**Apu Datta**  
Data Mining and Visualization  
Project: 2, Question 1: Solving a Classification Problem Using a Tabular (CSV) Dataset

**Project: Wine Quality Classification Using Ensemble Learning**

In this project, I built a machine learning system to predict the quality of wine Good, Average, or Poor based on its chemical properties. I used a well known Kaggle dataset (6,497 samples of red and white wines) containing measurements like alcohol content, acidity, sulphates, sugar, and more.

**Methodology:**

I have preprocessed the data by imputing missing values within wine type (red/white), label encoding the wine type, grouping the original 3 to 9 quality scores into three classes (Poor ≤4, Average 5–6, Good ≥7), and standardizing numeric features with StandardScaler. I have split the data with 70 to 30 stratified training and testing split. To address class imbalance, I have applied SMOTE on the training set and evaluated class-weighted models. I compared multiple algorithms Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree, Random Forest, Naive Bayes, plus K-Means as an unsupervised baseline using 5K fold cross validation for fair evaluation. For KNN, I built a validation curve (k from 1 to 29) and found k = 1 (Euclidean) to be optimal on this dataset.

**Model Evaluation:**

Across all models, Random Forest (Balanced) performed best with 82.2% test accuracy and strong, stable CV. KNN (k=1) reached 77.6% test accuracy. Feature importance from Random Forest highlighted alcohol, free SO₂, volatile acidity, total SO₂, and density as key predictors. Models were selected based on accuracy and macro level performance so that minority classes aren’t ignored.

**Deployment:**

The final model was serialized with Pickle (model, scaler, encoders) and deployed via Streamlit for real time predictions: <https://4okshqzpjpsapep79xtmbs.streamlit.app/>

**Real world Application:**

This system can help wineries and QA teams screen batches instantly before expert tasting, standardize quality checks, and surface which chemical factors are pushing a sample toward Poor vs. Good supporting faster, more consistent decisions in production.

**Dataset Reference:**  
Wine Quality Dataset: <https://www.kaggle.com/datasets/rajyellow46/wine-quality>

Github: <https://github.com/dattaBus-anls/-Professional-Wine-Quality-Classifier-.git>

**Appendix:**

1. **Testing Parameters**

**Test 1: Poor Quality Wine 🔴**

Fixed Acidity: 6.5

Volatile Acidity: 0.55

Citric Acid: 0.15

Residual Sugar: 8.0

Chlorides: 0.08

Free SO₂: 15.0

Total SO₂: 140.0

Density: 0.997

pH: 3.35

Sulphates: 0.4

Alcohol: 9.2

Wine Type: White

**Expected: 🔴 Poor Quality Wine**

**Test 2: Good Quality Wine 🟢**

Fixed Acidity: 7.2

Volatile Acidity: 0.25

Citric Acid: 0.35

Residual Sugar: 2.5

Chlorides: 0.045

Free SO₂: 35.0

Total SO₂: 120.0

Density: 0.993

pH: 3.15

Sulphates: 0.65

Alcohol: 12.5

Wine Type: White

**Expected: 🟢 Good Quality Wine**

**Test 3: Average Quality Wine 🟡**

Fixed Acidity: 8.60

Volatile Acidity: 0.63

Citric Acid: 0.35

pH Level: 3.15

Chlorides: 0.04

Sulphates: 0.65

Residual Sugar: 2.50

Free SO₂: 49.00

Total SO₂: 120.00

Density: 1.00

Alcohol Content: 8.39

**Expected: 🟡 Average Quality Wine**

1. **Visualizations**

***Image: alcohol\_vs\_volatile.png***  
A close-up of a graph

AI-generated content may be incorrect.

***Image: confusion\_matrix.png***

A graph with numbers and a blue square

AI-generated content may be incorrect.

***Image: cv\_results.png***

A graph of blue rectangular columns

AI-generated content may be incorrect.

***Image: feature\_distribution-1.png***

A group of graphs showing different sizes of data

AI-generated content may be incorrect.

***Image: feature\_distribution-2.png***

A close-up of a graph

AI-generated content may be incorrect.

***Image: feature\_distribution-2.png***

A diagram of different colors

AI-generated content may be incorrect.

***Image: feature\_distribution-4.png***

A screenshot of a graph

AI-generated content may be incorrect.

***Image: feature\_importance.png***

A graph of blue bars with white text

AI-generated content may be incorrect.

***Image: kmeans\_elbow.png***

A graph with a number of clusters

AI-generated content may be incorrect.

***Image: knn\_validation\_curve.png***

A graph of a graph

AI-generated content may be incorrect.

***Image: model vs weighted avg.png***

A graph of different colored lines

AI-generated content may be incorrect.