Description:

This is Dataset is about Wine Quality. In this we build a model which classify whether the wine taste is good or not

In this dataset label have multiple class like good,normal,bad. We are only interested in good and not good so make normal and bad as not good label. Now we have only two classes good and not good taste.

There is no missing value in this dataset.

I create multiple box plot to visualize outliers.

I use IQR technique to remove outliers from dataset. Because our dataset is skewed

After removing outliers we 5064 remaining instances.

Then I create relationship like pairplot matrix, scatter plot, bar chart and heatmap etc.

Afterward I create dummies of columns (type, pHValue)

Then I split the data into the train& test. I’m stratified sampling .I take 20% test set and 80% training set.

After splitting data I using standard scaling for normalizing the data.

Then I train data and test data on different model Logistic Regression, SVM, NaiveBayes, Decision Tree,

Random forest Algorithm. I also use ensemble method. Random forest f1-score and accuracy is highest in our case f1-score=65% for good class and 0.91 for not good class and our accuracy is 86%

Our Dataset is imbalanced mean we didn’t have equal samples for each class so we didn’t trust on accuracy measure we have to trust on f1-score for imbalanced dataset case.

Then I create accuracy and f1-score graphs

And answer the following questions based on what you observe from your model building process

For the best fitted model, does it work better for the Red wine or White wine or similar

Random Forest Classifier is best in our case

No if our Dataset is Balanced Dataset then it is possible the other models work better

Can ensemble methods improve your models?

Yes Random Forest is a Ensemble methods is improve our model

Whether the categorical variable pH Value is important or not. If it is important, and you need to make a prediction for one instance where the pH value is missing, how would you proceed?

Yes pH value is important for proceeding if pH value is missing let take example eg Random\_classifier.predict([0.17297758, -0.33190212, 0.24773165, 1.96154966, 1.21824058,0.19548849, 1.28638457, 1.61392447, 0.20503508, -1.22902996,-0.39133804, 0.39133804,0,0,0])