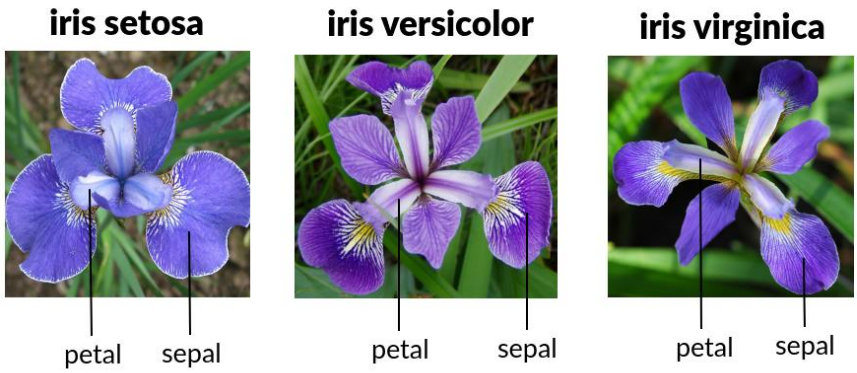
Classificador com árvode de decisão para dataset de flor de Iris

Antônio Lima Della Flora

Introdução

A análise e classificação de dados são etapas cruciais em diversas áreas do conhecimento, e o aprendizado de máquina (Machine Learning) oferece ferramentas poderosas para essa finalidade. Entre os conjuntos de dados clássicos e amplamente utilizados para ilustrar e testar algoritmos de classificação, destaca-se o Iris Flower Dataset. Introduzido por Ronald Fisher em 1936, este dataset icônico serve como um benchmark fundamental para o desenvolvimento e avaliação de técnicas de reconhecimento de padrões.

O Iris Flower Dataset consiste em 150 amostras de três espécies diferentes de flores de íris: *Iris setosa*, *Iris versicolor* e *Iris virginica*. Para cada amostra, são medidas quatro características: o comprimento da sépala, a largura da sépala, o comprimento da pétala e a largura da pétala. O objetivo principal ao trabalhar com este dataset é construir modelos capazes de classificar corretamente cada amostra em sua respectiva espécie com base nessas medidas.



Metodologia

Pré processamento

O primeiro passo foi importar o dataset. Por ser um dataset bem famoso, ele já está incluso dentro da scikit-learn. Após carregar o dataset foi feito o pré processamento dos dados, dividindo os dados em 2 conjuntos:

- Conjunto de treino e validação (80%): foi utilizado para treinar o modelo;
- Conjunto de testes (20%): Serviu para avaliar a performance final do modelo com a profundidade que entregou maior desempenho durante o treinamento. São dados que nunca foram utilizados para treinar o modelo.

A divisão dos conjuntos foi realizada utilizando o comando train_test_split da biblioteca sklearn. Para garantir uma divisão igual entre os 3 tipos de flores foi utilizada a estratificação com os tipos de flores do dataset. Também foi definido uma semente fixa para a divisão dos conjuntos, a fim de garantir a replicabilidade dos resultados.

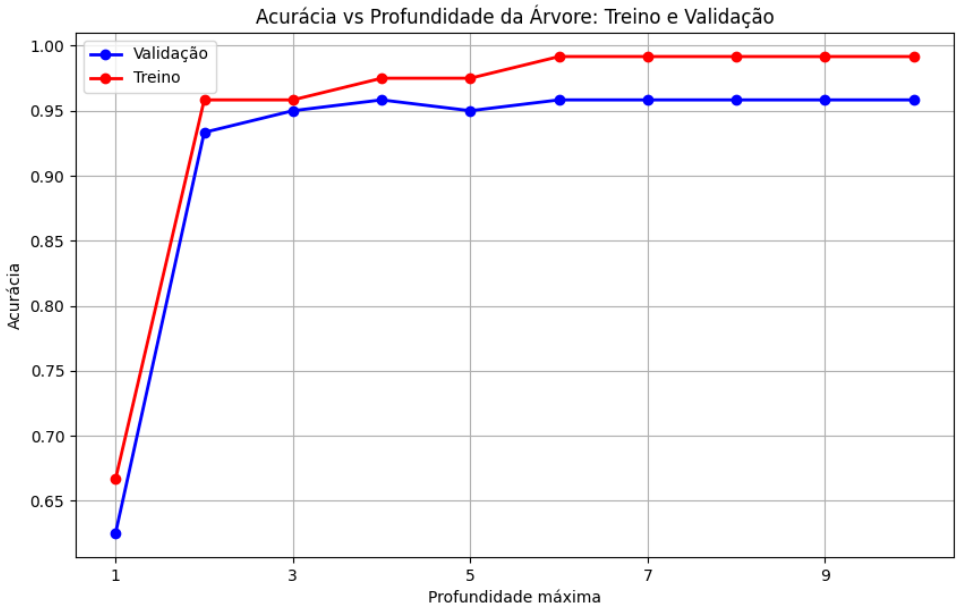
Determinando a melhor profundidade

Para determinar a profundidade ótima da árvore de decisão, foi implementada uma estratégia de validação cruzada com 5 folds, que divide o conjunto de treino e validação em cinco partes iguais. Para cada profundidade testada (variando de 1 a 10), o modelo é treinado e validado cinco vezes, com uma parte diferente dos dados servindo como conjunto de validação em cada iteração. Esta abordagem garante que cada amostra seja utilizada tanto para treino quanto para validação, proporcionando uma avaliação mais robusta do desempenho do modelo em diferentes configurações de profundidade.

Durante este processo, foram calculadas separadamente as taxas de acerto médias tanto no conjunto de treino quanto no conjunto de validação para cada profundidade.

A comparação dessas duas métricas é fundamental para identificar possíveis casos de overfitting, que ocorrem quando o modelo apresenta desempenho significativamente melhor no conjunto de treino em comparação com o conjunto de validação.

Os resultados obtidos podem ser vistos na imagem abaixo:



Avaliação do Modelo Final

Com base nos resultados da etapa anterior, foi escolhido a profundidade 4 como a melhor, tendo em vista que atingiu a melhor taxa de acerto. O modelo foi então treinado com o conjunto de treino e validação e sua performance foi comparada com o conjunto de testes.

Para avaliar o desempenho final do modelo foi utilizado o *classification_report,* que obtém métricas detalhadas de cada classe (cada tipo de flor) e a acurácia global.

As métricas obtidas para cada classe são as seguintes:

- Precisão (precision): é o número de verdadeiros positivos dividido pela quantidade de positivos detectados pelo modelo. É calculado com Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Positivos). Se o valor foi baixo indica que o modelo possui muitos falsos positivos.
- Revocação (recall): Indica quantos % o modelo conseguiu identificar corretamente. É calculado com Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Negativos). Um baixo número indica que o modelo classificou muitas amostras como sendo de outras classes.
- Pontuação F1 (F1-Score): é a média harmônica da precisão e a revocação.
- Suporte (support): representa o número de ocorrências reais da classe. Ajuda a entender o volume de cada classe que foi utilizado para calcular as demais métricas.

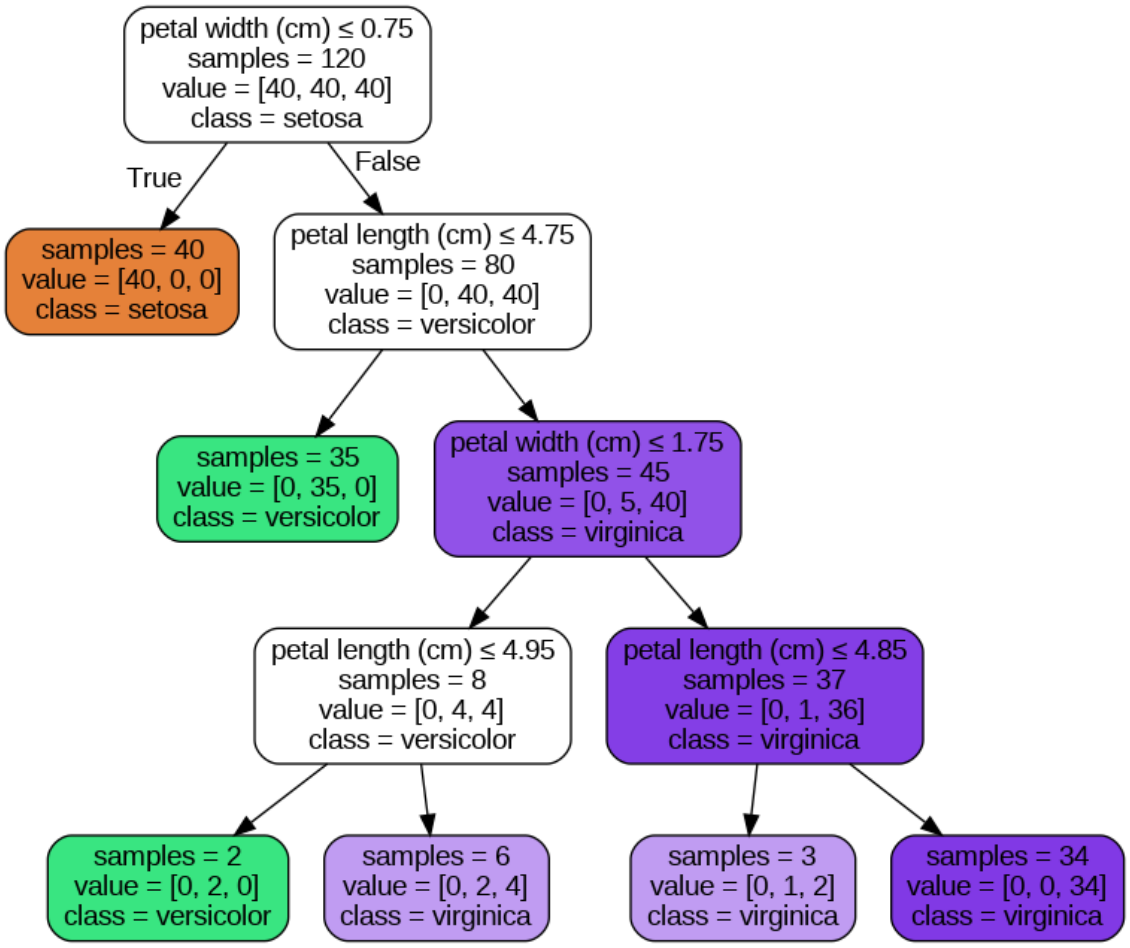
Os resultados foram os seguintes:

Modelo com dados normalizados:

Classe	precision	recall	1-score	support
setosa	1.00	1.00	1.00	10
versicolor	0.909	1.00	0.952	10
virginica	1.00	0.90	0.947	10

Acurácia global: 0.9667

Representação final da árvore:



Conclusão

Os resultados mostram que o modelo teve um bom desempenho. Ao analisar a tabela com as métricas, é possível perceber que o modelo classificou corretamente todas as 10 amostras da classe *setosa* e *versicolor*. Já a classe *virginica* teve um falso negativo em que a amostra foi classificada como *versicolor*.

A acurácia global de 0.9667 indica que a árvore de decisão conseguiu classificar de forma eficaz os diferentes tipos de flores.

Por fim, buscou-se alguns artigos que utilizaram a mesma base de dados e treinaram uma árvore para comparar a acurácia. Conforme a tabela abaixo, é possível observar que o presente trabalho performou dentro do esperado:

	Renas Rajab (2024)	Shivam (2019)	Projeto atual
Acurácia	97%	92.10%	96.67%

Referências

ASAAD, Renas Rajab; ABDULAZEEZ, Adnan M. Comprehensive Classification of Iris Flower Species: A Machine Learning Approach. **The Indonesian Journal of Computer Science**, v. 13, n. 1, 2024.

VATSHAYAN, Shivam. Performance evaluation of supervised learning for iris flower species. 2019.