

Lista 9 - Inteligência Artificial

Relatório de Pré-processamento e Agrupamento

Nome: Iasmin Oliveira

RA: 854946

1) Visualização da Base

A base foi carregada e inspecionada utilizando `head()`, `describe()` e análise da distribuição da variável alvo. Foram observadas fortes desproporções entre as classes (fraudes são raras).

2) Verificação e Tratamento de Valores Ausentes

A base não apresenta valores ausentes, porém utilizou-se `SimpleImputer(strategy='median')` para robustez.

3) Detecção e Eliminação de Redundância e Inconsistência

Foram removidas duplicatas. Valores negativos em 'Time' e 'Amount' foram corrigidos para zero.

4) Detecção e Tratamento de Outliers

Aplicou-se IQR clipping em 'Amount'. Também utilizou-se `IsolationForest` para sinalizar outliers nos componentes PCA.

5) Normalização e/ou Padronização

Foi utilizado `StandardScaler` para padronizar 'Time' e 'Amount'. Os atributos V1..V28 já são resultados de PCA e encontram-se normalizados.

6) Análise de Correlação e Multicolinearidade

Gerou-se a matriz de correlação e um heatmap. Também calculou-se VIF para identificar colinearidade entre atributos.

7) Codificação de Variáveis

A base contém apenas atributos numéricos, portanto não houve necessidade de One-Hot ou Label Encoding.

8) Balanceamento da Classe

Utilizou-se SMOTE apenas nos dados de treino para corrigir o desbalanceamento extremo entre as classes.

9) Divisão Treino–Teste (Estratificada)

O conjunto foi dividido em treino e teste usando stratify=y, preservando a proporção da classe fraudulenta.

Resultados do Modelo Antes e Depois do Pré-processamento

Antes do balanceamento (sem SMOTE): o modelo apresenta baixa revocação devido ao forte desbalanceamento. Após aplicar SMOTE, houve aumento significativo de F1, Recall e PR-AUC, mantendo estável o ROC-AUC. O aumento da sensibilidade demonstra que o pré-processamento foi eficaz para detectar fraudes.

Resultados de Agrupamento (KMeans, DBSCAN e SOM)

Foram aplicados KMeans (k=2), DBSCAN e SOM após remover o atributo 'Class'. As métricas avaliadas incluem Silhouette, Davies–Bouldin e Calinski–Harabasz. Além disso, ARI e NMI foram usados apenas como referência (comparação com as classes reais). KMeans e SOM geralmente formam dois grupos coerentes, enquanto DBSCAN pode gerar um único cluster dependendo dos parâmetros.

Fim do Relatório