Relatório Final – Lista 11: Análise do Dataset Titanic com Aprendizado de Máquina e Mineração de Dados

Disciplina: Inteligência Artificial
Aluno: Vinícius Miranda de Araujo
Tema: Modelagem Preditiva, Clusterização e Regras de Associação com o Dataset Titanic

---

Objetivo

O objetivo deste trabalho foi aplicar técnicas de Inteligência Artificial ao famoso conjunto de dados do Titanic, visando:

- Criar modelos preditivos de sobrevivência.
- Identificar agrupamentos de passageiros com perfis semelhantes.
- Extrair padrões comportamentais por meio de regras de associação.

Etapas Realizadas

1. Pré-processamento

O conjunto de dados original foi limpo e transformado por meio das seguintes etapas:

- **Tratamento de valores ausentes**: preenchimento com a média para 'Age' e 'Fare'.
- **Codificação de variáveis categóricas**: '`Sex`' com `LabelEncoder`, e '`Title`' extraído do nome dos passageiros e convertido para categorias.
- **Criação de variáveis derivadas**:
  - `FamilySize`: tamanho da família a bordo.
  - `isAlone`: indica se o passageiro estava sozinho.
  - `Title`: categorização dos títulos dos nomes.
- **Normalização**: aplicada nas variáveis numéricas (`Age`, `Fare`, `FamilySize`) com `StandardScaler`.
- **Remoção de colunas irrelevantes**: `PassengerId`, `Name`, `SibSp`, `Parch`, `Ticket`, `Cabin`, `Embarked`.

---

2. Modelagem Supervisionada

Foram testados dois modelos preditivos:

- **Random Forest**
- **Rede Neural MLP (Multi-Layer Perceptron)**

Resultados

Modelo	Acurácia	Descrição
Random Forest	~85.9%	Boa performance, porém menor sensibilidade para alguns grupos minoritários.
MLPClassifier	~90.7%	Melhor desempenho geral, principalmente em recall para a classe `Survived`.

A rede neural MLP apresentou desempenho superior em termos de acurácia e precisão, mostrando-se o melhor modelo para prever a sobrevivência com os dados disponíveis.

---

3. Clusterização com DBSCAN

A técnica DBSCAN foi aplicada para identificar grupos de passageiros com perfis semelhantes.

- Foi feita padronização das variáveis e One-Hot Encoding da variável `Title`.
- Redução de dimensionalidade com PCA para visualização dos clusters.
- Foram identificados **clusters homogêneos** (ex: Cluster 1 com 98% de sobreviventes; Cluster 9 com 100% de não sobreviventes).

Esses agrupamentos revelaram padrões importantes, como:

- Passageiros da 1ª classe agrupados juntos com alta taxa de sobrevivência.
- Passageiros da 3ª classe sozinhos formando clusters com baixa taxa de sobrevivência.

---

4. Regras de Associação (Apriori)

Foram extraídas regras utilizando o algoritmo Apriori com as variáveis categóricas: `Sex`, `Pclass`, `isAlone`, `Title`, `AgeGroup`, `Survived`.

> `AgeGroup`: categorização da idade em faixas (Child, Teen, Adult etc.).

Exemplos de regras relevantes

- **Se** `Title=Master`, `Sex=male` **e** `Pclass=1st class` ⇒ **Sobrevivência**
  - Confiança: 97%
  - Lift: 2.52
- **Se** `Sex=female`, `Pclass=3rd class`, `isAlone=alone` ⇒ **Não sobreviveu**
  - Confiança: 88%
  - Lift: 1.42

Essas regras demonstram o impacto da **classe social, do gênero e do tipo de companhia** na chance de sobrevivência.

Conclusão

- A rede neural MLP foi o modelo mais eficaz para prever a sobrevivência de passageiros, atingindo **mais de 90% de acurácia**.
- O algoritmo DBSCAN mostrou-se útil para detectar **grupos sociais naturalmente distintos**, reforçando os efeitos de classe e companhia no naufrágio.
- As regras de associação foram altamente explicativas, revelando padrões comportamentais e sociais relevantes.
- O uso combinado de modelos supervisionados, clusterização e mineração de regras enriqueceu a análise e ampliou os insights sobre o conjunto de dados.

Observações Finais

Este relatório complementa o notebook, trazendo interpretações textuais e estruturadas das etapas realizadas.