# Relatório de Classificação de Obesidade

Nome: Carla Reis (0205722)

Disciplina: Inteligência Artificial e Computacional

Data: 08/07/2024

## 1. Introdução

Este relatório apresenta a análise e os resultados de modelos de classificação aplicados ao conjunto de dados de risco de obesidade. O objetivo é prever o nível de obesidade com base no conjunto de dados baseado no domínio de estudos de saúde e estilo de vida.

## 2. Apresentação do dataset

O conjunto de dados selecionado contém informações detalhadas sobre diversos fatores associados ao risco de obesidade. Cada registro representa uma resposta de um indivíduo a uma série de perguntas sobre seus hábitos alimentares, histórico de saúde e atividades diárias. As respostas foram usadas para identificar o nível de obesidade do indivíduo.

O dataset possui 18 colunas: 17 com dados das entrevistas e uma coluna com o identificador do registro. No total, há 20.758 registros, todos completos, sem valores ausentes. A coluna que vamos considerar como nosso atributo classe é a coluna ‘0be1dad’, que indica o nível de obesidade do indivíduo. O dataset pode ser acessado através deste link.

## 3. Caracterização dos dados

### 3.1 Dados

A tabela abaixo apresenta os atributos do dataset, incluindo o tipo de dados, a quantidade de valores faltantes, o valor mais frequente, o intervalo de valores (mínimo e máximo) e a quantidade de valores únicos.

**Resultados:**

- **Quantidade de Atributos:** 18
- **Quantidade de Instâncias:** 20758

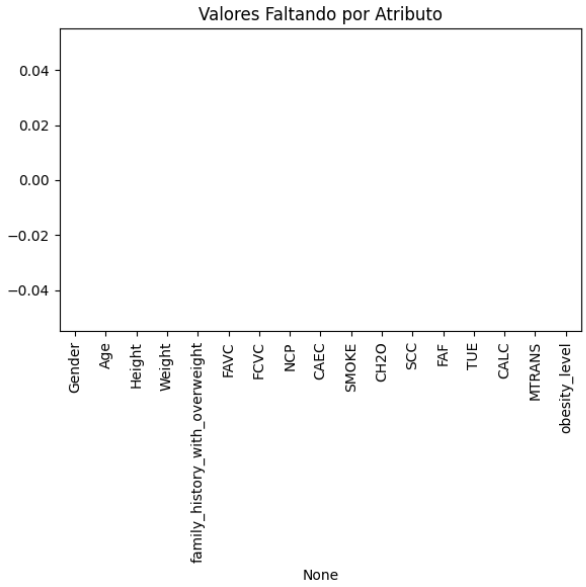
## Atributos:

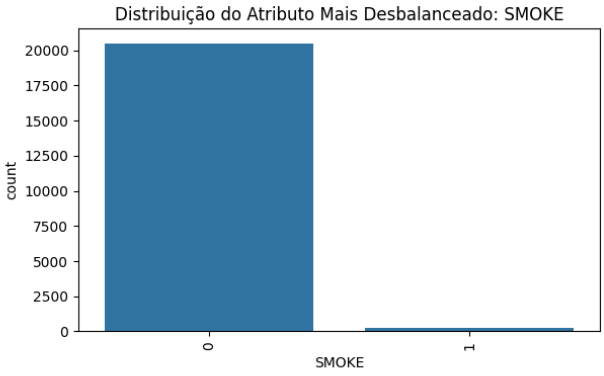
| Coluna | Significado | Tipo |
|---|---|---|
| id | Identificação do registro | int64 |
| Gender | Gênero do indivíduo | object |
| Age | Idade do indivíduo | float64 |
| Height | Altura do indivíduo | float64 |
| Weight | Peso do indivíduo | float64 |
| family_history_with_overweight | Histórico familiar de sobrepeso | int64 |
| FAVC | Frequência de consumo de alimentos calóricos | int64 |
| FCVC | Frequência de consumo de vegetais | float64 |
| NCP | Número de refeições principais por dia | float64 |
| CAEC | Consumo de alimentos entre as refeições | object |
| SMOKE | Hábito de fumar | int64 |
| CH2O | Consumo diário de água | float64 |
| SCC | Monitoramento das calorias consumidas | int64 |
| FAF | Frequência de atividade física | float64 |
| TUE | Tempo de uso de dispositivos eletrônicos | float64 |
| CALC | Consumo de álcool | object |
| MTRANS | Meio de transporte usado | object |
| 0be1dad | Nível de obesidade (rótulo) | object |

## Detalhes dos atributos:

| Atributo | Valores Faltando | Valor Mais Frequente (%) | Mínimo | Máximo | Únicos |
|---|---|---|---|---|---|
| id | 0 | 0.000048 | 0,0000000 | 20757,0000 | NaN |
| Gender | 0 | 0.502071 | NaN | NaN | 2 |
| Age | 0 | 0.092302 | 14,0000000 | 61,0000000 | NaN |
| Height | 0 | 0.064264 | 1,4500000 | 1,9756630 | NaN |
| Weight | 0 | 0.041719 | 39,0000000 | 165,0572690 | NaN |
| family_history_with_overweight | 0 | 0.819636 | 0,0000000 | 1,0000000 | NaN |
| FAVC | 0 | 0.914443 | 0,0000000 | 1,0000000 | NaN |
| FCVC | 0 | 0.368918 | 1,0000000 | 3,0000000 | NaN |
| NCP | 0 | 0.708450 | 1,0000000 | 4,0000000 | NaN |
| CAEC | 0 | 0.844446 | NaN | NaN | 4 |
| SMOKE | 0 | 0.988197 | 0,0000000 | 1,0000000 | NaN |
| CH2O | 0 | 0.318720 | 1,0000000 | 3,0000000 | NaN |
| SCC | 0 | 0.966904 | 0,0000000 | 1,0000000 | NaN |
| FAF | 0 | 0.242991 | 0,0000000 | 3,0000000 | NaN |
| TUE | 0 | 0.316312 | 0,0000000 | 2,0000000 | NaN |
| CALC | 0 | 0.725792 | NaN | NaN | 3 |
| MTRANS | 0 | 0.803883 | NaN | NaN | 5 |
| 0be1dad | 0 | 0.194913 | NaN | NaN | 7 |

### Gráficos:




### 3.2 Classificação dos dados

Os dados foram classificados conforme o Nível de obesidade do indivíduo. A tabela a seguir mostra cada uma

| | |
|---|---|
| **1** | Obesity_Type_III (Obesidade de Grau III) |
| **2** | Obesity_Type_II (Obesidade de Grau II) |
| **3** | 0rmal_Weight (Peso Normal) |
| **4** | Obesity_Type_I (Obesidade de Grau I) |
| **5** | Insufficient_Weight (Peso Insuficiente) |
| **6** | Overweight_Level_II (Sobrepeso Nível II) |
| **7** | Overweight_Level_I (Sobrepeso Nível I) |

### 3.3 Atributo Classe

Para este estudo, o atributo classe escolhido é o 0be1dad, que representa o nível de obesidade do paciente. Este atributo contém as seguintes categorias:

| **Categoria** | **Quantidade de registros** |
|---|---|
| Obesity_Type_III | 4046 |
| Obesity_Type_II | 3248 |
| Normal_Weight | 3082 |
| Obesity_Type_I | 2910 |
| Insufficient_Weight | 2523 |
| Overweight_Level_II | 2522 |
| Overweight_Level_I | 2427 |

## 4. PRÉ-PROCESSAMENTO E TREINAMENTO DE MODELOS

Para preparar os dados para os modelos de classificação, foram realizadas as seguintes etapas:

**Seleção de Atributos:** Os dados foram divididos em features (atributos preditores) e target (atributo a ser previsto). As features incluem informações como gênero, idade, altura, peso, histórico familiar de sobrepeso, entre outros.

**Codificação de Variáveis Categóricas:** Variáveis categóricas, como gênero, histórico familiar de sobrepeso, frequência de consumo de alimentos calóricos, hábito de fumar, entre outras, foram codificadas para que pudessem ser utilizadas nos modelos de aprendizado de máquina.

**Divisão em Conjuntos de Treino e Teste:** Os dados foram divididos aleatoriamente em conjuntos de treino e teste. O conjunto de treino foi utilizado para treinar os modelos, enquanto o conjunto de teste foi utilizado para avaliar o desempenho dos modelos em dados não vistos.

**Treinamento dos Modelos:** Foram utilizados quatro algoritmos de classificação: Árvore de Decisão, SVM (Support Vector Machine), Random Forest e KNN (K-Nearest Neighbors). Cada modelo foi treinado com o conjunto de treino para aprender padrões nos dados.

**Avaliação dos Modelos:** Após o treinamento, cada modelo foi avaliado utilizando métricas de desempenho, como acurácia (proporção de previsões corretas), precisão (proporção de previsões corretas entre as previsões positivas), recall (proporção de positivos reais corretamente previstos) e F1-Score (média harmônica entre precisão e recall). Essas métricas ajudam a entender como cada modelo se saiu na tarefa de prever o nível de obesidade com base nos atributos disponíveis.

Este processo de pré-processamento e treinamento é fundamental para desenvolver modelos de aprendizado de máquina capazes de fazer previsões precisas com base nos dados fornecidos.

## 5. RESULTADOS OBTIDOS

A análise comparativa dos algoritmos de aprendizado de máquina revelou insights significativos sobre sua performance na tarefa de classificação dos níveis de obesidade. Os modelos foram avaliados com base em métricas fundamentais como acurácia, precisão, recall, F1-Score e tempo de processamento. Essas métricas são cruciais não apenas para a precisão das previsões, mas também para a eficiência operacional dos modelos em cenários reais.

A seguir, são apresentados os resultados detalhados de cada algoritmo:

| Algoritmo | Acurácia | Precisão | Recall | F1-Score | Tempo de Processamento (segundos) |
|---|---|---|---|---|---|
| RandomForestClassifier | 0,8947 | 0,8846 | 0,8822 | 0,8832 | 1,95 |
| DecisionTreeClassifier | 0,8480 | 0,8330 | 0,8328 | 0,8328 | 0,12 |
| SVC | 0,7141 | 0,7114 | 0,6938 | 0,6908 | 3,97 |

O **RandomForestClassifier** apresentou a melhor performance geral entre os algoritmos avaliados. Com uma **acurácia de 89,47%**, **precisão de 88,46%**, **recall de 88,22%** e um **F1-Score de 88,32%**, o RandomForestClassifier superou os outros modelos em todas as principais métricas de desempenho. Isso indica que este modelo é o mais confiável para a tarefa de classificação dos níveis de obesidade, fornecendo as previsões mais precisas e equilibradas.

O **DecisionTreeClassifier** também apresentou um bom desempenho, com uma **acurácia de 84,80%**, **precisão de 83,30%**, **recall de 83,28%** e um **F1-Score de 83,28%**. Embora tenha sido um pouco inferior ao RandomForestClassifier, ainda ofereceu uma boa performance com um tempo de treinamento muito mais curto.

O **SVC** (Support Vector Classifier) teve o pior desempenho entre os algoritmos testados, com uma **acurácia de 71,41%**, **precisão de 71,14%**, **recall de 69,38%** e um **F1-Score de 69,08%**. Isso sugere que o SVC não é tão eficaz quanto os outros modelos para a classificação dos níveis de obesidade.

## 6. CONCLUSÕES

A acurácia indica a proporção de previsões corretas feitas pelo modelo. Os algoritmos RandomForestClassifier e DecisionTreeClassifier apresentaram acurácias superiores, indicando que eles tiveram a maior taxa de acertos na classificação dos níveis de obesidade. Isso sugere que esses modelos são eficazes em distinguir corretamente entre diferentes classes de obesidade, proporcionando previsões precisas.

A precisão mede a proporção de verdadeiros positivos entre as previsões positivas feitas pelo modelo. O RandomForestClassifier destacou-se com a maior precisão, indicando que, das instâncias previstas como pertencentes a uma determinada classe de obesidade, uma alta proporção realmente correspondia a casos verdadeiros dessa classe. Isso resultou em menos falsos positivos, mostrando que o modelo é confiável ao identificar corretamente os casos positivos de obesidade.

O recall, ou sensibilidade, mede a proporção de verdadeiros positivos identificados entre todas as instâncias positivas reais. O RandomForestClassifier apresentou o maior recall, demonstrando sua capacidade de identificar a maioria dos casos reais de obesidade entre todas as instâncias positivas existentes no dataset. Isso significa que o modelo conseguiu capturar a maior parte dos indivíduos que realmente pertencem às classes de obesidade, minimizando os falsos negativos.

O F1-Score é a média harmônica da precisão e do recall, proporcionando uma medida balanceada do desempenho do modelo. O RandomForestClassifier obteve o F1-Score mais elevado, indicando um bom equilíbrio entre precisão e recall. Isso é essencial para a qualidade geral do modelo, pois avalia o desempenho em situações em que tanto falsos positivos quanto falsos negativos são importantes.

O tempo de processamento refere-se ao tempo necessário para treinar o modelo. Embora o RandomForestClassifier tenha apresentado a melhor performance geral em termos de acurácia, precisão, recall e F1-Score, o DecisionTreeClassifier teve um tempo de treinamento significativamente menor. Isso pode ser uma vantagem em cenários onde o tempo de processamento é crítico.

O RandomForestClassifier se destacou como o modelo mais eficaz para prever níveis de obesidade, apresentando o melhor desempenho geral em termos de acurácia, precisão, recall e F1-Score. Embora o DecisionTreeClassifier tenha mostrado um desempenho ligeiramente inferior, seu tempo de treinamento mais rápido o torna uma alternativa viável, especialmente em cenários onde a eficiência do processamento é crucial. Esses resultados indicam que ambos os algoritmos são adequados para a classificação de obesidade, com o RandomForestClassifier sendo preferido para maior precisão e o DecisionTreeClassifier para maior rapidez.