

THỰC HÀNH HỌC MÁY CÓ GIÁM SÁT VỚI K-NN

Sinh viên tìm kiếm kênh Youtube **Nam Media Tivi** và sử dụng video: *Kỹ thuật Grid Search chọn Hyperparameters Tuning trong mô hình phân lớp với giải thuật K-NN*.

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.

- Tiến hành EDA dữ liệu:
 - Có bao nhiêu đặc trưng và tên đặc trưng
 - Có bao nhiêu chủng loài
 - Tổng quan dữ liệu kiểu dữ liệu và dữ liệu thiếu
 - Thống kê các đại lượng cơ bản cho các đặc trưng
 - Thống kê các đại lượng cơ bản cho các đặc trưng theo nhóm chủng loại
 - Trực quan phân phối các đặc trưng và nhận xét
 - Trực quan phân phối các đặc trưng theo nhóm chủng loại và nhận xét
 - Khám phá các giá trị bất thường của các đặc trưng theo nhóm chủng loại
- Tập dữ liệu sử dụng tỉ lệ train:test là 75:25 với hệ số ngẫu nhiên là 16
- 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.
- Sử dụng giá trị K tốt nhất để xây dựng mô hình
- Cho biết độ đo chính xác (accuracy) trên tập dữ liệu đánh giá (test data set)
- Xây dựng ma trận confusion matrix của mô hình trên tập dữ liệu test
- Với chủng loại: Iris-setosa
 - Hãy cho biết các giá trị: [tn, fp, fn, tp], [TPR, FNR, FPR, TNR], [precision, recall, F1]
 - Vẽ đồ thị AUC & ROC
- Lưu trữ mô hình với tên file iris_knn
- Xây dựng chương trình dự báo chủng loại dựa trên các đặc trưng đầu vào.
- 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?