

# Comparação de KFold e StratifiedKFold no SVM

Este projeto realiza uma comparação entre os métodos de validação cruzada KFold e StratifiedKFold utilizando o algoritmo SVM (Support Vector Machine).

O objetivo é demonstrar como a escolha da técnica de validação impacta os resultados de classificação em bases de dados biomédicas, onde o equilíbrio entre classes é essencial.

## Estrutura do Projeto

- `Comparacao_Completa_KFold_StratifiedKFold_SVM.ipynb` → Notebook com todo o código comentado e documentado.
- `README.md` → Este arquivo com explicações sobre o projeto.

## Tecnologias Utilizadas

- Python 3.10+
- Bibliotecas principais:
  - numpy
  - pandas
  - matplotlib
  - seaborn
  - scikit-learn

## Etapas do Projeto

### 1. Importação e exploração dos dados

- Utilização do Pima Indians Diabetes Dataset.
- Análise da distribuição das classes (diabético x não diabético).

### 2. Separação de Features e Target

- Variáveis explicativas (X) e variável alvo (y).

### 3. Treinamento com KFold (10 divisões)

- Avaliação da acurácia média e desvio padrão.

### 4. Treinamento com StratifiedKFold (10 divisões)

- Avaliação da acurácia média e desvio padrão.

### 5. Comparação gráfica dos métodos

- Boxplot comparando KFold e StratifiedKFold.

### 6. Treinamento do modelo final

- SVM treinado com StratifiedKFold (mais robusto).
- Avaliação em conjunto de teste separado.
- Relatório de classificação (Precisão, Recall, F1-Score).
- Matriz de confusão.

#### Resultados Obtidos

- **KFold (10):**
  - Acurácia média  $\approx 0.7643$
  - Desvio padrão  $\approx 0.0574$
- **StratifiedKFold (10):**
  - Acurácia média  $\approx 0.7708$
  - Desvio padrão  $\approx 0.0239$

👉 O StratifiedKFold apresentou resultados mais consistentes, já que mantém a proporção entre as classes em cada partição, evitando desequilíbrios.

#### Visualizações

- Distribuição das classes (0 = não diabético | 1 = diabético)
- Comparação gráfica das acurácias (Boxplot)
- Matriz de confusão do modelo final