APPLIED MACHINE LEARNING

Manuel Arambula, Quoc Dat Cao, Xavier Maravilla, Levi Valencia

1. INTRODUCTION

- Goal: Evaluate supervised ML model on 3 datasets.
- Tackled real-world challenges like imbalance datasets.
- Preprocessing: Standardization, outlier removal, polynomial features, hyperparameter tuning.
- Models: Logistic Regression, K-NN, Naïve Bayes, Decision Tree, Random Forest.

2. RELATED WORKS

- Wine Dataset:

Cortez et al. (2009) predicted wine quality with SVM (62.4% and 64.6% accuracy).

- Car Dataset:

Bohanec and Rajkovic (1988) used decision-making systems based on tree-structured criteria.

3. DATASETS OVERVIEW

Wine Datasets:

Red wine (1599 samples), White wine (4898 samples).

11 features + quality label (0-10).

Imbalanced (most samples quality 5-7).

Car Evaluation Dataset:

6 categorical features, 1728 samples.

Hierarchical decision model.

```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         6497 non-null   float64
 1   volatile acidity      6497 non-null   float64
 2   citric acid           6497 non-null   float64
 3   residual sugar        6497 non-null   float64
 4   chlorides             6497 non-null   float64
 5   free sulfur dioxide   6497 non-null   float64
 6   total sulfur dioxide  6497 non-null   float64
 7   density               6497 non-null   float64
 8   pH                    6497 non-null   float64
 9   sulphates             6497 non-null   float64
 10  alcohol               6497 non-null   float64
 11  quality               6497 non-null   int64
 12  wine_type             6497 non-null   object
```

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   buying    1728 non-null   object
 1   maint     1728 non-null   object
 2   doors     1728 non-null   object
 3   persons   1728 non-null   object
 4   lug_boot  1728 non-null   object
 5   safety    1728 non-null   object
 6   class     1728 non-null   object
```

4. PREPROCESSING STEPS

Wine Dataset

- Standardization: Rescaled features to mean=0, std=1.
- Polynomial Feature Transformation: Degree 3, to capture non-linear patterns.
- Z-score Outlier Removal: Threshold=3; removed extreme outliers to stabilize models.
- 80/20 Train-Test Split: Ensured balanced evaluation and enough training data.

Car Dataset

- Label Encoding: Converted categorical features into numbers.
- Standardization: Even categorical values were scaled for distance-based models like K-NN.
- Polynomial Features: Degree 3 added complexity
- Z-score Filtering: Minor improvements; categorical data had fewer outliers.
- GridSearchCV: Hyperparameter tuning for best model settings

5. METHODOLOGY - MODELS USED

Logistic Regression

- Strong baseline for multi-class classification.
- Simple, interpretable model - good for numeric features.
- Performed better after standardization and polynomial transformation.

K-Nearest Neighbors (K-NN)

- Simple, non-parametric model.
- Great for non-linear decision boundaries.
- Very sensitive to feature scaling - benefitted heavily from standardization and tuning.

Random Forest Classifier

- Ensemble of Decision Trees; reduces overfitting.
- Automatically handles feature interactions and outliers.
- Most robust model for the Wine dataset, high accuracy without needing heavy preprocessing.

5. METHODOLOGY - MODELS USED (CONT.)

K-Nearest Neighbors (K-NN)

- Same K-NN model; relied on distance between points.
- Needed careful scaling for encoded categorical features.

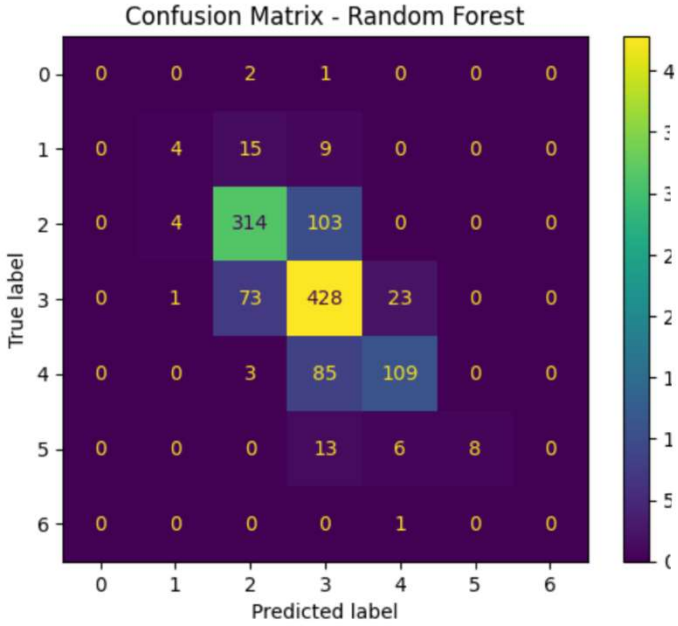
Decision Tree Classifier

- Naturally fits structured, categorical data.
- Fast to train, easy to interpret.
- Achieved near-perfect classification on Car dataset.

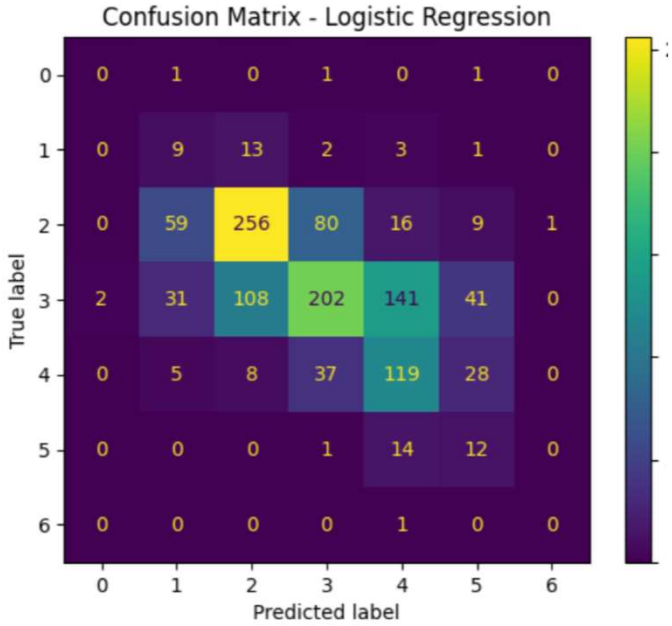
Categorical Naive Bayes

- Probabilistic model tailored for categorical features.
- Lightweight, fast, but assumes feature independence – limited improvement from feature engineering.

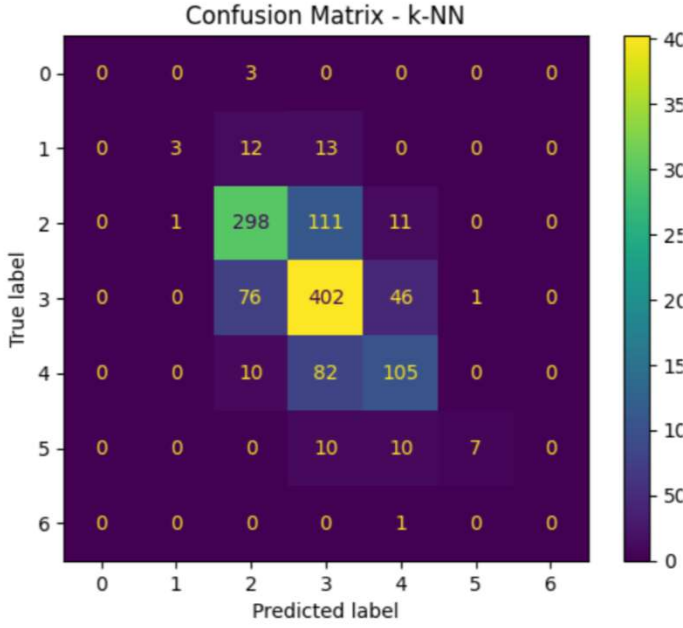
accuracy			0.72	1202
macro avg	0.52	0.36	0.40	1202
weighted avg	0.72	0.72	0.71	1202



accuracy			0.50	1202
macro avg	0.27	0.34	0.28	1202
weighted avg	0.58	0.50	0.52	1202



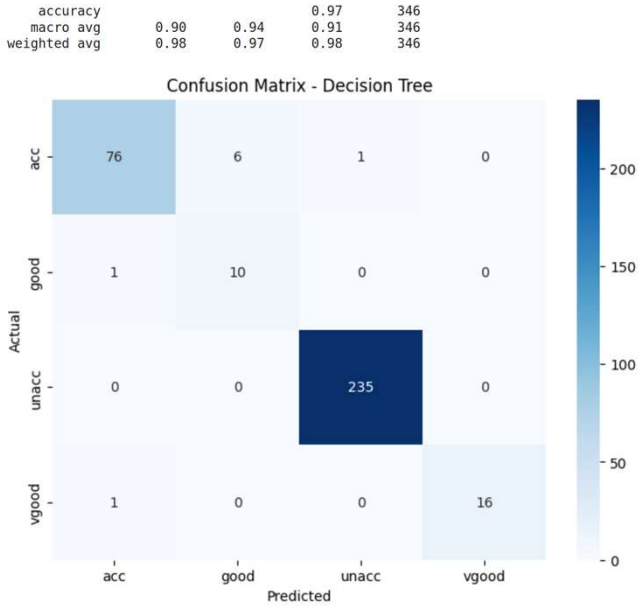
accuracy			0.68	1202
macro avg	0.52	0.34	0.37	1202
weighted avg	0.68	0.68	0.67	1202

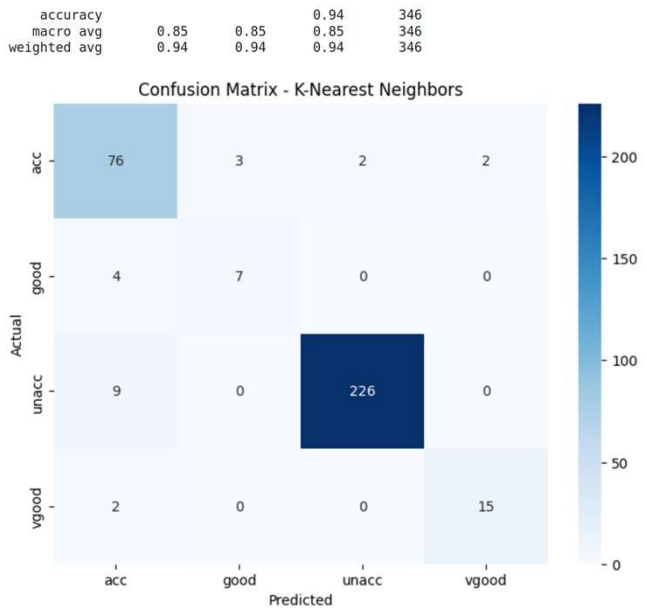


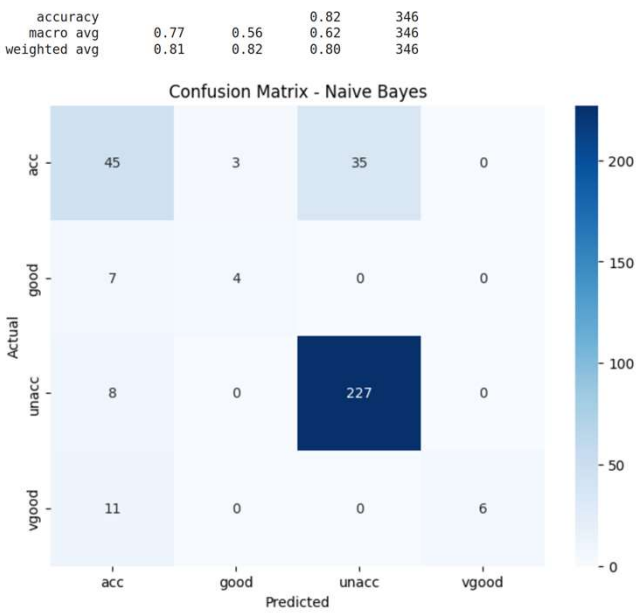
6. RESULTS - WINE DATASET

- Logistic Regression: 50% accuracy (up from 32% with polynomial features).
- K-NN: Improved to 68% with outlier removal and tuning (n=5, Manhattan distance).
- Random Forest: Best performance (72% accuracy).Naturally handled outliers and feature interactions.

6. RESULTS: CAR DATASET (CONT.)







- K-NN: 94% baseline accuracy, Polynomial Features worsened accuracy slightly.
- Categorical Naive Bayes: 82% accuracy.
- Decision Tree: Best model, 98% accuracy.

7. CONCLUSION

- Preprocessing and tuning greatly improved model performance.
- Random Forest excelled for wine dataset; Decision Tree excelled for car evaluation.
- Importance of choosing the right model and preprocessing based on dataset.