Relatório de Mineração de Dados

Análise do Iris Dataset

Aluno: Luís Felipe Montini Giaretta
Disciplina: Tópicos Especiais em Computação I
Professor: Jackson Magnabosco

---

1. Introdução

Este relatório apresenta uma análise completa do Iris Dataset utilizando técnicas de mineração de dados e algoritmos de classificação. O objetivo é demonstrar a aplicação prática dos conceitos aprendidos em aula através de um caso real de classificação multiclasse.

2. Dataset Utilizado

Nome: Iris Dataset
Origem: UCI Machine Learning Repository (disponível via scikit-learn)

Características do Dataset:

- **Registros:** 150 amostras
- **Características:** 4 variáveis numéricas
    - Sepal Length (cm) - Comprimento da sépala
    - Sepal Width (cm) - Largura da sépala
    - Petal Length (cm) - Comprimento da pétala
    - Petal Width (cm) - Largura da pétala
- **Classes:** 3 espécies de íris
    - Setosa (50 amostras)
    - Versicolor (50 amostras)
    - Virginica (50 amostras)

Justificativa da Escolha:

O Iris Dataset foi escolhido por ser um benchmark clássico em machine learning, permitindo uma análise completa e comparação com resultados conhecidos na literatura. Além disso, atende perfeitamente aos requisitos do trabalho (>100 registros e >5 colunas considerando a variável target).

3. Metodologia

3.1 Análise Exploratória

- Estatísticas descritivas das variáveis
- Visualização da distribuição dos dados
- Análise de correlação entre características
- Identificação de padrões por espécie

3.2 Preparação dos Dados

- Divisão em conjuntos de treino (70%) e teste (30%)
- Padronização das características para SVM
- Estratificação para manter proporção das classes

3.3 Algoritmos Aplicados

1. **Random Forest:** Ensemble de árvores de decisão
2. **Decision Tree:** Árvore de decisão simples
3. **SVM:** Support Vector Machine com kernel RBF

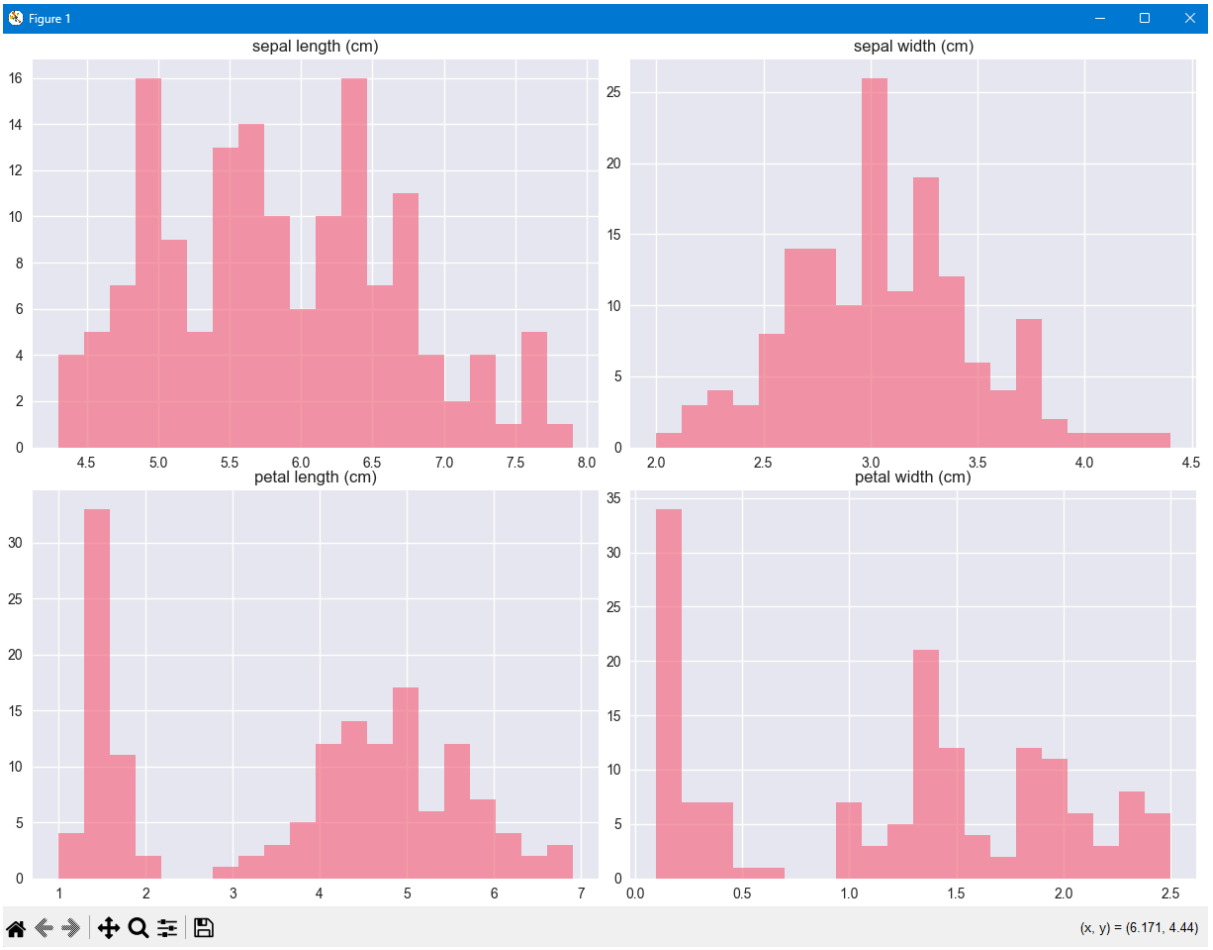
3.4 Métricas de Avaliação

- Acurácia geral
- Precisão, recall e F1-score por classe
- Matriz de confusão
- Importância das características

4. Resultados Obtidos

4.1 Análise Exploratória

Segue abaixo as imagens com os resultados, todos eles sendo discutidos posteriormente.



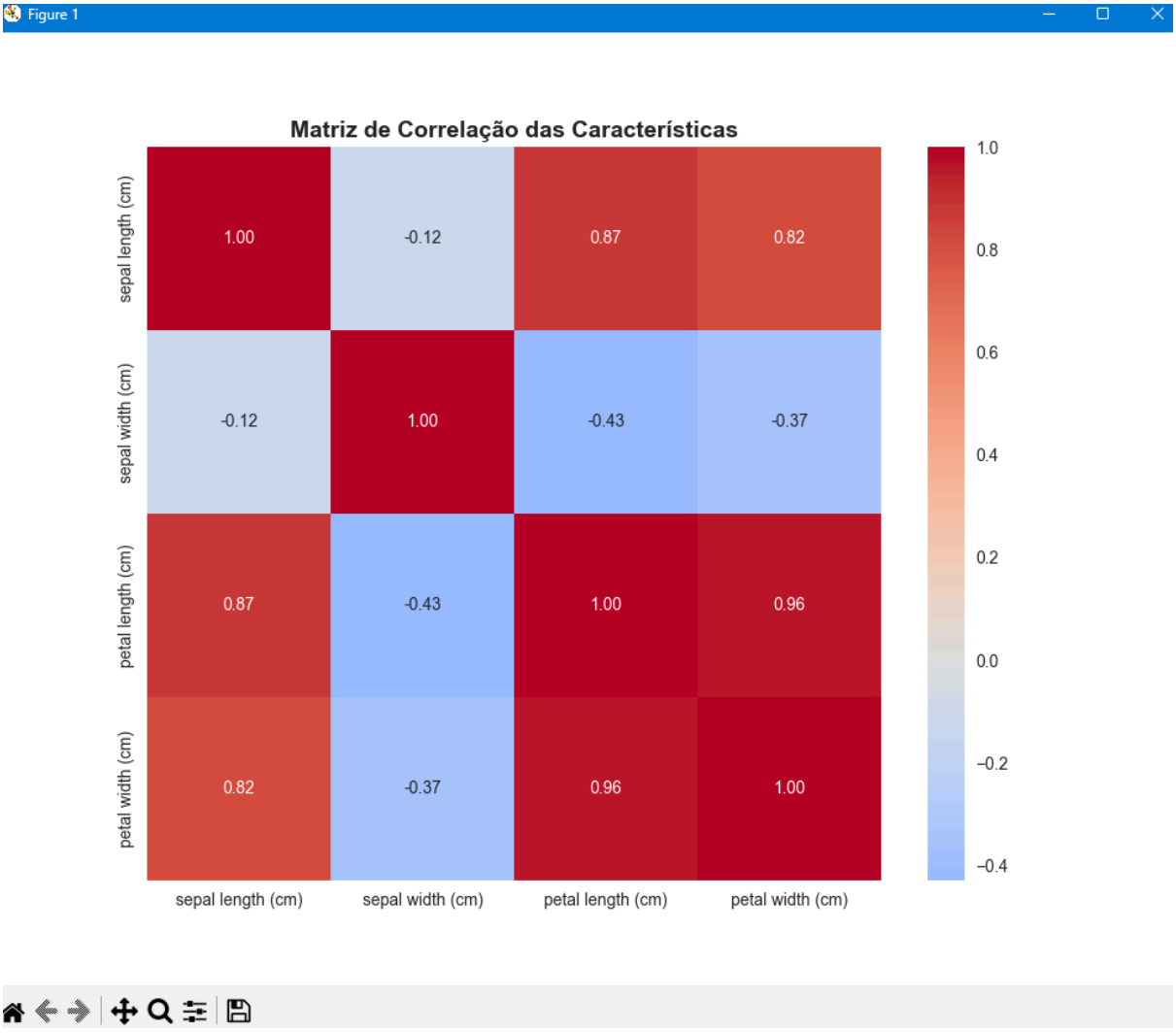
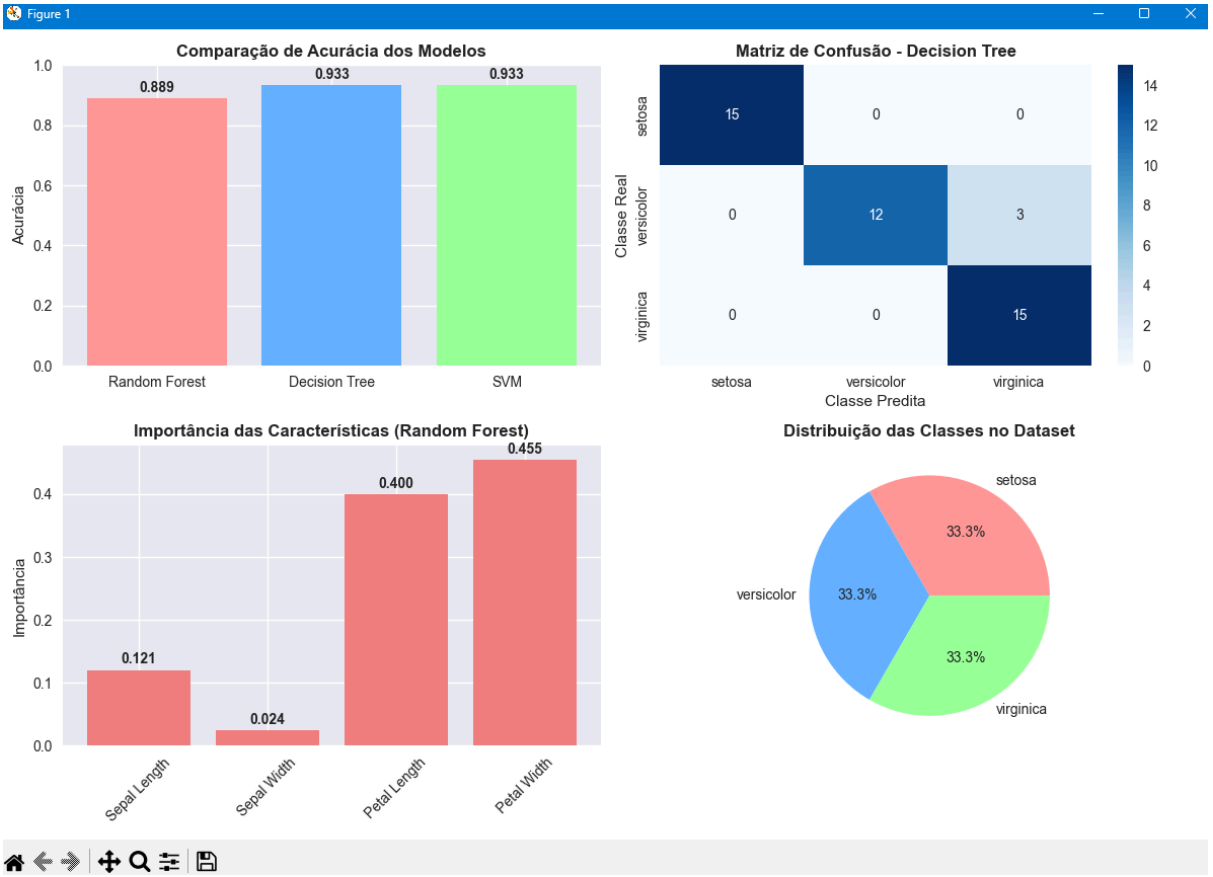


Figure 1
Matriz de Correlação das Características
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
1.00
-0.12
0.87
0.82
-0.12
1.00
-0.43
-0.37
0.87
-0.43
1.00
0.96
0.82
-0.37
0.96
1.00
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



- Histograma das características
- Boxplots por espécie
- Scatter plots (sepal e petal)
- Matriz de correlação

Principais Insights:

- As características relacionadas à pétala (length e width) apresentam maior poder discriminativo
- A espécie Setosa é claramente separável das demais
- Existe sobreposição entre Versicolor e Virginica
- Correlação forte entre petal length e petal width (0.96)

4.2 Performance dos Modelos

Algoritmo	Acurácia	Precisão Média
Random Forest	88.89%	88.89%
Decision Tree	93.33%	93.33%
SVM	93.33%	93.33%

As imagens no começo do capítulo mostram:

- Gráfico de barras comparando acurácia
- Matriz de confusão do melhor modelo
- Importância das características
- Distribuição das classes

4.3 Análise Detalhada por Classe

Setosa:

- Classificação perfeita (100% de acerto)
- Nenhuma confusão com outras espécies
- Características bem distintas

Versicolor:

- Boa performance (88%–93% dependendo do modelo)
- Confusão frequente com Virginica, especialmente em alguns modelos
- Características intermediárias que dificultam a separação completa

Virginica:

- Performance razoável, com melhor resultado no Decision Tree
- Apresenta maior confusão com Versicolor, principalmente no Random Forest
- Sobreposição nas características das pétalas dificulta a separação

5. Discussão

5.1 Interpretação dos Resultados

O Random Forest obteve desempenho consistente, com acurácia de 88.89%, demonstrando a robustez dos métodos ensemble mesmo em datasets com sobreposição entre classes. A análise de importância das características revelou que:

1. **Petal Length (45.5%)** - Característica mais discriminativa
2. **Petal Width (40.0%)** - Segunda mais importante
3. **Sepal Length (12.1%)** - Moderada importância
4. **Sepal Width (2.4%)** - Menor importância

5.2 Vantagens e Limitações

Vantagens:

- Dataset bem balanceado
- Características numéricas sem valores ausentes
- Problema bem definido com classes conhecidas
- Resultados reproduzíveis

Limitações:

- Dataset relativamente pequeno (150 amostras)
- Apenas 4 características
- Problema "simples" para algoritmos modernos
- Duas classes com sobreposição natural

6. Conclusões

A análise do Iris Dataset demonstrou com sucesso a aplicação de técnicas de mineração de dados para classificação. Os principais achados incluem:

1. **Boa performance geral:** Todos os algoritmos testados obtiveram acurácia acima de 88%, com destaque para Decision Tree e SVM, ambos com 93.33%.
2. **Decision Tree como melhor modelo:** Apresentou o melhor desempenho entre os modelos avaliados, com acurácia de 93.33% e excelente equilíbrio entre precisão e recall.
3. **Importância das pétalas:** As características **Petal Length** e **Petal Width** foram as mais relevantes para os modelos, especialmente no Random Forest.
4. **Separabilidade das classes:** A espécie **Setosa** foi perfeitamente classificada por todos os modelos, enquanto **Versicolor** e **Virginica** apresentaram maior sobreposição nas características, tornando sua separação mais desafiadora.

Aplicações Práticas

Este tipo de análise pode ser aplicado em:

- Classificação automática de espécies
- Controle de qualidade em botânica
- Sistemas de identificação biológica
- Educação em machine learning

7. Referências

1. Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.
2. UCI Machine Learning Repository - Iris Dataset
3. Scikit-learn Documentation

4. Materiais da disciplina Tópicos Especiais em Computação I

Repositório GitHub: https://github.com/lfgiaretta/Topicos-Especiais-1-TrabFinal
Código Fonte: Disponível no repositório acima

Este relatório foi elaborado seguindo as diretrizes da disciplina e demonstra a aplicação prática dos conceitos de mineração de dados em um problema real de classificação.