SPOT QUESTION

USING DECISION TREE AND K-NEAREST NEIGHBORS

Portuguese winery assesses 6497 wine samples (4898 red, 1599 white) using 11 physicochemical attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free/total sulfur dioxide, density, pH, sulphates, alcohol. Quality scores (0-10, mostly 3-8) guide blending/marketing decisions.

Scenarios include**:** High-Quality Wines (score ≥7): Balanced acidity (7-9), high alcohol (10-12%), low volatile acidity (<0.3). Medium-Quality (score 5-6): Average chemistry, common bulk wines. Low-Quality (score ≤4): High volatile acidity (>0.5), excess sugar (>10g/L).

Download from UCI ML Repository:

Red: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

White: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

Classify quality accurately to optimize production, experimenting with DT/KNN

Experiments

1. Merge datasets (add 'type' column: 0=red/1=white). Scale features and Stratified train/test split (80/20)
2. DT Hyperparameters: max_depth=, min_samples_split=, min_samples_leaf=, criterion=['gini','entropy']
3. KNN Hyperparameters: n_neighbors=, weights=['uniform','distance'], metric=['euclidean','manhattan']
4. Observe Key Challenges and Record the same under different Scenarios Test

Scenario	Samples	Classes	Key Challenge
1. Full	6497	3-9 (multi-class)	
2. Binary	6497	≥6 vs <6	
3. Red-only	1599	3-8	

5. Evaluation: Per-class Precision/Recall/F1, macro/micro-F1, confusion matrices; plot feature importances (DT), validation curves.
6. Inferences
    a. How does DT depth=10 perform on full vs red-only?
    b. How does KNN k=5 perform on full vs red-only?