# Relatório de Análise do Dataset Titanic: Um Pipeline Completo de Inteligência Artificial

Kaiky França

June 2, 2025

**Abstract**

Este relatório detalha o processo de análise do conjunto de dados do Titanic, utilizando um pipeline completo de Inteligência Artificial. O projeto abrange desde o pré-processamento dos dados, passando pela aplicação de algoritmos de classificação supervisionada para prever a sobrevivência dos passageiros, até a utilização de algoritmos de agrupamento para identificar perfis semelhantes e a extração de regras de associação para descobrir padrões de comportamento. Os resultados demonstram a eficácia das técnicas empregadas e fornecem insights valiosos sobre os fatores que influenciaram a sobrevivência no trágico evento.

# Contents

# 1 Introdução

O desastre do RMS Titanic é um dos eventos marítimos mais conhecidos da história. O conjunto de dados associado a este evento tornou-se um estudo de caso clássico em ciência de dados e aprendizado de máquina. Este projeto tem como objetivo aplicar um pipeline completo de Inteligência Artificial sobre o arquivo `train.csv`, contendo informações dos passageiros, para realizar as seguintes tarefas conforme solicitado na Lista #11:

- Prever a sobrevivência dos passageiros utilizando modelos de classificação supervisionada.
- Identificar grupos de passageiros com perfis semelhantes através de algoritmos de agrupamento.
- Extrair regras de associação para identificar padrões interessantes no comportamento dos passageiros.

O relatório segue a estrutura do notebook Jupyter desenvolvido, detalhando cada etapa do processo e as decisões tomadas.

# 2 Pré-processamento de Dados

A etapa de pré-processamento é crucial para garantir a qualidade dos dados que alimentarão os modelos de aprendizado de máquina, conforme descrito na Questão 01 da Lista #11.

## 2.1 Carregamento e Análise Exploratória Inicial

Os dados foram carregados a partir do arquivo `train.csv`. Uma análise exploratória inicial revelou a estrutura do dataset, os tipos de dados de cada coluna e a presença de valores ausentes. As colunas com valores nulos significativos identificadas foram `Age`, `Cabin`, e `Embarked`.

## 2.2 Tratamento de Valores Ausentes e Engenharia de Atributos

Para tratar os valores ausentes:

- A coluna `Age` teve seus valores nulos preenchidos pela mediana das idades.
- A coluna `Embarked` teve seus valores nulos preenchidos pela moda (local de embarque mais frequente).

Foram criadas as seguintes variáveis derivadas, conforme sugerido:

- `FamilySize`: Soma de `SibSp` (irmãos/cônjuges a bordo) e `Parch` (pais/filhos a bordo) mais 1 (o próprio passageiro). As colunas `SibSp` e `Parch` foram então removidas.
- `HasCabin`: Variável binária indicando se a informação da cabine (`Cabin`) estava disponível (1) ou não (0). A coluna `Cabin` original foi removida devido ao alto número de valores ausentes.

### 2.3 Transformação de Variáveis Categóricas

As variáveis categóricas foram convertidas para formato numérico:

- `Sex`: Mapeada para 0 (masculino) e 1 (feminino).
- `Embarked`: Convertida utilizando One-Hot Encoding, com a remoção da primeira categoria para evitar multicolinearidade (`drop_first=True`).

As colunas `Name` e `Ticket` foram removidas antes da modelagem de classificação por não serem diretamente utilizáveis em sua forma original para os algoritmos escolhidos. A normalização ou padronização das variáveis numéricas foi considerada e aplicada posteriormente na etapa de agrupamento.

## 3 Modelagem com Algoritmos de Classificação

Foram utilizados dois algoritmos de classificação supervisionada (Random Forest e Naive Bayes) para prever a sobrevivência dos passageiros. Os dados foram divididos em 80% para treino e 20% para teste. Os modelos foram avaliados com precisão, recall e F1-Score.

### 3.1 Random Forest Classifier

O Random Forest foi treinado com 100 estimadores e `class_weight='balanced'`. Os resultados para a classe "Sobreviveu" (classe 1) foram:

- Acurácia (geral do modelo): 0.8547
- Precisão (para classe 1): 0.8636
- Recall (para classe 1): 0.7703
- F1-Score (para classe 1): 0.8143

### 3.2 Naive Bayes (GaussianNB)

O modelo Gaussian Naive Bayes obteve os seguintes resultados para a classe "Sobreviveu" (classe 1):

- Precisão: 0.7407
- Recall: 0.8108
- F1-Score: 0.7742

### 3.3 Comparação dos Modelos de Classificação

O Random Forest apresentou um F1-Score superior para a classe de sobreviventes (0.8143) em comparação com o Naive Bayes (0.7742), indicando um melhor equilíbrio entre precisão e recall para este modelo. A precisão do Random Forest também foi maior, enquanto o Naive Bayes teve um recall ligeiramente superior para os sobreviventes.

## 4 Modelagem com Algoritmos de Agrupamento (K-Means)

Foi utilizado o algoritmo K-Means para identificar agrupamentos de passageiros com perfis semelhantes, com $k = 4$ clusters.

### 4.1 Preparação dos Dados e Aplicação

As features selecionadas para clusterização foram: `Sex`, `Age`, `FamilySize`, `HasCabin`, `Embarked_Q`, `Embarked_S`. Estas features foram padronizadas usando `StandardScaler`.

### 4.2 Visualização e Interpretação dos Clusters

Os clusters foram visualizados em 2D utilizando PCA. A análise das características médias e distribuições de variáveis chave por cluster revelou os seguintes perfis principais:

- **Cluster 0:** Principalmente mulheres jovens com famílias, da 3ª classe, embarcadas em Southampton, geralmente sem cabine registrada. Taxa de sobrevivência: 58%.
- **Cluster 1:** Passageiros de ambos os sexos, da 3ª classe, que todos embarcaram em Queenstown, geralmente sem cabine. Taxa de sobrevivência: 39%.
- **Cluster 2:** Passageiros mais velhos, de ambos os sexos, predominantemente da 1ª classe, todos com cabine registrada. Maior taxa de sobrevivência: 66.5%.

- **Cluster 3:** Quase exclusivamente homens, de idade média, viajando sozinhos ou em pequenas famílias, principalmente da 3ª e 2ª classe, sem cabine registrada. Menor taxa de sobrevivência: 14.7%.

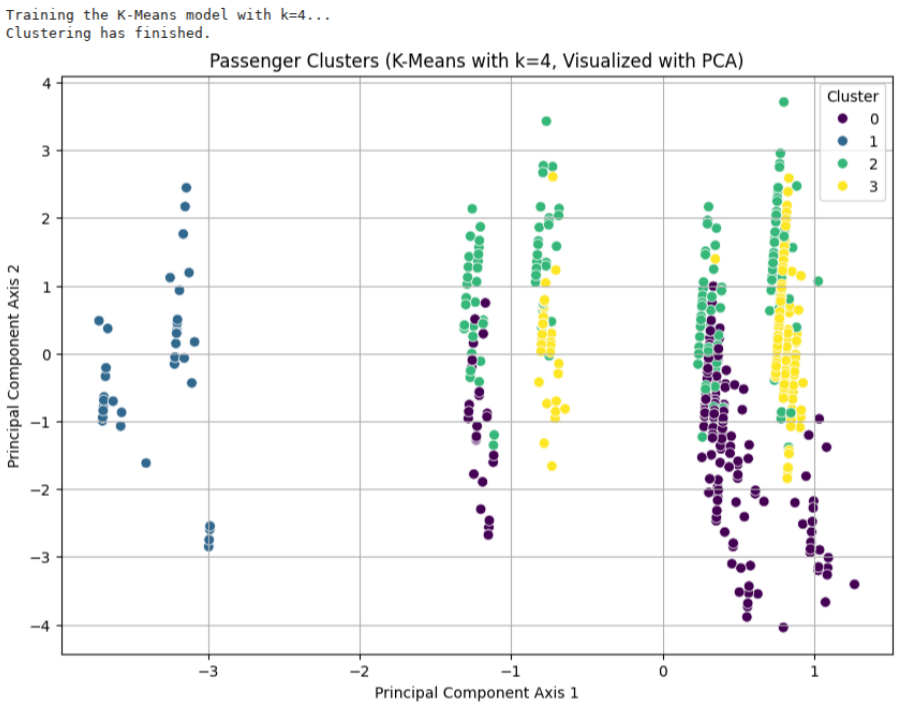


Figure 1: Clusters de Passageiros (K-Means com k=4, Visualizado com PCA)

# 5 Extração de Regras de Associação (Apriori)

O algoritmo Apriori foi aplicado para identificar padrões interessantes.

## 5.1 Preparação dos Dados

As variáveis numéricas `Age` e `Fare` foram discretizadas. As variáveis `Pclass`, `Sex`, `Survived` (como `Survived_Rule`) e `FamilySize` (como `FamilySize_Group`) foram formatadas como itens. A matriz transacional foi criada usando `pd.get_dummies`.

## 5.2 Aplicação do Apriori e Interpretação das Regras

Itemsets frequentes foram encontrados com `min_support=0.01`, e regras foram geradas com `metric="lift"` e `min_threshold=1.2`. Pelo menos 3 regras devem ser interpretadas com suporte, confiança e lift.

**Regra Exemplo 1 (Baseada na saída anterior):**

- *Antecedents:* {'Fare_Group_Fare_VeryHigh', 'Age_Group_Age_Child', 'Survived_Rule_Survived_No

- *Consequents:* {'FamilySize_Group_Family_Large', 'Pclass_Pclass_3'}
- *Support:* $\approx 0.0146$
- *Confidence:* $\approx 0.9286$
- *Lift:* $\approx 13.13$
- *Interpretação:* Crianças que pagaram tarifa muito alta e não sobreviveram têm uma probabilidade muito alta (92.86%) de também pertencerem a famílias grandes na 3ª classe. O Lift de aproximadamente 13.13 indica uma associação muito forte, ocorrendo 13 vezes mais do que o esperado caso fossem independentes.

*(O aluno deverá adicionar a interpretação de mais duas regras aqui, baseadas na saída do seu notebook.)*

A análise das regras pode revelar associações como a indagada no exercício: "A combinação de 'sexo feminino' e 'classe $= 1^a$' implica em alta probabilidade de sobrevivência?".

# 6 Conclusão

Este projeto demonstrou a aplicação de um pipeline de Inteligência Artificial no dataset Titanic, conforme os objetivos de aprendizagem da atividade. O pré-processamento adequado dos dados foi fundamental para o sucesso das etapas de modelagem. Os modelos de classificação supervisionada, especialmente o Random Forest, mostraram boa capacidade preditiva para a sobrevivência. A clusterização com K-Means permitiu a identificação de perfis de passageiros com características e taxas de sobrevivência distintas. Finalmente, as regras de associação extraídas com o Apriori revelaram padrões de coocorrência interessantes entre os atributos dos passageiros. A comparação dos modelos e os insights obtidos foram incluídos ao longo do desenvolvimento no notebook e resumidos aqui. Este relatório cumpre a exigência de entrega além do notebook.

# References


[1] Código Fonte. Disponível no meu GitHub: `https://github.com/GLKaiky/PucMinas/tree/master/Ia/Lista%2011`

[2] Titanic - Machine Learning from Disaster. Kaggle. Disponível em: `https://www.kaggle.com/c/titanic/data`

[3] Profª Cristiane Neri Nobre. Lista #11 - Pipeline e Titanic. Curso: Ciência da Computação, Disciplina: Inteligência Artificial.