FIAP - Faculdade de Informática e Administração Paulista

▲ Implementando algoritmos de Machine Learning com Scikit-learn

Nome do grupo: Grupo 13

# **♣** Integrantes:

- Fátima Vilela Candal

- Gabriel Viel dos Santos Delfino

- Guilherme Campos Hermanowski

- Jonathan Willian Luft

- Matheus Alboredo Soares

#### **A** Professores:

Tutor(a): Leonardo Ruiz Orabona

Coordenador(a): André Godoi Chiovato

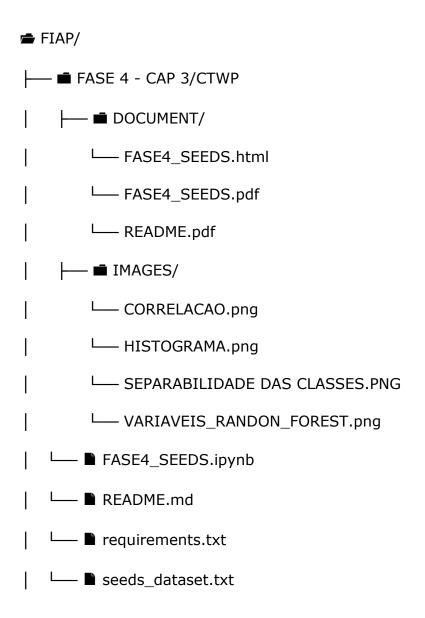
# ■ Descrição

Este projeto visa a classificação de variedades de grãos de trigo com base em suas características físicas, seguindo a metodologia CRISP-DM. O processo envolve um conjunto de tarefas bem detalhadas, incluindo análise exploratória dos dados, pré-processamento, implementação de algoritmos de classificação, otimização de modelos e interpretação dos resultados.

### **▶** Estrutura do Projeto

Os arquivos estão no GITHUB:

https://github.com/rm563003/FIAP/tree/main/FASE%204%20-%20CAP%203/CTWP



**≁** DATASET

Os dados utilizados (seeds\_dataset.txt) são provenientes do Seeds Dataset

do UCI Machine Learning Repository, que contém medições de 210 amostras

de grãos de trigo, divididas em três variedades: Kama, Rosa e Canadian. Os

atributos analisados incluem área, perímetro, compacidade, comprimento do

núcleo e coeficiente de assimetria, entre outros.

Para iniciar o estudo, os dados (seeds\_dataset.txt) devem ser carregados em

um ambiente de análise como Google Colab, permitindo a exploração da

estrutura, a identificação de padrões e a preparação dos dados para

modelagem.

O conjunto de dados é bem estruturado e adequado para análises de Machine

Learning. Ele contém três variedades de sementes e diversas características

físicas que podem ser usadas para classificá-las.

**☆** Como Executar Localmente

1. Clone o repositório:

git clone https://github.com/rm563003/FIAP/tree/main/FASE%204%20-

%20CAP%203/CTWP

cd CTWP

2. Instale os pacotes: pip install -r requirements.txt

3. Inicie o sistema: FASE4\_SEEDS.ipynb

Visão do código executado:

https://github.com/rm563003/FIAP/blob/main/FASE%204%20-

%20CAP%203/CTWP/DOCUMENT/FASE4 SEEDS.pdf

#### 🖈 Análise dos resultados dos modelos de classificação

- Melhor desempenho geral: Random Forest
- Alta acurácia (92,06%), precisão e revocação equilibradas.
- Melhor AUC-ROC (98,30%), indicando excelente distinção entre classes.
- Menor Log Loss (0,2246), o que significa que as previsões do modelo são confiáveis.
- Bom desempenho: KNN e Regressão Logística
- O KNN tem boa acurácia (87,30%) e AUC-ROC (94,97%), mas seu Log Loss (1,8758) é elevado, sugerindo previsões menos confiáveis.
- A Regressão Logística tem acurácia similar à SVM, mas com AUC-ROC alto (97,66%), o que indica bom poder de discriminação.
- ♦ Desempenho intermediário: SVM e Naive Bayes
- SVM apresenta acurácia de 85,71%, mas sem valores de AUC-ROC e
  Log Loss, dificultando uma avaliação completa.
- Naive Bayes tem a menor acurácia (82,54%) e um Log Loss maior (0,8231), o que pode indicar previsões menos certeiras.

# **№** Conclusão e recomendações

- ✓ Random Forest parece ser a melhor escolha para essa tarefa, pois
  equilibra bem todas as métricas.
- ✓ Se a interpretabilidade for importante, Regressão Logística pode ser uma opção sólida.
- ✓ Se quiser reduzir o custo computacional, KNN pode ser útil, mas pode sofrer com previsões menos confiáveis.

# Licença

[MODELO GIT FIAP](https://github.com/agodoi/template) por [Fiap](https://fiap.com.br/) está licenciado sobre [Attribution 4.0 International](http://creativecommons.org/licenses/by/4.0/?ref=chooser-v1).