Documentação do Projeto de Classificação com Comitê de Classificadores

Aluno: Pierry Boettscher

Curso: Engenharia de Software

Disciplina: Aprendizagem de Máquina

Data: 06/04/2024

Documentação do Projeto de Classificação com Comitê de Classificadores

Pierry Boettscher

SUMÁRIO

Data: 06/04/2024

Documentação do Projeto de Classificação com Comitê de Classificadores

Pierry Boettscher

1. ESCOLHA E ESTUDO DO DATA SET

Data Set Escolhido: Iris Dataset

Descrição:

O Iris Dataset é um conjunto de dados clássico e amplamente utilizado em aprendizado de máquina, que contém 150 instâncias divididas em 3 classes de íris (Setosa, Versicolor, Virginica), com 50 amostras cada. Cada instância é descrita por 4 atributos numéricos que representam as características físicas das flores de íris: comprimento e largura da sépala, e comprimento e largura da pétala.

2. INFORMAÇÕES DO DATA SET

Instâncias: 150 flores de íris.

Atributos:

- Comprimento da Sépala (cm)
- Largura da Sépala (cm)
- Comprimento da Pétala (cm)
- Largura da Pétala (cm)

Classes:

- Setosa
- Versicolor
- Virginica

3. VARIÁVEL TARGET

Target: Classe da íris (Setosa, Versicolor, Virginica)

A variável target representa a espécie da flor de íris, que é a conclusão que desejamos prever utilizando os atributos físicos.

Data: 06/04/2024

Documentação do Projeto de Classificação com Comitê de Classificadores

Pierry Boettscher

4. ALGORITMOS DE MACHINE LEARNING ESCOLHIDOS

Para o comitê de classificadores, foram escolhidos os seguintes algoritmos:

- KNN (K-Nearest Neighbors)
- Árvores de Decisão

5. RESULTADOS OBTIDOS EM CADA CLASSIFICADOR

Os resultados serão avaliados com base nas seguintes métricas:

- Acurácia (taxa de acerto)
- Taxa de erro
- Matriz de Confusão
- Precisão
- Sensibilidade/Recall

(Os resultados específicos serão detalhados após a implementação e execução dos algoritmos.)

6. COMPARAÇÃO DOS RESULTADOS DO COMITÊ DE CLASSIFICADORES

A comparação entre os classificadores será realizada utilizando a Curva ROC (Receiver Operating Characteristics), que é uma ferramenta visual para comparar a performance de modelos de classificação binária. Para adaptar à classificação multiclasse do Iris Dataset, será aplicada a técnica one-vs-rest (OvR) para gerar as curvas ROC para cada classe contra todas as outras.

7. CÓDIGO E EXPLICAÇÃO

Link do GitHub com o código desenvolvido em Python:
https://github.com/PierryB/N1-MachineLearning

O código realiza as seguintes operações:

- Carrega o Iris Dataset.
- Divide o dataset em conjuntos de treinamento e teste.
- Treina os classificadores KNN e Árvore de Decisão com os dados de treinamento.

Data: 06/04/2024

- Avalia os classificadores utilizando as métricas de acurácia, precisão e recall.
- Gera e exibe a Curva ROC para cada classe e classificador.

Com base nos resultados obtidos, podemos concluir o seguinte sobre o desempenho dos classificadores KNN e Árvore de Decisão no Iris Dataset:

KNN

- Acurácia: 95.56%
- Precisão: 96.08%
- Recall: 95.56%

Árvore de Decisão

- Acurácia: 93.33%
- Precisão: 94.44%
- Recall: 93.33%

8. ANÁLISE DOS RESULTADOS

Acurácia: A porcentagem de previsões corretas em relação ao total. KNN teve um desempenho ligeiramente superior à Árvore de Decisão.

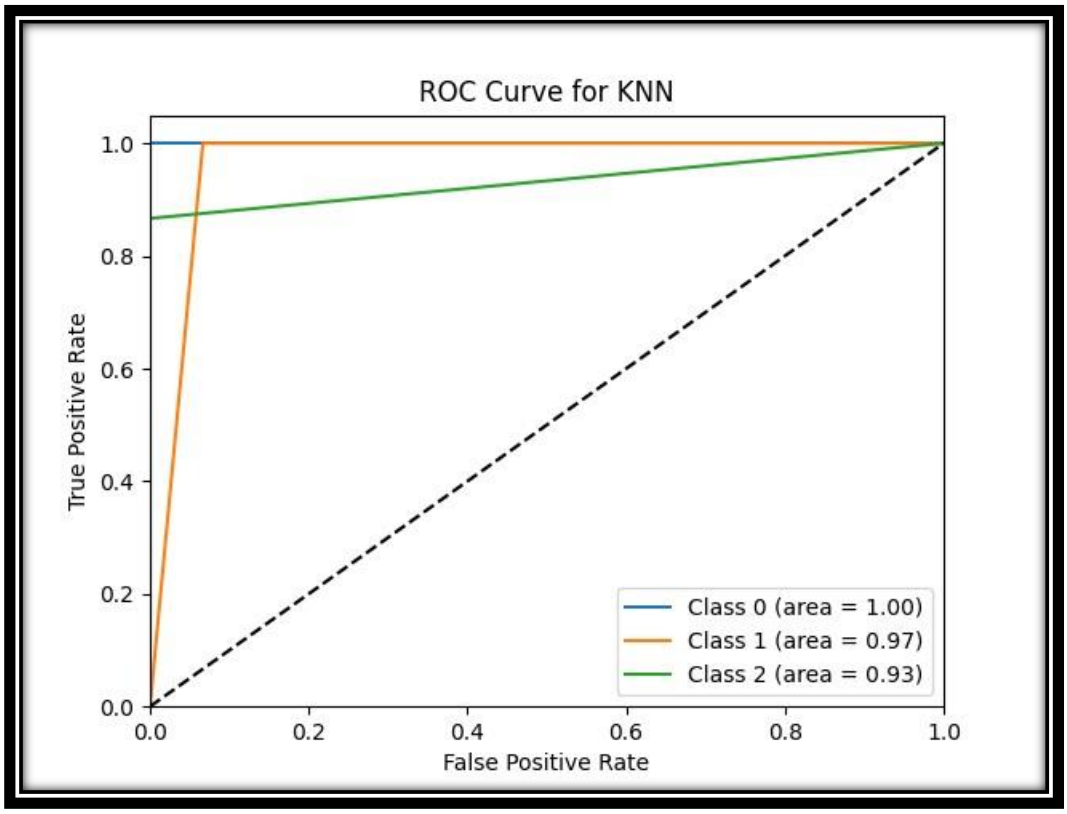
Precisão: A porcentagem de previsões positivas corretas em relação ao total de previsões positivas. KNN também teve uma precisão ligeiramente superior.

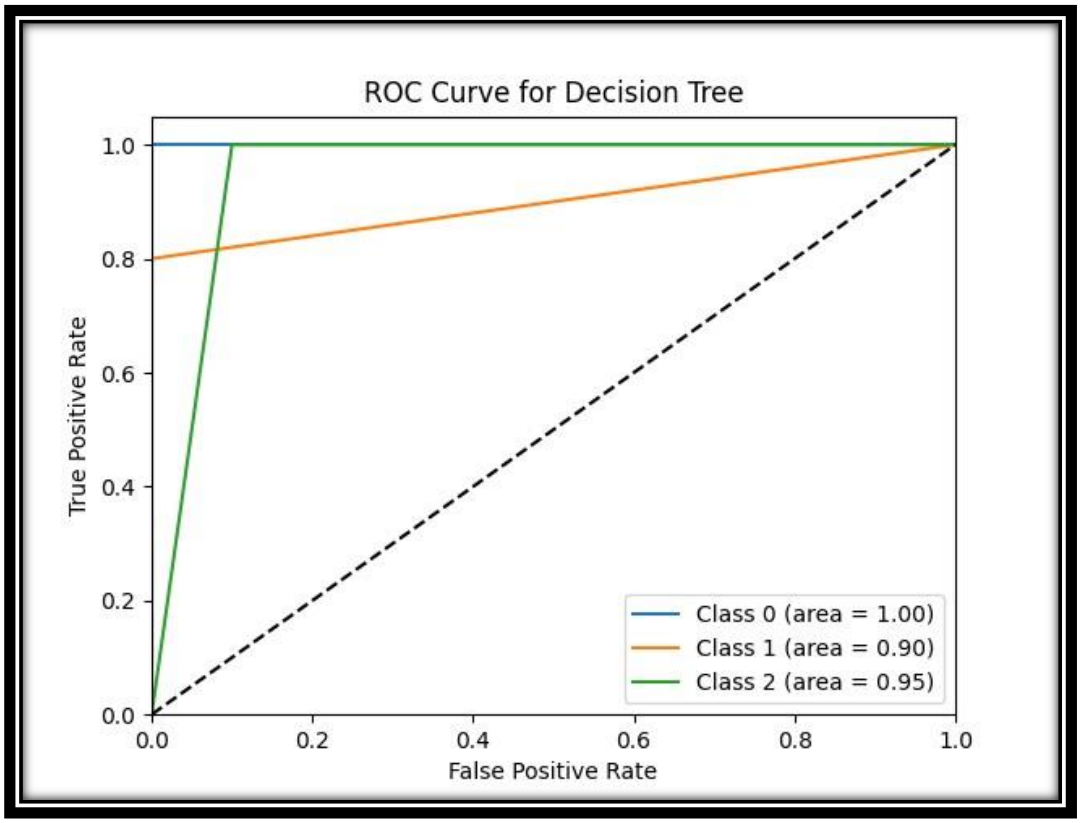
Recall: A porcentagem de positivos reais que foram corretamente identificados. Novamente, o KNN superou a Árvore de Decisão, embora por uma margem estreita.

Ambos os classificadores demonstraram alto desempenho no conjunto de dados Iris, o que é esperado dada a simplicidade relativa do problema de classificação. O KNN mostrou-se um pouco mais eficaz em todos os aspectos medidos. Este resultado talvez esteja relacionado à natureza do dataset Iris, que é bem separado em termos das

características das flores, beneficiando-se do método de votação baseado em vizinhança do KNN.

9. CONSIDERAÇÕES SOBRE OS GRÁFICOS ROC





Os gráficos ROC gerados nos permite ter uma representação visual da capacidade de cada classificador em distinguir entre as classes. Uma área sob a curva (AUC) mais próxima de 1 indica um modelo com melhor desempenho. As diferenças nos desempenhos dos modelos também podem ser refletidas aqui, com o KNN apresentando curvas ROC com áreas maiores em comparação à Árvore de Decisão.

10. CONCLUSÃO

Ambos os classificadores são adequados para o problema em questão, com o KNN apresentando uma pequena vantagem em todos os aspectos. A escolha entre eles pode depender de outros fatores, como a complexidade do modelo, tempo de treinamento e predição, e a interpretabilidade do modelo.