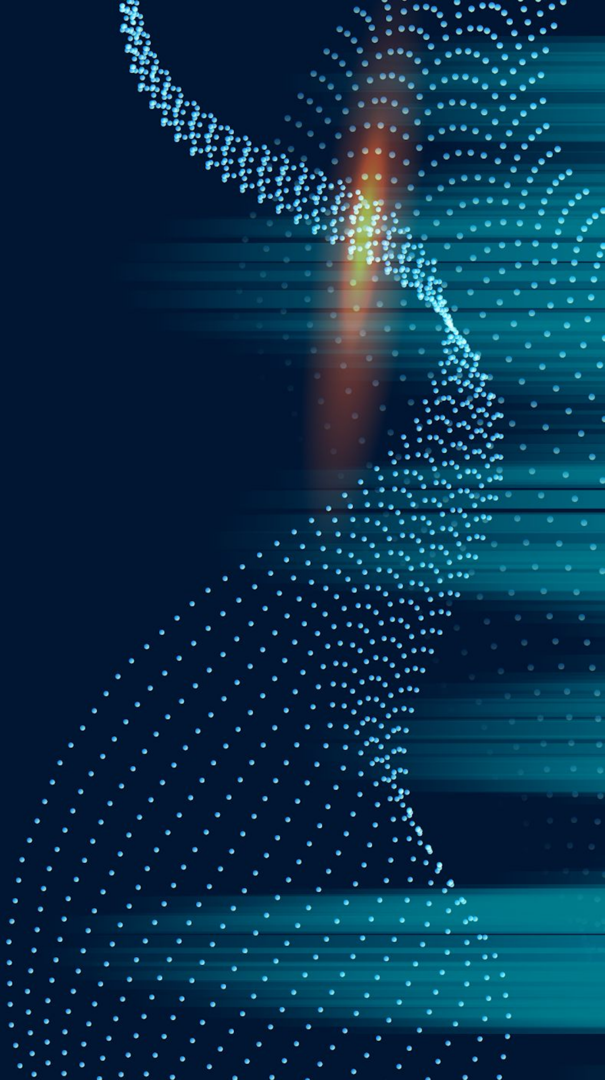
Prediction of wine quality using ML algorithms

ML Mini Project

By - Vaibhav Kesarwani (1901212)
vaibhav.kesarwani@iiitg.ac.in

CONTENT

- Problem Definition
- Dataset Description
- Literature Survey
- Results and Discussion
- Accuracy Comparison
- Conclusion
- References



PROBLEM DEFINITION

In recent years there is a modest increase in the wine consumption as it has been found that wine consumption has a positive correlation to the heart rate variability. With the increase in the consumption wine industries are looking for alternatives to produce good quality wine at less cost.

In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analyses as it takes lot of time and also need more money.

These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analyses parameters in the wine and which one to ignore for reduction of cost.

DATASET DESCRIPTION

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.

Link for Dataset - (https://archive.ics.uci.edu/ml/datasets/wine+quality)

These datasets can be viewed as classification or regression tasks.

Features:

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

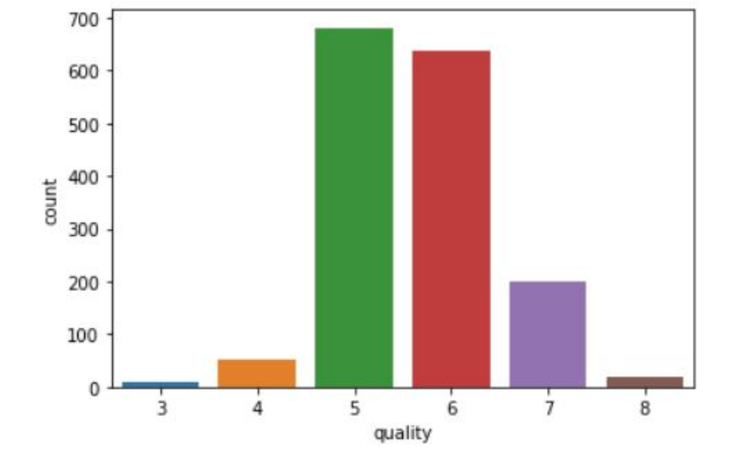
Output variable (based on sensory data):

quality (score between 0 and 10)

DATASET DESCRIPTION

- **Number of Instances:**
  - red wine - 1599
  - white wine - 4898
- **Number of Attributes:** 11 + output attribute
- **Number of Features:** 11
- **Pattern:** Wine
- **Missing Attribute Values:** None
- **Attribute Characteristics:** Real

DATASET DESCRIPTION



Class Imbalance Problem:

As we can see, wine with quality 5 and 6 are more in number and wine of quality 1, 2, 9, 10 are no in the dataset. So there are no training samples for the best and worst wines.

I have handled this by doing binary classification, i.e. assigned wine quality >=7 in positive class (good wine) and wine quality <=6 in negative class (bad wine). This will reduce the imbalance problem.

There are many other methods to solve Class Imbalance Problem.

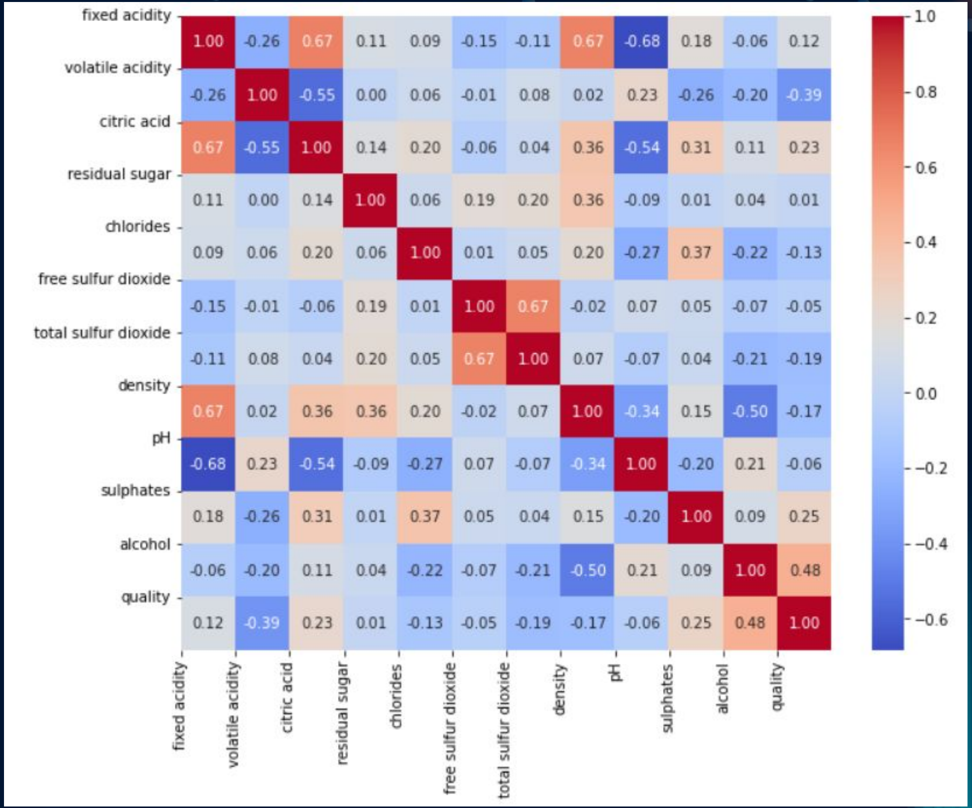
DATASET DESCRIPTION

Correlation Matrix:

Determines the correlation between every feature.

Positive value means inc in 1 feature will lead to inc in other feature.(Direct relation)

Negative value means inc in 1 feature will lead to dec in other feature.(Inverse relation)



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1.00

LITERATURE SURVEY

1) Er and Atasoy (2016) proposed a method to classify the quality of wines using three different classifier such as support vector machines, Random forest and k-nearest neighbourhood. They have used principal component analysis for feature selection and they found good result using Random forest algorithm.

2) Paulo Cortez ,Juliana Teixeira,António CerdeiraFernando AlmeidaTelmo MatosJosé Reis wrote a paper on wine Quality assessment using Data Mining techniques.In this paper,they proposed a data mining approach to predict wine preferences that is based on easily available analytical tests at the certification step. A large dataset was considered with white vinho verde samples from the Minho region of Portugal. Wine quality is modeled under a regression approach, which preserves the order of the grades. 95% accuracy was obtained using these data mining techniques.

3) Chen et al proposed an approach that will predict the grade of wine using the human savory reviews. They have used hierarchical clustering approach and association rule algorithm to process the reviews and predict the wine grade and they found an accuracy of 85.25% while predicting the grade.

4) Thakkar et al (2016) used analytical hierarchy process (ahp) to rank the attributes and then used different machine learning classifiers such as support vector machine and random forest and they found accuracy of 70.33% using random forest and 66.54% using SVM.

5) Reddy and Govndarajulu (2017) used a user centric clustering approach to recommend the product. They have used red wine data set for the survey purpose. They have allocated relative voting to the attributes based on the literature review. Then they assigned weight to the attributes using Gaussian Distribution Process. They judged the quality based on the user preference group

6) 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 .

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.

7) A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality
Mohit Gupta, Vanmathi C. (May 2021)

(https://www.ijrte.org/wp-content/uploads/papers/v10i1/A58540510121.pdf)

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.

Results and Discussion

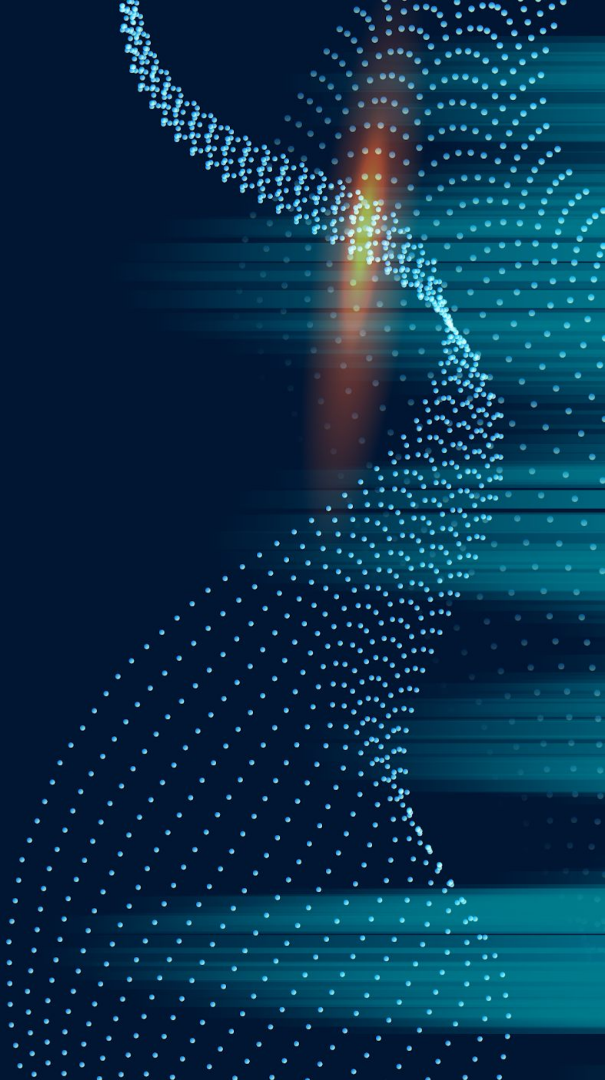
For Red Wine Dataset

RESULTS AND DISCUSSION

Divided the data into three groups such as train data, validation data and test data.

Then trained each classifier based on the trained data and predict the power of classifier on the test data. So, each classifier able to show all the performance metrics such as accuracy, recall and precision. Also build few graphs so that it will be easy to understand it in a better way. Observed the data types present in the dataset and also analysed the data for null values.

Also done the K-Fold Cross Validation (k=5) on different models to check the robustness of the model and overfitting issue. Calculated Accuracy, Precision and Recall for every folds.



RESULTS AND DISCUSSION

Algorithms used for classification are :

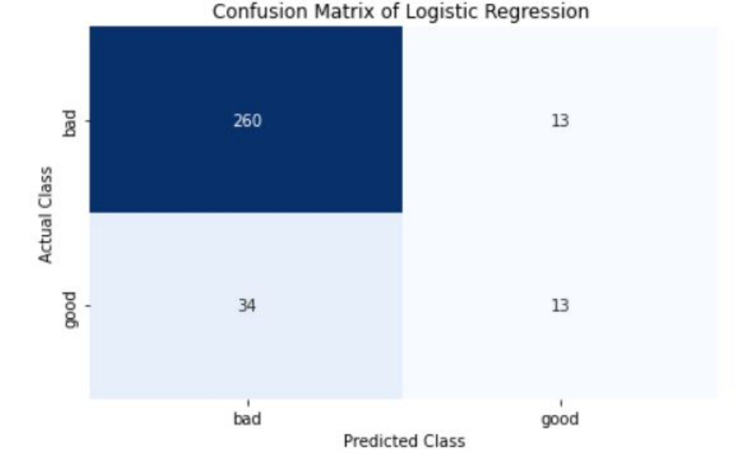
- Logistic Regression
- Single Perceptron Model
- MultiLayer Perceptron

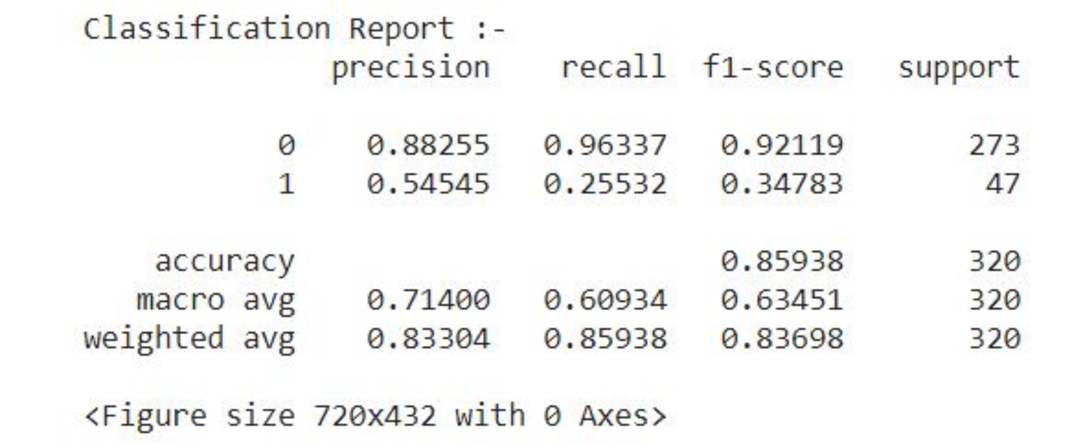
Applied and Compared the results of the above three models .

Performed hyper-parameter tuning using inbuilt method (GridSearchCV()) to get the best set of hyperparameters among the given param_grid (The parameter grid to explore, as a dictionary mapping estimator parameters to sequences of allowed values.)

RESULTS AND DISCUSSION

Logistic Regression : It gave us an accuracy of 85.938

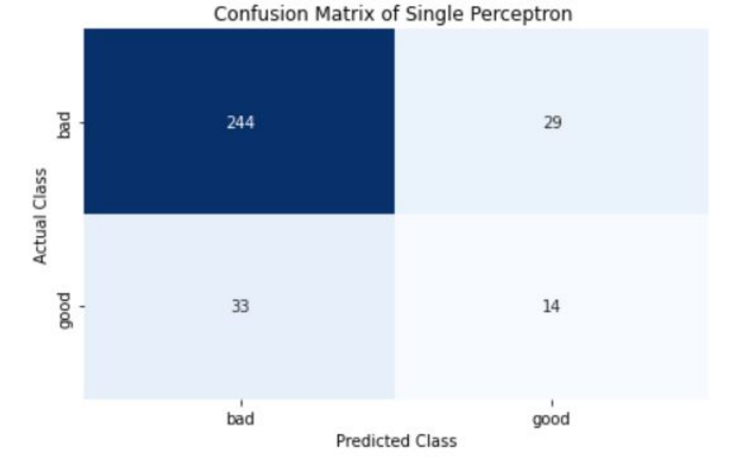




```
Classification Report :-
              precision    recall  f1-score   support

           0    0.88255   0.96337   0.92119       273
           1    0.54545   0.25532   0.34783        47

    accuracy                        0.85938       320
   macro avg    0.71400   0.60934   0.63451       320
weighted avg    0.83304   0.85938   0.83698       320

<Figure size 720x432 with 0 Axes>
```

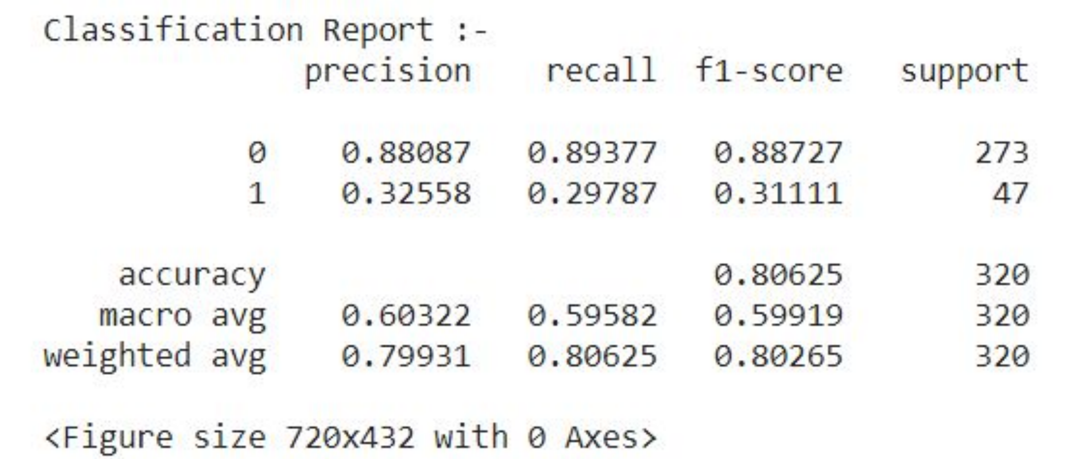
RESULTS AND DISCUSSION

Single Perceptron : It gave us an accuracy of 80.625



Confusion Matrix of Single Perceptron

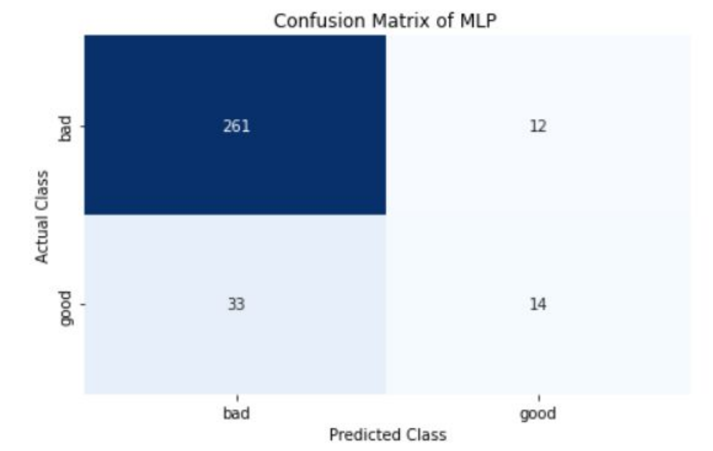
	bad	good
bad	244	29
good	33	14

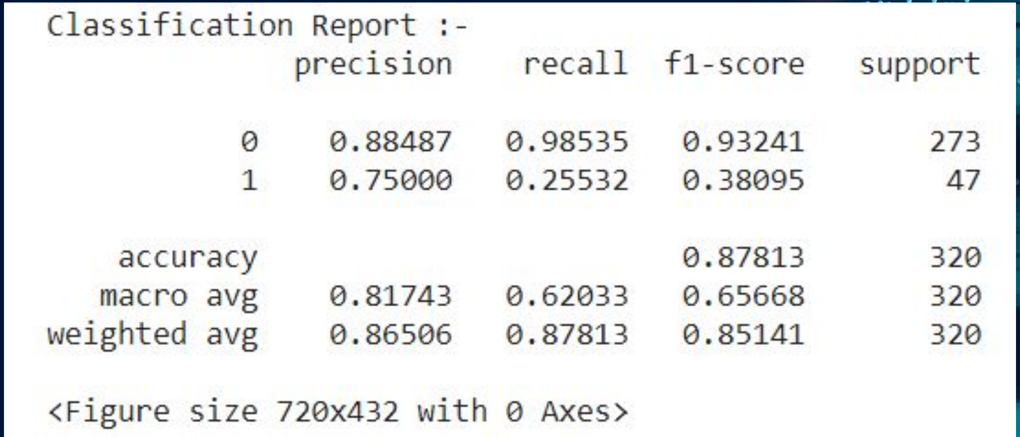


```
Classification Report :-
              precision    recall  f1-score   support

           0    0.88087   0.89377   0.88727       273
           1    0.32558   0.29787   0.31111        47

    accuracy                        0.80625       320
   macro avg    0.60322   0.59582   0.59919       320
weighted avg    0.79931   0.80625   0.80265       320

<Figure size 720x432 with 0 Axes>
```

RESULTS AND DISCUSSION

MultiLayer Perceptron : It gave us an accuracy of 87.813





```
Classification Report :-
              precision    recall  f1-score   support

           0    0.88487   0.98535   0.93241       273
           1    0.75000   0.25532   0.38095        47

    accuracy                        0.87813       320
   macro avg    0.81743   0.62033   0.65668       320
weighted avg    0.86506   0.87813   0.85141       320

<Figure size 720x432 with 0 Axes>
```

THESE ARE THE ACCURACY COMPARISONS!

MLP has the best accuracy among all 3 models.

85.938%

Logistic Regression

87.813%

Multi Layer Perceptron (MLP)

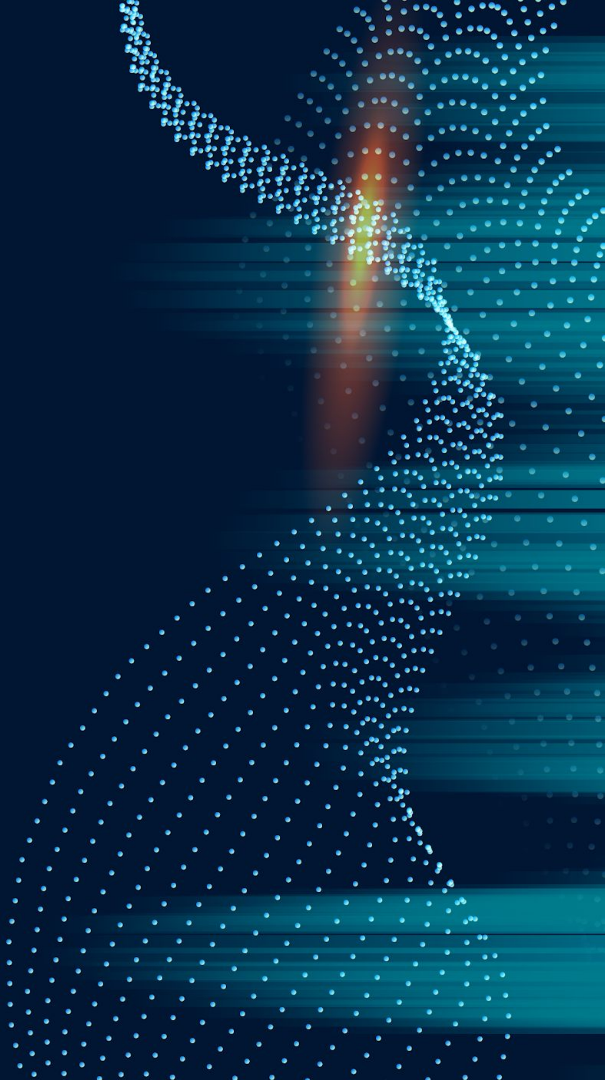
80.625%

Single Perceptron

CONCLUSION

This Mini Project is about predicting the quality of Red Wine using various machine learning techniques. We have compared the accuracy of each technique used in prediction of quality and it was found that these classifiers performed well. We have also found that the MLP classifier performed better compared to all other classifiers for red wine data set.

In future we can try other performance measures and other machine learning techniques for better comparison on results, to improve class-wise accuracy and select a robust model to predict the quality of red wine. This analysis will help the industries to predict the quality of the different type of wines based on certain attributes and also it will helpful for them to make good product in the future.



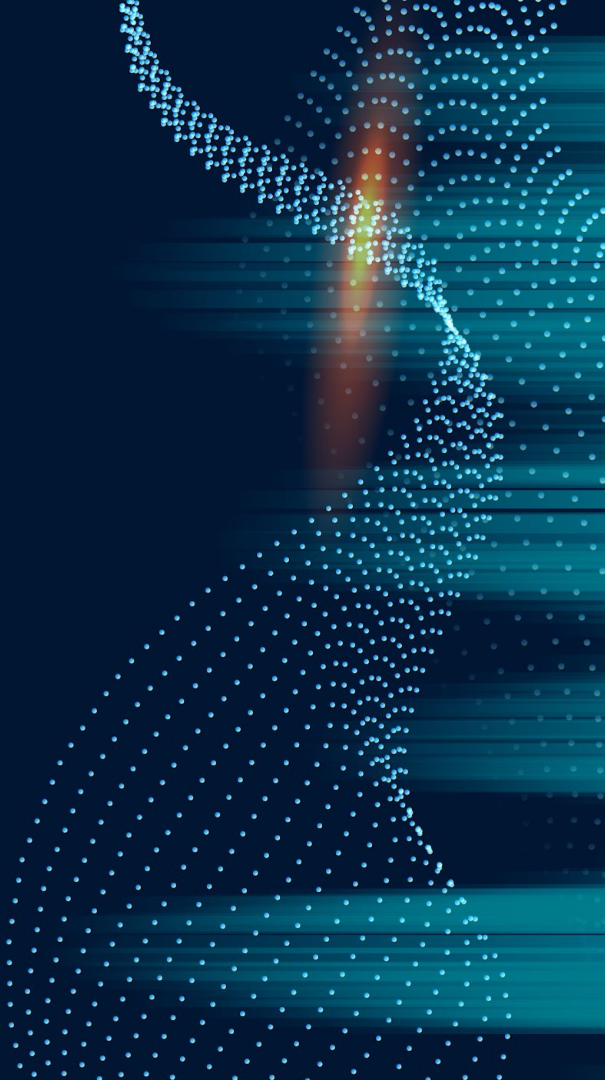
CONCLUSION

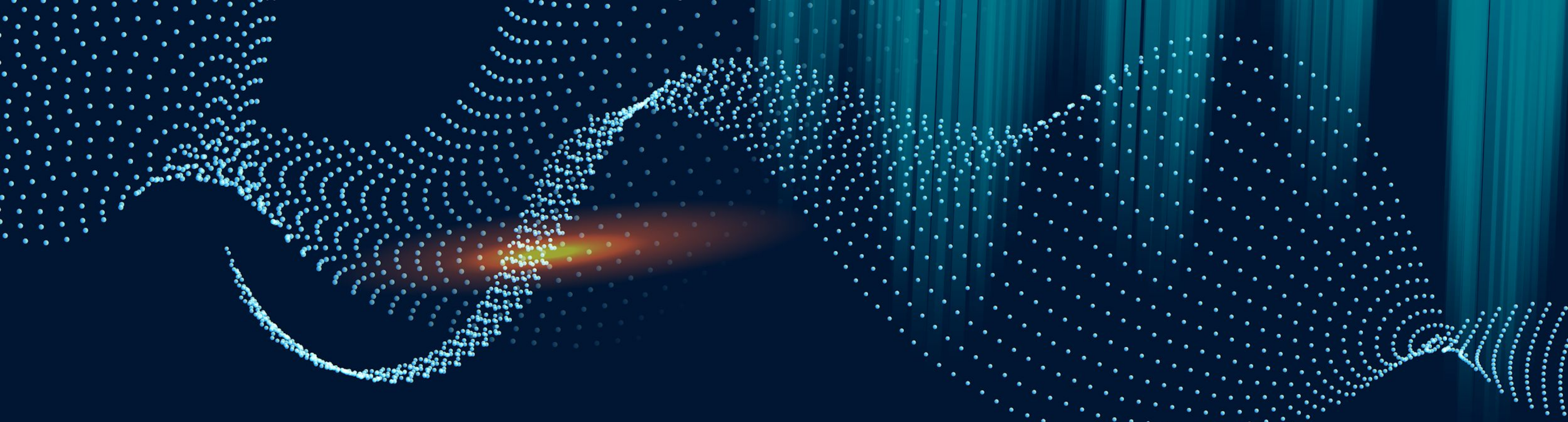
Based on the bar plots plotted I come to an conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not have change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection.

Also, like this we can drop some more features, which do not affect the quality, to get the accurate quality of wine using different models.

REFERENCES

- https://www.diva-portal.org/smash/get/diva2:1574730/FULLTEXT01.pdf
- https://www.ijrte.org/wp-content/uploads/papers/v10i1/A58540510121.pdf
- http://ijcsit.com/docs/Volume%207/vol7issue5/ijcsit20160705044.pdf
- https://ijcat.com/archieve/volume8/issue9/ijcatr08091010.pdf
- https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine/data





THANK YOU!

To my instructor (Dr. Moumita Roy)
for giving us this ML project.
I have learned many new things
while doing this Project.