Exercício

Análise Exploratória de Dados

Objetivo: Utilizando o dataset **Wine**, realizar uma análise exploratória completa com o objetivo de identificar possíveis características irrelevantes e redundantes. O aluno deverá usar ferramentas como matriz de correlações, scatterplots e boxplots.

1. Carregamento dos Dados

O dataset **Wine** pode ser carregado diretamente da biblioteca `sklearn`. Para iniciar, carregue e visualize as primeiras linhas do dataset:

```
from sklearn.datasets import load_wine
import pandas as pd

# Carregar o dataset
wine_data = load_wine()
df = pd.DataFrame(data=wine_data.data, columns=wine_data.feature_names)

# Adicionando a variável alvo
df['target'] = wine_data.target

# Exibir as primeiras linhas do dataframe
df.head()
```

- Procure as informações do banco de dados **Wine** na documentação da biblioteca `sklearn`. Informe no relatório o nome das variáveis de entrada e saida.

2. Descrição Estatística dos Dados

Antes de qualquer análise visual, é importante obter uma visão geral estatística das variáveis.

- Qual o resumo estatístico das features? Use o comando `.describe()` para explorar a média, desvio padrão, valores mínimos e máximos de cada variável.

```
# Resumo estatístico
df.describe()
```

3. Matriz de Correlações

Construa uma matriz de correlações para identificar relações lineares entre as variáveis. Quais variáveis apresentam alta correlação? Variáveis altamente correlacionadas podem ser redundantes.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Matriz de correlação
plt.figure(figsize=(12,8))
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Matriz de Correlação')
plt.show()
```

Análise da matriz de correlação:

- Identifique variáveis com correlação forte (próxima de 1 ou -1).
- Variáveis com alta correlação entre si (acima de 0.8) podem ser redundantes e consideradas para remoção.

4. Matriz de Scatterplots

Construa uma matriz de scatterplots para visualizar graficamente as relações entre as variáveis. Esse gráfico ajuda a identificar padrões ou redundâncias visuais que podem não ser evidentes na matriz de correlação.

```
# Matriz de scatterplots
sns.pairplot(df, hue="target", diag_kind="kde")
plt.show()
```

Análise dos scatterplots:

- Identifique pares de variáveis que mostram padrões semelhantes ou superposição de classes. Isso pode indicar redundância.
- Explore se alguma variável claramente separa as classes ou se algumas variáveis não fornecem informações adicionais para distinguir as classes.

5. Boxplots

Os boxplots permitem observar a distribuição de cada variável em relação às classes da variável alvo. Isso ajuda a identificar variáveis que não variam muito entre as classes (potencialmente irrelevantes) ou que têm distribuições similares entre elas.

Outliers são valores que se distanciam significativamente do restante dos dados. Use boxplots para tentar identificá-los.

O código abaixo plota todos os boxplots em um gráfico. Melhore o código plotando cada boxplot individualmente por gráfico. Pra você poder observar melhor.

```
# Boxplots das features
plt.figure(figsize=(15,10))
df_melt = pd.melt(df, id_vars='target', var_name='features', value_name='values')
sns.boxplot(x='features', y='values', hue='target', data=df_melt)
plt.xticks(rotation=90)
plt.show()
```

Análise dos boxplots:

- Variáveis cujas distribuições entre as classes são muito similares podem ser consideradas irrelevantes para o problema de classificação.
- Observe se alguma variável apresenta pouca ou nenhuma variação entre as classes (indicando baixa relevância).
- Quais variáveis apresentam outliers significativos analisando os boxplots?

6. Conclusão: Identificação de Features Irrelevantes e Redundantes

Com base na análise exploratória, elabore um relatório respondendo às perguntas anteriores e às seguintes:

1. **Quais variáveis apresentam alta correlação entre si?** Explique por que você acredita que são redundantes.
2. **Há variáveis que, com base nos scatterplots e boxplots, parecem não ajudar a distinguir as classes?** Quais você considera irrelevantes?
3. **Quais variáveis você consideraria remover para otimizar o modelo de classificação, baseado nas observações feitas?**