Análise Detalhada do Dataset Titanic

Prof. Dr. Fabiano B. Menegidio

1 Introdução

O **dataset Titanic** é um dos mais conhecidos e amplamente utilizados em análises de dados, aprendizado de máquina e visualizações. Ele contém informações sobre os passageiros do famoso navio RMS Titanic, que afundou durante sua viagem inaugural em 1912. O dataset é frequentemente usado para demonstrar técnicas de análise e predição, devido à sua simplicidade e estrutura bem definida. A principal tarefa é prever se um passageiro sobreviveu ou não ao desastre, com base nas informações disponíveis.

2 Estrutura do Dataset

O dataset geralmente possui uma tabela com várias colunas (features) que descrevem as características dos passageiros e uma coluna de resultado que indica se o passageiro sobreviveu ao naufrágio. A seguir, uma descrição detalhada das colunas mais comuns encontradas no dataset Titanic:

- **PassengerId**:
  - **Descrição**: Um identificador numérico único para cada passageiro.
  - **Tipo de dado**: Inteiro.
  - **Uso**: Apenas para identificação; normalmente não é usado para análise ou modelagem.
- **Survived**:
  - **Descrição**: Indica se o passageiro sobreviveu ou não ao desastre.
    - 0: Não sobreviveu.
    - 1: Sobreviveu.
  - **Tipo de dado**: Binário (0 ou 1).
  - **Uso**: Esta é a **variável-alvo** quando utilizamos o dataset em algoritmos de predição. Nosso objetivo é prever essa variável com base nas outras características do passageiro.
- **Pclass**:

  - **Descrição**: Classe de viagem do passageiro.
    * 1: Primeira classe.
    * 2: Segunda classe.
    * 3: Terceira classe.
  - **Tipo de dado**: Categórico (ordinal).
  - **Uso**: Representa o status socioeconômico do passageiro, o que se mostrou um fator importante na taxa de sobrevivência. Passageiros da primeira classe tinham uma probabilidade muito maior de sobrevivência do que os da terceira classe.

- **Name**:
  - **Descrição**: Nome completo do passageiro.
  - **Tipo de dado**: Texto.
  - **Uso**: O nome pode fornecer informações adicionais, como o título social do passageiro (Sr., Sra., Srta., etc.), que pode ser usado para inferir informações sobre gênero, status social e até mesmo idade.

- **Sex**:
  - **Descrição**: Sexo do passageiro (male ou female).
  - **Tipo de dado**: Categórico (nominal).
  - **Uso**: O gênero foi um fator importante na taxa de sobrevivência. As mulheres tinham uma probabilidade significativamente maior de sobreviver.

- **Age**:
  - **Descrição**: Idade do passageiro em anos.
  - **Tipo de dado**: Numérico (contínuo).
  - **Uso**: A idade também teve um impacto na taxa de sobrevivência. Crianças, especialmente, tinham uma chance maior de sobrevivência.

- **SibSp**:
  - **Descrição**: Número de irmãos/esposas a bordo do Titanic.
  - **Tipo de dado**: Numérico (discreto).
  - **Uso**: Esta variável ajuda a identificar se o passageiro estava viajando em grupo ou sozinho.

- **Parch**:
  - **Descrição**: Número de pais/filhos a bordo do Titanic.
  - **Tipo de dado**: Numérico (discreto).

  - **Uso**: Assim como *SibSp*, esta variável mostra se o passageiro estava acompanhado de familiares.

- **Ticket**:
  - **Descrição**: Número do bilhete do passageiro.
  - **Tipo de dado**: Texto.
  - **Uso**: O número do bilhete pode conter padrões que indicam o tipo de ticket comprado ou o grupo em que o passageiro estava.

- **Fare**:
  - **Descrição**: Tarifa paga pelo passageiro.
  - **Tipo de dado**: Numérico (contínuo).
  - **Uso**: O valor da tarifa está correlacionado com a classe do passageiro (Pclass). Passageiros que pagaram mais estavam em classes mais altas.

- **Cabin**:
  - **Descrição**: Número da cabine do passageiro.
  - **Tipo de dado**: Texto (muitas vezes incompleto).
  - **Uso**: A localização da cabine no navio pode ser um fator importante para a sobrevivência.

- **Embarked**:
  - **Descrição**: Porto de embarque do passageiro.
    * C: Cherbourg.
    * Q: Queenstown.
    * S: Southampton.
  - **Tipo de dado**: Categórico (nominal).
  - **Uso**: O porto de embarque pode fornecer informações adicionais sobre a origem geográfica e socioeconômica dos passageiros.

3 Análise Descritiva do Dataset

O dataset Titanic é bastante utilizado para demonstrar técnicas de pré-processamento de dados, como tratamento de valores ausentes, imputação de dados e codificação de variáveis categóricas. Vamos analisar algumas características importantes com base nas colunas acima.

3.1 Valores Ausentes

Há uma quantidade significativa de dados ausentes em colunas como *age*, *cabin* e *embarked*. Essas colunas geralmente precisam ser tratadas antes da modelagem. Por exemplo:

- A coluna *age* pode ter seus valores ausentes preenchidos com a média ou mediana.
- A coluna *cabin* tem uma alta taxa de valores ausentes, o que pode tornar seu uso inviável sem um pré-processamento adequado.
- A coluna *embarked* tem poucos valores ausentes e pode ser preenchida com a moda (valor mais frequente).

3.2 Correlação entre Variáveis

Algumas variáveis estão altamente correlacionadas com a sobrevivência. Por exemplo:

- **Sexo**: As mulheres tinham uma probabilidade muito maior de sobreviver.
- **Classe**: Passageiros da primeira classe tinham uma taxa de sobrevivência significativamente maior.
- **Idade**: Crianças tinham uma chance relativamente maior de sobreviver em comparação aos adultos.

3.3 Variáveis Categóricas

As variáveis categóricas, como *Pclass*, *Sex*, *Embarked*, são frequentemente transformadas em variáveis numéricas usando técnicas como *One-Hot Encoding* para serem utilizadas em algoritmos de aprendizado de máquina.

3.4 Distribuição de Variáveis Numéricas

As variáveis numéricas, como *fare* e *age*, costumam ter distribuições assimétricas (não normal). Isso pode afetar a performance de alguns algoritmos de aprendizado de máquina e pode exigir transformações (como log ou normalização).

4 Análise de Sobrevivência

O principal foco das análises que utilizam o dataset Titanic é a previsão de sobrevivência dos passageiros. Alguns insights gerais incluem:

- **Sexo**: Cerca de 74% das mulheres sobreviveram, enquanto apenas 19% dos homens sobreviveram.

- **Classe**: Passageiros da primeira classe tinham uma taxa de sobrevivência de aproximadamente 62%, enquanto passageiros da terceira classe tinham apenas 24% de chance de sobreviver.
- **Idade**: Crianças tinham uma probabilidade maior de sobreviver, refletindo a prioridade dada a crianças durante a evacuação do navio.

5 Aplicações do Dataset

O dataset Titanic é amplamente utilizado para:

- **Aprendizado de Máquina**: Exemplo clássico de problemas de classificação, onde o objetivo é prever se o passageiro sobreviveu ou não com base nas suas características.
- **Análise Exploratória de Dados (EDA)**: Ideal para explorar técnicas de visualização de dados, correlações e padrões de sobrevivência.
- **Ensino de Pré-Processamento de Dados**: O dataset contém exemplos de dados ausentes, variáveis categóricas e variáveis numéricas, o que o torna útil para ensinar e aplicar técnicas de limpeza e transformação de dados.

6 Conclusão

O dataset Titanic oferece uma excelente oportunidade para praticar análises de dados, visualizações e técnicas de aprendizado de máquina. Suas variáveis bem definidas, com padrões interessantes de sobrevivência, tornam-no um exemplo perfeito para explorar o comportamento de modelos preditivos em um problema clássico de classificação.