NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering

Academic Year: 2017-2018

System Requirements Specification Report on

'Wine quality Prediction'

Submitted by

ABHAY NAVADA	**1NT15CS007**
ANKIT DATTA	**1NT15CS028**
HARSHITH NARAHARI	**1NT15CS064**
ROSHAN BADRINATH	**1NT15CS140**

Under the able guidance of

Mrs. Manasa Gowda

Assistant Professor, Dept. of CSE

Table of Contents

Revision History

Name	**Date**	**Version**
Wine Quality Predictor	4th April 2018	1.0.0

1. Introduction

1.1 Purpose

We propose a machine learning approach to predict human wine tasting preferences. Our project focuses on some of the physiochemical properties that will be used to predict the quality of wine. A large data set, from Portugal, with white *Vinho Verde* wine data sample is considered for training and testing. We have used a Fully Connected Multi Perceptron Neural Network to classify the wine data. Our project focuses on some of the physiochemical properties that will be used to predict the quality of wine.

1.2 Document Conventions

This Document was created based on the IEEE template for System Requirement Specification Documents.

1.3 Intended Audience and Reading Suggestions

- Oenologist can use our product for analysis of wine.
- Programmers who are interested in working on the project by further developing it or fix existing bugs.

1.4 Product Scope

Our model helps in supporting the oenologist in wine tasting evaluation and production of wine. Similar techniques can help in target marketing by modeling consumer tastes from niche markets.

1.5 References

Data set is obtained from:
https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub

IEEE Template for System Requirement Specification Documents:
https://goo.gl/nsUFwy

2. Overall Description

2.1 Product Perspective

Human wine tasting is a sensory examination and evaluation of wine. There are many properties that decide the quality of wine such as color, swirl, smell and savor. There are also various physiochemical properties that decide the quality. A large data set, from Portugal, with white *Vinho Verde* wine data sample is considered for training and testing. R programming language is used for implementation of our project. R provides wide variety of tools for statistical and data analytics.

2.2 Product Functions

Tensorflow: Tensorflow is an open-source software library for numerical communication and data flow programming using data flow graphs. It is used for machine learning applications such as neural networks. It was developed by Google Brain team for internal use and was later launched under Apache 2.0 open source license for common use.

Reticulate: Interface to Python modules, classes and functions are provided by the Reticulate function. Reticulate acts as a wrapper package for R as Tensorflow and Keras are written in Python.

Keras: Keras is a high level neural network API. It enables fast experimentation of both convolutional and recurrent neural network. It’s written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It runs seamlessly on CPU as well as and CUDA supported Nvidia graphic cards.

caTools: caTools provides basic utility functions like moving window statistic function, read and write for GIF,ENVI binary files, etc and functionality for splitting data set into training and testing set.

ggplot2: ggplot2 is a data visualization package for R. It was created by Hadley Wickham in 2005. In contrast to R base graphics, ggplot2 provides functionality to add, remove and alter components in a plot with a high level of abstraction. It contains a number of defaults for web and print display of common scales and can serve as a replacement for base graphics in R

corrplot: corrplot provides graphical display of correlation matrix. Calculation of correlation between attributes of the data set can be done with the help of corrplot and the same can be used for plotting correlation plots.

2.3 User Classes and Characteristics

- Oenologist can use our product for analysis of wine.
- Programmers who are interested in working on the project by further developing it or fix existing bugs.

2.4 Operating Environment

Implementation platform:

- Windows 10 or Linux Operating System
- R-base
- R Studio

2.5 Design and Implementation Constraints

R programming language is used for implementation of our project. R provides wide variety of tools for statistical and data analytics. R Studio, an open source IDE for R programming, provides powerful coding tools and an interacting graphic environment. Tensorflow and Keras API, written in Python, are used for the implementation of the Neural Network. A package called reticulate is used which acts as an interface to Python modules. A Fully Connected Neural Network model is created using the Keras API. User Documentation.

2.6 Assumptions and Dependencies

R Packages necessary:

- Tensorflow

- Keras
- Reticulate
- corrPlot
- caTools
- ggplot2

Software to be installed:

- Tensorflow library for R

  Reference: https://tensorflow.rstudio.com/
- Keras API

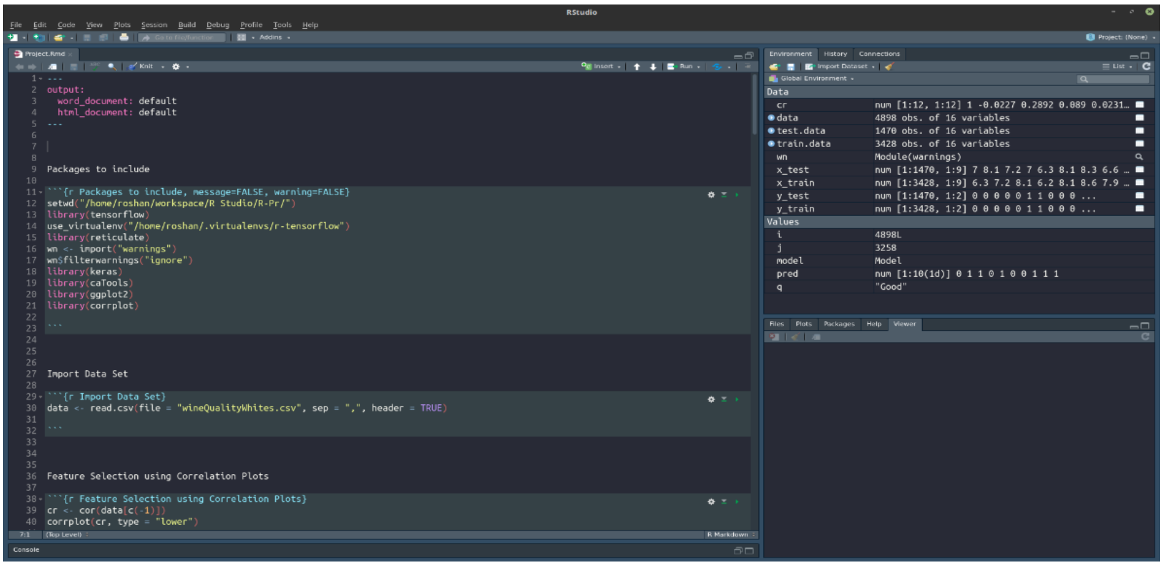
  Reference: https://keras.rstudio.com/

3. External Interface Requirements

3.1 User Interfaces

R Studio IDE:

Use the shortcut key Ctrl + O to select and open the R Markup Document.



~~~
---
output:
  word_document: default
  html_document: default
---



Packages to include

```{r Packages to include, message=FALSE, warning=FALSE}
setwd("/home/roshan/workspace/R Studio/R-Pr/")
library(tensorflow)
use_virtualenv("/home/roshan/.virtualenvs/r-tensorflow")
library(reticulate)
wn <- import("warnings")
wn$filterwarnings("ignore")
library(keras)
library(caTools)
library(ggplot2)
library(corrplot)

~~~

Import Data Set

```{r Import Data Set}
data <- read.csv(file = "wineQualityWhites.csv", sep = ",", header = TRUE)

```

Feature Selection using Correlation Plots

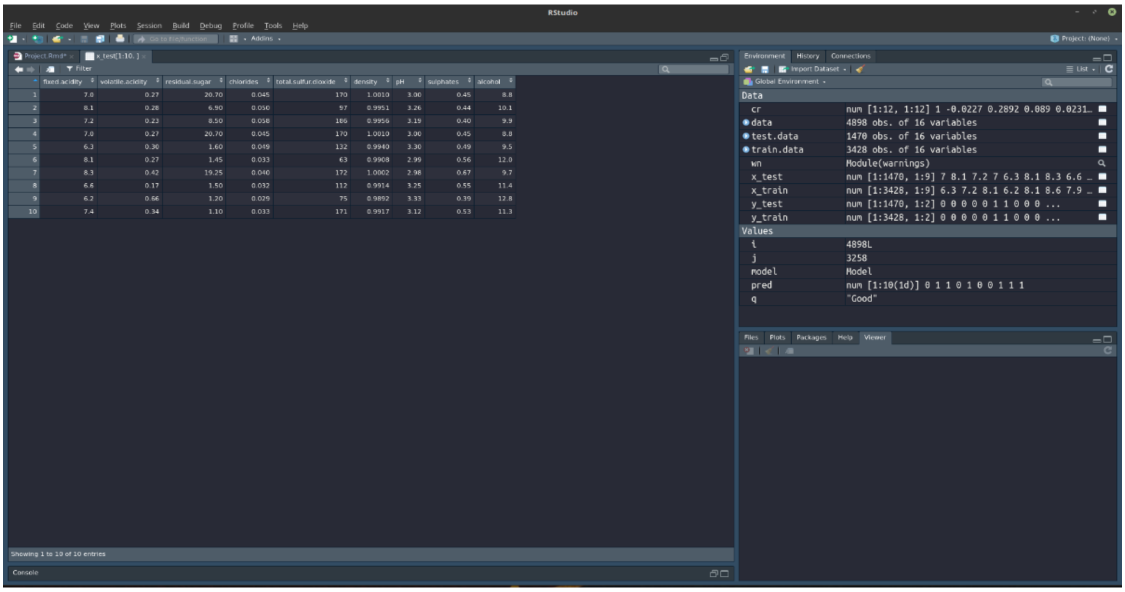
```{r Feature Selection using Correlation Plots}
cr <- cor(data[c(-1)])
corrplot(cr, type = "lower")
```

```
Data
  cr          num [1:12, 1:12] 1 -0.0227 0.2892 0.089 0.0231…
  data        4898 obs. of 16 variables
  test.data   1470 obs. of 16 variables
  train.data  3428 obs. of 16 variables
  wn          Module(warnings)
  x_test      num [1:1470, 1:9] 7 8.1 7.2 7 6.3 8.1 8.3 6.6 …
  x_train     num [1:3428, 1:9] 6.3 7.2 8.1 6.2 8.1 8.6 7.9 …
  y_test      num [1:1470, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
  y_train     num [1:3428, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
Values
  i           4898L
  j           3258
  model       Model
  pred        num [1:10(1d)] 0 1 1 0 1 0 0 1 1 1
  q           "Good"
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Project.Rmd
Knit
Insert
Run
Project: (None)
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
7:1 (Top Level)
R Markdown
Console

Overview of CSV file:

The csv file must contain the fields specified in the figure below. Name of the csv file must be ‘test.csv’.



	fixed.acidity	volatile.acidity	residual.sugar	chlorides	total.sulfur.dioxide	density	pH	sulphates	alcohol
1	7.0	0.27	20.70	0.045	170	1.0010	3.00	0.45	8.8
2	8.1	0.28	6.90	0.050	97	0.9951	3.26	0.44	10.1
3	7.2	0.23	8.50	0.058	186	0.9956	3.19	0.40	9.9
4	7.0	0.27	20.70	0.045	170	1.0010	3.00	0.45	8.8
5	6.3	0.30	1.60	0.049	132	0.9940	3.30	0.49	9.5
6	8.1	0.27	1.45	0.033	63	0.9908	2.99	0.56	12.0
7	8.3	0.42	19.25	0.040	172	1.0002	2.98	0.67	9.7
8	6.6	0.17	1.50	0.032	112	0.9914	3.25	0.55	11.4
9	6.2	0.66	1.20	0.029	75	0.9892	3.33	0.39	12.8
10	7.4	0.34	1.10	0.033	171	0.9917	3.12	0.53	11.3

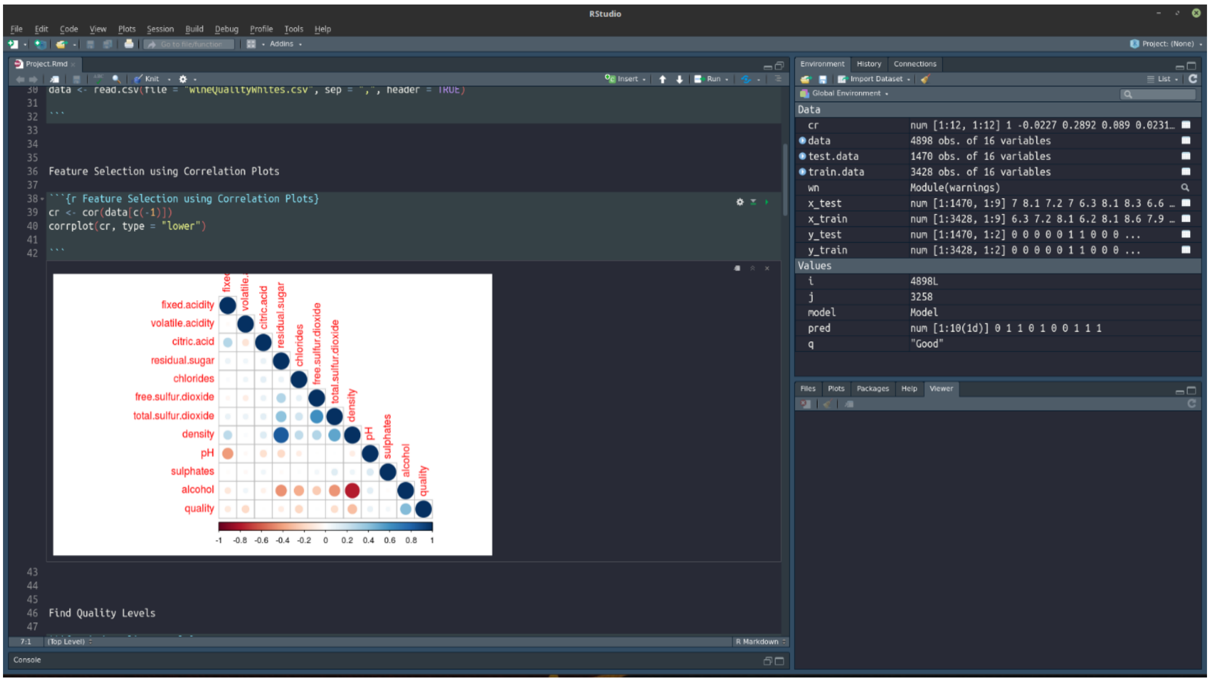
Showing 1 to 10 of 10 entries

```
Data
  cr          num [1:12, 1:12] 1 -0.0227 0.2892 0.089 0.0231…
  data        4898 obs. of 16 variables
  test.data   1470 obs. of 16 variables
  train.data  3428 obs. of 16 variables
  wn          Module(warnings)
  x_test      num [1:1470, 1:9] 7 8.1 7.2 7 6.3 8.1 8.3 6.6 …
  x_train     num [1:3428, 1:9] 6.3 7.2 8.1 6.2 8.1 8.6 7.9 …
  y_test      num [1:1470, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
  y_train     num [1:3428, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
Values
  i           4898L
  j           3258
  model       Model
  pred        num [1:10(1d)] 0 1 1 0 1 0 0 1 1 1
  q           "Good"
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Project.Rmd*
x_test[1:10,]
Filter
Project: (None)
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
Console

Correlation Plots:

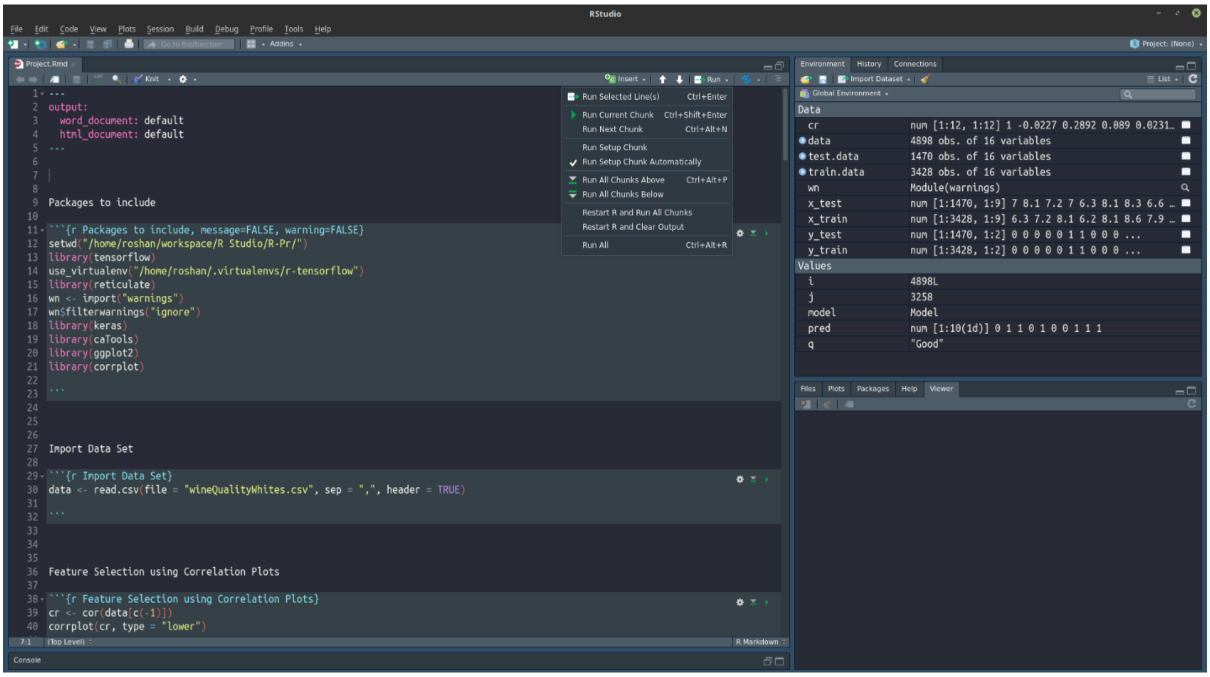
Feature selection is done using correlation plots. Only the attributes correlated to quality attribute are considered for training.



```
data <- read.csv(file = "wineQualityWhites.csv", sep = ",", header = TRUE)

```

Feature Selection using Correlation Plots

```{r Feature Selection using Correlation Plots}
cr <- cor(data[c(-1)])
corrplot(cr, type = "lower")

```

Find Quality Levels
```

```
Data
  cr           num [1:12, 1:12] 1 -0.0227 0.2892 0.089 0.0231…
  data         4898 obs. of 16 variables
  test.data    1470 obs. of 16 variables
  train.data   3428 obs. of 16 variables
  wn           Module(warnings)
  x_test       num [1:1470, 1:9] 7 8.1 7.2 7 6.3 8.1 8.3 6.6 …
  x_train      num [1:3428, 1:9] 6.3 7.2 8.1 6.2 8.1 8.6 7.9 …
  y_test       num [1:1470, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
  y_train      num [1:3428, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
Values
  i            4898L
  j            3258
  model        Model
  pred         num [1:10(1d)] 0 1 1 0 1 0 0 1 1 1
  q            "Good"
```

Execution of RMD Script in R Studio:

Click the Run button and select Run All below Chunks option to execute the code.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Project.Rmd
Knit
Insert
Run
Run Selected Line(s) Ctrl+Enter
Run Current Chunk Ctrl+Shift+Enter
Run Next Chunk Ctrl+Alt+N
Run Setup Chunk
✓ Run Setup Chunk Automatically
Run All Chunks Above Ctrl+Alt+P
Run All Chunks Below
Restart R and Run All Chunks
Restart R and Clear Output
Run All Ctrl+Alt+R
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
7:1
(Top Level)
R Markdown
Console

```
---
output:
  word_document: default
  html_document: default
---

Packages to include

```{r Packages to include, message=FALSE, warning=FALSE}
setwd("/home/roshan/workspace/R Studio/R-Pr/")
library(tensorflow)
use_virtualenv("/home/roshan/.virtualenvs/r-tensorflow")
library(reticulate)
wn <- import("warnings")
wn$filterwarnings("ignore")
library(keras)
library(caTools)
library(ggplot2)
library(corrplot)

```

Import Data Set

```{r Import Data Set}
data <- read.csv(file = "wineQualityWhites.csv", sep = ",", header = TRUE)

```

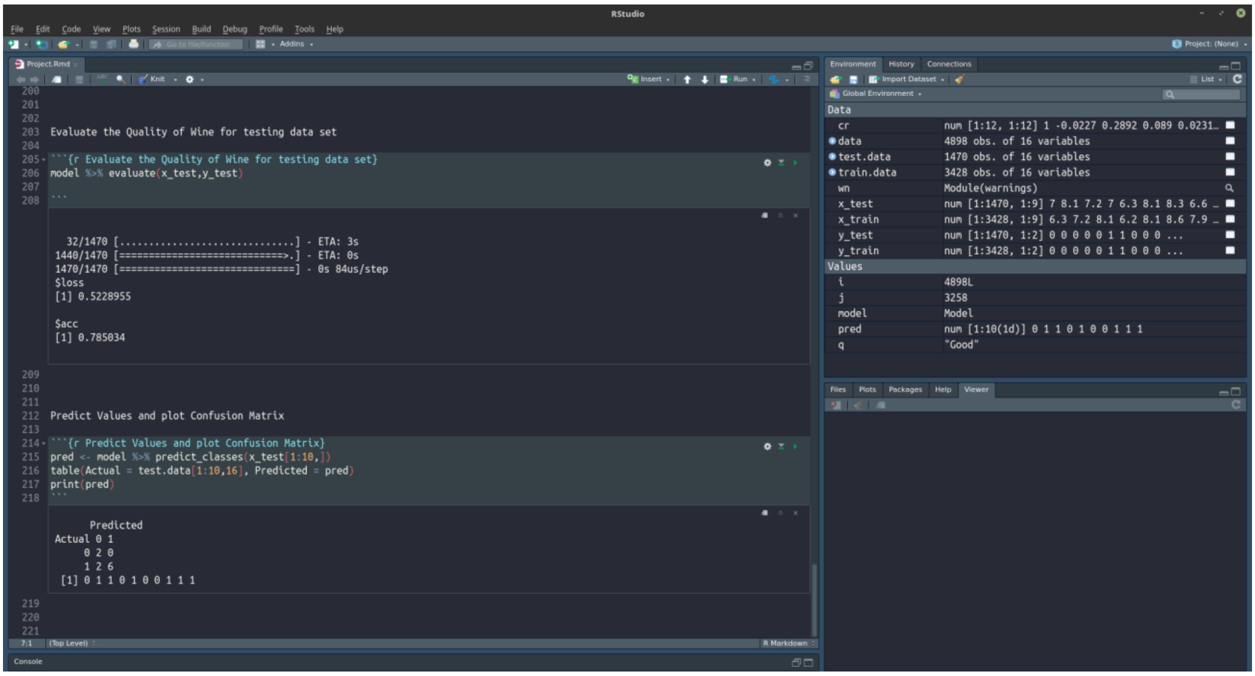
Feature Selection using Correlation Plots

```{r Feature Selection using Correlation Plots}
cr <- cor(data[c(-1)])
corrplot(cr, type = "lower")
```

```
Data
  cr           num [1:12, 1:12] 1 -0.0227 0.2892 0.089 0.0231…
  data         4898 obs. of 16 variables
  test.data    1470 obs. of 16 variables
  train.data   3428 obs. of 16 variables
  wn           Module(warnings)
  x_test       num [1:1470, 1:9] 7 8.1 7.2 7 6.3 8.1 8.3 6.6 …
  x_train      num [1:3428, 1:9] 6.3 7.2 8.1 6.2 8.1 8.6 7.9 …
  y_test       num [1:1470, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
  y_train      num [1:3428, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
Values
  i            4898L
  j            3258
  model        Model
  pred         num [1:10(1d)] 0 1 1 0 1 0 0 1 1 1
  q            "Good"
```

Prediction:

The last chunk shows the predicted values of the test data and the confusion matrix. Value 1 stands for Good quality and value 0 stands for Bad quality.



~~~
Evaluate the Quality of Wine for testing data set

```{r Evaluate the Quality of Wine for testing data set}
model %>% evaluate(x_test,y_test)

~~~
```

```
  32/1470 [..............................] - ETA: 3s
1440/1470 [============================>.] - ETA: 0s
1470/1470 [==============================] - 0s 84us/step
$loss
[1] 0.5228955

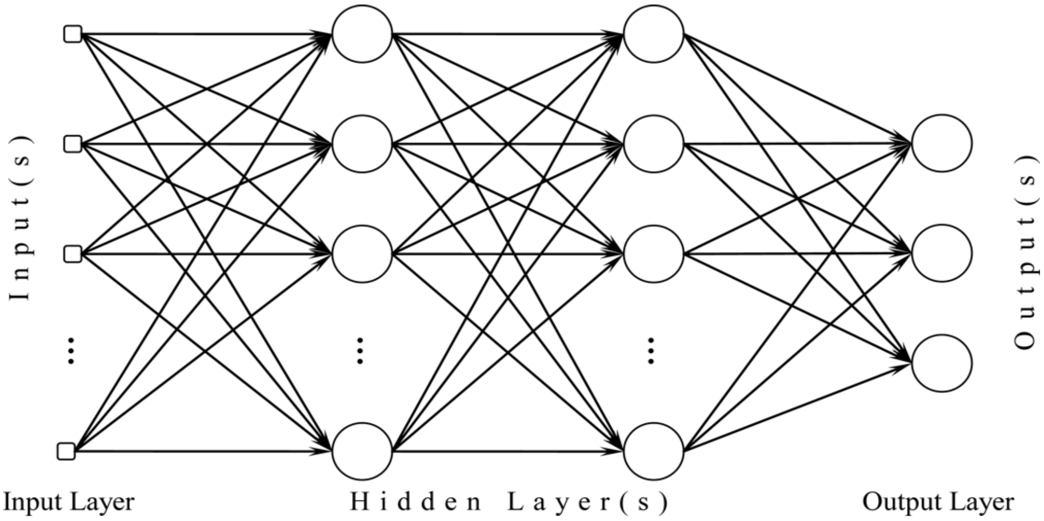
$acc
[1] 0.785034
```

```
Predict Values and plot Confusion Matrix

```{r Predict Values and plot Confusion Matrix}
pred <- model %>% predict_classes(x_test[1:10,])
table(Actual = test.data[1:10,16], Predicted = pred)
print(pred)
```
```

```
      Predicted
Actual 0 1
     0 2 0
     1 2 6
 [1] 0 1 1 0 1 0 0 1 1 1
```

```
Data
  cr          num [1:12, 1:12] 1 -0.0227 0.2892 0.089 0.0231…
  data        4898 obs. of 16 variables
  test.data   1470 obs. of 16 variables
  train.data  3428 obs. of 16 variables
  wn          Module(warnings)
  x_test      num [1:1470, 1:9] 7 8.1 7.2 7 6.3 8.1 8.3 6.6 …
  x_train     num [1:3428, 1:9] 6.3 7.2 8.1 6.2 8.1 8.6 7.9 …
  y_test      num [1:1470, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
  y_train     num [1:3428, 1:2] 0 0 0 0 0 1 1 0 0 0 ...
Values
  i           4898L
  j           3258
  model       Model
  pred        num [1:10(1d)] 0 1 1 0 1 0 0 1 1 1
  q           "Good"
```

3.2 Hardware Interfaces

The minimum hardware requirements are Intel i5 or i7 processor (CPU) or NVIDIA CUDA supported graphic card (GPU). A minimum of 8GB RAM is recommended for smooth performance.

3.3 Software Interfaces

R Studio is required for the execution and visualization. Tensorflow CPU or GPU (only for CUDA supported graphic cards) version has to be installed. Tensorflow acts as a backend tool, upon which Keras API has to be installed.
```

4. System Features

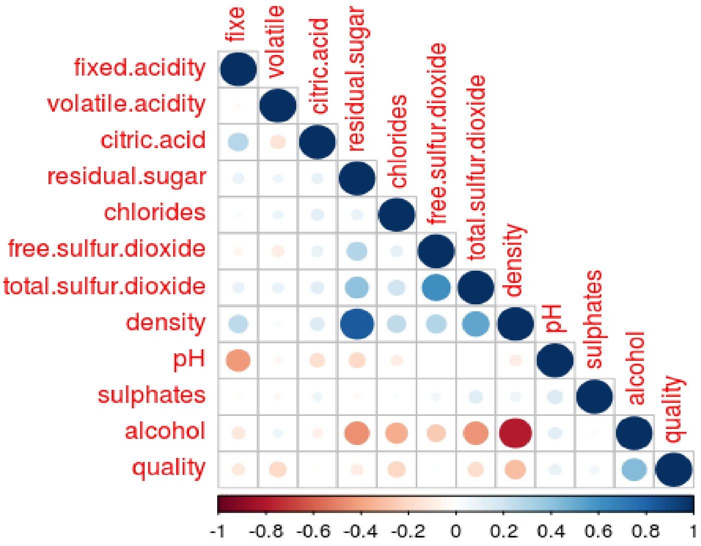
4.1 Use of Artificial Neural Network for prediction

Tensorflow and *Keras* API, written in Python, are used for the implementation of the Neural Network. A package called *reticulate* is used which acts as an interface to Python modules. A *Fully Connected Neural Network* model is created using the Keras API.

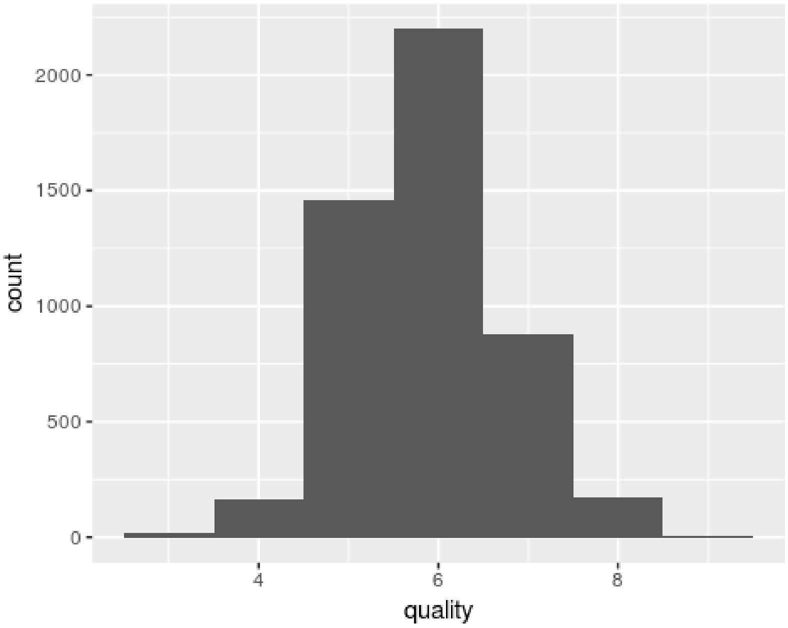


4.2 Data Visualization

- Feature Selection using Correlation Plots:



- Histogram of Quality:



4.3 Test data as CSV

The test data is obtained from a CSV file. CSV (Comma Separated Values) is a file format for data storage which looks like a text file. The information is organized with one record on each line and each field is separated by comma. The file can be managed via Microsoft Excel (or similar programs) and lists the merchant products, codes, image links, etc.

5. Other Nonfunctional Requirements

5.1 Performance Requirements

A minimum of 8GB RAM is recommended for smooth performance. Tensorflow can be installed to work on CPU or GPU. Installing GPU version increases the performance. Only CUDA supported NVIDIA graphic cards can be used to install the GPU version.

5.2 Safety Requirements

Executing multiple instances of the project is not recommended. GPU Tensorflow version throws out of memory error if multiple instances are executed.

5.3 Security Requirements

Our Project does not have any security requirements and thus any type of user can use it without any additional privileges.

5.4 Software Quality Attributes

The users must have basic knowledge of R programming and the use of R Studio. It is recommended for the users to have basic knowledge of the physiochemical properties of wine. Minimum knowledge for creating CSV file is needed.

Glossary

Related to Artificial Neural Network:

- **Neuron**: A neuron forms the basic structure of a neural network. It receives an input, processes it, and generates an output. It is sometimes called as a node.
- **Layer**: A neural network consists of an input layer, set of hidden layers, and an output layer. Each layer consists of set of nodes. The number of attributes in the data set used for prediction, i.e. set of explanatory variables, decide the number of nodes in the input layer. Number of nodes in the output layer depends on the values of the predictor variable.
- **Weight**: Synaptic weight refers to the strength or amplitude of a connection between two nodes. It is considered as a linear component.
- **Bias:** Another linear component is applied to the input, called as the bias. Bias is added to the product of the weights of the input
- **Activation Function**: It is a non-linear component applied to the linear combination of the input. It converts the input signal to an output signal based on the function. Various activation function we have used are:

  ReLU: It stands for *Rectified Linear Unit*. It is usually applied to the *hidden layers*. The function is defined as:

  $$f(x) = \max(x,0)$$

  Softmax: It is generally applied to the *output layer* for classification problem.

- **Optimizers**: Optimization algorithms help in minimizing the loss in the training process. One of the optimizer is *Adam*, which stands for *Adaptive Moment Estimation.*
- **Dropout**: Dropout is a technique used to prevent over-fitting of the network. In this technique, certain weights are randomly frozen in the hidden layers during the training process. In other words, certain neurons are dropped. Hence, it allows the neural network to find new paths.
- **Epochs**: Single iteration of all the batches of the training process is called as epochs. The number of epochs to train can be chosen. As the number of epochs increases, the accuracy increases, provided over-fitting does not occur.

Related to the data set:

- **fixed.acidity** (Tartaric acid – g/dm^3): Most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
- **volatile.acidity** (Acetic acid – g/dm^3): The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
- **citric.acid** (g/dm^3): Found in small quantities, citric acid can add 'freshness' and flavor to wines.
- **residual.sugar** (g/dm^3): The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.
- **chlorides** (Sodium Chloride – g/dm^3): The amount of salt in wine.
- **free.sulfur.dioxide** (mg/dm^3): The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.
- **total.sulphur.dioxide** (mg/dm^3): Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.
- **density** (g/cm^3): The density of water is close to that of water depending on the percent alcohol and 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** (Potassium Sulphate – g/dm^3): A wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant.
- **alcohol** (% by Volume): The percent alcohol content of wine.
- **quality**: It is the response variable, scores between 0 to 10.