REPORT ON

Wine Dataset Classification

and

Regression

Using Machine learning and Deep learning algorithms

By

Sushan Shankar

Introduction:

Wine quality classification is a typical duty in the wine industry, to classify wines based on their chemical features in order to anticipate their quality. For this study, we used a collection of wine samples with 6497 samples and 12 attributes, including density, pH, sulphates, alcohol, quality, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, and total sulphur dioxide. Our target variable, which goes from 1 to 10, is wine quality.

Results:

Using Z-scores and subplots, we first explored the dataset to look for any outliers. The correlations between various attributes and the objective variable were then clarified using a variety of visualisation techniques, such as histograms, scatter plots, and subplots. We divided the data into training and testing sets after analysis.

Then, we applied four different machine learning algorithms—Linear regression, XGBoost, ANN, Logistic regression, Random Forest, and Decision Tree—to predict wine quality. These algorithms were developed using the training set, and their effectiveness was assessed using the test set.

The confusion matrix for the logistic regression method, which had an accuracy score of 0.824 and correctly predicted 1004 low-quality wines and 68 high-quality wines, but incorrectly identified 184 low-quality wines and 44 high-quality wines.

With an accuracy score of 0.893, the Random Forest algorithm performed better than all other algorithms. It properly forecasted 1015 low-quality wines and 146 high-quality wines, according to its confusion matrix, while misclassifying only 33 low-quality wines and 106 high-quality wines.

The Decision Tree algorithm achieved an MSE of 0.136, which indicates that it performed relatively well but not as well as the Random Forest algorithm.

Finally, we gave some sample data to our model, and it correctly predicted the quality of the wine in most cases.

Two additional models, LSTM and MLP, have also been implemented. The LSTM model has a test loss of 0.5073338150978088, and the MLP model has a test

MSE of 0.4930566425509396. These models seem to perform well, and they are relatively new models that are being used in various fields of study.

	RMSE	MAE	MSE	R-squared Score	Accuracy Score	Type
Linear Regression	**0.73939**	**0.56587**	**0.5467**	**0.25976731**		Regression
XGBoost	**0.62903**	**0.46277**	**0.39568**	**0.4642409**		
ANN	**0.75841**	**0.55905**	**0.57519**	**0.22118286**		
Logistic Regression					**0.82461538**	Classification
Random Forest					**0.89307692**	
LSTM			**0.50733**			DL Model
MLP			**0.49306**			

Figure 1. Result table

Conclusion:

In conclusion, the quality of wine can be predicted using various machine learning algorithms based on a variety of chemical characteristics. Although The LSTM and MLP models are relatively new models that are being used in various fields of study, and they seem to perform well in this context as well.

With an accuracy score of 0.893, we discovered that the Random Forest method performed better than all other algorithms. In the wine sector, this experiment shows how machine learning algorithms can accurately predict wine quality based on chemical characteristics.

However, by utilising more sophisticated techniques and taking into account additional elements like the region, climate, and type of soil in which the grapes were grown, more research can be done to increase the accuracy of these algorithms.