November 8, 2021

CS306 : Project- Work Report , Sem: V, 2021

Red Wine Quality Prediction Using Machine Learning Techniques

Aman Kushwaha
Department of Computer Science and Engineering
Indian Institute of Information Technology, Guwahati
Guwahati, Assam
amankushwaha008@gmail.com

Abstract

These days the consumption of red wine is very common to all. So it became important to analyze the quality of red wine before its consumption to preserve human health. Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Hence this project is a step towards the quality prediction of the red wine using its various attributes. Dataset is taken from the sources and the techniques Logistic Regression, Stochastic Gradient Descent Classification, Perceptron Model and artificial neural network (ANN) Models are applied. Various measures are calculated and the results are compared among training set and testing set and accordingly the best out of these techniques depending on the training set results is predicted. We used the performance measurement matrices such as accuracy recall, precision, and f1 score for comparison of the machine learning algorithm.

Introduction

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increases their value in the market. Previously, testing of product quality used to be done at the end of the production, which is time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion about the test, so identifying the quality of the wine based on humans experts it is a challenging task. The aim of this project is to predict the quality of wine on a scale of 0–10 given a set of features as inputs. The dataset used is Wine Quality Data set from UCI Machine Learning Repository. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol. And the output variable is quality (score between 0 and 10). The value of pH depicts the acidity and basicity of the wine. Consumable wines have their pH scale between 3-4..We are dealing only with red wine.We have quality being one of these values: [3, 4, 5, 6, 7, 8]. The higher the value the better the quality.

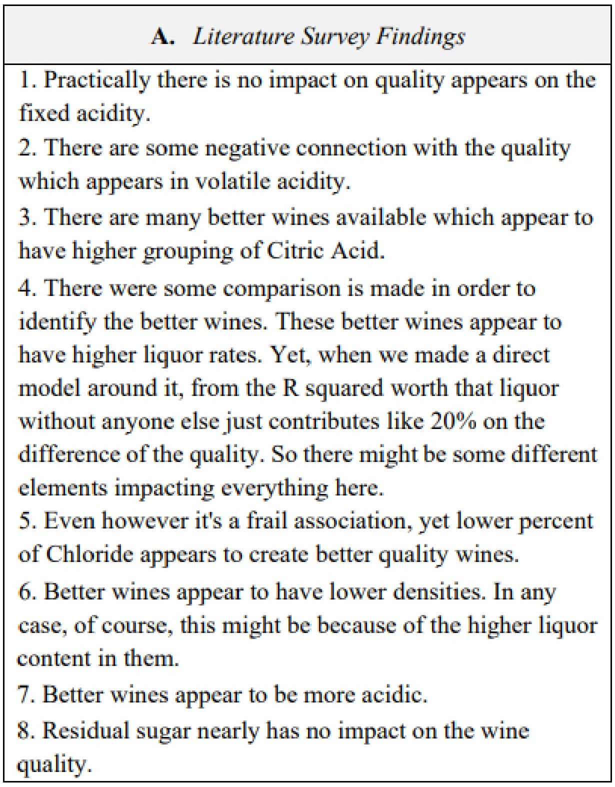
LITERATURE SURVEY

Well, for research purposes, this could be described as the process of extracting secret information from the loads of databases. Information Discovery in Databases (KDD) is also known as Data Mining. Data Mining is commonly used in a variety of applications, like e-commerce, stock, product analysis, including understanding customer research marketing and real estate investment pattern etc.

Data Mining is based on the mathematical algorithm required to drive the preferred results from the enormous collection of databases. Business Intelligence (BI) can be used for the analysis of pricing, market research, economic indicators, behavior use, industry research, geographic information analysis, and so on. Data mining technologies are commonly used in the fields of Customer Relationship Management, direct marketing, healthcare, e-commerce, telecommunications, and finance. This could also be likely you need to contact outsourcing companies for help. Such outsourcing firms are experienced in processing or scraping the data, filtering it out, and then keeping it for examination. Usually, data mining involves collecting information and analyzing the data and to search for more details etc.

Data mining helps to forecasts accurate and reliable historical data that we have at our fingertips and guessing about future outcomes. Businesses may use data mining to assess why it is necessary to find the information they need to use business intelligence to analytics. They considered the gut micro biome of red wine drinkers to be more diverse than the non-red wine drinkers. He was not found with the consumption of white wine, beer, or spirits.

Tim et. al discussed the effects of red wine on the guts of nearly three thousand people in three different countries , they found that polyphenols in grape skin could be responsible for much of the controversial health benefits when used in moderation." The study also discovered that lower levels of obesity and 'bad' cholesterol were associated with red wine consumption, partly due to gut micro biota. "Even though we have established a correlation between the consumption of red wine and the diversity of gut micro biota, drinking red wine rarely, like once every two weeks, seems sufficient to detect an impact. Though, alcohol consumption with moderation is still advisable, "Dr. Le Roy added.



A. *Literature Survey Findings*

1. Practically there is no impact on quality appears on the fixed acidity.

2. There are some negative connection with the quality which appears in volatile acidity.

3. There are many better wines available which appear to have higher grouping of Citric Acid.

4. There were some comparison is made in order to identify the better wines. These better wines appear to have higher liquor rates. Yet, when we made a direct model around it, from the R squared worth that liquor without anyone else just contributes like 20% on the difference of the quality. So there might be some different elements impacting everything here.

5. Even however it's a frail association, yet lower percent of Chloride appears to create better quality wines.

6. Better wines appear to have lower densities. In any case, of course, this might be because of the higher liquor content in them.

7. Better wines appear to be more acidic.

8. Residual sugar nearly has no impact on the wine quality.

Dataset Description

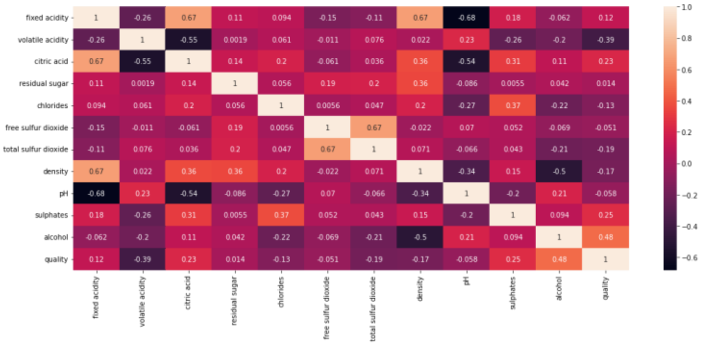
Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. For more details, consult: [Web Link] . Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). From the given dataset we are using only red wine dataset to predict the quality of red wine.
These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). So to get rid of class imbalance we classified the quality [0,1,2,3,4,5,6] as bad quality and [7,8,9,10] as good wine quality. For this dataset qualities are predicted between the range 3-

8, where ‘3’ predicts poor quality of red wine and ‘8’ predicts excellent quality of red wine. The features includes

1)fixed acidity
2) volatile acidity
3) citric acid
4) residual sugar
5) chlorides
6)free sulfur dioxide
7)total sulfur dioxide
8)density
9)pH
10) sulphates
11) alcohol

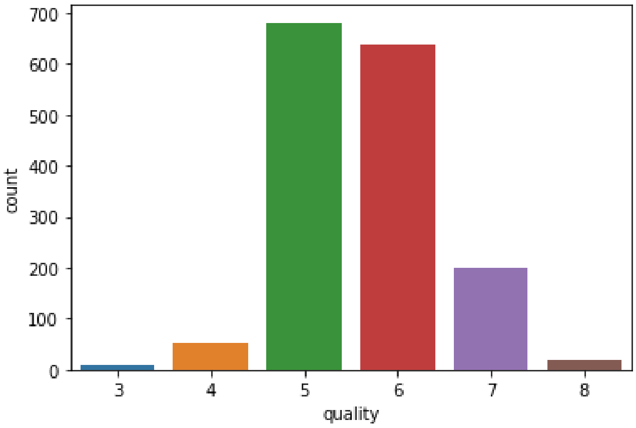
Output variable (based on sensory data)
12)quality (score between 0 and 10)

To predict the quality of the wine it is very important to analyse the data and check for the features affecting the quality of the wine. In order to understand the effect of every feature on the quality of the wine correlation matrix is a good option. It shows the relation between each of the features and output too. The positive value in the table shows the positive correlation between the features whereas negative value shows the negative correlation.



Data Preprocesing

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. Label Encoding is used to convert the labels into numeric form so as to convert it into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning. We have used label encoding to label the quality of data as good or bad. Assigning 1 to good and 0 to bad.. Wine Dataset was visualizes and checked for any missing data, the dataset does not contain any missing data. Then Class balancing is checked and found that there is class imbalance which can be seen by the below figure.



```
5    681
6    638
7    199
4     53
8     18
3     10
```

The data for each class can be seen here, and this can be found that class 8 and class 3 has very low data so we need to balance it because with this data we are getting low accuracy which was around 57%. To get rid of this problem Label Encoding is used in which the classes 3, 4, 5 and 6 were considered as bad quality of the wine and class 7 and 8 are considered as good quality wine.

EXPERIMENT DESIGN

In the field of machine learning, a confusion matrix is a table that is frequently used to depict the presentation of a grouping model on a lot of test information for which the genuine qualities are known. It permits the perception of the presentation of a calculation. This research basically uses the red wine data set and then calculates the confusion matrix, relevant performance measures and finally compares the different machine learning algorithms on the basis of accuracy predicted on this dataset

A. Measures to Calculate Performance in Research

Performance measures are the measures that are used in the research so as calculate and evaluate the techniques to detect the effectiveness and efficiency of the techniques. Some of them are listed below:

- Accuracy: The value predicted when the sum of True Positive and True Negative is divided by the sum of True Positive, False positive, False Negative and True Negative values of a confusion matrix.

Accuracy=(True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)

- Precision: The value obtained when True Positive is divided by the sum of True Positive and False Positive values of a confusion matrix

Precision = True Positive / (True Positive + False Positive)

- Recall: Sensitivity sometimes also known as Recall. It is the value obtained when True Positive is divided by the sum of True Positive and False Negative values of a confusion matrix.

Recall= True Positive / (True Positive + False Negative)

- F-Measure: F1 Score is obtained by multiplying Recall and Precision divided by sum of Recall and precision of a confusion matrix. Result is then multiplied by two.

*F1 Score = 2 * (Recall * Precision) / (Recall + Precision)*

B. Techniques Involved

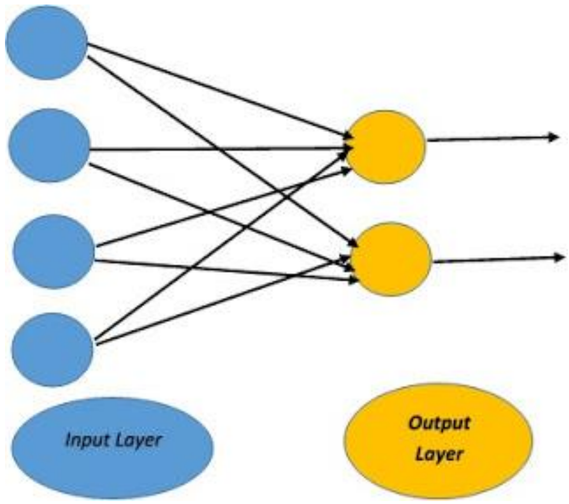
Techniques used are as given below. These are:

- Logistic regression: It is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, Logistic regression can be divided into following types –
    - Binary or Binomial
    - Multinomial
    - Ordinal

- Stochastic Gradient Descent (SGD) : SGD Classifier is an optimization algorithm used to find the values of parameters of a function that minimizes a cost function. The algorithm is very much similar to the traditional Gradient Descent. However, it only calculates the derivative of the loss of a single random data point rather than all of the data points (hence the name, stochastic). This makes the algorithm much faster than Gradient Descent.

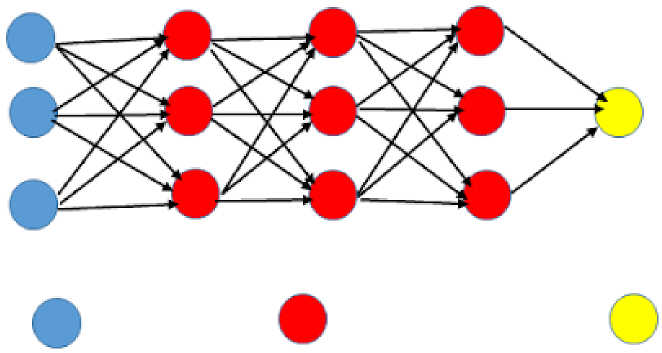
- Perceptron Algorithm: The Perceptron is a linear machine learning algorithm for binary classification tasks. It may be considered one of the first and one of the simplest types of artificial neural networks. Like logistic regression, it can quickly learn a linear separation in feature space for two-class classification tasks, although unlike logistic regression, it learns using the stochastic gradient descent optimization algorithm and does not predict calibrated probabilities.

- *Single Layer Perceptron :* A single layer perceptron (SLP) is a feed-forward network based on a threshold transfer function. SLP is the simplest type of artificial neural networks and can only classify linearly separable cases with a binary target (1 , 0).



Single Layer Perceptron has just two layers of input and output. It only has single layer hence the name single layer perceptron. It does not contain Hidden Layers as that of Multilayer perceptron. Input nodes are connected fully to a node or multiple nodes in the next layer. A node in the next layer takes a weighted sum of all its inputs.

- Multi-Layer Perceptron (MLP): A multilayer perceptron is a type of feed-forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. The MLP network consists of input, output, and hidden layers. Each hidden layer consists of numerous perceptron’s which are called hidden layers or hidden unit.



🔵 : Input

🔴 : Hidden Layers

🟡 : Output Layer

Related Work

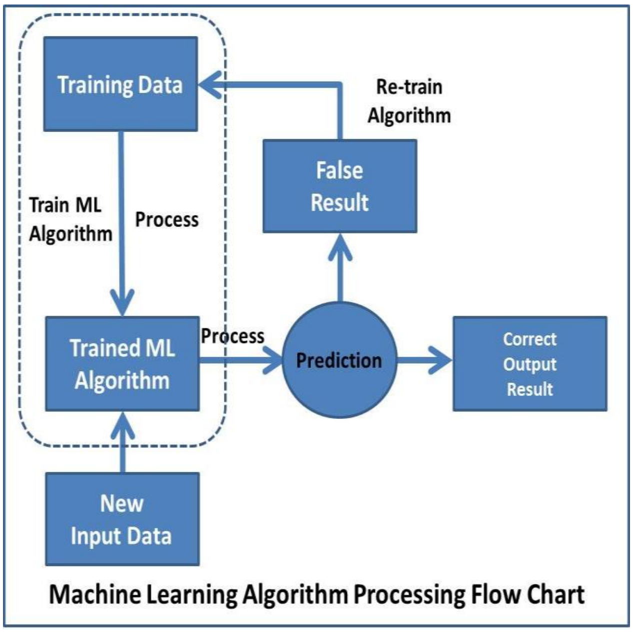
Kumar et al. (2020) have used prediction of red wine quality using its various attributes and for the prediction, they used random forest, support vector machine, and naive Bayes techniques (Kumar et al., 2020). They have calculated the performance measurement such as precision, recall, f1-score, accuracy, specificity, and misclassification error. Among these three techniques, they achieved the best result from the support vector machine as compare to the random forest and naive Bayes techniques. They achieved the accuracy of the support vector machine technique is 67.25%.

Dahal et al., (2021) has predicted the wine quality based on the various parameters by applying various machine learning models such as rigid regression, support vector machine, gradient boosting regressor, and multi-layer artificial neural network. They compare the performance of the models to predict wine quality and from their analysis, they found gradient boosting regressor is the best model to other model performances with the MSE, R, and MAPE of 0.3741, 0.6057, and 0.0873 respectively(Dahal et al., 2021).

IMPLEMENTATION

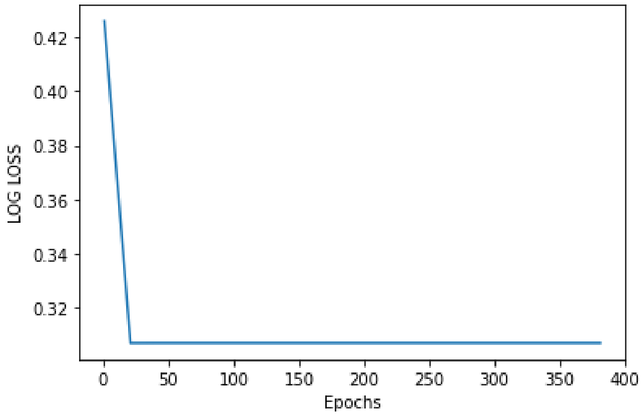
An analysis is done on the redwine.csv dataset extracted from huge database [11] that contains the details of Red Wine. The datasets contain 1599 observation and have 12 attributes such as fixed acidity, sugar, sulphates, chlorides, volatile acidity, citrus acid, residual, free sulphur dioxide, absolute sulphur dioxide, thickness, pH and alcohol. All these attributes are used from the dataset. The Dataset was divided into train set , validation set and test set in the ratio 0.6, 0.2 and 0.2 respectively. Validation set is used to tune the best hyperparameters and to test the overfitting of the model. Libraries such as Logistic Regression, KFold , SGDClassifier, Perceptron, MLPClassifier are imported. After calculating the summeries confusion matrix is created and accuracy of the model is calculated. The flow of the algorithm is shown in the figure. Further various performance measures such as precision, f-score, recall and accuracy are calculated using these algorithms. Results were predicted on the basis of these measures. Hyperparameter tunning is also done using the inbuild function so that we can get best result using the best hyperparameters.

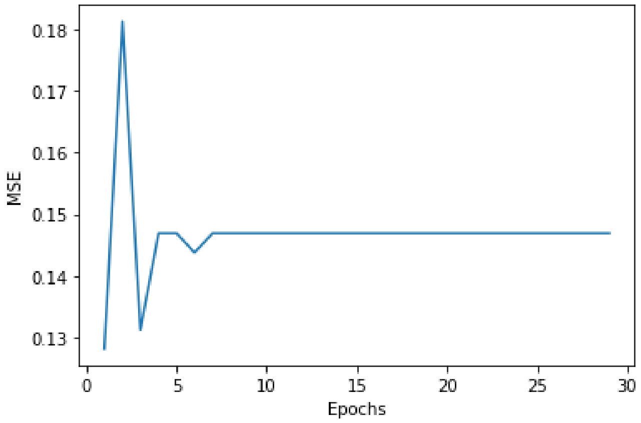
After Encoding the data set Data Normalization is done to using MinMaxScaler algorithm. After normalization the dataset, it is divided into training , validation and test data set. Validation set is used to validate the set and check for overfitting the dataset. After which Hyperparameter tunning is done using the validation set and testing on test set is done.



Machine Learning Algorithm Processing Flow Chart

For Logistic regression, mse vs epoch and logloss vs epoch graph are ploted.

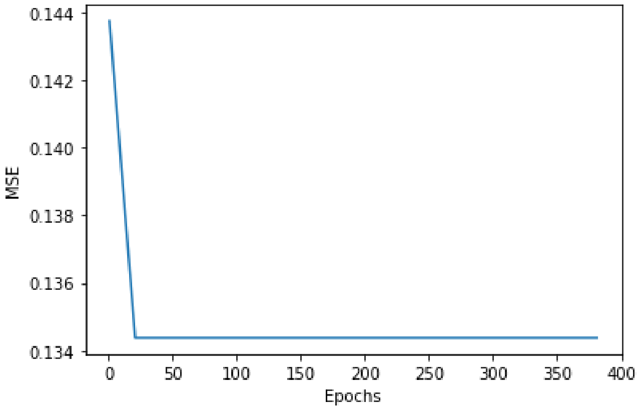




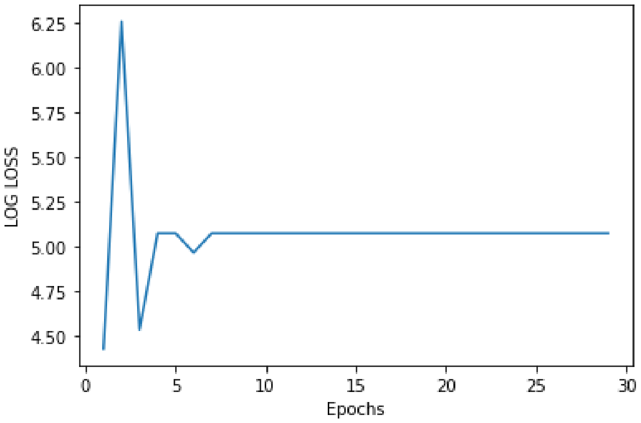
Best Hyperparameters for Logistic Regression were found as C=10, max_itr=100, solver='lbfgs', penalty='l2'.

For Stochastic Gradient Descent Classifier Hyperparameter tunning was done and found best hyperparameters for it.

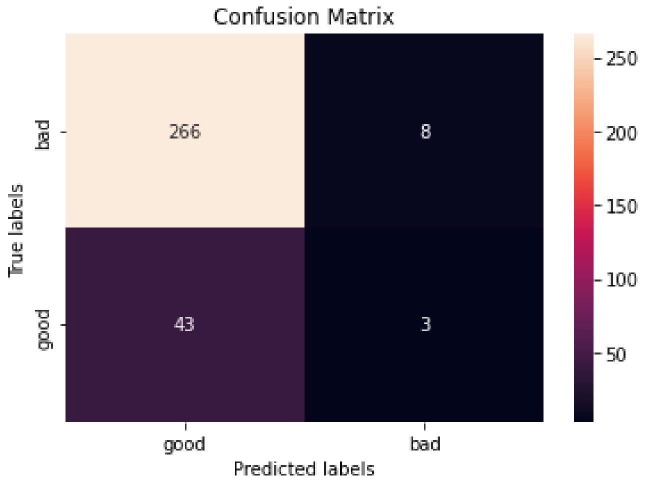
KFold crossvalidation for fold=5 was used on Logistic Regression and fold wise and overall accuracy for training and test set is found.



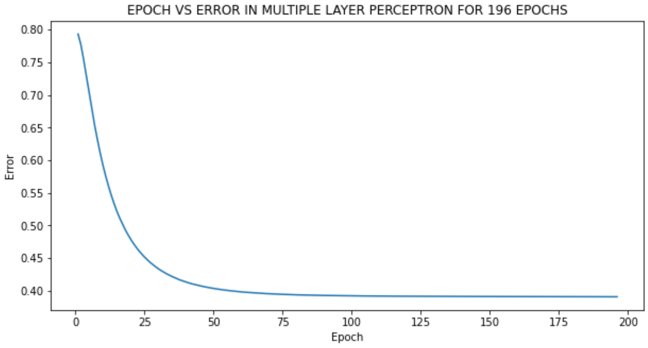
Perceptron Model is converging for the epoch 7 for which MSE vs epoch and LOGLOSS vs epoch graph are printed and are as follows:



Single Layer Perceptron model is implemented and trained using the train data set which is tested using test data set. The confusion matrix for the output data set is shown below:



Multi Layer Perceptron model is implemented using inbuilt library function. For which while training the dataset error vs epoch graph was calculated as is as follows:



Result and Discussion

Nowadays people used to consume red wine either as a necessity or for show off. All this results in health loss. Hence to preserve human health it became essential to predict the red wine quality before its consumption. So, in this project the dataset taken contains the information related to red wine extracted from the database which is used to predict the wine quality. Different machine learning algorithms are executed on the dataset.

Accuracy is calculated and the best algorithm is predicted for a given dataset. During the usage, the data is separated into testing set , validation set and training set each with probability of 0.2, 0.2 and 0.8 respectively. The result shows that, accuracy obtained for training set and testing set using Logistic Regression 88.11% and 86.56% respectively, using Logistic Regression after hyperparameter tunning are 88.73% and 88.75% respectively, while using Stochastic Gradient Descent Classifier without hyperparameter tunning it were 88.42% and 88.12% respectively which after hyperparameter tunning became 87.79% and 88.125%, for KFold cross validation the overall accuracy are 86.42% and 86.42% respectively, while using Perceptron model the accuracy are 86.96% and 85.62% respectively For Single Layer Perceptron Model the scores are 86.23% and 84.06% respectively, for our last model that is Multi Layer Perceptron model the accuracy are

86.96% and 85.62% respectively. As there is high probability of division for training set hence taking the accuracy of training sets for examination shows that Logistic Regression after hyperparameter tunning gives the more accuracy.

Above results for Logistic Regression and SGD Classifier shows that Hyperparameter tunning can improve the algorithm and results can be more accurate. Hence Hyperparameter tunning for better result prediction is a good thing to be done.

CONCLUSION

Some features affects the quality of the wine which can be seen in the correlation table. So by just studying that features wine quality can be improved. Using all these algorithms it can be said that Logistic Regression Algorithm give the best result after hyperparameter tunning.

References

1. https://ijcat.com/archieve/volume8/issue9/ijcatr08091010.pdf
2. https://www.diva-portal.org/smash/get/diva2:1574730/FULLTEXT01.pdf
3. https://sci-hub.hkvisa.net/10.1109/iccci48352.2020.9104095
4. https://www.ijrte.org/wp-content/uploads/papers/v10i1/A58540510121.pdf