Project Report

RED Wine Prediction

*ANISHA GOYAL*

*1 SEP 2020*

**Contents**

**1. Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Exploratory Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

**2. Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.1.1 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.1.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.3 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2.1 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2.2 KNN . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2.3 SVM . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . 9

2.2.4 Naïve Bayes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2.5 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2.6 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .11

2.2.7 Artifical Neural Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

This datasets is related to red variants of the Portuguese "Vinho Verde" wine.

1. What changes company should bring to make the quality of red wine better?
   1. **Data**

There are 12 variables in our data in which 20 are independent variables and 1 (Quality) is dependent variable. These data can be viewed as classification tasks.

**Variables Information:**

**1.** Fixed acidity

**2.** Volatile acidity

**3.** Citric Acid

**4.** Residual Sugar

**5.** Chlorides

**6.** Free sulphur dioxide

**7.** Total sulphur dioxide

**8.** Density

**9.** PH

**10.** Sulphates

**11.** Alcohol

**12.** Quality

**1.3 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 12 variables and data types of all variables are either float64 or int64. There are 1600 observations and 12 columns in our data set. No missing value is present in our data.

**List of columns and their number of unique values** -

Fixed Acidity 96

Volatile Acidity 143

Citric Acid 80

Residual Sugar 91

Chlorides 153

Free sulphur dioxide 60

Total sulphur dioxide 144

Density 436

PH 89

Sulphates 96

Alcohol 65

Quality 2

**From EDA we have concluded that there are 11 continuous variable and 1 categorical variable in nature.**

**Chapter 2**

**Methodology**

Before fitting the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

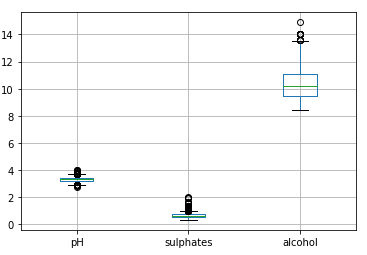
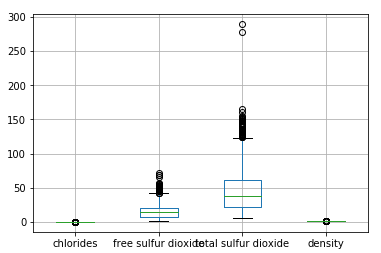
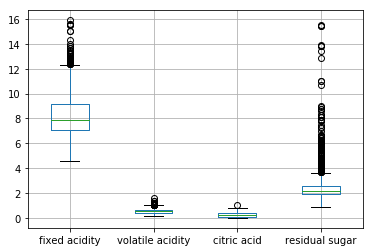
**2.1 Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

**2.1.1 Outlier Analysis**

The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

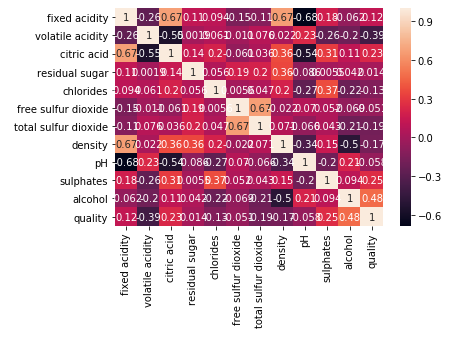
In figure we have plotted the boxplots of the 11 predictor variables. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.



From the boxplot almost all the variables consists of outliers. We have replaced the outliers with median.

**2.1.2 Feature Selection**

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable and **chi-square test** for categorical variable.



**2.2.3 Feature Scaling**

**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

**2.2 Modeling**

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

**2.2.1 Logistic Regresssion**

**Logistic regression** is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**).The confusion matrix and the accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 276 | 12 |
| 25 | 7 |

Accuracy is **88.44** percent

**2.2.2 KNN**

K-nearest neighbors (**KNN**) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 280 | 8 |
| 22 | 10 |

Accuracy is **90.625**  percent

**2.2.3 SVM**

A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an **SVM** model sets of labeled training data for each category, they're able to categorize new text.

SVM using kerenal as linear function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 288 | 0 |
| 32 | 0 |

Accuracy is **90** percent

SVM using kerenal as radial base function then the confusion matrix and accuracy is

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 281 | 7 |
| 25 | 7 |

Accuracy is **90** percent

SVM using kerenal as polynomial function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 283 | 5 |
| 26 | 6 |

Accuracy is **90.3125** percent

SVM using kerenal as sigmoid function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 256 | 32 |
| 18 | 14 |

Accuracy is **84.375** percent

**2.2.4 Naïve Bayes**

**Naive Bayes** is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a **Bayesian** setting. It can also be represented using a very simple **Bayesian** network .below is confusion matrix and accuracy of the model.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 240 | 48 |
| 10 | 22 |

Accuracy is **81.875** percent

**2.2.5 Decision Tree**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 260 | 28 |
| 12 | 20 |

Accuracy is **87.5** percent

**2.2.6 Random Forest**

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual valve | 285 | 3 |
| 17 | 15 |

Accuracy is **93.75** percent

**2.2.7 ANN** 

An artificial neural network (**ANN**) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual valve | 272 | 16 |
| 16 | 16 |

Accuracy is **90** percent

**Chapter 3**

**Conclusion**

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

**3.1 Model Evaluation**

In the previous chapter we have seen the **Confusion matrix** and **Accuracy** Value of different models. A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The **confusion matrix** gives you a lot of information, but sometimes you may prefer a more concise metric. TP is the number of true positives, and FP is the number of false positives. A trivial way to have perfect precision is to make one single positive prediction and ensure it is correct (precision = 1/1 = 100%).Higher is the accuracy better is the model.

**3.2 Model Selection**

From the observation of all **confusion matrix** and **Accuracy** Value we have concluded that **Naïve Bayes**  has minimum value of accuracy (i.e 81.875%). And **Random forest classifier**  has Maximum value of accuracy (i.e 93.75%)