THỰC HÀNH HỌC MÁY CÓ GIÁM SÁT VỚI K-NN

Sinh viên tìm kiếm kênh Nam Media Tivi và sử dụng video: Kỹ thuật Grid Search trong Hyperparameter Tuning cho mô hình phân lớp với giải thuật K-nearest neighbors.

Sử dụng dữ liệu: heart.csv

BÀI TẬP VỀ NHÀ

Sinh viên viết code và trả lời câu hỏi trên file .ipynb

Tập dữ liệu iris.csv.

Mô tả dữ liệu iris: *The Iris dataset consists of 150 samples of iris flowers from three different species: Setosa, Versicolor, and Virginica. Each sample includes four features: sepal length, sepal width, petal length, and petal width. It was introduced by the British biologist and statistician Ronald Fisher in 1936 as an example of discriminant analysis.*

Yêu cầu: Xây dựng mô hình dự báo chủng loại Species dựa trên các đặc trưng SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm.

1. Tiến hành EDA dữ liệu:
    a. Có bao nhiêu đặc trưng và tên đặc trưng
    b. Có bao nhiêu chủng loài
    c. Tổng quan dữ liệu kiểu dữ liệu và dữ liệu thiếu
    d. Thống kê các đại lượng cơ bản cho các đặc trưng
    e. Thống kê các đại lượng cơ bản cho các đặc trưng theo nhóm chủng loại
    f. Trực quan phân phối các đặc trưng và nhận xét
    g. Trực quan phân phối các đặc trưng theo nhóm chủng loài và nhận xét
    h. Khám phá các giá trị bất thường của các đặc trưng theo nhóm chủng loài
2. Tập dữ liệu sử dụng tỉ lệ train:test là 75:25 với hệ số ngẫu nhiên là 16
3. Sử dụng biểu đồ để tìm giá trị K tốt nhất cho mô hình K-NN dựa theo độ đo chính xác để đánh giá.
4. Sử dụng giá trị K tốt nhất để xây dựng mô hình
5. Cho biết độ đo chính xác (accuracy) trên tập dữ liệu đánh giá (test data set)
6. Xây dựng ma trận confusion matrix của mô hình trên tập dữ liệu test
7. Với chủng loại: Iris-setosa
    a. Hãy cho biết các giá trị: [tn, fp, fn, tp], [TPR, FNR, FPR, TNR], [precision, recal, F1]
    b. Vẽ đồ thị AUC & ROC
8. Lưu trữ mô hình với tên file iris_knn
9. Xây dựng chương trình dự báo chủng loại dựa trên các đặc trưng đầu vào.
10. Với các đặc trưng SepalLengthCm=4.5, SepalWidthCm=2.7, PetalLengthCm=2.0, PetalWidthCm=0.24 thì mô hình phân lớp sẽ dự báo là chủng loại nào?