

Wine Quality Prediction

By Anushree Anil(1901039)

Problem Definition:

Given two multivariate datasets for red and white wine , from the north of Portugal we have to predict the quality of the wine based on physicochemical tests.

The datasets can be viewed as regression or classification problem.

Attribute Information:

Features:

fixed acidity,volatile acidity,citric acid,residual sugar,chlorides,free sulfur dioxide,total sulfur dioxide,density,pH,sulphates,alcohol

Output variable (based on sensory data):

quality (score between 0 and 10)



Literature Survey:

1.Dahal, K.R., Dahal, J.N., Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. <https://doi.org/10.4236/ojs.2021.112015>

DATA PARTITION:

The data was split into training data set and testing data set in the ratio 3:1. They trained the data and is used to find the relationship between target and predictor variables

FEATURE SCALING:

They used the standardization method. The formula they used was:


$z = \frac{x - \text{mean}}{\text{std}}$ where z , x , mean, and std are standardized input, input, mean and standard deviation of the feature, respectively

ALGORITHMS USED:

- 1.Ridge Regression
- 2.Support Vector Machine
3. Gradient Boosting Regressor
- 4.Artificial Neural Network (ANNs)

CONCLUSION:

This work demonstrated that various statistical analysis can be used to analyze the parameters in the existing dataset to determine the wine quality. Based on their analysis, Gradient Boosting performs best to predict the wine quality. The prediction of ANN lies behind other mathematical models because the dataset is small and heavily skewed.If the datasets were large enough then ANN could render better predictions.



2.Wine Quality Prediction Using Different Machine Learning Techniques Bhavya AG

Department of Computer Science and Engineering Cambridge Institute of Technology K.R. Puram Bangalore,India

DATA PRE-PROCESSING:


The dataset was divided into train and test set.

CONCLUSION:

For classification models Random Forest performed better than others.

ALGORITHMS USED:

- 1.Naïve Bayes algorithm
- 2.SVM Classifier Algorithm
- 3.K Nearest Neighbors (KNN)
- 4.Random Forest



3.A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality

Mohit Gupta, Vanmathi C

DATA PRE-PROCESSING:

In this particular paper the datasets they have used does not have any outliers or missing values so they kept the datasets unchanged.

FEATURE SELECTION:

Volatile Acidity shares a correlation with quality. The concentration of acid contributes the higher wine quality that direct correlation. An honest range against alcohol expected for quality wine. The lower concentration of Chloride appears to provide higher quality wines. Better wines used to be more acidic. The important attributes are as follows acidity, sugar content, chlorides, sulfur, alcohol, pH and density

DRAWBACK:

Dataset was not broad.

ALGORITHMS USED:

- 1.Random Forest
- 2.Support Vector Machine
3. Decision Tree
- 4.K Nearest Neighbors (KNN)
- 5.MP5 Model(combination of data classification and regression)

CONCLUSION:

Using white wine samples only one variant, the Random Forest variant, performed better. K-nearest neighbors performed statistically worse.

While in the case of samples in red wine, only Random Forest performed well.



Result Analysis



Regression Models:

Both the red wine and white wine datasets were divided into train and test set, with train set as 60% of the whole dataset and remaining 40% of the dataset as test set.

After training the regression model with the train set, Coefficients for both types of wine and Mean squared error for both train and test set was calculated.

Linear Regression

For Red Wine:

Training MSE: 0.40916942858811073

Test MSE: 0.43470158338896586

For white wine:

Training MSE: 0.569490575930567

Test MSE: 0.5553310548807744

Stochastic Gradient Descent Regression

For Red Wine:

Training MSE: 0.4098304278073065

Test MSE: 0.4340388665924884

For white wine:

Training MSE: 0.5714168967386914

Test MSE: 0.5605763606058174



Classification Models:

As the quality of the wine in the original dataset was in the range 1-10 where 1 is 'very bad' and 10 as 'very good',
So for classification models the quality was replaced by two values 0 and 1 ,
where 1 means "good quality" and 0 means "bad quality".

If the quality values are greater than 6.5 then we change it to 1 and if it is less than 6.5 then it is 0.

The datasets were normalized.

After the data preprocessing the datasets were divided into train, validation and test sets.

60% were given to train sets and 20% for validation sets and 20% for test sets.



Classification Models Used in this project:

1. Logistic Regression
2. Perceptron Model
3. Sigmoid Neuron
4. Multi Layer Perceptron

After training all the classification models with the train set , following things were calculated for both types of wine:

1. Confusion Matrix
2. Class-wise Accuracy, precision, recall
3. Train and Test Accuracy
4. K-Fold cross validation accuracies



Logistic Regression:

Red Wine:

Training Accuracy : 89.25964546402503

Testing Accuracy : 87.1875

White Wine:

Training Accuracy : 80.428863172226

Testing Accuracy : 80.40816326530611

Sigmoid Neuron :

Red Wine:

Training Accuracy : 87.69551616266945

Testing Accuracy : 86.875

White Wine:

Training Accuracy : 79.88427501701838

Testing Accuracy : 78.57142857142857

Perceptron :

Red Wine:

Training Accuracy : 86.02711157455683

Testing Accuracy : 84.0625

White Wine:

Training Accuracy : 76.2083049693669

Testing Accuracy : 77.9591836734694

Multi-Layer Perceptron :

Red Wine:

Training Accuracy : 99.7914494264859

Testing Accuracy : 89.65517241379311

White Wine:

Training Accuracy : 94.17971409121851

Testing Accuracy : 83.46938775510205



Conclusion

After analysing all the algorithms that were mentioned earlier,

it can be stated that among regression models both linear and stochastic gradient descent regression gives almost same MSE , so either of them can be used to predict the quality.

When it comes to classification models, for both the wines , Multi Layer Perceptron gives higher accuracies among all.