Predicting RED Wine Quality Using K-Nearest Neighbours

Team Name: CRACK CODE

Tarun Kumar (2025JRB2025)

KNN Regressor Analysis

1 Introduction

This report presents a regression-based analysis of the Red Wine Quality dataset using:

- Custom KNN Regressor
- Linear Regression
- Ridge Regression
- Random Forest Regressor

The objective is to predict wine quality scores (0–10) using physicochemical properties.

2 Data Preprocessing

- Original samples: 1599
- Duplicates removed: 240
- Final dataset size: 1359
- Train–Test split: 70:30
- Feature transformation: PowerTransformer (Yeo–Johnson)

Why PowerTransformer?

It reduces skewness, stabilizes variance, and improves model convergence and performance.

3 Models Implemented

3.1 Custom KNN Regressor

- Distance metrics: Euclidean / Manhattan / Minkowski
- Weighting schemes: Uniform / Distance
- Number of neighbors: $k = 14$ (varied during tuning)

3.2 Linear Regression

A baseline Ordinary Least Squares (OLS) regression model.

3.3 Ridge Regression

- Regularization parameter: $\alpha = 12.5$
- Controls coefficient magnitude to reduce overfitting

3.4 Random Forest Regressor

- Ensemble model using multiple decision trees
- Captures nonlinear relationships effectively
- Reduces variance compared to single-tree models

4 Evaluation Metrics

- R^2 Score
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- 5-Fold Cross-Validation Score

5 Results: 70:30 Train-Test Split

Table 1: **Regression Model Performance**

Model	R^2	**MSE**	**RMSE**	**CV Score**
KNN Regressor	0.4003	0.4059	0.6371	0.3048
Linear Regression	0.3826	0.4178	0.6464	0.3272
Ridge Regression	0.3834	0.4173	0.6460	0.3280
Random Forest Regressor	**0.4189**	**0.3921**	**0.6264**	**0.3417**

Random Forest Insight

Random Forest Regressor achieved the highest R^2 score and the lowest error values, demonstrating superior nonlinear modeling capability.

6 Additional Experiment: 80:20 KNN Split

KNN Results (80:20 Split)

R^2: 0.4070

CV Scores: [0.3592, 0.3450, 0.3079, 0.2388, 0.3446]

Mean CV R^2: 0.3191

7 Hypertuning of k

$$k \in [1, 5, 9, 13, 17, 21, 25, 29, 33, 37]$$

GridSearchCV Result

Best k: 25
Best CV Score: 0.32827

8 Conclusion (Regression)

- **Best Overall Model:** Random Forest Regressor
- **Best Distance-Based Model:** KNN Regressor
- **Most Stable Linear Model:** Ridge Regression

KNN Classification Analysis

1 Introduction

Wine quality is converted into a binary classification problem:

$$\text{Good (1) if quality} \geq 7, \quad \text{else Poor (0)}$$

2 Custom KNN (No PCA)

Accuracy: 0.9007
Precision: 0.6296
Recall: 0.5000
F1 Score: 0.5574
CV Score: 0.8666
ROC-AUC: 0.8151

3 KNN with PCA

PCA Components: 8

Accuracy: 88.60%

4 Other Classifiers

4.1 Logistic Regression

Accuracy: 0.9081
Precision: 0.6667
Recall: 0.5294
F1 Score: 0.5902
CV Score: 0.8685
ROC-AUC: 0.8914

4.2 Decision Tree

Accuracy: 0.8860
Precision: 0.5405
Recall: 0.5882
F1 Score: 0.5634
CV Score: 0.8270
ROC-AUC: 0.7584

4.3 Random Forest

Accuracy: 0.9007
Precision: 0.6842
Recall: 0.3824
F1 Score: 0.4906
CV Score: 0.8712
ROC-AUC: 0.8967

4.4 Support Vector Classifier

Accuracy: 0.9081
Precision: 0.8000
Recall: 0.3529
F1 Score: 0.4898
CV Score: 0.8620
ROC-AUC: 0.8739

5 Handling Class Imbalance (SMOTE + ENN)

Accuracy: 0.8713
Pipeline CV Score: 0.8491

6 Conclusion

- Logistic Regression and SVC achieved the highest overall accuracy on the original imbalanced dataset, indicating strong global classification performance.
- Random Forest produced the best ROC-AUC score, demonstrating superior ranking capability and probability estimation.
- The custom KNN classifier showed competitive performance on the imbalanced dataset but remained sensitive to skewed class distributions.
- After applying **SMOTE + ENN**, the model achieved an **overall accuracy of 71.32%** and an **macro F1-score of 0.76**, reflecting improved balance between precision and recall.
- The recall for the minority (good-quality wine) class increased significantly after resampling, enabling better detection of rare positive samples.
- This improvement in recall came at the cost of reduced overall accuracy and F1-score compared to the imbalanced setting, highlighting the trade-off between global performance and minority-class sensitivity.
- The imbalanced-data experiment confirms that a **combined evaluation of accuracy, F1-score, recall, and ROC-AUC** provides a more reliable assessment than accuracy alone for skewed datasets.