FIAP - Faculdade de Informática e Administração Paulista

Grupo 24

Professores:

Tutor

- Leonardo Ruiz Orabona

Coordenador

- André Godoi

Cap 3 - (IR ALÉM) Implementando algoritmos de Machine Learning com Scikit-learn

Nomes:

- Felipe Sabino da Silva **RM:** rm563569
- Juan Felipe Voltolini **RM:** rm562890
- Luiz Henrique Ribeiro de Oliveira **RM:** rm563077
- Marco Aurélio Eberhardt Assimpção **RM:** rm563348
- Paulo Henrique Senise **RM:** rm565781

Fase: 4
Capítulo: 3

Objetivo

Aplicar a metodologia CRISP-DM para desenvolver um modelo de aprendizado de máquina que classifique variedades de grãos de trigo (Kama, Rosa e Canadian) com base em suas características físicas.

Dataset

Seeds Dataset - UCI Machine Learning Repository

- 210 amostras de grãos de trigo
- 3 variedades: Kama, Rosa e Canadian
- 7 atributos físicos por amostra

Características do Projeto

- **Análise Exploratória de Dados (EDA)** completa com visualizações estatísticas
- **Comparação de 5 algoritmos** de Machine Learning:
  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
  - Random Forest
  - Naive Bayes
  - Logistic Regression
- **Otimização automática** de hiperparâmetros usando Grid Search
- **Métricas detalhadas** de desempenho (acurácia, precisão, recall, F1-Score)
- **Visualizações interativas** dos resultados

Dataset

O projeto utiliza o dataset `seeds_dataset.txt` que contém:

- **210 amostras** de grãos de trigo
- **7 atributos** físicos por amostra:
  - Área
  - Perímetro
  - Compacidade
  - Comprimento do kernel
  - Largura do kernel
  - Coeficiente de assimetria
  - Comprimento do sulco do kernel
- **3 classes** de trigo: Kama, Rosa e Canadian

Estrutura do Projeto

```
.
├── JuanFelipeVoltolini_rm562890_fase4_cap3.ipynb  # Notebook principal com toda a anál
├── seeds_dataset.txt                               # Dataset com os dados dos grãos
├── requirements.txt                                # Dependências do projeto
├── README.md                                       # Este arquivo
├── LICENSE                                         # Licença do projeto
└── .gitignore                                      # Arquivos ignorados pelo git
```

Instalação

1. Clone o repositório:

```
git clone [URL_DO_REPOSITORIO]
cd [NOME_DO_REPOSITORIO]
```

2. Crie um ambiente virtual (recomendado):

```
python -m venv venv
source venv/bin/activate  # No Windows: venv\Scripts\activate
```

3. Instale as dependências:

```
pip install -r requirements.txt
```

Como Usar

1. Abra o Jupyter Notebook:

```
jupyter notebook
```

2. Navegue até o arquivo `JuanFelipeVoltolini_rm562890_fase4_cap3.ipynb`
3. Execute as células sequencialmente para:
    - Carregar e explorar os dados
    - Visualizar distribuições e correlações
    - Treinar e comparar os modelos
    - Analisar os resultados

Resultados

O projeto demonstra excelentes resultados de classificação:

- **Random Forest**: ~97% de acurácia
- **SVM**: ~96% de acurácia
- **Logistic Regression**: ~95% de acurácia
- **KNN**: ~94% de acurácia
- **Naive Bayes**: ~92% de acurácia

Todos os modelos são otimizados automaticamente usando Grid Search para encontrar os melhores hiperparâmetros.

Visualizações

O projeto inclui diversas visualizações para melhor compreensão dos dados:

- Histogramas de distribuição das características
- Boxplots para análise de outliers
- Matriz de correlação entre atributos

- Gráficos de comparação de desempenho dos modelos
- Matrizes de confusão para cada classificador

Tecnologias Utilizadas

- **Python 3.11+**
- **Pandas**: Manipulação de dados
- **NumPy**: Operações numéricas
- **Scikit-learn**: Algoritmos de Machine Learning
- **Matplotlib/Seaborn**: Visualizações
- **Jupyter**: Ambiente de desenvolvimento

Licença

Agradecimentos

- FIAP pela estrutura do curso
- UCI Machine Learning Repository pelo dataset
- Comunidade open-source pelas bibliotecas utilizadas