CAPSTONE PROJECT

MODEL TO PREDICT QUALITY OF WINE

INTRODUCTION

- Wine quality depends on the vinification process, geographical origin of grapes and varietal composition of grape must.
- The assessment of wine quality and authenticity can be considered as a social demand since consumers require high quality products.
- The wine chemistry and several features influencing the complexity of chemical composition turns its quality analysis highly challenging.
- Effective tools for wine classification and detection of adulterants is required.

OBJECTIVE

- The aim is to create a prediction model in Python language for the quality analysis of wine
- A large dataset is considered which shows the quality of wine for several chemical components such as pH,density,alcohol etc.
- Multiple classification algorithms will be applied on the dataset after data preparation.
- The accuracy will be analysed for each algorithm and optimised to choose a final model for the quality prediction of wine.

DATA GATHERING

IMPORTING LIBRARIES

- The initial step before importing dataset is to import various libraries in Python.
- To name a few, pandas is used in converting csv format to dataframe and for exploratory data analysis
- Numpy is used for mathematical operations.
- Matplotlib and seaborn is used for data visualisation.
- Scikit learn features various algorithms, data preprocessing and model evaluation matrices.

READ DATASET

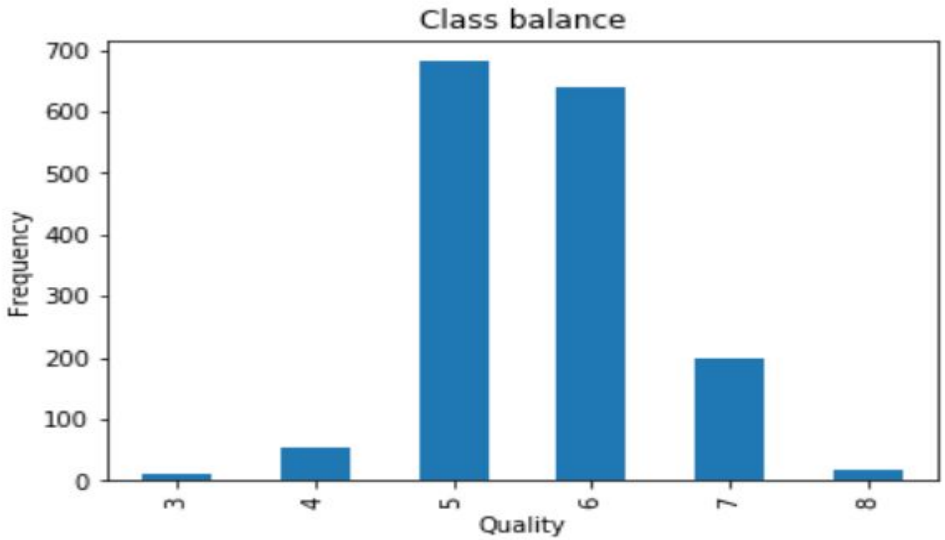
- The dataset which is in csv format is extracted to a dataframe ‘wine’.
- The independent variables are ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’, ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘density’, ‘pH’, ‘sulphates’ and ‘alcohol’.All are continuous variables.
- The target variable is ‘quality’, which is a categorical variable and rates the wine quality on a scale of 1 to 10.

EXPLORATORY DATA ANALYSIS

UNDERSTANDING DATA

- Checking null values : To find missing values which should be treated.Our dataset has no null values.
- Dimension : The dataset has total 1599 rows and 12 columns.
- Statistical summary : describe() is used to analyse mean , median , count and five-point summary.Variation in median and mean was observed which hints the presence of outliers.
- Target variable is extracted to ‘Y’ and the predictor variables to ‘X’

- Heat map:To visualise correlation between variables
- Visualising the balance of classes : A bar graph was plotted to see the contribution towards each output classes.



UNDERSAMPLING

- Here, we observe data imbalance. A model fed with such a data could give biased decision and so less accuracy in prediction.It also affects the correlation between features.To avoid this,we do resampling.
- Re-sampling: Undersampling and oversampling.

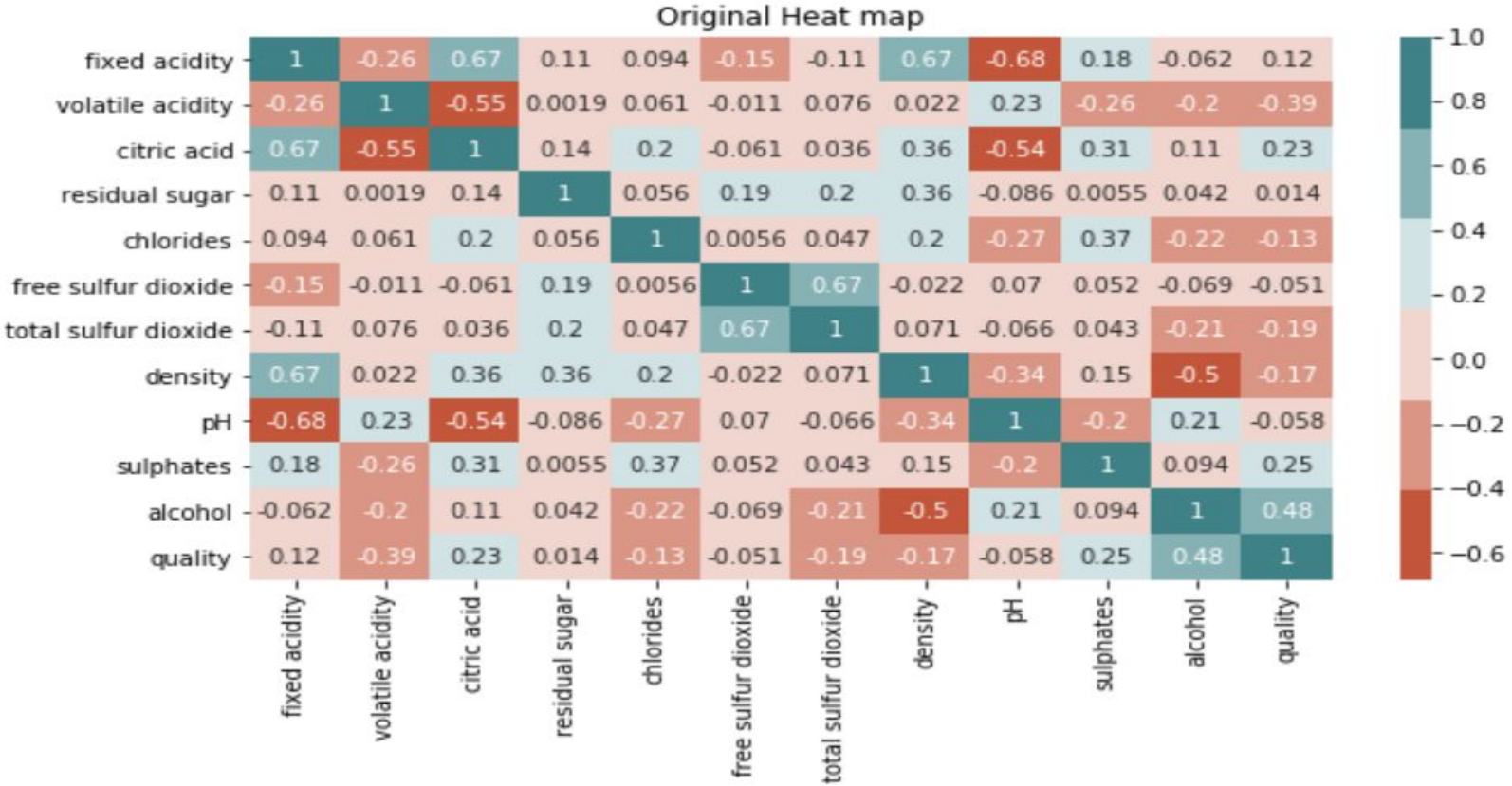
  1. Undersampling : Removes some observations from majority class.

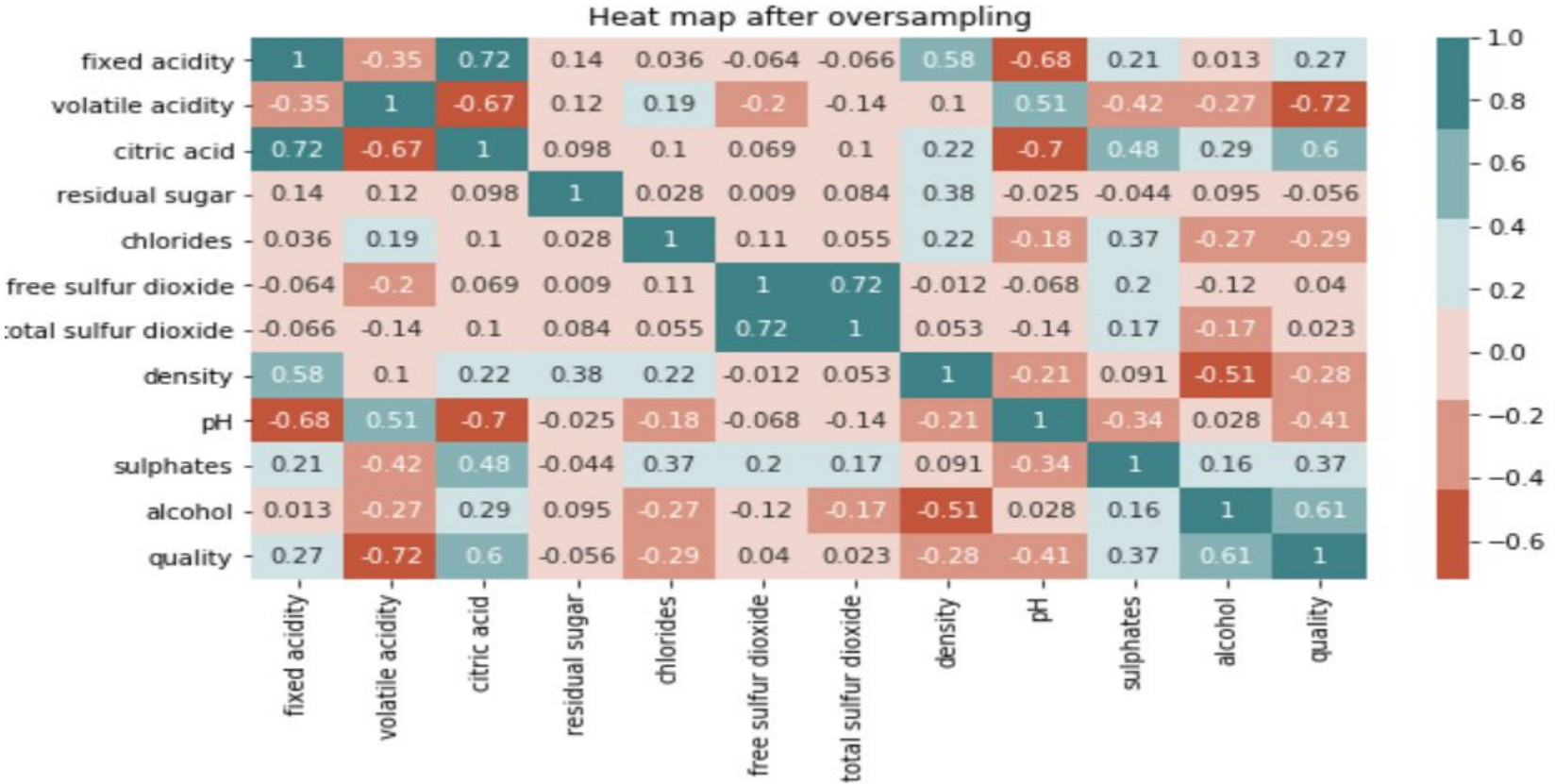
  2.Oversampling : Adds data to generate output from minority class-SMOTE

- From the bar graph ‘Class balance’,majority classes are 5,6 and minority classes are 3,4,7,8.
- SMOTE : To oversample the data in minority classes to the class with most number.Alternatively, each class could be assigned with required number of samples.This resulted in strengthening the relationship between variables, ie. , an increase in correlation coefficients.
- This transformation should be performed on training data only.

SPLITTING DATASET

- First the target variable and predictor variables are separated.
- The entire dataset is splitted into training set and testing set- x_train,y_train,x_test,y_test at 25% split.
- Random_state : The parameter that is used to reproduce the result.


Original Heat map
fixed acidity 1 -0.26 0.67 0.11 0.094 -0.15 -0.11 0.67 -0.68 0.18 -0.062 0.12
volatile acidity -0.26 1 -0.55 0.0019 0.061 -0.011 0.076 0.022 0.23 -0.26 -0.2 -0.39
citric acid 0.67 -0.55 1 0.14 0.2 -0.061 0.036 0.36 -0.54 0.31 0.11 0.23
residual sugar 0.11 0.0019 0.14 1 0.056 0.19 0.2 0.36 -0.086 0.0055 0.042 0.014
chlorides 0.094 0.061 0.2 0.056 1 0.0056 0.047 0.2 -0.27 0.37 -0.22 -0.13
free sulfur dioxide -0.15 -0.011 -0.061 0.19 0.0056 1 0.67 -0.022 0.07 0.052 -0.069 -0.051
total sulfur dioxide -0.11 0.076 0.036 0.2 0.047 0.67 1 0.071 -0.066 0.043 -0.21 -0.19
density 0.67 0.022 0.36 0.36 0.2 -0.022 0.071 1 -0.34 0.15 -0.5 -0.17
pH -0.68 0.23 -0.54 -0.086 -0.27 0.07 -0.066 -0.34 1 -0.2 0.21 -0.058
sulphates 0.18 -0.26 0.31 0.0055 0.37 0.052 0.043 0.15 -0.2 1 0.094 0.25
alcohol -0.062 -0.2 0.11 0.042 -0.22 -0.069 -0.21 -0.5 0.21 0.094 1 0.48
quality 0.12 -0.39 0.23 0.014 -0.13 -0.051 -0.19 -0.17 -0.058 0.25 0.48 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Heat map after oversampling
fixed acidity 1 -0.35 0.72 0.14 0.036 -0.064 -0.066 0.58 -0.68 0.21 0.013 0.27
volatile acidity -0.35 1 -0.67 0.12 0.19 -0.2 -0.14 0.1 0.51 -0.42 -0.27 -0.72
citric acid 0.72 -0.67 1 0.098 0.1 0.069 0.1 0.22 -0.7 0.48 0.29 0.6
residual sugar 0.14 0.12 0.098 1 0.028 0.009 0.084 0.38 -0.025 -0.044 0.095 -0.056
chlorides 0.036 0.19 0.1 0.028 1 0.11 0.055 0.22 -0.18 0.37 -0.27 -0.29
free sulfur dioxide -0.064 -0.2 0.069 0.009 0.11 1 0.72 -0.012 -0.068 0.2 -0.12 0.04
total sulfur dioxide -0.066 -0.14 0.1 0.084 0.055 0.72 1 0.053 -0.14 0.17 -0.17 0.023
density 0.58 0.1 0.22 0.38 0.22 -0.012 0.053 1 -0.21 0.091 -0.51 -0.28
pH -0.68 0.51 -0.7 -0.025 -0.18 -0.068 -0.14 -0.21 1 -0.34 0.028 -0.41
sulphates 0.21 -0.42 0.48 -0.044 0.37 0.2 0.17 0.091 -0.34 1 0.16 0.37
alcohol 0.013 -0.27 0.29 0.095 -0.27 -0.12 -0.17 -0.51 0.028 0.16 1 0.61
quality 0.27 -0.72 0.6 -0.056 -0.29 0.04 0.023 -0.28 -0.41 0.37 0.61 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

FEATURE SELECTION

- Heatmap : The heatmap is plotted for the dataset to find the strength of correlation between variables.Some analysis:

1. Relatively strong positive correlation : ‘fixed acidity’ and ‘citric acid’,’fixed acidity’ and ‘density’,’free sulfur dioxide’ and ‘total sulfur dioxide’.
2. Strong negative correlation : ‘pH’ and ‘fixed acidity’,’pH’ and ‘citric acid’

- Total sulfur dioxide’ has strong positive correlation with ‘free sulfur dioxide’,and its contribution towards ‘quality’ variable is very less.So we drop that column.

DETECTING OUTLIERS

- There are various methods to detect and remove outliers.These are:

1. Visualising through boxplot
2. Visualising through scatterplot
3. Z-score method
4. Interquartile Range method(IQR)

- Here both box plot and IQR method is chosen , where upper and lower threshold is found out and any data points not in the range are removed.

BUILDING MACHINE LEARNING MODEL

DECISION TREE ALGORITHM

- Different models were built for different values of max_depth and splitting criteria(‘gini’ and ‘entropy’).
- For each model,training score and testing score are evaluated to choose the better model out of them, to find a balance between better accuracy and reducing overfitting problem.
- The best out of the lot is decision tree with max_depth=4 and criteria=gini.Accuracy=0.61
- For each of the algorithms,SMOTE was avoided since it lowered accuracy.

NAIVE BAYES ALGORITHM

- It is a supervised learning algorithm which applies Bayes' Theorem with 'naive' assumptions such as every feature is independent and every feature has equal contribution to the outcome.
- The basic logic is the comparison of probabilities of a data point falling into one of the output classes.
- Here,GaussianNB() classifier is chosen where the likelihood of features is assumed to be Gaussian.GaussianNB() showed better accuracy and F1 score than ComplementNB() ,which is used for imbalanced datasets.

ENSEMBLE ALGORITHMS

- Various Ensemble methods are chosen for the same processed data.
- AdaBoostClassifier() : This uses Adaptive boosting technique, which assigns higher weight on weak learners and lower weights on strong learners.Accuracy=0.5675 and F1 score= 0.524
- GradientBoostingClassifier() : This boosting method supports both binary and multi-class classification. This works by optimising loss functions.Accuracy=0.615 and F1 score= 0.611

- RandomForestClassifier() : Uses bagging method and uses different decision trees having random subset of dataset(sampling with replacement).Accuracy= 0.64, F1 score=0.62
- It is clear that Random Forest showed better result.

RANDOM FOREST

- Hyperparameter tuning for Random Forest is done using GridSearchCV() function.It gave parameters to be : criterion='gini', max_depth= 6, max_features= 'auto'.
- These parameters is fit into the model.
- OOB error rate: It is the misclassification rate and is used to find the optimum value of n_estimators(no of decision trees) by analysing the graph.Optimum value of n_estimators is found to be 300.
- The model is fitted and trained using these parameters.

MODEL EVALUATION

- The model is evaluated using Accuracy,Precision,Recall,F1 score.
- F1 score: It is the harmonic mean of Recall and precision.This metric is chosen for imbalanced multi class classification.
- The results are :
  - Accuracy=0.6375,
  - Precision=0.60,
  - Recall=0.6375,
  - F1 score=0.617

CONCLUSION

DECISION

- Although oversampling increased correlation between variables,it was found to decrease the accuracy and F1 scores.In the final model,it was avoided.
- With removal of column and outliers, the evaluation scores decreased slightly,but the training score decreased had this step been omitted.So,we followed the conventional method.
- The final results doesn’t improve due to the highly imbalanced dataset.

THANK YOU !