

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY

BACHELOR OF COMPUTER SCIENCE (HONS)

Session: 201905

**UCCD2063 ARTIFICIAL INTELLIGENCE TECHNIQUES**

**Classification of Red Wine Quality using Neural Networks**

|  |  |  |
| --- | --- | --- |
| Group Members | Student ID | Contribution |
| Leong Khei Sheng | 1403785 | 34% |
| Wan Kar Hou | 1607221 | 33% |
| Leong Khai Siang | 1803387 | 33% |

Table of Contents

[Introduction 1](#_Toc16988088)

[Methodology 2](#_Toc16988089)

[Experiments and Analysis 5](#_Toc16988090)

[Description of the dataset 5](#_Toc16988091)

[Details of the experiment 9](#_Toc16988092)

[Results and discussion 10](#_Toc16988093)

[Comparisons 15](#_Toc16988094)

[Conclusion 17](#_Toc16988095)

# Introduction

Wine is an alcoholic beverage made from grapes, fermented without the addition of sugars, acids, enzymes, water, or other nutrients, it could be categorized into two main groups, which were red wine and white wine. Red-wine production process involves extraction of color and flavor components from the grape skin and made from dark-colored grape varieties. The actual color of the wine can range from violet, typical of young wines, through red to mature wines, to brown for older red wines. Wine, and especially red wine Evidence suggests that moderate consumption may help people live longer, protect against certain cancers, improve mental health, and enhance heart health.

Quality is always the main component for improvement of wine making in wine industries .The quality of wine is difficult to define, as it is a multi-faceted construct, and lacking of generally accepted definition, yet this is most certainly accredited to everyone’s different perception of quality. As technology advanced, a more scientific method used to determine the quality of wine was introduced by scientist which was physicochemical. Physicochemical lab test used to characterize wine into the determination of density, alcohol or pH values and others. Nonetheless, the relationship between the physicochemical and sensory analysis is complex and still not fully understood.

In the project , we investigate how physicochemical are related to the quality of wine based on analytical data .Such study is useful for wine producers to potentially improve the revenue, marketing strategy, and the decision process ,most important, it would standardized the way of determine the quality of wine.

# Methodology



The data sets of red wine quality was obtained from UCI Machines Learning Repository. During the initial observation of the data, this dataset can be viewed as classification tasks or regression tasks. In this project, classification tasks needed to be performed to predict the quality of red wine given the physicochemical (input) variables. The data was explored and visualized after imported into Jupyter Notebook. Irrelevant columns such as ID must not be included in the features as these columns do not affect the quality of red wine. The data should checked with any empty or null values and have to be replaced with median values.

Besides that, the important information that can extracted from the dataset such as mean, standard deviation, and correlation have to be calculated and visualized to identify further preprocessing actions needed to be performed. These actions are important as the learning algorithms can handle all the features well and perform better accuracy. Generating correlation helps to identity the features that contribute much to the output. The histogram of each columns in the dataset should be plotted to check whether any visible issue to the data. The data should then split into 80% of training dataset and 20% of testing dataset.

To increase the accuracy of the model, the training dataset should have consistency amount of each output values. If the amount of each output values have much difference within each other, the learning model and predicted output will be highly affected. For this situation, the output values needed to be categorized into other forms that have balanced number of values. For both training dataset and testing dataset, standardization have to be performed so that the features have same scaled and standard deviation of 1. This was done to make sure that the learning algorithms can handle the data in the same scaled.

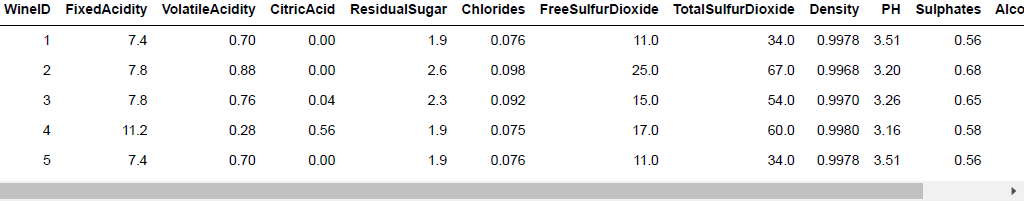
The classification algorithm that used in this project was MLPClassifier in Neural Networks. The classifier was fitted with training dataset and validated by predicting the same dataset. Features selection should be performed by using PCA to investigate that the model can works better by reducing the number of features. Both the training accuracy were compared to see reducing number of features can improve the training accuracy. After getting the better model, the model was performed cross validation to check the accuracy of the training output. A total of 4-folds cross validation needed to be performed and the average accuracy can be obtained. A confusion matrix will be generated to should the value of true positive, true negative, false positive and false negative.

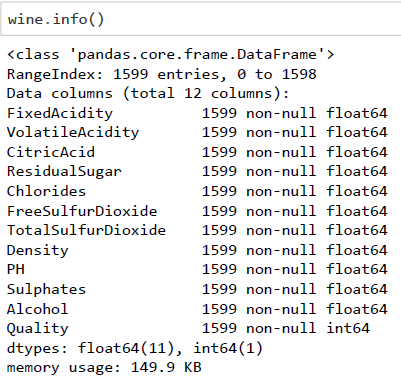
After validate the model with few steps, the model should be fined tune using GridSearchCV. This was done to obtain the best hyperparameters for the estimator with higher training accuracy. Lastly, the best model was then evaluated using the testing dataset to predict the quality of red wine. The accuracy of evaluation mode was obtained by comparing the actual output and predicted output. If the accuracy is too low, the model may experience overfitting where the estimator needed to adjust its hyperparameters to solve this issue.

If the evaluation part have high accuracy, the best model can be used for further prediction of red wine quality. Other classifiers such as RandomForestClassifier and Support Vector Classifier were also used to generate the models to test whether can obtain higher accuracy of predicted values. These was done to compare the predicted accuracy with MLPClassifier and check that other classifiers can fit the model better on this dataset.

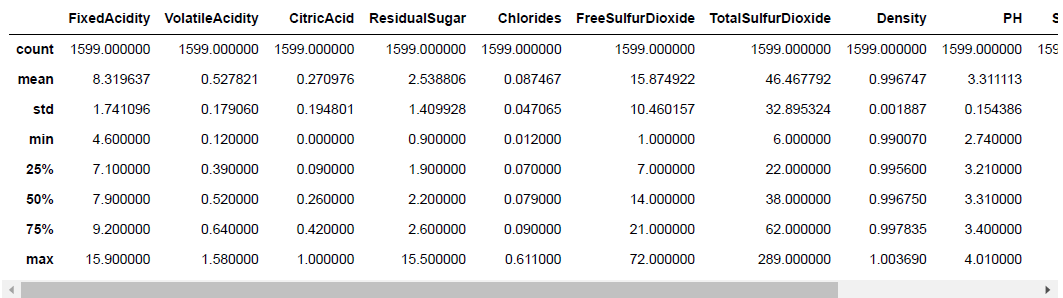
# Experiments and Analysis

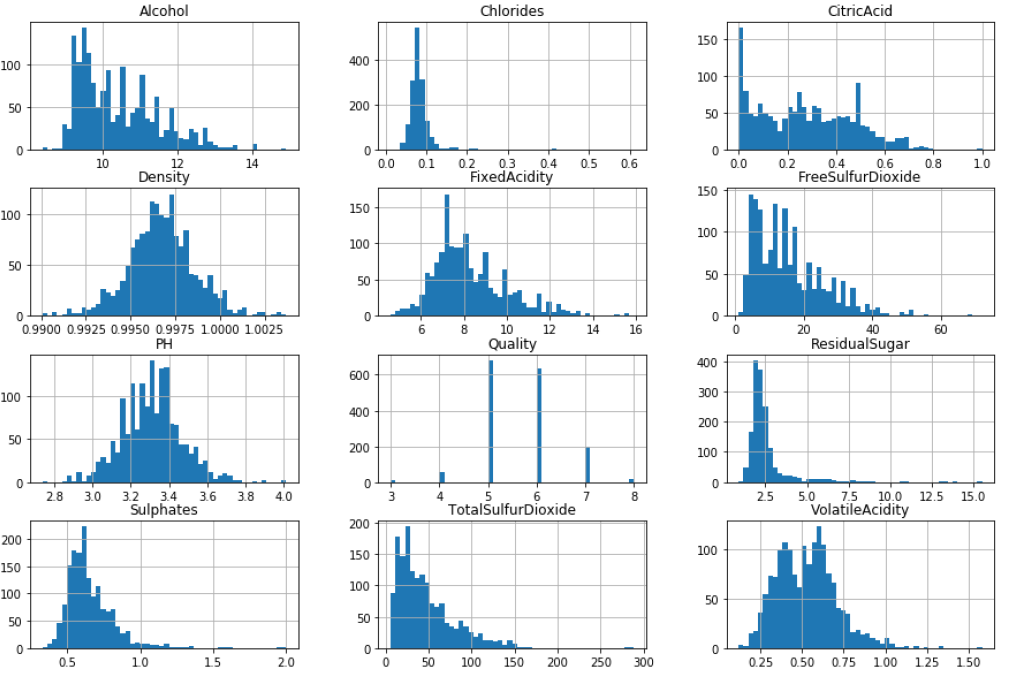
## Description of the dataset



In this assignment, the dataset that we used, Wine Quality Data Set was obtained from UCI Machine Learning Repository which uploaded on 2009. By analyzed the dataset using Jupyter Notebook, we found that there were total of 1599 entries and 12 columns in the dataset. The types of each columns was float and no null or empty values found in the dataset. There were 11 features in the dataset and the last column which referred to quality was the output of the dataset.

According to the references from the contributors, every feature in the dataset played as an important role which would highly affected the quality of red wines. Therefore, the amount of each feature should be controlled within a range values to ensure the good quality of red wines. The figure below shown the statistics and the graphs of all the columns that existed in the dataset.

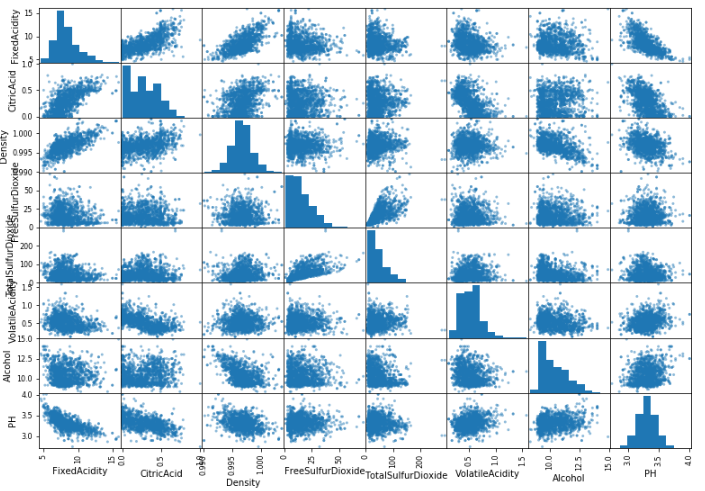


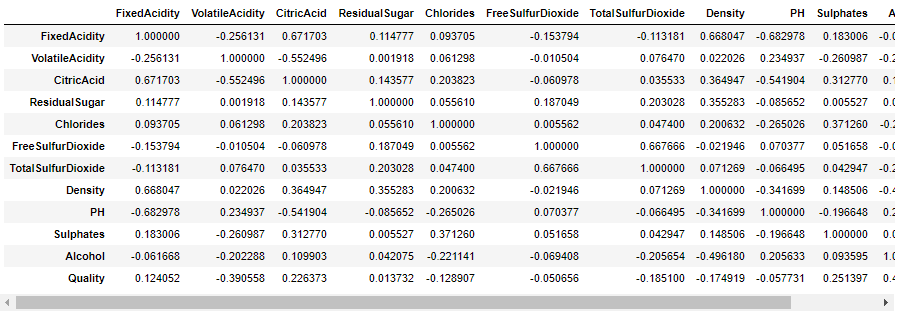


1. Alcohol – Refers to the percentage of alcohol content of the red wines. The graph shown was tail heavy and the mean value was around 10.4.
2. Chlorides – Stated the amount of salt in the red wines. There was a capped values found in the dataset.
3. Citric Acid – Added in the red wines to affect the ‘freshness’ and flavor of the red wines, usually was added in small quantities.
4. Density – The density was affected by the percentage of alcohol and sugar content in the red wines.
5. Fixed Acidity – Capped values was found in the graph.
6. Volatile Acidity – the amount of acetic acid that existed in the red wines. Too high of levels could lead to an unpleasant, vinegar taste.

According to the graph, the range of quality was between rating 3 to 8 and majority of the quality of red wines have either rating of 5 and 6. The values were not balanced to each other thus would highly affected the accuracy of prediction.

Since the attributes of all columns have very different scales, feature scaling should be done to allow the optimizer to find the best model easier.



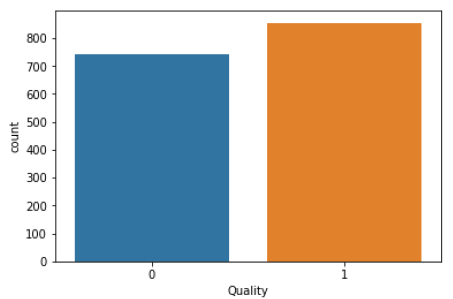


Correlation coefficient could guide us to analyze the most useful feature in the dataset. From the above figure, fixed acidity shown string positive correlation with citric acid and density.

The overall graph generated with the dataset shown most of the features were tail heavy and skewed toward the end. Some graph occurred capped values but none of these was the output values.

## Details of the experiment

In order to obtain higher accuracy of prediction, the output values were categorized into 2 groups where the values from 3 to 5 grouped into 0 (Low quality) while the values from 6 to 8 grouped into 1 (High quality). After the output values been categorized, the total amount of each category was more balanced to each other.

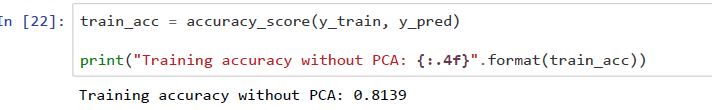


WineID was dropped during the initial phase as it was not one of the features and did not contributed at all the quality of red wines. It was used to labelled the datasets and for reference only. Besides that, the dataset was split into 80% of training set and 20% to testing set. The training set was used to generate the model while the testing set was used to validate the model. Prediction of values in testing set could let us know whether the model have underfit the data or overfit the data.

Standardization on the features was performed to ensure all the attributes have same scales so that the learning algorithms could handle the data well.

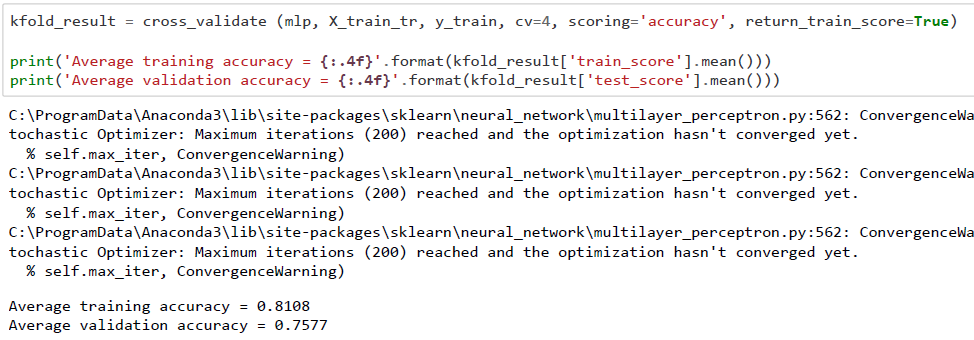
By referring to the above visualization, the data was not a linear model therefore linear model classifier should not be used in this assignment. In this assignment, MLPClassifier in Neural Network was chosen as the classifier to predict on the quality of red wines.

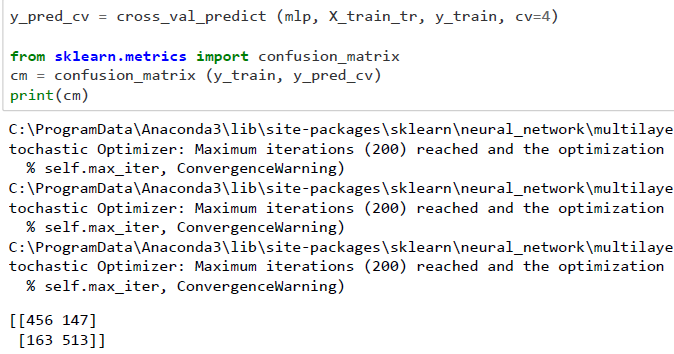
## Results and discussion

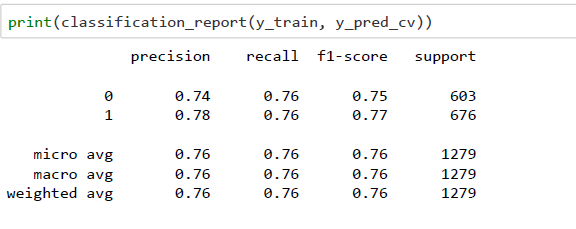
Before the training dataset was used to fit the model, the training was performed with PCA to compare the training accuracy with full features model. Principal Component Analysis, or PCA, is a dimensionality-reduction method that is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. In our coding, we try to reduce the dimensional of data into three dimensional, as a result, the accuracy of model with PCA is around 75% which lower than accuracy of model without PCA 81%. In the comparison of training’s accuracy between model with and without PCA, it is showed that the reduction of features or dimension of data would cost the loss significant of information as each features are highly dependent on each other.

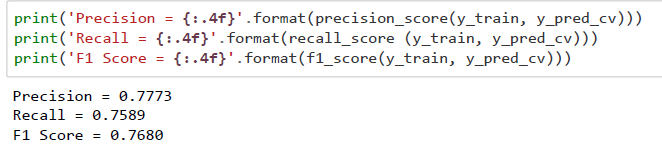
Therefore, the model was with full feature of dataset instead of selected features of the dataset.

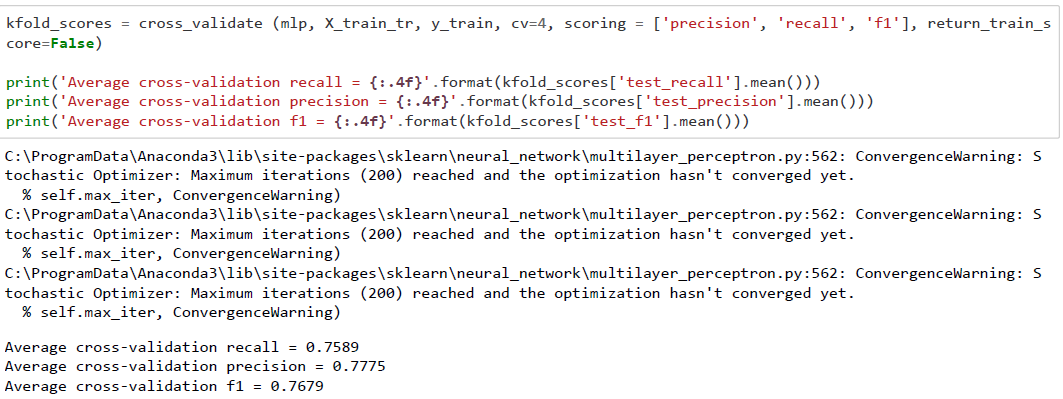
There were few validation steps performed to verify the accuracy of the training accuracy predicted using MLPClassifier with default hyperparameters.

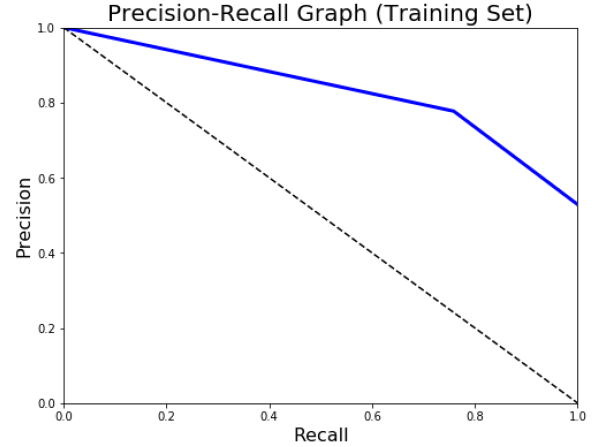


Cross validation mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. By applying cross validation to the model, we obtained that average training accuracy is around 81% which is 5% higher than average validation accuracy, 76%. Thus, the training model was quite accurate but we expected higher accuracy could be achieved. The differences between both validation accuracy and training accuracy did not exceeded 10%. With the validation of accuracy of 76%, it is estimated that around 76% accuracy will obtained when model is test with real data.

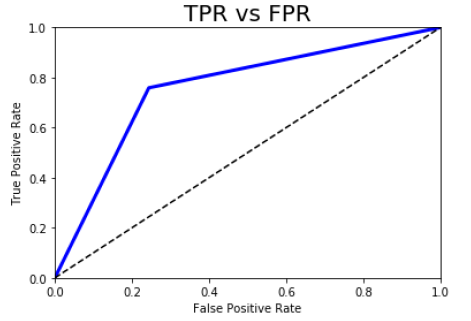
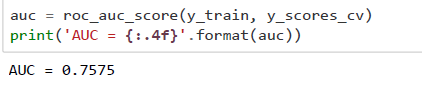
Confusion Matrix gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Precision measures the accuracy of positive prediction while recall measures the ratio of positive samples that are correctly detected. For Precision obtained is 0.773 (78%), which most of the samples predicted by the model is indeed positive while the Recall obtained is 0.759 (76%). Both precision and recall is above 0.5 and considered above average. The high precision and recall related to low false positive rate and low false negative rate , high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results. As the recall and precision is high, which resulted in high F1 score with 0.7680 (77%).



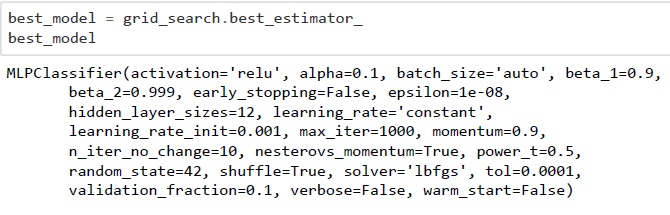


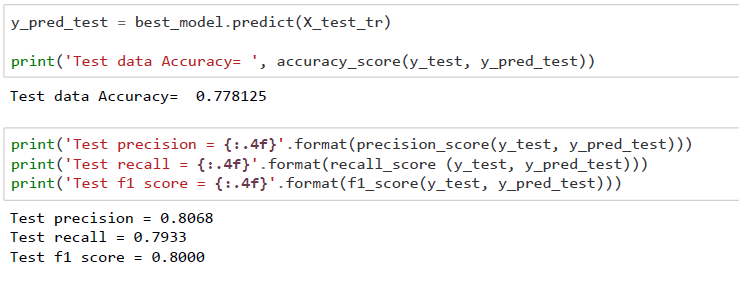


Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. From the Precision Recall graph of the model , we could observed that it stays a distance from the dotted line with represents P-R curve of a purely random classifier, thus, the classifier is considered as a good classifier. In ROC curves, the true positive rate (TPR, y-axis) is plotted against the false positive rate (FPR, x-axis), as for ROC curves of our model, the plotted line between TPR and FPR stays away from the dotted line which represent the ROC curve of a purely random classifier, thus, the classifier is considered as a good classifier. An ideal classifier does not make any prediction errors. This means that the classifier can perfectly separate the two classes such that the model achieves a true positive rate of 100% before producing any false positives. Thus, the AUC of such a classifier is 1. The AUC score for our classifier is 0.77575 , it means there is around 80% chance that model will be able to distinguish between positive class and negative class. 80 % is considered above average which not low enough to recognize as a bad classifier but still considered a good classifier.

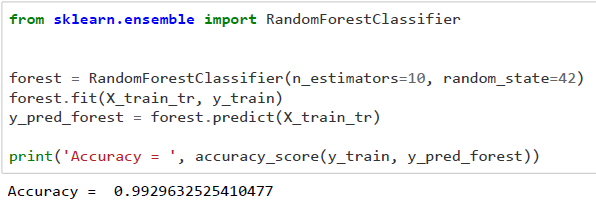
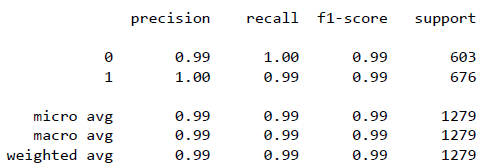
In order to obtained higher accuracy of training model, fine tuning was done to obtained the best hyperparameters. In this case, GridSearchCV was recommended and implemented to obtaine a best hyperparameters. Grid search builds a model for every combination of hyperparameters specified and evaluates each model which find the best model with suitable hyper parameters. By using GridSearchCV, the accuracy of the best model is 0.750586, when we used the best model to predict, it resulted a high accuracy of 0.778125 (78%), which is much higher than average validation accuracy of 0.76%, prediction on testing data during cross validation.

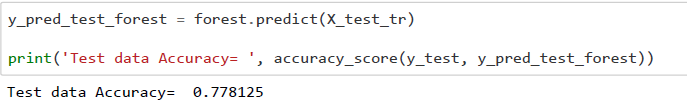




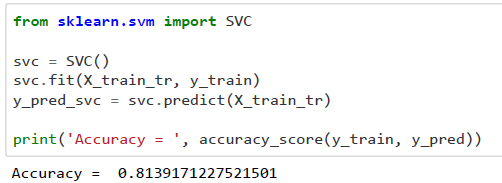
The best model was then used to validate with the testing dataset. The testing accuracy obtained was around 78%. The precision obtained was 80% while recall of 79% and F1 score of 80%. The predicted values were quite accurate with this hyperparameters and settings.

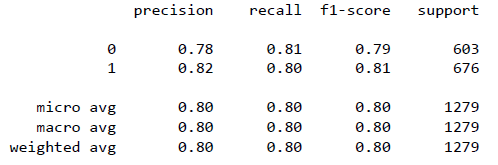
## Comparisons

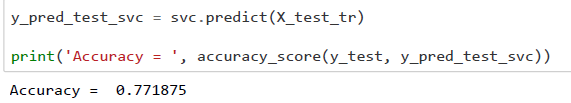
  




Random Forests Classifier (RFC) perform implicit feature selection and provide a pretty good indicator of feature importance and quick to train. The precision, recall and F1 score scored by RFG is around perfect score, 1, higher than score by MLP classifier, which around 0.8, Precision- Recall graph of RFC would be an ideal graph. However, RFC scored accuracy of 78 % in testing set, the accuracy scored in training set is almost 21% higher than the accuracy score in testing set, which show that the model tends to slightly overfitting. By comparing RFC with MLP classifier, the accuracy score by MLP classifier in training set is 81% and accuracy score in testing is around 78%, difference with only nearly 3%, show that the model is less overfitting compared to RFG, thus, MLP classifier is more suitable than Random Forest Classifier to apply as classifier for the datasets.







The precision, recall and F1 score scored by SVClassifier (SVC), which around 0.8, similar with score of MLP classifier. When using cross validation to evaluate the performance scored by RFC, average training accuracy is 80.30% while average validation accuracy is 75.46%, predicted that almost 75% accuracy for new data. SVClassifier performed well in training set with an accuracy score of 81.39%. However, the model scored accuracy of 77.18% in testing set, the accuracy scored in training set is almost 4% slightly higher than the accuracy score in testing set, which show that the model tends to slightly overfitting. By comparing SVC with MLP classifier, the accuracy score by MLP classifier network in training set is 81% and accuracy score in testing is around 78%, difference with only nearly 3%, SVClassifier have slightly worse performance during prediction of test data compared to MLP classifier. Nevertheless, neural networks typically perform better on very large datasets. On smaller application domains and data set, SVClassifier will proved to be perform better. Thus, MLP classifier is more suitable than the SVClassifier to apply as classifier for the datasets.

# Conclusion

In conclusion, from this assignment, we could conclude that MLPClassifier in neural networks was the most accurate model to fit and predict the red wine quality dataset as compared to Random Forest Classifier and Support Vector Classifier. The accuracy of training dataset obtained was around 81% while the testing accuracy obtained was around 78%. The precision, recall and f1 scores that obtained for training dataset were higher than 75% while testing dataset obtained around 80%. The AUC score generated from cross validation was 75%. The model didn’t not overfit the training data which delivered high accuracy of testing data. In this case, this model can be used in future industry so that higher quality of red wines can be produced. Future improvement can be made on the range of physiochemical used to produce the red wines so that the possibility of producing lower quality red wines can be decreased. The industry can also use this model to check the quality of red wines during production for quality check of red wines.