CGAS MINI PROJECT

Title:

Cuisine prediction and cuisine similarity using yummly dataset and predicting wine quality using wine dataset.

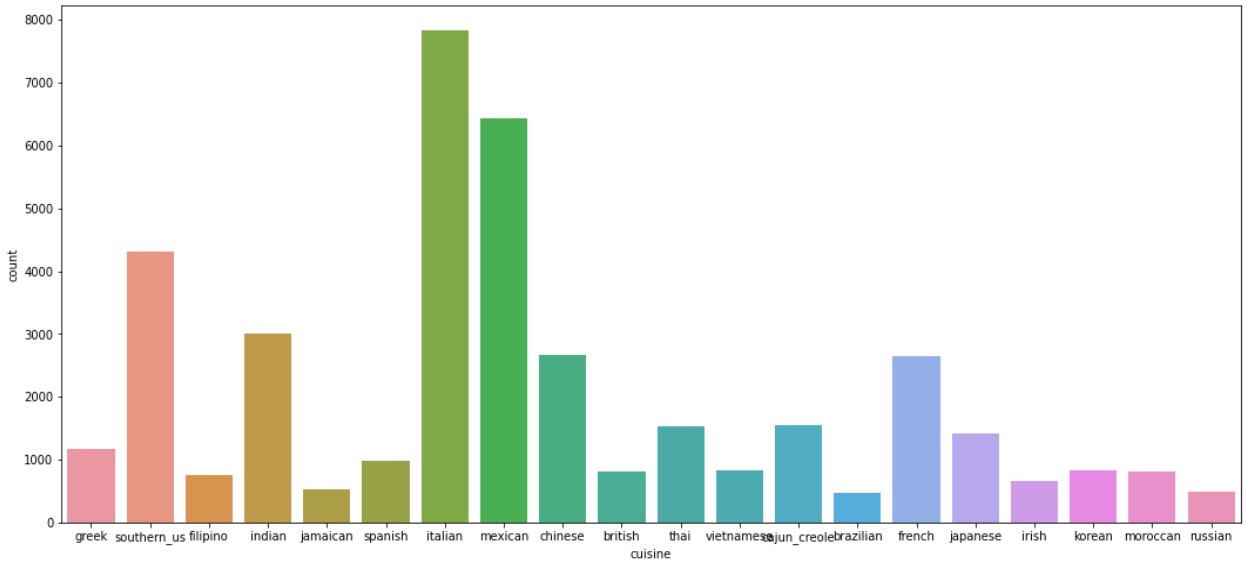
Submitted by:
Abhishek Vyas(MT19086),
Era Sharma (MT19121),
Saumya Jain(MT19098)

Sub Project 1: Cuisine prediction using yummly dataset.

Introduction:

Problem Statement: Prediction of the cuisines based on the ingredients using the yummly dataset. In this dataset, there are 3 columns: 'ID', 'Ingredients' and 'Cuisine'. There are 39774 different recipes, each recipe is associated with a cuisine. Each recipe is the set of its ingredients. There are 20 different cuisines in total. We need to predict the cuisine based on the ingredient set given from the data.

Analysis:



The above graph shows Number of Recipes vs cuisine graph. It tells that cuisine italian has the maximum number of recipes followed by mexican.

Work Done:

To predict the cuisines from the ingredient set we have used different preprocessing steps, Features are extracted and prediction is done using different machine learning models.

Preprocessing:

- Duplicate recipes are checked: No duplicate recipe is present.
- NaN values are checked: No nan values are there.
- Data is tokenized and re is used.
- Label Encoder is used.

Feature Extraction:

Technique 1

- Tfidf based features are extracted using tfidfVectorizer.
- Sparse Matrix is made

Technique 2

- A dataframe of all 39974 rows and 6715 columns is created
- 0/1 is filled if that ingredient is present in that cuisine or not.

Data is split into training and testing.

Different Machine Learning Models are applied:

1. Logistic Regression
2. K Nearest Neighbor
3. Cosine based Similarity
4. Decision Tree
5. Random Forest

Results:

The accuracy for 80:20 split for different models comes out to be:

Technique 1 Results:

Models	Accuracy
Logistic Regression	0.7812
K Nearest Neighbor	0.7355
Cosine based Similarity	0.6926
Decision Tree	0.6080
Random Forest	0.7406

Logistic regression performed better with 78.12% accuracy.

Technique 2 Results:

Models	Accuracy
Logistic Regression	0.692
K Nearest Neighbor	0.602
Decision Tree	0.6921
Random Forest	0.69

Logistic regression and decision tree performed better with 69.2% accuracy.

Out of two techniques technique 1 (tfidf vectorization)performed better than technique 2 (using ingredients presence or absence).

Sub Project 2: **Cuisine similarity using yummly dataset.**

Introduction:

Problem Statement: Predicting Cuisine similarity based on the ingredients using the yummly dataset. In this dataset, there are 3 columns: 'ID', 'Ingredients' and 'Cuisine'. There are 39774 different recipes, each recipe is associated with a cuisine. Each recipe is the set of its ingredients. There are 20 different cuisines in total. We need to predict the most similar cuisine based on the ingredient set given from the data.

Work Done:

We have applied below two techniques to perform cuisine similarity.

Preprocessing:

- Duplicate recipes are checked: No duplicate recipe is present.
- NaN values are checked: No nan values are there.
- Data is tokenized and re is used.
- Label Encoder is used.

Technique 1: Based on matching ingredients

We created a list of unique ingredients in all 20 cuisines and matched the number of ingredients shared between every two cuisines ,the cuisine with maximum ingredients shared is considered as similar to that cuisine.

Technique 2: TF-IDF based cosine similarity

TF-IDF based cosine similarity is calculated using the recipes(set of ingredients).

Features are extracted and most similar cuisine out of all is calculated by majority voting of the cuisines, based on the training and testing data.

Results:

Technique 1 Results:

For cuisine	Similar Cuisine
brazilian	mexican
british	italian
cajun_creole	italian
chinese	mexican
filipino	mexican
french	italian
greek	italian
indian	italian
irish	italian
italian	mexican
jamaican	mexican
japanese	chinese
korean	chinese
mexican	italian
moroccan	italian
russian	italian
southern_us	italian
spanish	italian
thai	chinese
vietnamese	chinese

Technique 2 Results:

Cuisine	Similar Cuisine
brazilian	italian
british	southern_us
cajun_creole	southern_us
chinese	italian
filipino	italian
french	italian
greek	italian
indian	italian
irish	french
italian	mexican
jamaican	italian
japanese	chinese
korean	chinese
mexican	italian
moroccan	italian
russian	italian
southern_us	italian
spanish	italian
thai	chinese
vietnamese	thai

Both techniques matched 70-80% in results and as there was no ground truth so after searching on the internet we can say that both experiments were successful in predicting 70% correct similar cuisines.

Sub Project 3: **Red wine Quality dataset analysis and quality prediction**

Objectives:-

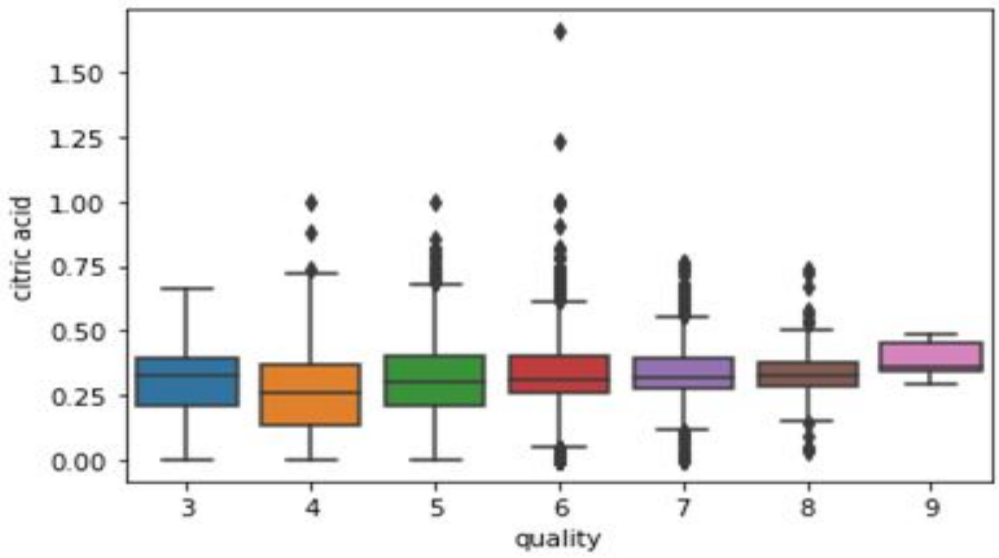
1. Analysis of Red wine quality dataset including outlier detection and removal.
2. Prediction of wine quality based on different features using regression model and show betterment of result after outlier removal.
3. Predicting whether a particular red wine “good quality” or not using various classification models.
   The quality of is determined by 11 input attributes which are as:-
   Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides,Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulfates, Alcohol

Preprocessing:

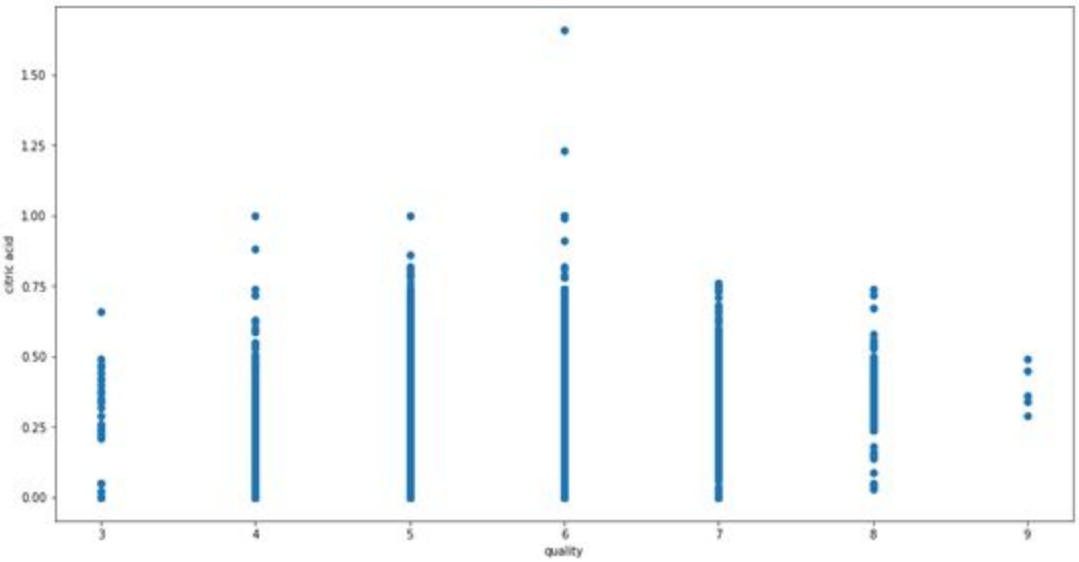
- Remove duplicate rows : (before:6497 , after: 5329)
- Check if any row have all Nan : no such row found.
- Check % of NULL values in each column : as most of the columns have less than 1% of NULL values so we remove all such row in which any column have NULL values (0.53 % of data is deleted)
- Make heatmap and remove column ID as have all unique values.
- Scale the data in range of 0 to 1 using MinMaxScaler

Various plots used to detect outliers

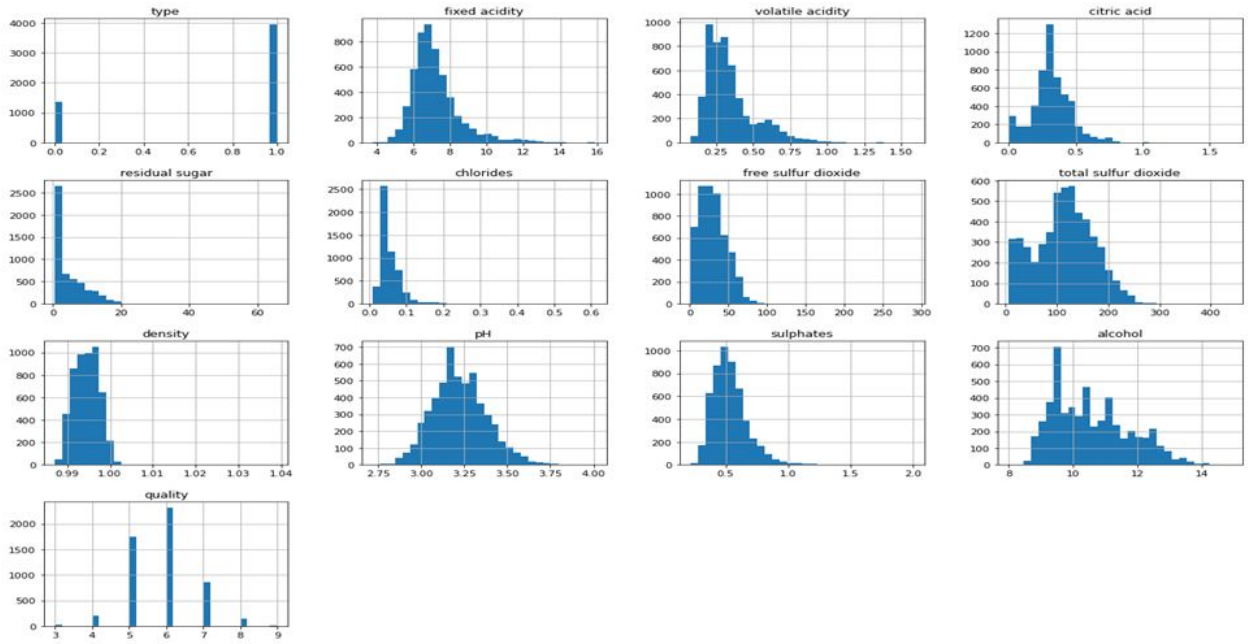
- Plot *Box Plot* and *Scatter Plot* to detect outliers, as target variable is Quality, so make plots for each column against target variable quality.
- For example, for feature “Citric acid” plots are as:-
  - *Box Plot for* “Citric acid” vs “quality”



❖ Scatter plot for “Citric acid” vs “quality”:-



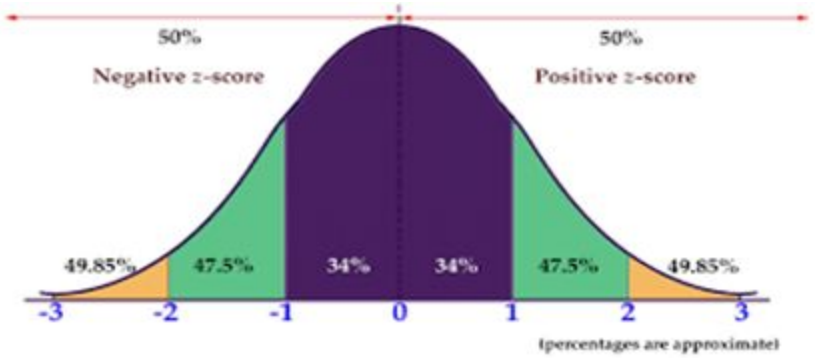
❖ Histogram plot for all attributes:-



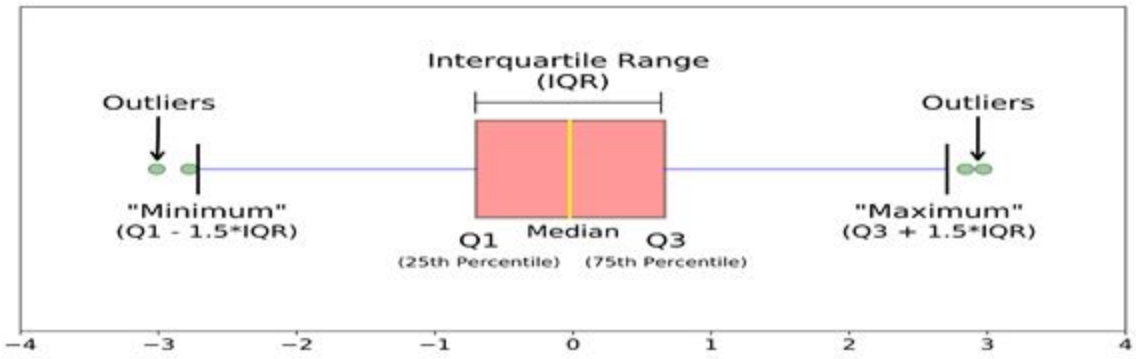
Removing outliers

- **Z-Score method** =>
    1. We can find out numbers of standard deviation values away from mean.
    2. 68% of data lies b/w (-1sd to +1sd) , 95% (-2sd to +2sd), 99.7% (-3sd to +3sd)
    3. Here in our case we take data which is in range between -3sd to +3sd.
    4. Other remaining values are treated as outliers.
    5. Formula to calculate the Z-Score and distribution curve

$$z = \frac{X - \mu}{\sigma}$$



- **IQR(Inter Quartile Range) method** =>
    1. In this method we detect outliers using IQR(inter quartile range) which tell us the variation in the dataset.
    2. Any value which is beyond the range for -1.5 * IQR to +1.5 * IQR is treated as outlier.
    3. Q1-1.5 * IQR is min and Q3+1.5 * IQR is the max value in the Selected dataset.
    4. Here
        a) Q1 (quartile 1)=> 25% of data bw min to Q1
        b) Q3 (quartile 3) => 75 % of data b/w min and Q3.
    5. Figure to show different quartiles is as:-

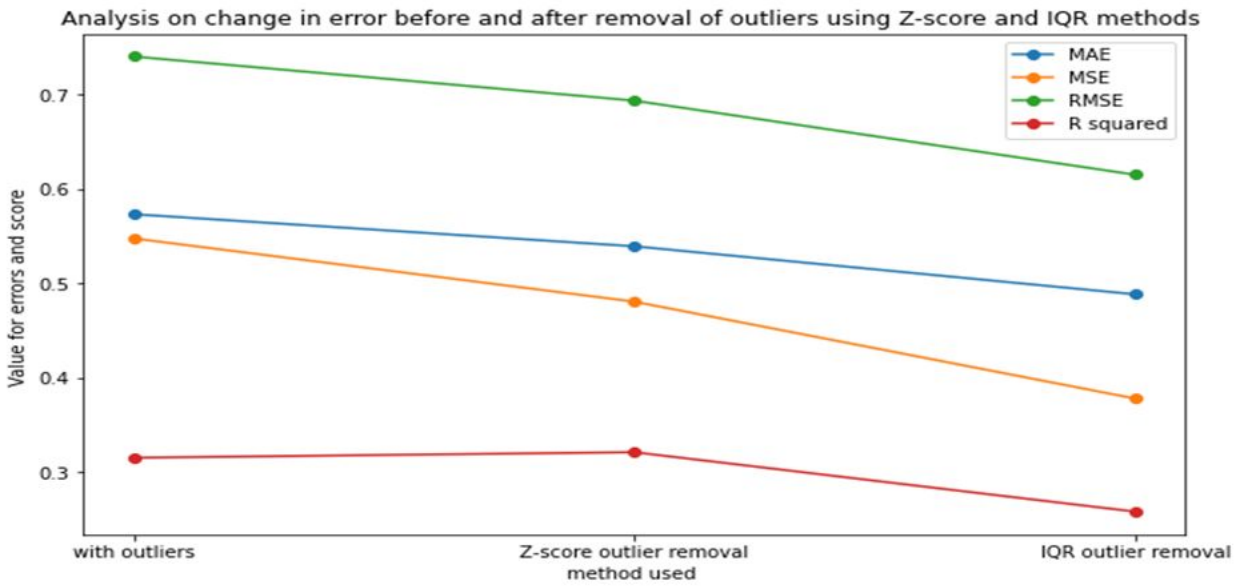


Prediction Using Linear Regression and Betterment of result

- We implement Linear Regression using sklearn on dataset before removing the outliers and after removing the outliers and compare various evaluation matrics results in different cases which are as:-
- The Result of different matrices are as:-

	With Outliers	**After Z-score**	**After IQR**
MAE	0.5731	0.5392	0.4884
MSE	0.5476	0.4806	0.3781
RMSE	0.7399	0.6932	0.6149
R squared	0.3154	0.3213	0.2821

Plot for betterment analysis

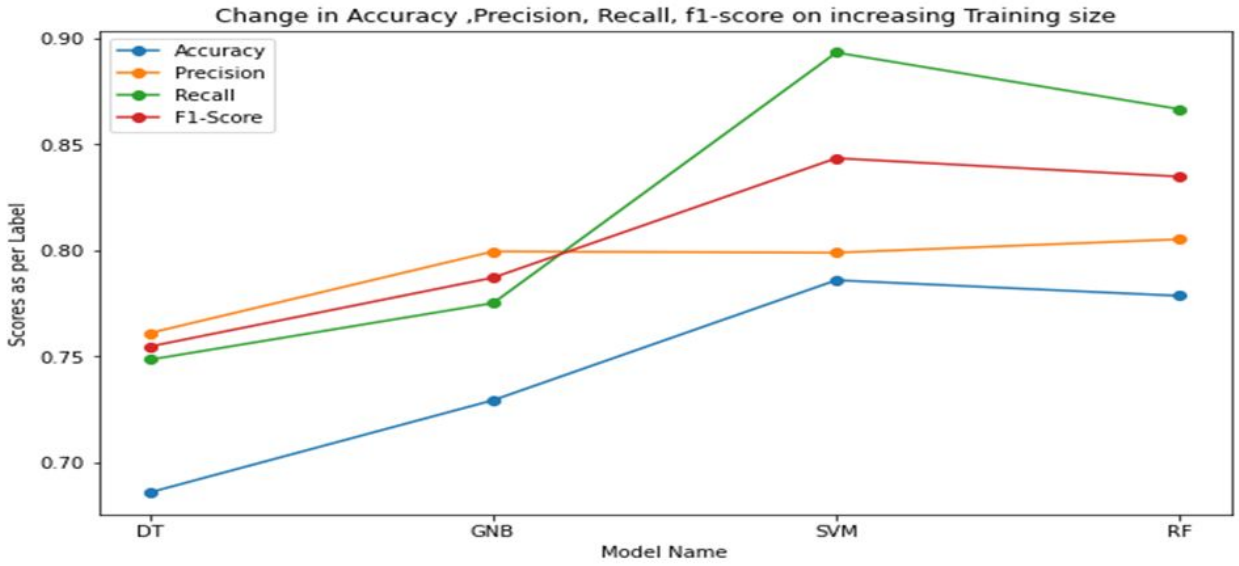


Wine Quality prediction using Classification

- **Various models and Metric Results:**

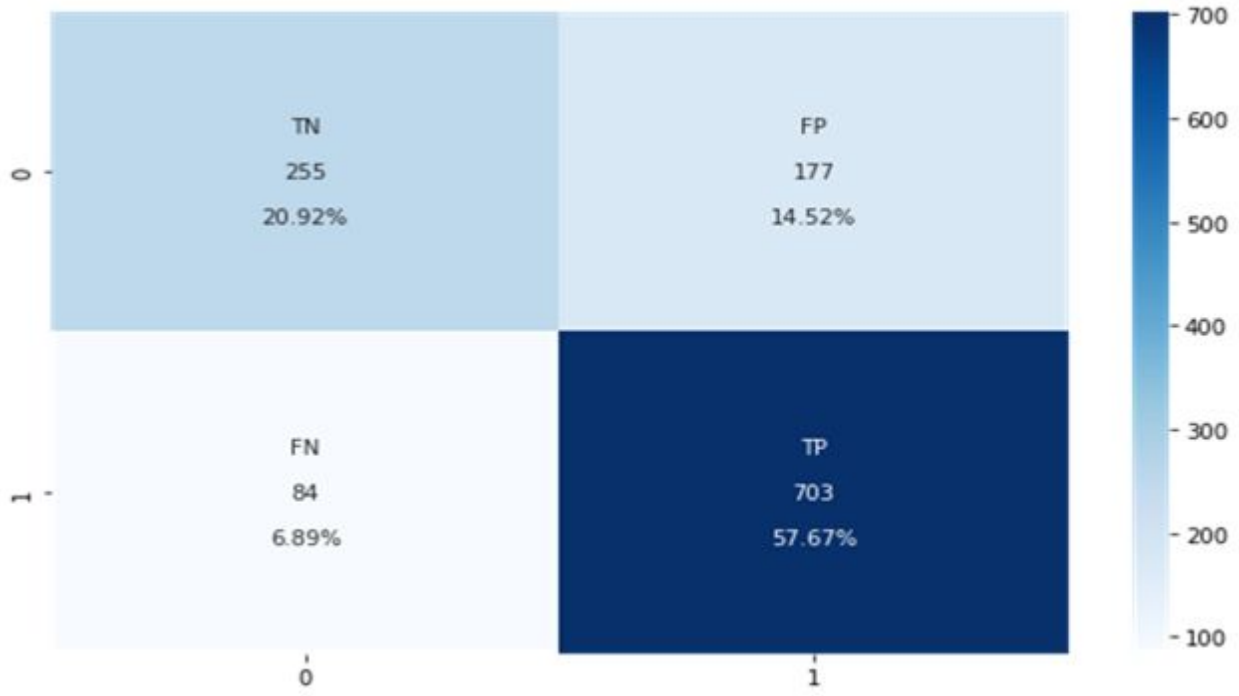
	Accuracy	Precision	Recall	F1_score
Decision Tree	0.6858	0.7609	0.7484	0.7546
GaussianNB	0.7292	0.7994	0.7750	0.7870
SVM	**0.7858**	**0.7988**	**0.8932**	**0.8434**
Random Forest	0.7785	0.8051	0.8665	0.8347

Graphical representation for Results of Classification



Confusion matrics

- We get best results using SVM.
- We also make confusion matrics for all models with various details.
- For example, here confusion matrics for SVM is as :-



References:

- https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
- https://www.kaggle.com/c/whats-cooking/overview
- https://www.whatissixsigma.net/box-plot-diagram-to-identify-outliers/
- https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
- https://www.pluralsight.com/guides/cleaning-up-data-from-outliers