

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

PROJECT REPORT

**R.SIVAKUMAR- 121012012767**

**B.TECH-CSE (2 ND YEAR)**

**TOPIC :**

**RED WINE QUALITY USING IN R- PROGRAMMING LANGUAGE**

TEAM MEMBERS :

* + **C.NATARAJ – 121012012756**
  + **D.ARUN PRASATH – 121012012721**
  + **A.B.KEERTHANA – 121011012743**

**Project Description :**

**It shows relation between quality and other variables of wine. We want to do transformation to see if we can increase correlation coefficient between them. Usedstepwise variable selection method to choose best predictor of wine quality.**

**Goal :**

• **Our focus is to see how each chemical component influences the quality of wine (0'very bad' to 10 'very excellent'). The usage of this analysis will help to understandwhether by modifying the variables, it is possible to increase the quality of the wineon the market.**

• **In this project we do Analysis ofRed Wine Datawhich contains 1,599 red wines with 12variables on the chemical properties of the wine.**

**RED WINE DATASET INFORMATION**

**Input Variable :**

**Fixed acidity:** Most acids involved with wine or fixed or nonvolatile (do not evaporatereadily)

**Volatile acidity**: The amount of acetic acid in wine, which at too high of levels can lead toan unpleasant, vinegar taste

**Citric acid:** Found in small quantities, citric acid can add 'freshness' and flavor to wines

**Residual sugar:** The amount of sugar remaining after fermentation stops, it's rare to findwines with less than 1 gram/liter and wines with greater than 45 grams/liter areconsidered sweet

**Chlorides:** The amount of salt in the wine

**Free sulfur dioxide:** The free form of SO2 exists in equilibrium between molecular SO2 (asa dissolved gas) and bisulfite ion; - it prevents microbial growth and the oxidation of wine

**Total sulfur dioxide:** Amount of free and bound forms of S02; in low concentrations, SO2is mostly undetectable in wine, but at - free SO2 concentrations over 50 ppm, SO2becomes evident in the nose and taste of wine

**Density:** The density of water is close to that of water depending on the percent alcoholand sugar content

**pH:** Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic);most wines are between 3-4 on the pH scale

**Sulphates:** A wine additive which can contribute to sulfur dioxide gas (S02) levels, wichacts as an antimicrobial and antioxidant

**Alcohol:** The percent alcohol content of the wine

**Wine Quality Exploration with R :**

**In this post, I will highlight the exploratory data analysis (EDA) with R to explore relationships in one variable to multiple variables and to discover for distributions, outliers, and anomalies.**

**The data set it consists of 12 variables:**

1.Fixed acidity (tartaric acid — g / dm³)

2. Volatile acidity (acetic acid — g / dm³)

3. Citric acid (g / dm³)

4. Residual sugar (g / dm³)

5. Chlorides (sodium chloride — g / dm³)

6. Free sulfur dioxide (mg / dm³)

7. Total sulfur dioxide (mg / dm³)

8. Density (g / cm³)

9. pH

10. Sulphates (potassium sulphate — g / dm3)

11. Alcohol (% by volume)

12. Quality (score between 0 and 10)

**Red wine model:**

The most important variables are: volatile.acidity, sulphates and alcohol.

fixed.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide and density doesn’t seems to be relevant in this model.

Adjusted R2 = 0.389

MSE = 0.3274646

**Red wine :**

TestRegsubsets(dataset = cleanRedDat, yColName = "quality", methods = c("exhaustive", "backward", "forward"), metrics = c("rsq", "rss", "adjr2", "cp", "bic"))

All the variables in the model are important predictors.

Adjusted R2 = 0.3873 (0.389 in the previous model)

MSE = 0.3300581 (doesn’t improve)

**SOURCE CODE FOR RED WINE QUALITY USING R :**

#importing library

library("ggplot2")

library("dplyr")

library("gridExtra")

library(GGally)

library(memisc)

library(pander)

library(corrplot)

**#Loading the csv file**

wine= read.csv("C:\\Users\\Arun\\Documents\\winequality-red.csv")

View(wine)

**#Transforming Quality from an Integer to a Factor**

wine$quality <- factor(wine$quality, ordered = T)

**#Creating a new Factored Variable called 'Rating'**

wine$rating <- ifelse(wine$quality < 5, 'bad', ifelse(

wine$quality < 7, 'average', 'good'))

wine$rating <- ordered(wine$rating,levels = c('bad', 'average', 'good'))

**Structure and summary of the Dataframe :**

## 'data.frame': 1599 obs. of 14 variables:

## $ X : int 1 2 3 4 5 6 7 8 9 10 ...

## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...

## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...

## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...

## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...

## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...

## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...

## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...

## $ density : num 0.998 0.997 0.997 0.998 0.998 ...

## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...

## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...

## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...

## $ quality : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...

## $ rating : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2 ...

## X fixed.acidity volatile.acidity citric.acid

## Min. : 1.0 Min. : 4.60 Min. :0.1200 Min. :0.000

## 1st Qu.: 400.5 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090

## Median : 800.0 Median : 7.90 Median :0.5200 Median :0.260

## Mean : 800.0 Mean : 8.32 Mean :0.5278 Mean :0.271

## 3rd Qu.:1199.5 3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420

## Max. :1599.0 Max. :15.90 Max. :1.5800 Max. :1.000

## residual.sugar chlorides free.sulfur.dioxide

## Min. : 0.900 Min. :0.01200 Min. : 1.00

## 1st Qu.: 1.900 1st Qu.:0.07000 1st Qu.: 7.00

## Median : 2.200 Median :0.07900 Median :14.00

## Mean : 2.539 Mean :0.08747 Mean :15.87

## 3rd Qu.: 2.600 3rd Qu.:0.09000 3rd Qu.:21.00

## Max. :15.500 Max. :0.61100 Max. :72.00

## total.sulfur.dioxide density pH sulphates

## Min. : 6.00 Min. :0.9901 Min. :2.740 Min. :0.3300

## 1st Qu.: 22.00 1st Qu.:0.9956 1st Qu.:3.210 1st Qu.:0.5500

## Median : 38.00 Median :0.9968 Median :3.310 Median :0.6200

## Mean : 46.47 Mean :0.9967 Mean :3.311 Mean :0.6581

## 3rd Qu.: 62.00 3rd Qu.:0.9978 3rd Qu.:3.400 3rd Qu.:0.7300

## Max. :289.00 Max. :1.0037 Max. :4.010 Max. :2.0000

## alcohol quality rating

## Min. : 8.40 3: 10 bad : 63

## 1st Qu.: 9.50 4: 53 average:1319

## Median :10.20 5:681 good : 217

## Mean :10.42 6:638

## 3rd Qu.:11.10 7:199

## Max. :14.90 8: 18

**Conclusion :**

The use of regularized approaches to model quality of white wine lead us to a new model containing 7 variables instead of the 9 variable model previously found. This new model contains volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, pH, sulphates, alcohol.Density and fixed.acidity are left out.