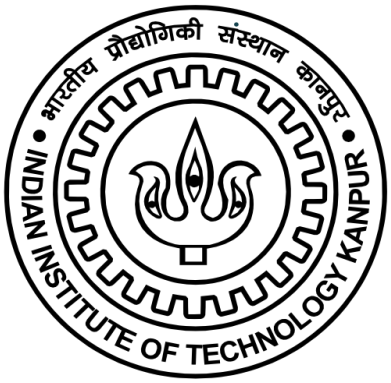

IIT Kanpur

Wine Quality Prediction
Project Report

By Kartikeya Kesharwani

Dataset link- https://archive.ics.uci.edu/ml/datasets/wine+quality

'winequality-red.csv' file has been used in this project

GitHub: https://github.com/KartikKesharwani/Red-wine-quality-prediction.git

Introduction:

The dataset contains the entries for chemical composition of different wines and the quality based on sensory input.

It has 1599 rows.

Input variables are:

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable:

12 - quality (score between 0 and 10)

This goal of this project is to predict the quality of the wine based on the various input features.

Train test split used is 80:20. The random_state used is 42.

Feature scaling technique is Standardization since there are outliers present in the data.

10 Questions to answer:

1. What is the distribution of the wine quality scores?

   The quality scores of the wine in the data range from 3 to 8. Each quality level has different number of samples. The distribution is highly skewed.
   The number of samples in each class are as follows:
   3 – 10
   4 – 53
   5 – 681
   6 – 638
   7 – 199
   8 – 18
   We can see that the dataset is highly imbalanced.
   Number of samples of quality 5 and 6 are way more (~1300) compared to the other qualities (~300).
   This imbalance will lead the model to be biased towards samples of average quality wine and hence make inaccurate predictions.
   To solve this issue, we can use SMOTE library to synthetically create samples of the minority classes.

2. What are the relationships between the different features?

   Relationship of features with ‘quality’:

   a. **Fixed acidity** - All qualities of wine have similar fixed acidity values.
   b. **Volatile acidity** - Higher quality wines have lower volatile acidity(inversely proportional).
   c. **Citric acid** - Higher quality wines have higher citric acid content (directly proportional).
   d. **Residual sugar** - All qualities of wine have similar residual sugar values.
   e. **Chlorides** - Higher quality wines have lower chlorides(inversely proportional).
   f. **Free sulfur dioxide** - Free sulfur dioxide does not show a linear relationship with the quality of the wine.
   g. **Total sulfur dioxide** - Total sulfur dioxide does not show a linear relationship with the quality of the wine.
   h. **Density** - All quality of wines seem to have the same density.
   i. **pH** - All quality of wines seem to have the same pH.
   j. **Sulphates** - Higher quality wines have higher level of sulphates (directly proportional).
   k. **Alcohol** - Higher quality wines have higher level of alcohol(directly proportional).

   Correlations from heatmap:

   a. Citric acid – fixed acidity : 0.7
   b. Density – fixed acidity : 0.7
   c. pH – fixed acidity : -0.7
   d. Citric acid – volatile acidity : -0.6
   e. Citric acid – pH : -0.5
   f. Total sulfur dioxide – free sulfur dioxide : 0.7
   g. Alcohol – density : -0.5
   h. Alcohol – quality : 0.5

3. Are there any outliers in the data?

   Yes, Outliers are present in all the features in large quantities. However it is difficult to ascertain whether these are errors in measurement or a characteristic of the data, hence we will not remove them.

4. What is the accuracy of the linear regression model?

   I have trained three different models for the dataset and calculated the metrics MSE, RMSE and R2-score.

   Model 1:
   Using the data as is with test size = 0.2
   MSE: 0.3900
   RMSE: 0.6245
   R2-Score: 0.4031

   Model 2:
   Using the data after applying SMOTE to solve the problem of class imbalance.
   MSE: 0.7957
   RMSE: 0.8920
   R2-Score: 0.7162

   Model 3: (Best R2-Score)
   Using the data after applying SMOTE and using Polynomial Features(degree=2) to get a better result than previous models.
   MSE: 0.5947
   RMSE: 0.7712
   R2-Score: 0.7879

5. What are the most important features for the linear regression model?

   After training models 1 and 2, I have printed the weights of each feature by using the model.coef_ method.
   The feature **'alcohol'** had the highest weight and hence can be considered as the most important feature.

6. What is the MSE of the linear regression model?

   Model 1:
   Using the data as is with test size = 0.2
   MSE: 0.3900

   Model 2:
   Using the data after applying SMOTE to solve the problem of class imbalance.
   MSE: 0.7957

   Model 3: (Best R2-Score)
   Using the data after applying SMOTE and using Polynomial Features(degree=2) to get a better result than previous models.
   MSE: 0.5947

7. What is the R-squared of the linear regression model?

   Model 1:
   Using the data as is with test size = 0.2
   R2-Score: 0.4031

   Model 2:
   Using the data after applying SMOTE to solve the problem of class imbalance.
   R2-Score: 0.7162

   Model 3: (Best R2-Score)
   Using the data after applying SMOTE and using Polynomial Features(degree=2) to get a better result than previous models.
   R2-Score: 0.7879

8. How can you improve the performance of the linear regression model?

   The performance of a linear regression model in my project was improved by using **Polynomial Features** and **SMOTE.**

   Polynomial features add polynomials of a certain degree to increase the complexity of the input feature vector. This helps the model to find patterns in the data which does not have a linear relation between the input and output features.
   SMOTE is used to synthetically create samples of the minority classes so that the model does not develop a bias towards the majority class and gets to learn from all the samples equally.

   Other methods to improve accuracy can be:
   a. Feature Engineering
   b. Data Cleaning and Preprocessing
   c. Feature Scaling
   d. Hyperparameter tuning
   e. Outlier Handling
   f. Expanding database

9. What are the limitations of the linear regression model?

   Limitations of Linear regression include:

   a. **Outliers**: Linear regression is sensitive to outliers which can influence the model's predictions. Outliers can lead inaccurate model predictions and low accuracy.
   b. **Limited to Linear Relationships**: Linear regression is not suitable for modelling complex, nonlinear relationships between input and output features.
   c. **Less accuracy metrics are available**: Accuracy metrics for linear regression are limited to MSE, RMSE and R2-Score. These can sometimes be insufficient to gauge a model performance. Classification models have a variety of metrics like recall, precision, f1-score, confusion matrix and more which help us effectively evaluate the model.

10. What are the implications of your findings for the real-world problem?

    Predicting wine qualities can be of great use in the real world. It can help wine companies to determine the quality of their product and hence improve it to gain business advantage. It can help them to price various wines according to its quality. It can help in quality control of the wines produced.
    It can help consumers to make more informed choices about the type of wine they are buying. Investors can better asses which types of wine they should invest in.

Conclusion:

In this project I have created models to predict the quality of red wine. There are a lot of outliers present in the data but I believe that they are not misrepresentations but a part of the data itself. Three models have been trained with data that has been processed with different techniques. The best model of the three obtained R2-score of 0.7879 and RMSE of 0.7712. I feel that linear regression is not the best model for this dataset and with a different algorithm we can achieve higher accuracy.