Atividade Prática

CIN0208 - Ciência de Dados

Objetivo

Explorar a codificação de variáveis categóricas e o uso de técnicas de normalização/padronização no desempenho do algoritmo **KNN (K-Nearest Neighbors)**.

Conjunto de Dados

- Cada grupo deve acessar a planilha compartilhada e **escolher um dataset disponível** para trabalhar.
- Cada grupo deve registrar na planilha qual dataset escolheu, para evitar duplicidade.

Encoding e Scaling

(a) Se o dataset possuir colunas nominais, criar versões codificadas:

- One-hot encoding
- Dummy coding
- Effect coding

(b) Aplicar diferentes técnicas de normalização/padronização nas features numéricas:

- StandardScaler
- Min–Max Scaler
- MaxAbs Scaler
- Robust Scaler
- QuantileTransformer (uniforme e normal)

(c) Separar os dados em treino e teste (70% treino / 30% teste) utilizando `train_test_split`.

(d) Treinar e avaliar o KNN em cada cenário utilizando o **valor padrão de k do `scikit-learn` (`n_neighbors=5`)**, sem alterar o hiperparâmetro.

- Para classificação: acurácia e F1-score
- Para regressão: MAE e RMSE

(e) Gerar gráficos para interpretar os resultados do KNN:

- Para classificação: gráfico de barras comparando acurácia e F1-score entre os diferentes scalings e codificações

- Para regressão: gráfico de barras comparando MAE e RMSE entre os diferentes scalings e codificações
- Os gráficos devem permitir identificar facilmente qual técnica de scaling proporcionou melhor desempenho do KNN
- Discutir possíveis motivos para as diferenças observadas entre os scalings

Entrega e Estrutura do Relatório

Cada grupo deve entregar um relatório e o link para o notebook Python utilizado na atividade. O relatório deve seguir a estrutura abaixo:

1. **Capa:**
   - Nome dos integrantes
   - Dataset utilizado
   - Link para o notebook Python
2. **Introdução:**
   - Breve descrição do objetivo da atividade
   - Características do dataset escolhido
   - Explicação sobre o tipo de problema (classificação ou regressão)
3. **Encoding e Scaling:**
   - Descrever as colunas nominais e as codificações aplicadas
   - Apresentar tabelas separadas de resultados do KNN para cada métrica. Exemplo:

Classificação – Acurácia

Dataset	Encoding	Scaling	Acurácia KNN
Nome_DS	One-hot	Standard	0.72
Nome_DS	Dummy	Min–Max	0.75
Nome_DS	Effect	Robust	0.79

   - Gerar gráficos comparativos de desempenho do KNN para diferentes scalings e codificações, permitindo identificar qual combinação apresentou melhor resultado
   - Discutir observações sobre o comportamento do KNN frente às diferentes técnicas
4. **Conclusão:**
   - Resumo dos principais achados
   - Lições aprendidas sobre encoding, scaling e KNN
5. **Anexos:**
   - Inserir gráficos adicionais ou observações importantes