WINE DATASET

WINE QUALITY PREDICTION

By Navya & Vaishnavi

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Dataset Description

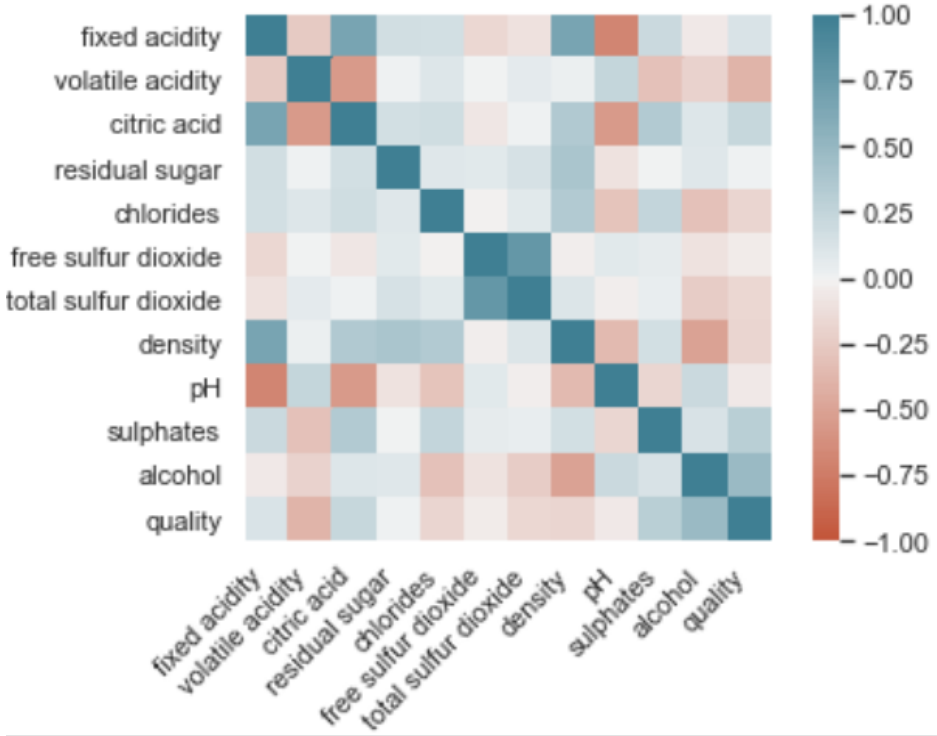
- Number of independent variables are 11
- Target variable is “quality” with six unique integer values
- The highly skewed features are ‘chlorides’, ‘residual sugar’, ‘sulphates’, ‘total sulfur dioxide’, and ‘free sulfur dioxide’. Hence log transformation is applied for all these features.
- Fixed acidity and density are the highest positively correlated features.
- Citric acid and volatile acidity are the highest negatively correlated features.
- There is a class imbalance in the dataset, and it was categorized by grouping the quality 3,4,5 into one group and 6,7,8 into another group.

Main Statistics

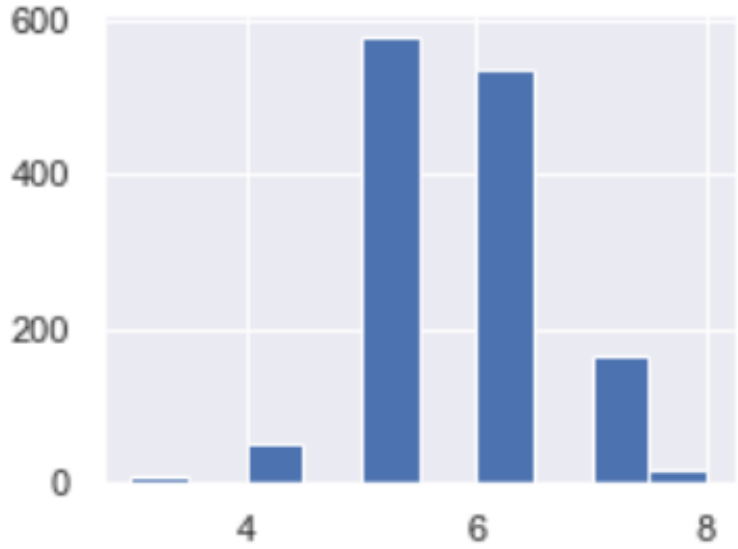
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.14	-0.10	0.67	-0.69	0.19	-0.06	0.12
volatile acidity	-0.26	1.00	-0.55	-0.00	0.06	-0.02	0.07	0.02	0.25	-0.26	-0.20	-0.40
citric acid	0.67	-0.55	1.00	0.14	0.21	-0.05	0.05	0.36	-0.55	0.33	0.11	0.23
residual sugar	0.11	-0.00	0.14	1.00	0.03	0.16	0.20	0.32	-0.08	-0.01	0.06	0.01
chlorides	0.09	0.06	0.21	0.03	1.00	0.00	0.05	0.19	-0.27	0.39	-0.22	-0.13
free sulfur dioxide	-0.14	-0.02	-0.05	0.16	0.00	1.00	0.67	-0.02	0.06	0.05	-0.08	-0.05
total sulfur dioxide	-0.10	0.07	0.05	0.20	0.05	0.67	1.00	0.08	-0.08	0.04	-0.22	-0.18
density	0.67	0.02	0.36	0.32	0.19	-0.02	0.08	1.00	-0.36	0.15	-0.50	-0.18
pH	-0.69	0.25	-0.55	-0.08	-0.27	0.06	-0.08	-0.36	1.00	-0.21	0.21	-0.06
sulphates	0.19	-0.26	0.33	-0.01	0.39	0.05	0.04	0.15	-0.21	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.06	-0.22	-0.08	-0.22	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.40	0.23	0.01	-0.13	-0.05	-0.18	-0.18	-0.06	0.25	0.48	1.00

Dataset Visualization

Correlation Plot:



Class Imbalance:



ML Models

Logistic Regression:

It is a statistical method used to predict binary classes and to explain the relationship between the independent and dependent variables.

Hyper parameters used in the model implementation are C of range 0.01 to 1 and lbfgs optimizer.

KNN Algorithm:

It is a supervised machine learning algorithm used to solve both classification and regression tasks. It determines the class of the data point by majority voting principle.

Hyperparameter used in this model implementation is k of range 1 to 40.

ML Models

Support Vector Machine

It is a supervised Learning technique, which tries to find a hyper plane that maximizes the margin between the two classes in training data.

For wine data set, we used Linear kernel with penalty term C in the range 0.1 to 1000.

Random Forest:

It is an ensemble method, which makes predictions based on multiple decision trees by using bootstrapping technique.

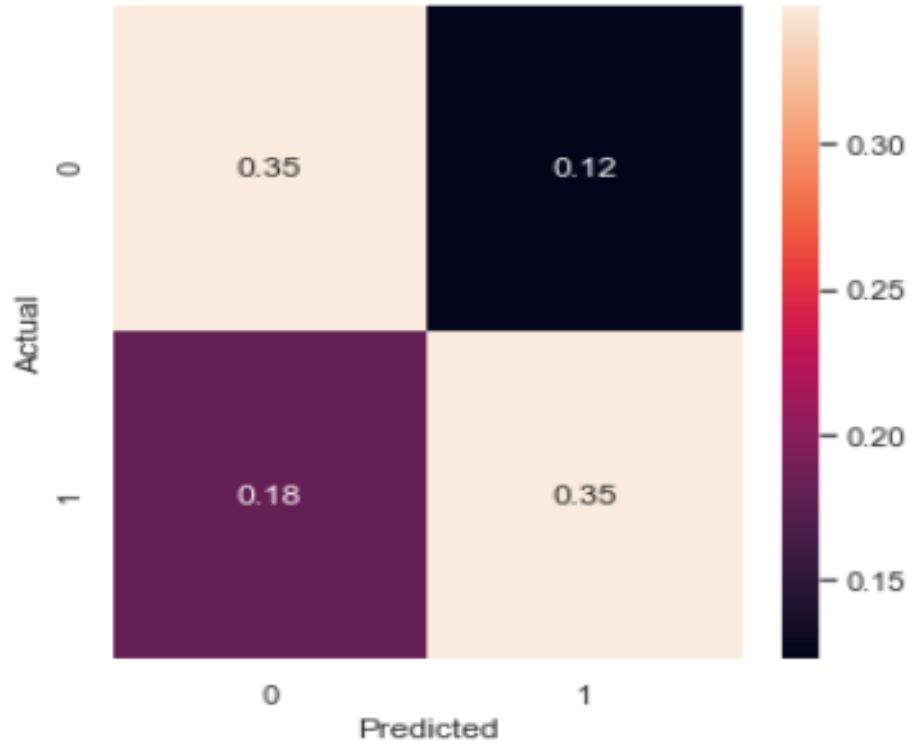
We used grid search method to find the best parameters (tree depth, split criteria etc.)

Results of Logistic Regression

1. **Accuracy on validation data - 69.6%**
2. **Confusion Matrix:**

Predicted	0	1
Actual		
0	71	25
1	37	71

3. **Precision is 73.95%**

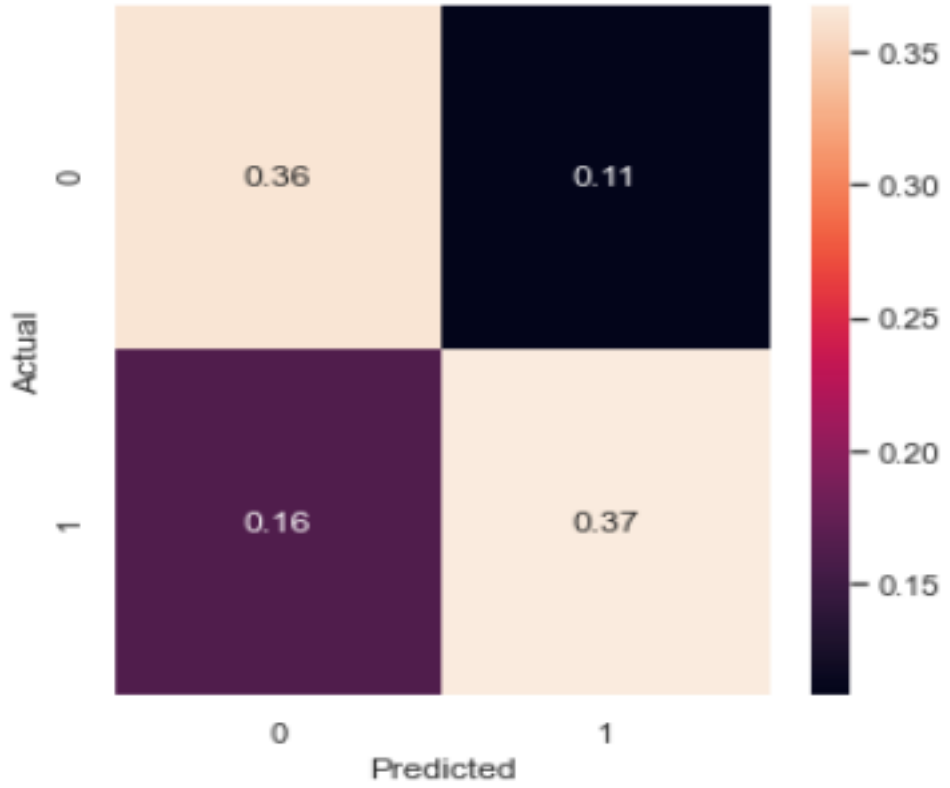


Results of KNN Algorithm

1. **Accuracy on validation data - 73.03%**
2. **Confusion Matrix:**

Predicted	0	1
Actual		
0	74	22
1	33	75

3. **Precision on validation data - 77.31%**

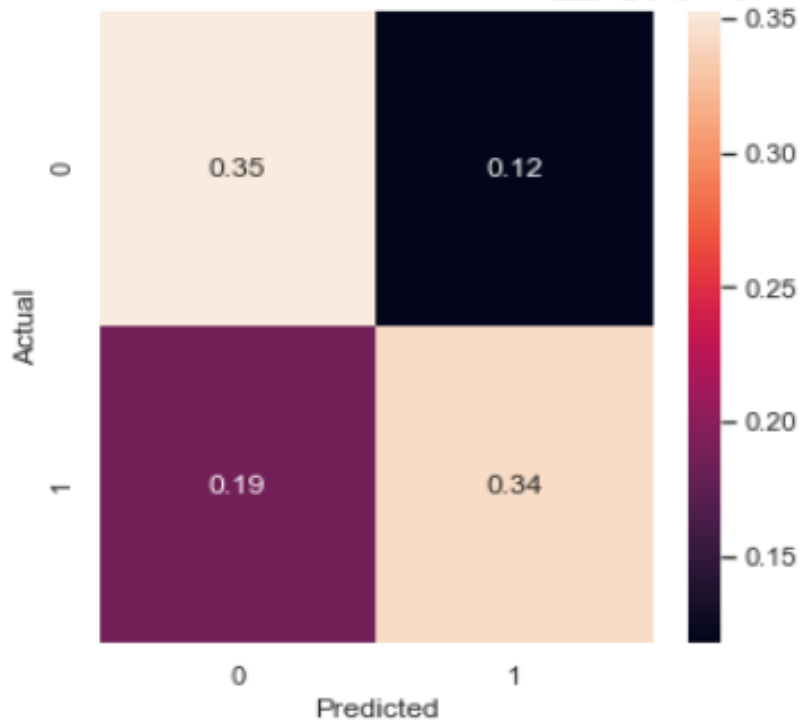


Results of SVM

1. **Accuracy on validation data – 69.60%**
2. **Confusion Matrix:**

Predicted	0	1
Actual		
0	72	24
1	38	70

3. **Precision on validation data – 74.46%**

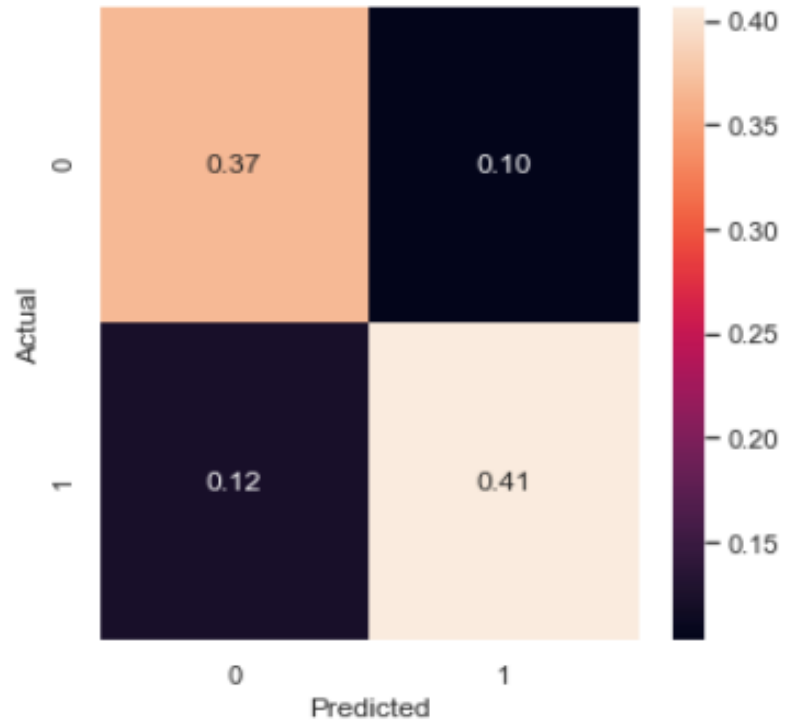


Results of Random Forest

1. **Accuracy on validation data – 77.45%**
2. **Confusion Matrix**

Predicted	0	1
Actual		
0	75	21
1	25	83

3. **Precision on validation data - 79.80%**



Neural Network (Bonus)

A neural network is a series of algorithms used to recognize hidden pattern and correlation in the raw data. Hyperparameters used in the implementation are the number of hidden layers, learning rate, activation functions and the optimizers.

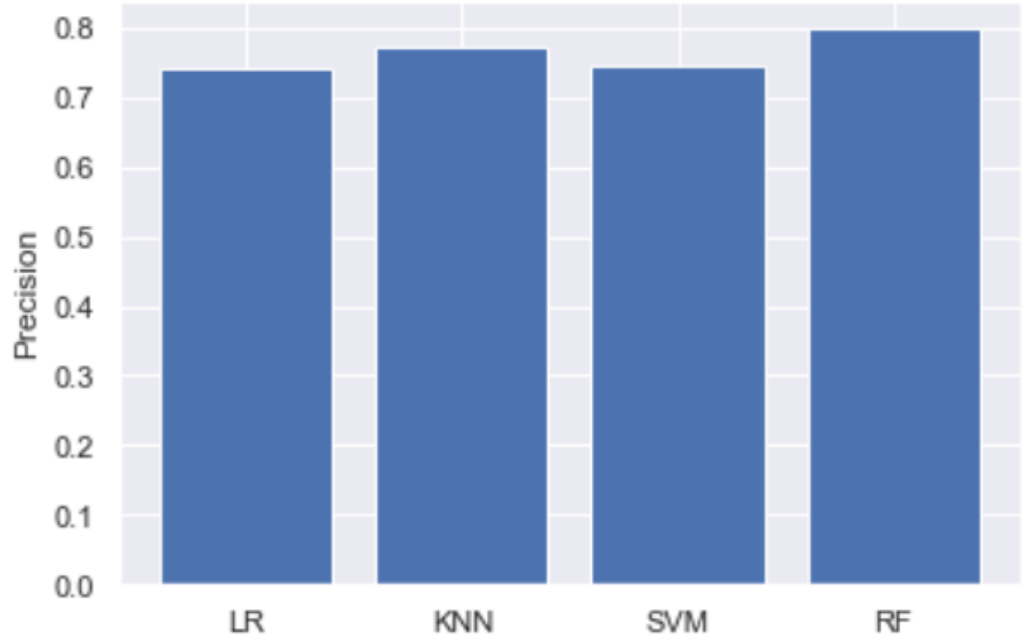
Accuracy is 76.25% for the neural network with 2 hidden layers, relu activation function and Adam optimizer with learning rate of 0.01.

```
model = Sequential()
model.add(Dense(150, input_dim=11, activation="relu"))
model.add(Dense(150, activation="relu"))
model.add(Dense(150, activation="relu"))
model.add(Dense(1, activation="sigmoid"))

adam = Adam(learning_rate = 0.01)
model.compile(loss = 'binary_crossentropy',optimizer=adam, metrics=['acc'])
```

Key Observations

Random Forest algorithm performed better than the logistic regression, KNN, and SVM with precision of 72.46% on the test data.



Team Contribution

Team Member	Assignment Part	Contribution
Navya	Data Analysis	50%
Vaishnavi	Data Analysis	50%
Vaishnavi	Model Implementation – LR, KNN	100%
Navya	Model Implementation – SVM, RF	100%
Navya	Neural Network Implementation	50%
Vaishnavi	Neural Network Implementation	50%

THANK YOU