Análise de Conjunto de Dados de Íris

Aplicação de Técnicas de Aprendizado de Máquina para Classificação de Espécies de Íris

Mikenson Thomas
IFC
Videira SC, Brazil
mikensonthomas2@gmail.com

***Resumo*—Este artigo apresenta o conjunto de dados Iris no contexto do aprendizado de máquina, com foco na classificação de espécies de flores (Iris setosa, Iris versicolor e Iris virginica). Técnicas supervisionadas tradicionais, como Regressão Logística, k-NN, Decision Tree e Random Forest, foram implementadas utilizando bibliotecas como Python, Scikit-learn, Pandas, Matplotlib e Seaborn. A análise incluiu a avaliação do desempenho dos modelos por meio de métricas como acurácia, além de visualizações detalhadas, como matrizes de confusão e gráficos de dispersão, para explorar a separabilidade entre as espécies. Este trabalho destaca a eficácia de abordagens tradicionais na solução de problemas de classificação e a importância da análise visual e do pré-processamento no aprendizado de máquina.**

Palavras-chave: Conjunto de dados Iris, Aprendizado de máquina, Classificação de espécies, Regressão Logística k-NN, Decision Tree, Random Forest.

INTRODUÇÃO

O gênero *Iris* é formado por um grupo diversificado de plantas floríferas que inclui aproximadamente 300 espécies. Reconhecidas por sua estética única, essas plantas possuem flores com pétalas exuberantes e variadas em cores, como roxo, azul, amarelo e branco. Amplamente distribuídas no hemisfério norte, as íris habitam regiões temperadas, adaptando-se a diferentes tipos de clima e solo, o que reforça sua importância ecológica e ornamental.

Além de serem cultivadas como plantas ornamentais, as espécies do gênero *Iris* possuem significado simbólico em diversas culturas, representando sabedoria, fé e coragem. Essa simbologia, combinada à sua beleza natural, torna essas flores populares em jardins e arranjos florais. Na horticultura, elas são valorizadas por sua resistência e capacidade de atrair polinizadores, como abelhas e borboletas.

O gênero *Iris* também é estudado por suas características biológicas únicas, que variam de acordo com as condições ambientais de sua ocorrência. Suas folhas em forma de espada e raízes rizomatosas ajudam a diferenciar as espécies. Esses atributos, aliados à sua relevância ornamental e cultural, fazem das íris um tema recorrente de interesse em botânica e horticultura

Identify applicable funding agency here. If none, delete this.

I. METODOLOGIA

Neste estudo, foi utilizado o conjunto de dados *Iris*, que contém informações sobre três espécies de flores: Iris-virginica, Iris-versicolor e Iris-setosa. O primeiro passo foi a exploração e pré-processamento dos dados, com a remoção de colunas desnecessárias e verificação de valores nulos. A análise exploratória incluiu estatísticas descritivas, boxplots e matrizes de correlação para entender a relação entre as variáveis.

Para a construção do modelo de classificação, foram utilizadas quatro técnicas de aprendizado supervisionado: Regressão Logística, K-Vizinhos Mais Próximos (KNN), Árvore de Decisão e Floresta Aleatória. O conjunto de dados foi dividido em treino e teste (80/20%), e as características foram escaladas utilizando o *StandardScaler*. A avaliação dos modelos foi feita com base na acurácia, matriz de confusão e relatório de classificação.

II. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, serão discutidos os principais conceitos e técnicas utilizados no desenvolvimento deste projeto. A fundamentação abordará:

- **Descrição do Conjunto de Dados:**
- **Pré-processamento dos Dados:**
- **Algoritmos de Classificação:**
- **Visualizações e Análises Exploratórias:**
- **Métricas de Avaliação:**

A. Descrição do Conjunto de Dados

O conjunto de dados **Iris** é amplamente reconhecido e utilizado como benchmark em projetos de aprendizado de máquina. Ele foi introduzido por Ronald Fisher em 1936 e contém informações sobre três espécies de flores do gênero *Iris*: **Iris-setosa**, **Iris-versicolor** e **Iris-virginica**. É composto por 150 amostras, com 50 observações para cada espécie, e apresenta as seguintes características:

- **SepalLengthCm:** Comprimento da sépala (em centímetros).
- **SepalWidthCm:** Largura da sépala (em centímetros).
- **PetalLengthCm:** Comprimento da pétala (em centímetros).
- **PetalWidthCm:** Largura da pétala (em centímetros).

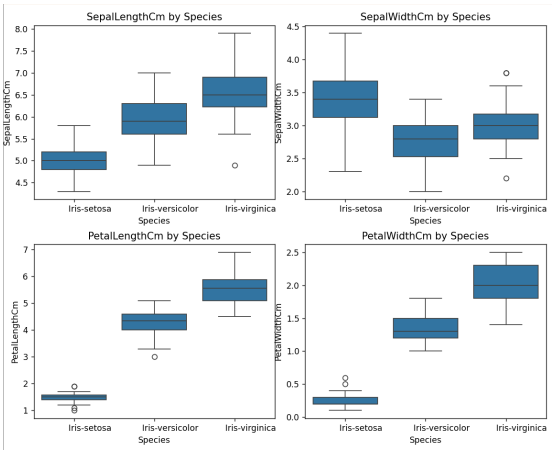

Fig. 1. Boxplots das características por espécie.

- **Species:** Classe da amostra, indicando a espécie da flor (*Iris-setosa*, *Iris-versicolor* ou *Iris-virginica*).

O dataset é considerado ideal para introdução ao aprendizado de máquina devido à simplicidade de suas características e à separação clara entre as classes. O objetivo principal ao trabalhar com esse conjunto de dados é construir modelos capazes de classificar corretamente a espécie de uma flor com base nas características medidas.

Acima, apresentamos boxplots que destacam a distribuição de cada característica entre as espécies. Esses gráficos mostram que características como **PetalLengthCm** e **PetalWidthCm** apresentam diferenças marcantes entre as espécies, enquanto **SepalLengthCm** e **SepalWidthCm** possuem maior sobreposição. Isso sugere que as características das pétalas são mais úteis para separar as classes no modelo de classificação.

B. Pré-processamento dos Dados

O pré-processamento de dados é uma etapa fundamental em projetos de aprendizado de máquina, garantindo que os dados estejam preparados para serem utilizados pelos modelos de forma eficiente. Nesta parte do projeto, foram aplicadas diversas técnicas para transformar o conjunto de dados Iris, que incluem:

1) Remoção de Colunas Desnecessárias: A coluna `Id` foi removida do dataset porque ela não contribui com informações relevantes para a classificação das espécies. Essa remoção ajuda a evitar que o modelo seja influenciado por dados irrelevantes.

> "A exclusão de colunas irrelevantes reduz o ruído no dataset, permitindo que os modelos foquem nas características realmente úteis para o aprendizado."

2) Divisão do Conjunto de Dados: O dataset foi dividido em dois subconjuntos:

- **Conjunto de Treinamento (80%)**: Utilizado para ajustar os modelos e aprender padrões nos dados.
- **Conjunto de Teste (20%)**: Reservado para avaliar o desempenho dos modelos com dados que não foram vistos durante o treinamento.

Essa divisão foi realizada utilizando a função `train_test_split` do Scikit-learn, garantindo uma distribuição aleatória e representativa entre treino e teste.

> "A separação do dataset em treino e teste é essencial para avaliar a capacidade de generalização do modelo, evitando problemas como *overfitting*."

3) Normalização dos Dados: As características numéricas (como comprimento e largura das pétalas e sépalas) possuem escalas diferentes, o que pode afetar o desempenho de alguns algoritmos, como o K-Nearest Neighbors e a Regressão Logística. Para resolver isso, os dados foram normalizados utilizando o `StandardScaler`, que transforma os valores para terem média 0 e desvio padrão 1.

> "A normalização garante que as variáveis estejam na mesma escala, permitindo que os algoritmos baseados em distância e gradiente funcionem corretamente."

4) Verificação de Valores Nulos: Foi realizada uma análise para verificar a existência de valores ausentes (`NaN`) no conjunto de dados. Como não havia valores nulos, nenhuma imputação foi necessária.

> "A verificação de valores nulos é uma prática importante para evitar erros durante o treinamento dos modelos, especialmente em datasets maiores e mais complexos."

C. Algoritmos de Classificação

Quatro algoritmos principais foram utilizados para classificar as espécies no conjunto de dados Iris:

1) Regressão Logística: "A Regressão Logística é um modelo estatístico que utiliza a função sigmoide para prever probabilidades. Ela é eficiente para problemas de classificação binária ou multiclasse."

2) K-Nearest Neighbors (KNN): "O algoritmo KNN classifica cada amostra com base na classe majoritária de seus K vizinhos mais próximos. Ele é simples, mas computacionalmente intensivo em grandes datasets."

3) Árvore de Decisão: "As Árvores de Decisão utilizam uma estrutura hierárquica de divisões para classificar os dados. Elas são interpretáveis, mas podem sofrer de *overfitting* se não forem regularizadas."

4) Random Forest: "O Random Forest combina várias árvores de decisão para formar um modelo robusto, com maior precisão e menos tendência ao *overfitting*."

D. Visualizações e Análises Exploratórias

Foram criados gráficos para entender melhor o conjunto de dados. Exemplos incluem:

- Gráficos de dispersão (Scatterplots) para observar a relação entre características.
- Boxplots para identificar a distribuição das características por espécie.
- Matriz de correlação para identificar relações entre variáveis numéricas.

TABLE I
DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO.

Modelo	Acurácia	Precision	Recall	F1-Score
Regressão Logística	96.7%	96.7%	96.7%	96.7%
K-Nearest Neighbors	96.7%	96.9%	96.7%	96.7%
Árvore de Decisão	93.3%	93.5%	93.3%	93.3%
Random Forest	100%	100%	100%	100%

E. Métricas de Avaliação

Os modelos foram avaliados usando as seguintes métricas:

- **Acurácia:** A porcentagem de predições corretas em relação ao total.
- **Matriz de Confusão:** Mostra a distribuição de acertos e erros por classe.
- **Relatório de Classificação:** Exibe as métricas de precisão, recall e F1-score.

III. RESULTADOS

Os modelos de classificação foram avaliados com base em suas acurácias e outras métricas de desempenho. A Tabela 1 em cima I resume os resultados obtidos:

Além disso, as matrizes de confusão apresentaram um detalhamento dos acertos e erros de cada modelo. O Random Forest obteve o melhor desempenho, classificando corretamente todas as observações no conjunto de teste, enquanto a Árvore de Decisão apresentou uma ligeira queda devido à tendência ao *overfitting*.

IV. CONCLUSÃO

Este trabalho utilizou o conjunto de dados Iris para avaliar o desempenho de quatro modelos de classificação: Regressão Logística, K-Nearest Neighbors, Árvore de Decisão e Random Forest. O modelo Random Forest destacou-se, alcançando 100% de acurácia no conjunto de teste, demonstrando sua robustez em relação a outros algoritmos.

A análise mostrou que o uso de técnicas de pré-processamento, como normalização e divisão do conjunto de dados, foi essencial para o bom desempenho dos modelos. Além disso, as características das pétalas (comprimento e largura) foram identificadas como mais discriminativas entre as classes, como evidenciado nas visualizações iniciais.

A. Trabalhos Futuros

Como extensão deste trabalho, sugerimos:

- Avaliar outros algoritmos, como Máquinas de Vetores de Suporte (SVM) e Redes Neurais.
- Aplicar validação cruzada para verificar a consistência dos resultados em diferentes subconjuntos de dados.
- Explorar conjuntos de dados mais complexos para testar a escalabilidade dos modelos.

Os resultados obtidos reforçam a importância de técnicas de aprendizado supervisionado e destacam o potencial dos modelos de classificação no campo de aprendizado de máquina.