# Análise de Dados Médicos: Pré-processamento e Modelagem para Diagnóstico de Doenças

Aluna: Bruna Mayumi Hori

Professor: Emmanuel Andrade

# Introdução

Nossa análise se concentra em compreender a relação entre doenças e seus sintomas por meio de técnicas de análise de dados. O objetivo deste projeto é explorar um conjunto de dados que contém informações sobre várias doenças e os sintomas associados a elas. Ao entender melhor essas relações, podemos desenvolver modelos que auxiliam na identificação e diagnóstico preciso de doenças com base nos sintomas apresentados pelos pacientes.

# Pré-processamento de Dados:

Tratamento de Valores Ausentes:

Os valores ausentes são valores em branco ou faltantes em um conjunto de dados.

O tratamento de valores ausentes é importante para garantir a qualidade dos dados e evitar distorções nos resultados da análise.

Existem várias maneiras de lidar com valores ausentes, como preenchê-los com valores médios, medianos ou moda, ou até mesmo remover as linhas ou colunas que contêm valores ausentes.

A escolha do método de tratamento de valores ausentes depende do contexto do problema e da natureza dos dados.

Remoção de Duplicatas:

Duplicatas são linhas repetidas em um conjunto de dados.

A remoção de duplicatas é importante para garantir a integridade dos dados e evitar viés nos resultados da análise.

Geralmente, as duplicatas são identificadas com base em todas as colunas ou em uma chave única específica.

A remoção de duplicatas pode ser realizada antes ou depois de outras etapas de pré-processamento, dependendo do contexto do problema.

Normalização de Dados:

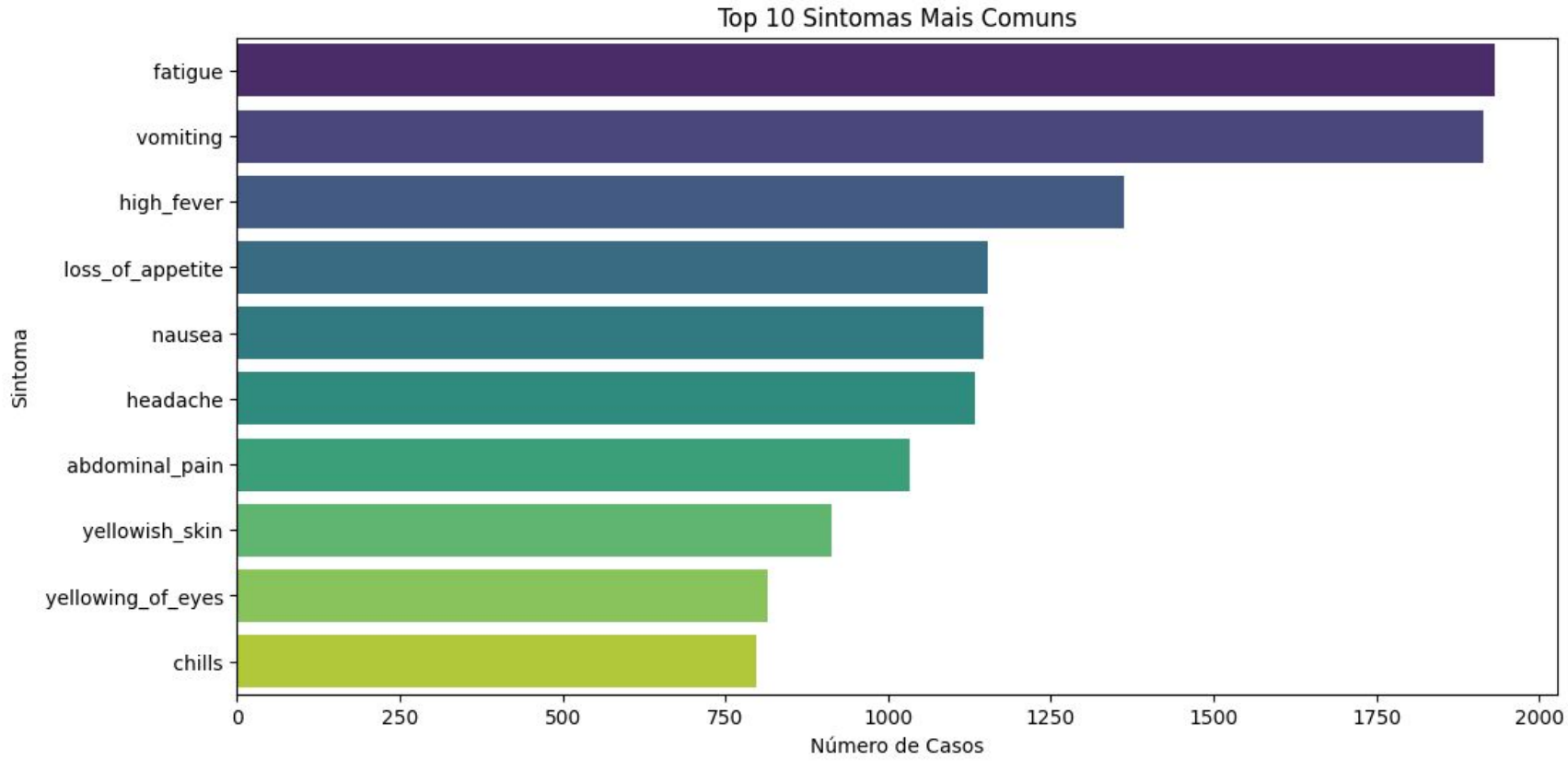
A normalização de dados é o processo de dimensionar os valores das características para uma escala específica, geralmente entre 0 e 1.

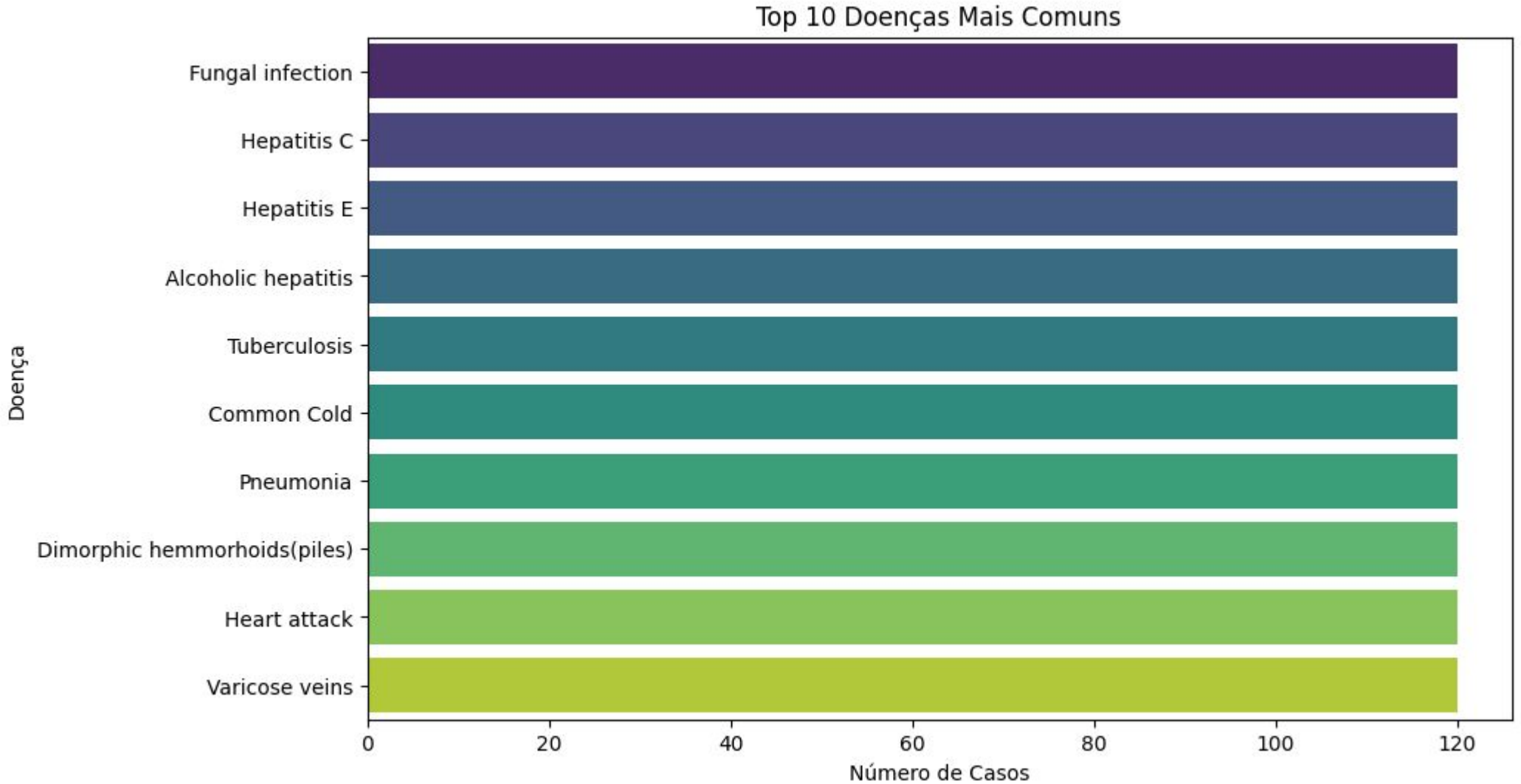
A normalização é importante quando as características do conjunto de dados têm escalas diferentes e podem distorcer os resultados dos modelos de machine learning.

Existem várias técnicas de normalização, como Min-Max Scaling, Z-score Normalization e Normalização por Escala Logarítmica.

A escolha da técnica de normalização depende do tipo de dados e dos requisitos do modelo de machine learning.

# A análise de dados de doenças e sintomas.




Top 10 Doenças Mais Comuns
Fungal infection
Hepatitis C
Hepatitis E
Alcoholic hepatitis
Tuberculosis
Common Cold
Pneumonia
Dimorphic hemmorhoids(piles)
Heart attack
Varicose veins
Doença
0
20
40
60
80
100
120
Número de Casos

**Precision (Precisão):** A precisão é a proporção de verdadeiros positivos (casos corretamente classificados como pertencentes à classe) em relação a todos os casos classificados como pertencentes à classe. Por exemplo, para a classe 'migraine', a precisão é de 1.00, o que significa que todos os casos classificados como 'migraine' estão corretos.

**Recall (Revocação):** A revocação é a proporção de verdadeiros positivos em relação a todos os casos que realmente pertencem à classe. Por exemplo, para a classe 'psoriasis', a revocação é de 0.00, o que indica que nenhum caso verdadeiro de 'psoriasis' foi corretamente identificado pelo modelo.

**F1-score (F1-Score):** O F1-score é a média harmônica da precisão e da revocação. É uma métrica útil para resumir o desempenho do modelo em uma única medida. Quanto mais próximo de 1.00, melhor.

**Support (Suporte):** O suporte é o número de ocorrências reais de cada classe no conjunto de dados de teste.

**Acurácia (Accuracy):** A acurácia é a proporção de todos os casos corretamente classificados em relação ao total de casos. Por exemplo, neste relatório, a acurácia é de 0.72, o que significa que 72% dos casos foram corretamente classificados pelo modelo.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| (vertigo) paroymsal positional vertigo | 1.00 | 1.00 | 1.00 | 2 |
| acne | 0.50 | 0.67 | 0.57 | 3 |
| aids | 1.00 | 1.00 | 1.00 | 1 |
| alcoholic hepatitis | 1.00 | 1.00 | 1.00 | 1 |
| allergy | 1.00 | 0.33 | 0.50 | 3 |
| bronchial asthma | 0.33 | 0.33 | 0.33 | 3 |
| cervical spondylosis | 1.00 | 1.00 | 1.00 | 1 |
| chicken pox | 1.00 | 0.67 | 0.80 | 3 |
| chronic cholestasis | 0.00 | 0.00 | 0.00 | 2 |
| dengue | 1.00 | 1.00 | 1.00 | 3 |
| diabetes | 1.00 | 1.00 | 1.00 | 2 |
| dimorphic hemmorhoids(piles) | 1.00 | 1.00 | 1.00 | 2 |
| drug reaction | 1.00 | 1.00 | 1.00 | 1 |
| fungal infection | 0.00 | 0.00 | 0.00 | 0 |
| hepatitis a | 0.50 | 1.00 | 0.67 | 2 |
| hepatitis b | 1.00 | 1.00 | 1.00 | 2 |
| hepatitis c | 0.00 | 0.00 | 0.00 | 1 |
| hepatitis d | 0.50 | 1.00 | 0.67 | 1 |
| hepatitis e | 1.00 | 1.00 | 1.00 | 2 |
| hyperthyroidism | 0.50 | 0.50 | 0.50 | 2 |
| hypoglycemia | 1.00 | 0.67 | 0.80 | 3 |
| hypothyroidism | 0.67 | 1.00 | 0.80 | 2 |
| impetigo | 0.00 | 0.00 | 0.00 | 1 |
| jaundice | 1.00 | 1.00 | 1.00 | 1 |
| malaria | 1.00 | 0.50 | 0.67 | 2 |
| migraine | 1.00 | 1.00 | 1.00 | 4 |
| paralysis (brain hemorrhage) | 1.00 | 1.00 | 1.00 | 1 |
| peptic ulcer diseae | 0.00 | 0.00 | 0.00 | 1 |
| pneumonia | 1.00 | 0.50 | 0.67 | 2 |
| psoriasis | 0.00 | 0.00 | 0.00 | 1 |
| tuberculosis | 1.00 | 1.00 | 1.00 | 2 |
| urinary tract infection | 0.25 | 1.00 | 0.40 | 1 |
| varicose veins | 1.00 | 0.67 | 0.80 | 3 |
| accuracy | | | 0.72 | 61 |
| macro avg | 0.70 | 0.69 | 0.67 | 61 |
| weighted avg | 0.78 | 0.72 | 0.72 | 61 |