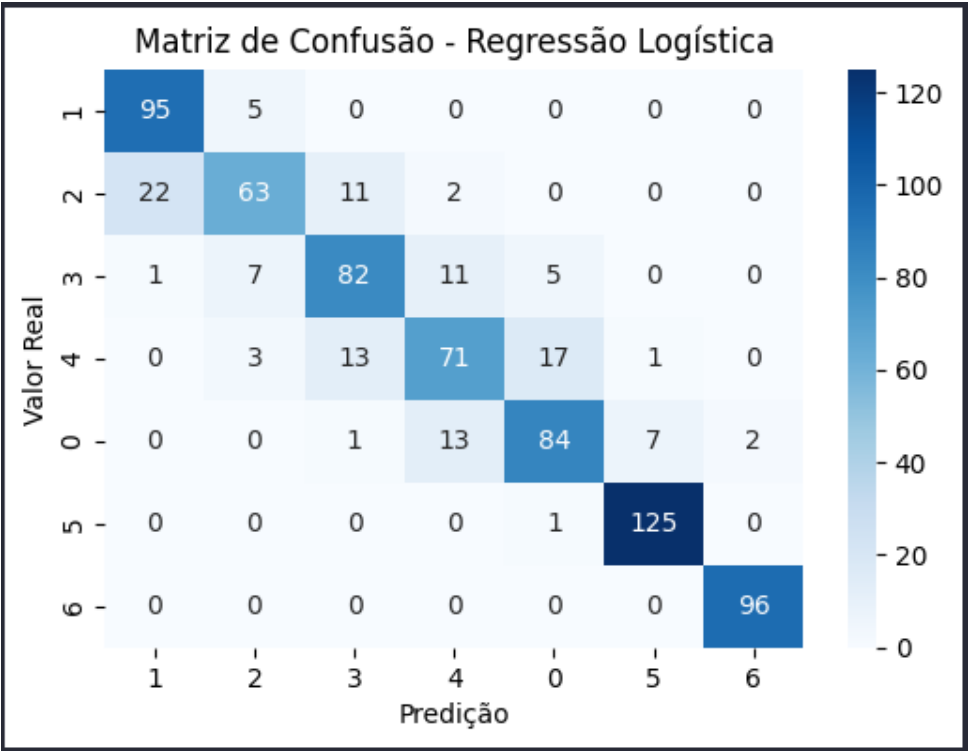
4. Metodologia

4.1. Modelos de IA

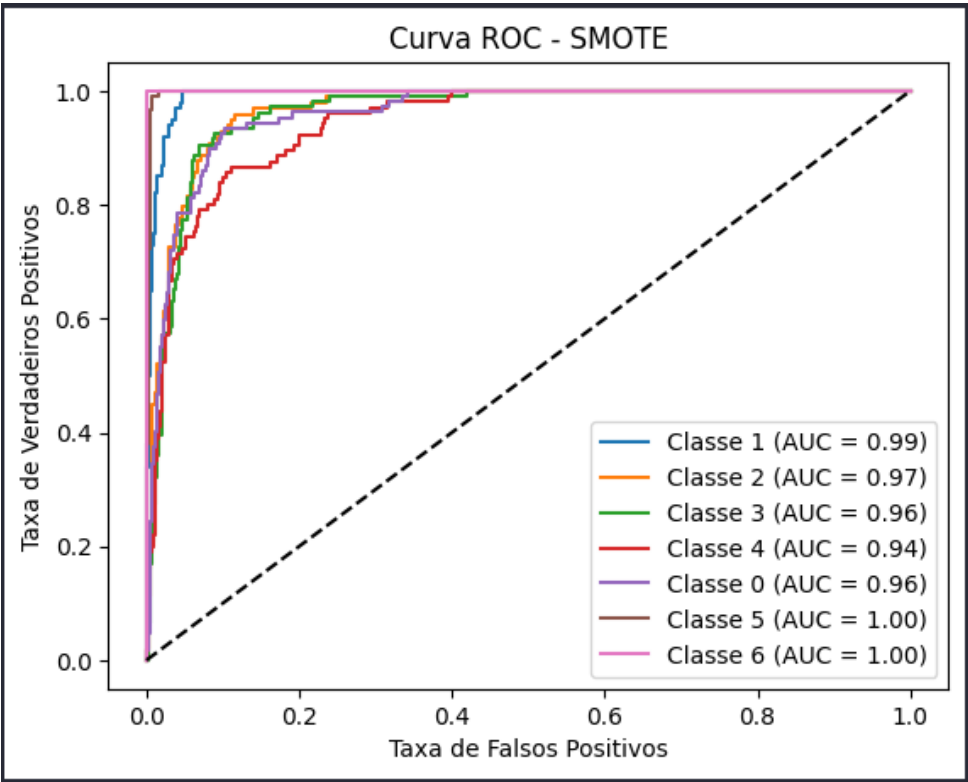
1. **Regressão Logística:** Modelo estatístico clássico para classificação binária ou multiclasse.
2. **Árvore de Decisão:** Algoritmo baseado em regras de decisão para classificação.
3. **Rede Neural Artificial:** Arquitetura com múltiplas camadas para aprendizado não linear.
4. **Comitê de Avaliação:** Abordagem que combina os três modelos, avaliando-os com base em métricas de desempenho para selecionar o mais eficaz na classificação dos níveis de obesidade.

4.2. Avaliação e Comparação

- Regressão Logística:
 - Gráfico Matriz de Confusão:



- Gráfico Curva ROC:



Relatório de Classificação:

nivel_obesidade	precision	recall	f1-score	support
1	0.81	0.95	0.87	100
2	0.81	0.64	0.72	98
3	0.77	0.77	0.77	106
4	0.73	0.68	0.70	105
0	0.79	0.79	0.79	107
5	0.94	0.99	0.97	126
6	0.98	1.00	0.99	96
accuracy			0.83	738
macro avg		0.83	0.83	738
weighted avg		0.83	0.83	738

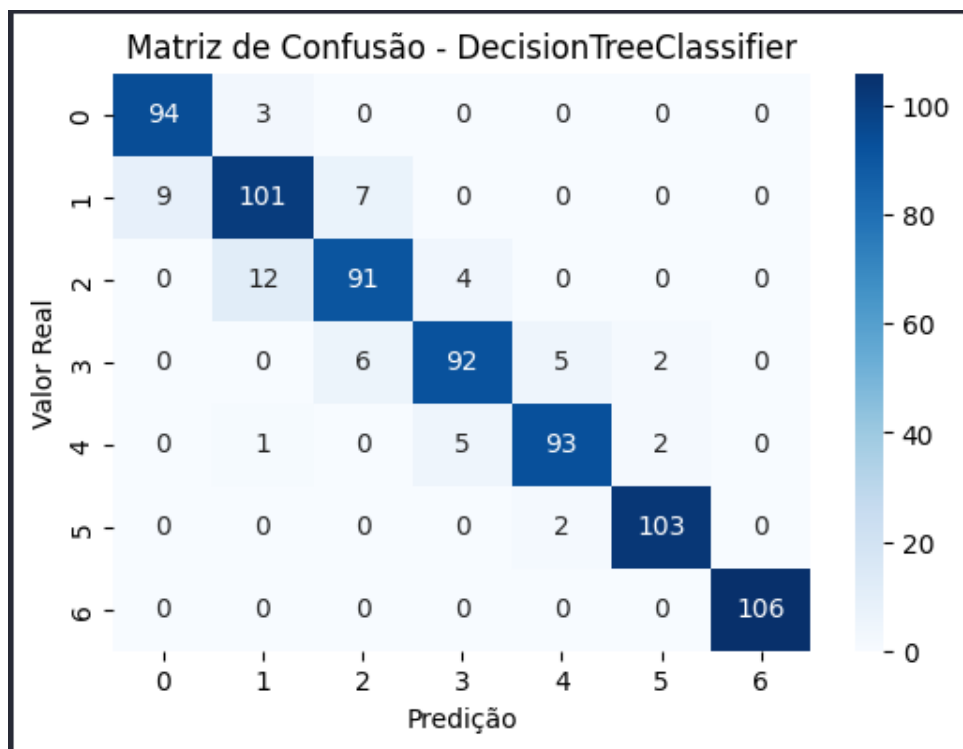
Legenda:

- Abaixo do Peso: 0
- Peso Normal: 1
- Excesso de Peso Nível I: 2
- Excesso de Peso Nível II: 3
- Obesidade Tipo I: 4
- Obesidade Tipo II: 5
- Obesidade Tipo III: 6

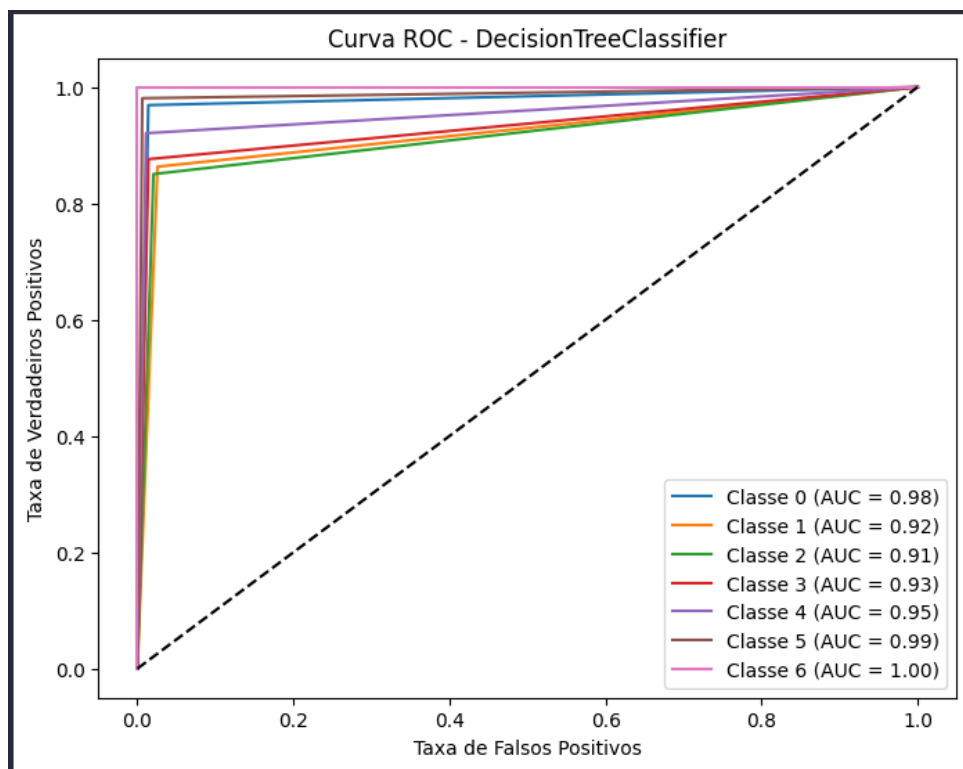
AUC-ROC = 97.44%

F1-score Minor Class = 87.16%

- **Árvore de Decisão:**
 - **Gráfico Matriz de Confusão:**



- **Gráfico Curva ROC:**



Relatório de Classificação:

nivel_obesidade	precision	recall	f1-score	support
0	0.91	0.97	0.94	97
1	0.86	0.86	0.86	117
2	0.88	0.85	0.86	107
3	0.91	0.88	0.89	105
4	0.93	0.92	0.93	101
5	0.96	0.98	0.97	105
6	1.00	1.00	1.00	106
accuracy				0.92
macro avg				0.92
weighted avg				0.92

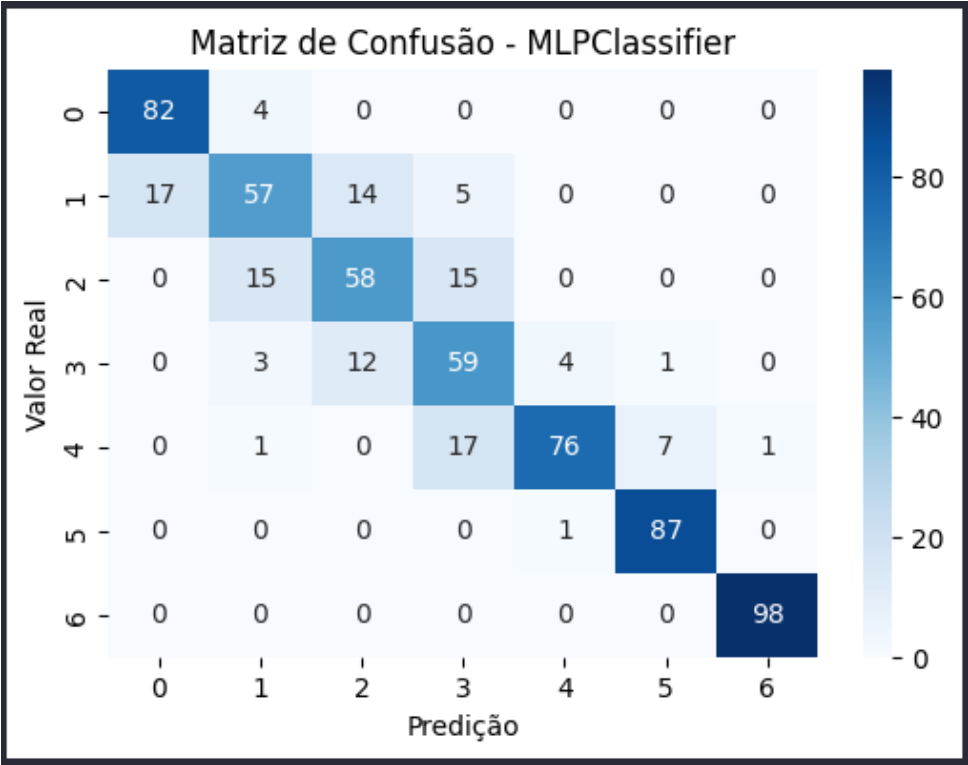
Legenda:

- Abaixo do Peso: 0
- Peso Normal: 1
- Excesso de Peso Nível I: 2
- Excesso de Peso Nível II: 3
- Obesidade Tipo I: 4
- Obesidade Tipo II: 5
- Obesidade Tipo III: 6

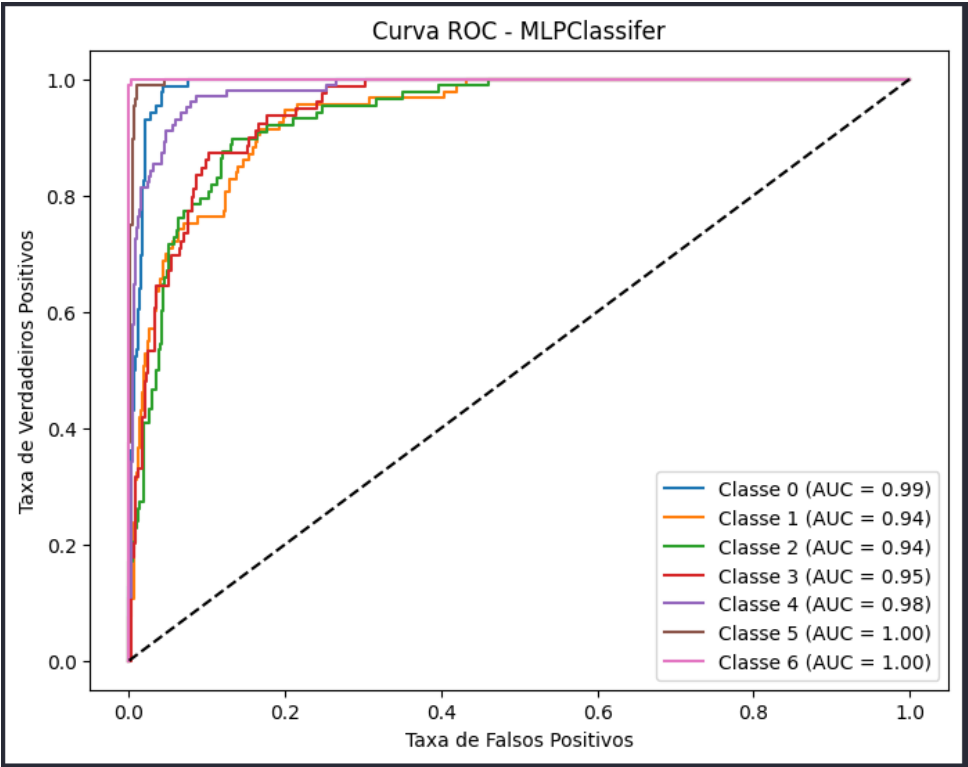
AUC-ROC = 95,49%

F1-score Minor Class = 94%

- Rede Neural Artificial:
 - Gráfico Matriz de Confusão:



- Gráfico Curva ROC:



Relatório de Classificação:

nivel_obesidade	precision	recall	f1-score	support
0	0.83	0.95	0.89	86
1	0.71	0.61	0.66	93
2	0.69	0.66	0.67	88
3	0.61	0.75	0.67	79
4	0.94	0.75	0.83	102
5	0.92	0.99	0.95	88
6	0.99	1.00	0.99	98
accuracy			0.82	634
macro avg			0.81	634
weighted avg			0.82	634

Legenda:

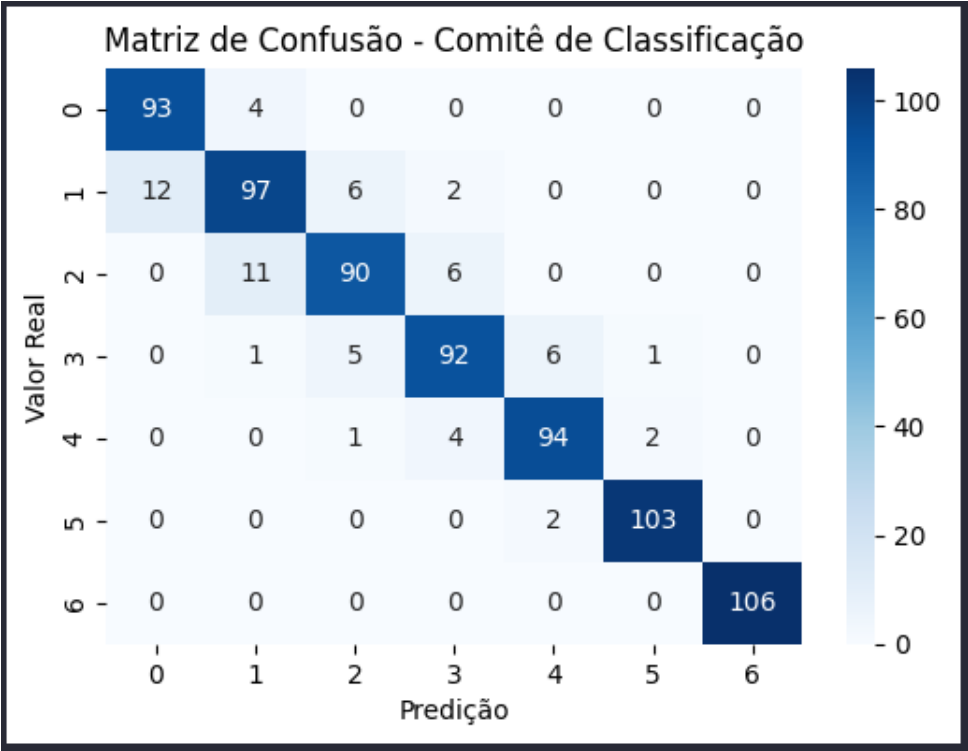
- Abaixo do Peso: 0
- Peso Normal: 1
- Excesso de Peso Nível I: 2
- Excesso de Peso Nível II: 3
- Obesidade Tipo I: 4
- Obesidade Tipo II: 5
- Obesidade Tipo III: 6

AUC-ROC = 97%

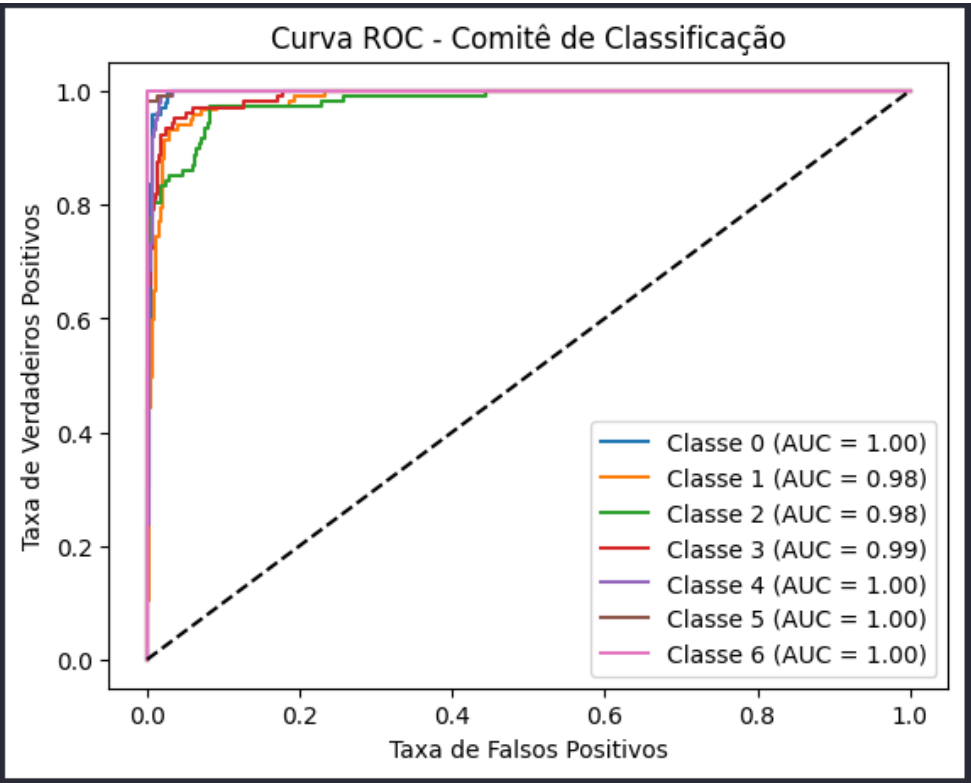
F1-score Minor Class = 99,49%

5. Avaliação do Comitê: O comitê de avaliação proposto neste projeto tem como objetivo comparar os três modelos principais de Machine Learning para identificar o mais eficaz na classificação dos níveis de obesidade. Além de avaliar o desempenho individual de cada modelo, o comitê busca compreender como cada técnica se comporta.

○ **Gráfico Matriz de Confusão:**



○ **Gráfico Curva ROC:**



Relatório de Classificação:

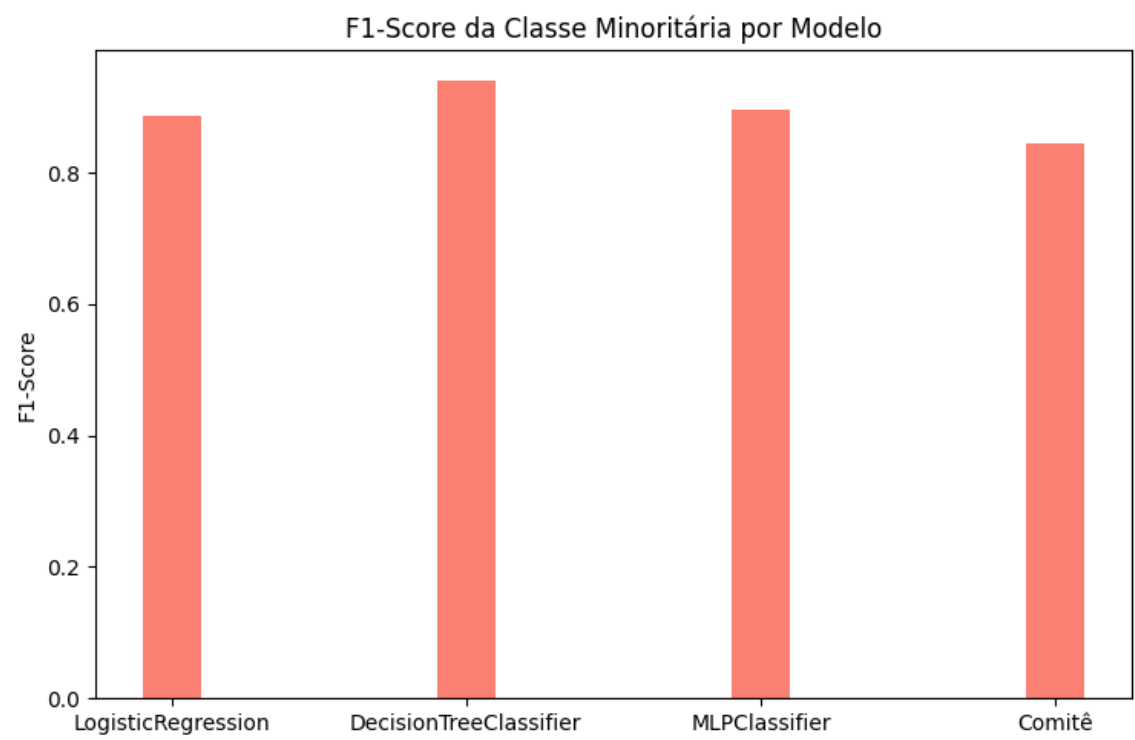
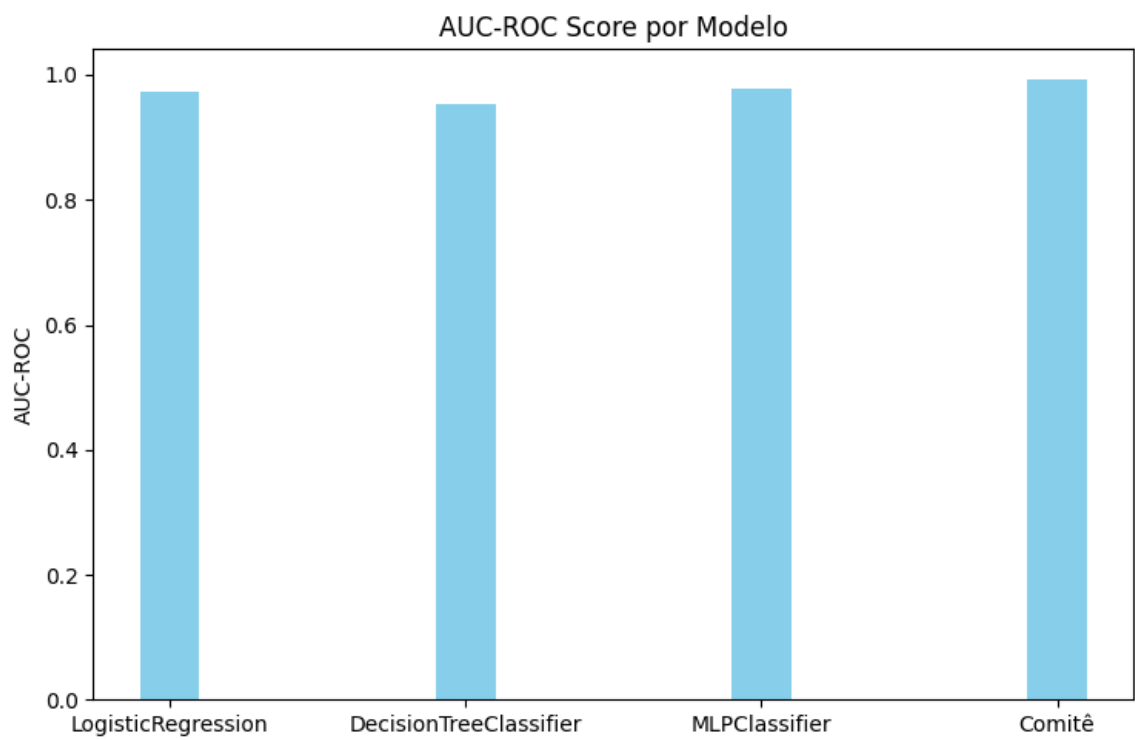
nivel_obesidade	precision	recall	f1-score	support
0	0.89	0.96	0.92	97
1	0.86	0.83	0.84	117
2	0.88	0.84	0.86	107
3	0.88	0.88	0.88	105
4	0.92	0.93	0.93	101
5	0.97	0.98	0.98	105
6	1.00	1.00	1.00	106
accuracy				0.91
macro avg				0.91
weighted avg				0.91

Legenda:

- Abaixo do Peso: 0
- Peso Normal: 1
- Excesso de Peso Nível I: 2
- Excesso de Peso Nível II: 3
- Obesidade Tipo I: 4
- Obesidade Tipo II: 5
- Obesidade Tipo III: 6

Resultados Comparativos entre Modelos:

	AUC-ROC	F1-score Minor Class	Peso Aplicado
Regressão Logística	97,04%	88,68%	33,46
Árvore de Decisão	95,40%	94%	32,85
Rede Neural	97,83%	89,62%	33,69
Comitê	97,04%	99,49%	



8. Conclusão

Este projeto mostrou como técnicas de Machine Learning podem ser aplicadas para classificar níveis de obesidade a partir de dados sobre hábitos alimentares e condições físicas. Com o uso de modelos como Regressão Logística, Árvore de Decisão e Rede Neural, foi possível avaliar o desempenho de diferentes abordagens e identificar a mais eficiente para esse tipo de problema.

A implementação do comitê de avaliação foi essencial para comparar as estratégias de aprendizado e entender melhor como cada modelo se comporta diante de dados balanceados e multiclases. Isso trouxe mais clareza sobre qual método é mais adequado em termos de precisão e confiabilidade.

Os resultados destacam a importância de uma análise cuidadosa, desde o pré-processamento até a validação final, reforçando que não existe uma solução única, mas sim escolhas que dependem do contexto e dos objetivos.

Integrantes do Projeto:

Gustavo Rodrigues – RA 822125117

Juan Souza – RA 822138724

João Pedro Silva – RA 822153960

Marcio Faria – RA 824219962

Professores:

José Carmino Gomes Junior

Bruno Silveira de Lima Honda