Build a Model To Predict Quality Of Red Wine Using Its Attributes

TASK 1 : Convert data in excel file

At first convert data from text to column in excel sheet before importing the data. Because, unstructured data were found in the dataset. Then, save the data in csv format.

TASK 2 : Import data

Import data to R from csv format. No clean up is required as the data has already been cleaned previously.

```
wine quality <- read.csv(file. Choose(), header=T, sep=",", check.names=TRUE)
```

TASK 3: Check characteristics and data relationship

- **Check data characteristics with following codes -**

```
 head(winequality)
str(winequality)
summary(winequality)
class(winequality)
nrow(winequality)
ncol(winequality)
```

✓ **Findings** : Datatype is dataframe. Number of attributes is 12.

- **Find out missing values from dataset**

```
sapply(winequality, function(x) sum(
is.na(x))) sum(!complete.cases(winequality))
```

✓ **Findings :** There is no missing values exist in the dataset as set has already been cleaned

- **Perform following correlation**

1. Check correlation among variables

```
cor(winequality[,unlist(lapply(winequality, is.numeric))])
```

2. Check correlations and significance levels with pearson and spearman. Correlation coefficients for all possible pairs of columns of a matrix.

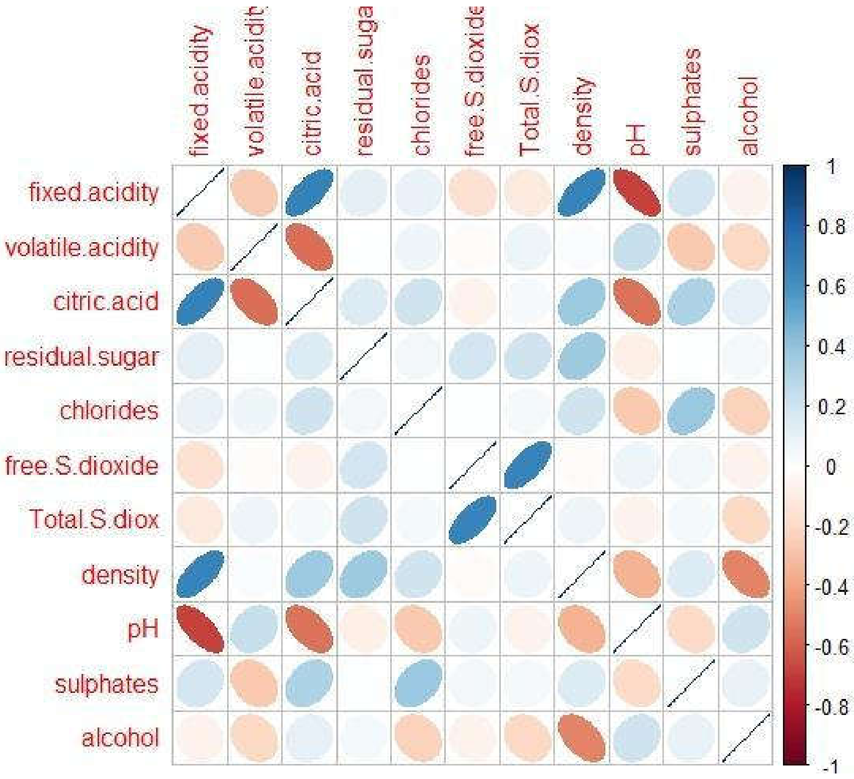
```
library(Hmisc)

rcorr(as.matrix(winequality), type=c("pearson", "spearman"))
```

3. **Visualize data**

```
library(corrplot)
library(ggplot2)

corrplot(cor(winequality[,c(1,2,3,4,5,6,7,8,9,10,11)]), method = "ellipse")
```



- **Preprocessing data**

1. **Convert characteristics of Predicted variable from integer to numeric**

```
winequality$qualiy <-
as.numeric(winequality$qualiy) str(winequality)
redwine<-winequality
```

2. **Convert Predicted variable from multi-class to binary class**

```
redwine$qualiy[redwine$qualiy <= 5] <- 1
redwine$qualiy[redwine$qualiy > 5] <- 0
str(redwine$qualiy)
```

Model perform better if convert data from multi-class to binary. Here, create 2 classes high=1 (score: <=5)& low=0(score>5)

3. **Change class of Predicted variable from numeric to factor**

```
redwine$qualiy <- as.factor(redwine$qualiy) str(redwine)
```

TASK 4: Reason behind selecting the Model and Define Task, Experience and Performance Criteria

Considering the given dataset, logistic regression model has been selected to predict. Here, model is using to find out the relationship among attributes of red wine data. i.e. which are the attributes influenced to increase the quality of red wine. Also, here is a column named “qualiy” in the dataset with ratings which also leads to select logistic regression model. To avoid the complexity of the model, multiclass of “qualiy” variable has been converted to binary class. Also, choose the model as classification. Supervised learning algorithm deals with structured data and classification model works on structured data.

In wine industry two elements always consider to conduct to assess the quality of wine:

1. **Physicochemical test (lab based test ;i.e. Ph level, % of alcohol):**
   Based on the information on available data, we have to predict quality of wine.

2. **Sensory test (taste preference performed by human experts):**
   Quality determine by customer satisfaction.
   Customer grades wine using blind test method.

Here, Physicochemical test has been considered as from given dataset it is possible to fine out the pattern in attribute that affect the quality of wine.

Define followings -

Task : classify the attributes of red wine that affects quality of red wine or not(high/low).

Experience : Collection of previous data where already measuring the quality of wine with classes.

Performance Criteria : Use for classification accuracy. Find out how many correctly indentify that attributes are affecting in quality of alcohol.

TASK 5 : Algorithms and Experiment

Perform Logistic Regression, “rWeka” package has been using here.

- **Use R Formula before run the model, no need to write variables. names every time**

```
formula_text <- paste(names(redwine)[12] ,"~",
                      paste(names(redwine[1:11]), collapse="+"))
formula <- as.formula(formula_text)
```

- **Fit checks if model on the same data**

```
fit <- glm(formula,data=redwine,family=binomial())
summary(fit)

library("RWeka")
library("ROCR")
library("caret")
library("e1071")
library("rJava")

weka_fit <- Logistic(formula, data=wine)
evaluate_Weka_classifier(weka_fit, numFolds = 10)
```

Findings:

=== Confusion Matrix ===

```
   a    b  <-- classified as
 639  216 |  a = 0
 199  545 |  b = 1
```

From the output it has been concluded that 545 is the corrected number out of 1599 instances. It means only 34% has correctly identified the attributes that influenced the quality of Alcohol.

TASK 6. Comparing with previous study (i.e. Modeling wine preferences by data mining from physicochemical properties)

In the study conducted by Cortez et al. (2009) the same dataset (red wine) was used for comparing three regression techniques: multiple regression, artificial neural network (ANN) and support vector machine (SVM) methods. The methods were applied for modeling taste preferences (i.e. quality) based on analytical data. The methods were applied under a computationally efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods.

They used sensitivity analysis to extract knowledge from the NN/SVM models, given in terms of relative importance of the inputs and reduced the number of inputs. Simultaneous variable and model selection scheme was also used, where the variable selection was guided by sensitivity analysis and the model selection was based on parsimony search that starts from a reasonable value and is stopped when the generalization estimate decreases.

They used Kappa statistic as a performance criteria. The Kappa statistic measures the accuracy when compared with a random classifier (which presents a Kappa value of 0%). The higher the statistic, the more accurate the result. The most practical tolerance values are T = 0:5 and T = 1:0. The former tolerance rounds the regression response into the nearest class, while the latter accepts a response that is correct within one of the two closest classes (e.g. a 3.1 value can be interpreted as grade 3 or 4 but not 2 or 5). For T = 0:5, Kappa values for the SVM, NN, and MLR were 32.2, 32.5, and 38.7 pp. The NN is quite similar to MR in the red wine modeling, thus similar performances were achieved. In our assignment, the calculated Kappa statistics was 0.48 which shows more accuracy of our model comparing the developed model by Cortez et al. (2009).

Regarding the variable selection, the average number of deleted inputs ranges from 0.9 to 1.8, showing that most of the physicochemical tests used are relevant. A detailed analysis of the SVM classification results is presented by the average confusion matrixes for T = 0:5. Most of the values are close to the diagonals (in bold), denoting a good fit by the model. The true predictive accuracy for each class is given by the precision metric. This statistic is important in practice, since in a real deployment setting the actual values are unknown and all predictions within a given column would be treated the same. For a tolerance of 0.5, the SVM red wine accuracies are around 57.7 to 67.5% in the intermediate grades (5 to 7) and very low (0%/20%) for the extreme classes (3, 8 and 4), which are less frequent.

In our study, the confusion matrix showed that correctly classified instances were around 74%. It should be noted that the whole 11 inputs are relevant, since in each simulation different sets of variables can be selected. In several cases, the obtained results confirm the oenological theory. For instance, an increase in the alcohol (4th and 2nd most relevant factor) tends to result in a higher quality wine. The volatile acidity has a negative impact, since acetic acid is the key ingredient in vinegar. The most intriguing result is the high importance of sulphates . Ontologically this result could be very interesting. An increase in sulphates might be related to the fermenting nutrition, which is very important to improve the wine aroma. The relative rank was as follows:

Sulphates>pH> total sulfur dioxide> alcohol> volatile acidity> free sulfur dioxide> fixed acidity> residual sugar> chlorides> density> citric acid In our study, the ranks were different and were as follows regarding the weights developed by logistic model:

volatile acidity>chlorides>sulphates>citric acid>alcohol>free sulfur dioxide>total sulfur dioxide.