Predicting wine quality using K-nearest neighbours

Github URL https://github.com/scholar-lab/ML_Project
Student Angom Umakanta Singh
Entry No. 2025EEZ8518
Course AIL7024

Introduction

Following experiments were completed as per the given instructions.

1. Regression of wine quality.
2. Binary classification of wine quality as good and bad.

The experiment was carried out with python 3.10.16 and Nvidia RTX5000 Quadro with 7 CPUs, 32 GB RAM and 16 GB VRAM.

Preprocessing

The wine classification data provided in the instruction is loaded. The dataset is split in two ways 70:30 and 80:20 train test sets.

Regression of wine quality

For each of the split. Three distance metrics are used in sequence to find the best k using grid search and KNN regression is performed. Finally the baseline is calculated for each of the split. The result is as shown in Table 1.

Experimental Results

Split	Model	Distance	Best k	MSE	RMSE	R2
80:20	KNN	Euclidean	7	0.48	0.70	0.34
80:20	KNN	Manhattan	11	0.50	0.71	0.33
80:20	KNN	Minkowski	7	0.47	0.69	0.36
80:20	Linear Regression	NA	NA	0.55	0.74	0.26
70:30	KNN	Euclidean	7	0.48	0.69	0.35
70:30	KNN	Manhattan	11	0.49	0.70	0.33
70:30	KNN	Minkowski	7	0.48	0.69	0.34
70:30	Linear Regression	NA	NA	0.53	0.73	0.27

Table 1: Result for Regression of Wine quality using KNN

Classification

For each of the split type the performance metrics is computed and compared with the baseline as shown in the table 2.

Experimental Results

Split	Model	Distance	Best k	Accuracy	Precision	Recall	F1	ROC_AUC
80:20	KNN	Euclidean	1	0.86	0.64	0.63	0.63	0.77
80:20	KNN	Manhattan	1	0.86	0.65	0.65	0.65	0.78
80:20	KNN	Minkowski	1	0.86	0.64	0.63	0.63	0.77
80:20	1R	NA	NA	0.80	0.00	0.00	0.00	0.72
80:20	Decision Tree	NA	NA	0.85	0.60	0.63	0.61	0.76
80:20	Random Forest	NA	NA	0.88	0.78	0.55	0.65	0.92
80:20	SVM	NA	NA	0.83	0.67	0.28	0.40	0.84
70:30	KNN	Euclidean	1	0.86	0.64	0.64	0.64	0.78
70:30	KNN	Manhattan	1	0.86	0.64	0.64	0.64	0.78
70:30	KNN	Minkowski	1	0.85	0.63	0.63	0.63	0.77
70:30	1R	NA	NA	0.80	0.00	0.00	0.00	0.72
70:30	Decision Tree	NA	NA	0.84	0.59	0.63	0.61	0.76
70:30	Random Forest	NA	NA	0.88	0.79	0.55	0.65	0.91
70:30	SVM	NA	NA	0.83	0.66	0.28	0.40	0.84

Table 2: Result for classification of Wine quality using KNN

Conclusion

For the regression task, it is observed that the both the split give similar results. Manhattan distance perform the worst for the MSE and RMSE, but slightly better in R2 metrics. In general the implemented model performs better than the baseline model.

For the binary classification using KNN, The accuracy is similar for both data splits. The classifier performs better than baseline classification using 1R, Decision tree and SVM . However Random Forest classifier performs better than the implemented model.