NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

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

Department of Computer Science and Engineering

Academic Year: 2017-2018

Data Warehouse and Mining Project on

‘WINE QUALITY PREDICTION USING DECISION TREES’

Submitted by

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

Under the able guidance of

Mrs. Chaitra H V

Associate Professor, Dept. of CSE

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION)

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

YELAHANKA, BANGALORE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the course project in DWM in VI semester report entitled

Wine Quality Prediction using Decision Trees

is an authentic record of the project carried out by

TEAM MEMBERS:

Abhay Navada (1NT15CS007)
Ankit Datta (1NT15CS028)
Harshith Narahari (1NT15CS064)
Roshan Badrinath (1NT15CS140)

SIGNATURE OF GUIDE	**SIGNATURE OF HOD**
…………………………….	……………………………...
Mrs. CHAITRA H V	**DR. THIPPESWAMY M N**
ASSOCIATE PROFESSOR	**HEAD OF DEPARTMENT**
CSE DEPT, NMIT	**CSE DEPT, NMIT**

ACKNOWLEDGEMENT

We are extremely grateful to our HOD, **Dr. Thippeswamy** who extended his support towards our project. We remain indebted to our lecturer **Mrs. Chaitra HV** for her constant support in the Design, Implementation and Evaluation of the project. We are thankful to her for constructive criticism and valuable suggestions, which benefited us a lot while developing the project, also without whom we might not have been able to accomplish this project. Finally, we gratefully acknowledge the support, encouragement and patience of our friends.

ABSTRACT

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.

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 Decision Trees to classify the wine data.

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.

CONTENTS OF THE PROJECT

INTRODUCTION

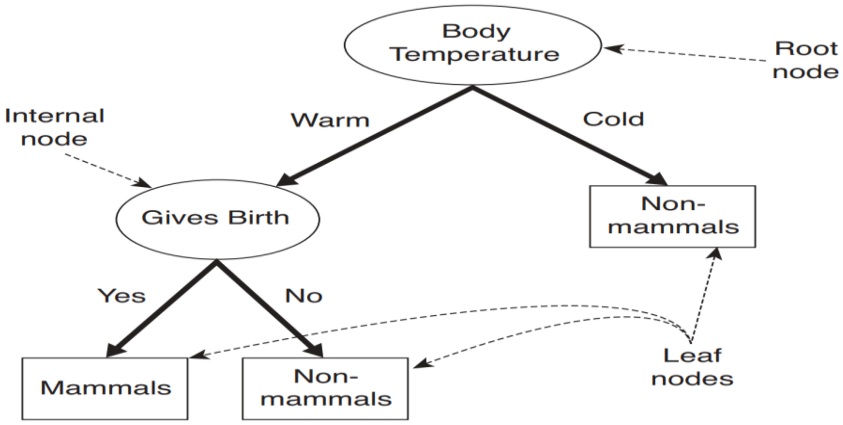
R programming language is used for implementation of our project. R provides wide variety of tools for statistical and data analytics. *R Studio*, an open source IDE for R programming, provides powerful coding tools and an interacting graphic environment.

A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it's also widely used in machine learning.

There are three types on nodes:

1. A root node that has no incoming edges but has zero or more outgoing edges.
2. Internal nodes, each of which have exactly one incoming edge and two outgoing edges.
3. Leaf or terminal nodes, which have exactly one incoming node and no outgoing nodes.



The above figure shows a decision tree to find out whether given creature is a mammal or not. If flamingo is taken as the test data, it will be observed that it has warm body temperature but it doesn't give birth (it lays eggs). Hence it is a non-mammal.

DESCRIPTION OF PACKAGES

Different packages that have been used are:

party: Party is a computational toolbox for recursive partitioning. The core of the package is ctree(), an implementation of conditional inference trees which embed tree-structured regression models into a well-defined theory of conditional inference procedures. This non-parametric class of regression trees is applicable to all kinds of regression problems, including nominal, ordinal, numeric, censored as well as multivariate response variables and arbitrary measurement scales of the covariates.

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. It also provides functionality for fast calculation of AUC, LogitBoost classifier, base64 encoder/decoder, round-off-error-free sum and cumulative sum, etc.

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.

DESCRIPTION OF THE DATA SET

A large data set, from Portugal, with white *Vinho Verde* wine data sample is considered for training and testing. The data set consists of 12 attributes, out of which 11 are considered as explanatory variables. The 12th attribute, quality, is considered as the response variable. Various attributes available in the data set are:

- *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.

CODE DESCRIPTION AND OUTPUT

Packages to include

```
setwd("/home/roshan/workspace/R Studio/DataMining/")
library(party)
library(corrplot)
library(caTools)
library(ggplot2)
```

Import Data Set

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

Feature Selection using Correlation Plots

```
cr <- cor(data[c(-1)])
corrplot(cr, type = "lower")
```

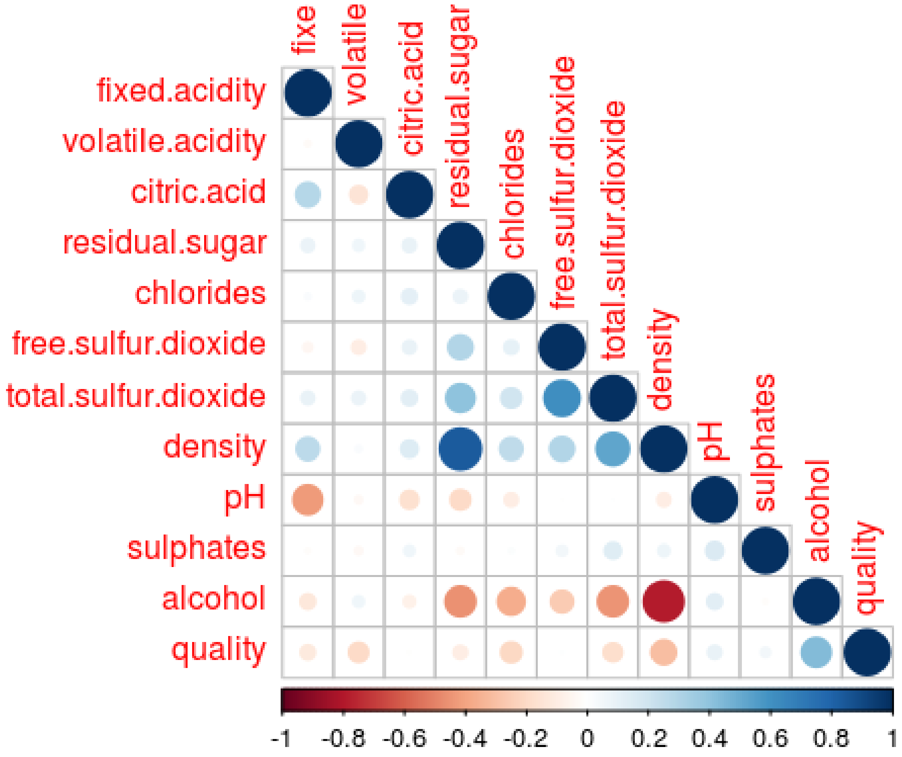


Fig 1: Correlation Plot

Find Quality Levels

```
data$quality.factor <- as.factor(data$quality)

levels(data$quality.factor)

## [1] "3" "4" "5" "6" "7" "8" "9"
```

Histogram of Quality

```
ggplot(data = data, aes(x = quality)) +
  geom_histogram(binwidth = 1)
```

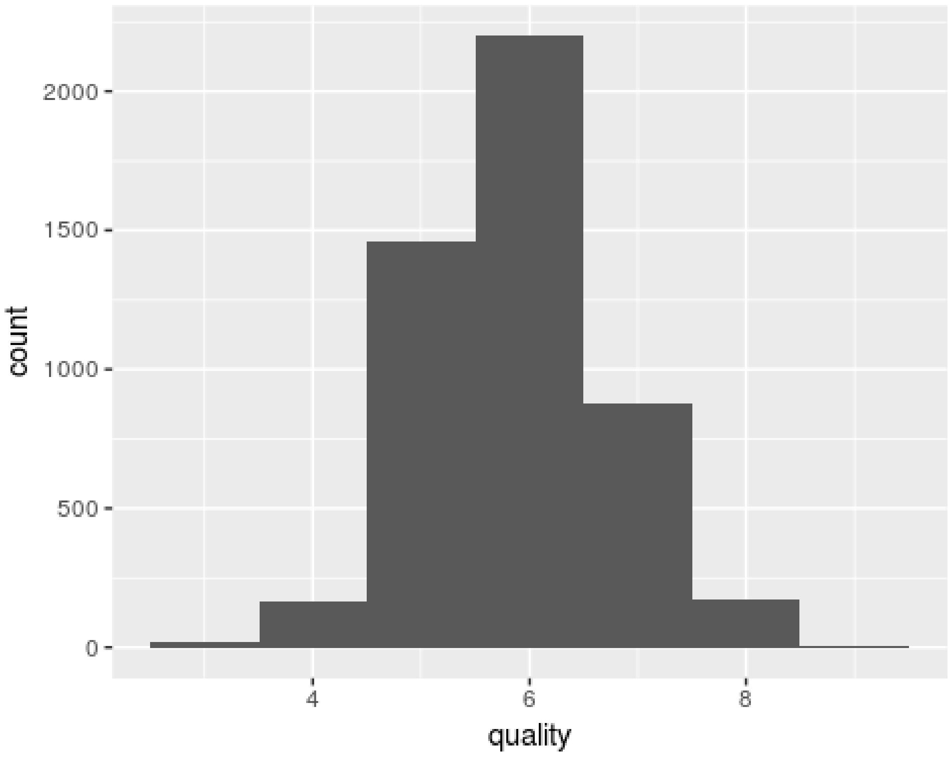


Fig 2: Histogram

Add Binary Ordered Quality Attribute

```
for(i in 1:nrow(data)){
  if(data$quality[[i]] < 6){
    data$quality.order[[i]] <- "Bad"
  }else{
    data$quality.order[[i]] <- "Good"
  }
}
```

Count

```
i <- j <- 0
for(q in data$quality.order){
  if(q == "Bad"){
    i <- i+1
  }else{
    j <- j+1
  }
}
print(i)

## [1] 1640
```

```
print(j)

## [1] 3258

data$quality.order <- factor(factor(data$quality.order), levels = c("Bad",
"Good"))
```

Convert Quality to Binary Attribute

```
for(i in 1:nrow(data)){
  if(data$quality.order[[i]] == "Bad"){
    data$quality.num[[i]] <- 0
  }else{
    data$quality.num[[i]] <- 1
  }
}
```

Train and Testing Data Sets

```
value <- sample.split(data$X, SplitRatio = 0.7)
train.data <- subset(data, value == TRUE)
test.data <- subset(data, value == FALSE)

write.csv(train.data, file = "train_data.csv")
write.csv(test.data, file = "test_data.csv")
```

Loading existing training and testing data sets

```
train.data <- read.csv(file = "train_data.csv", sep = ",", header = TRUE)
test.data <- read.csv(file = "test_data.csv", sep = ",", header = TRUE)
```

Fitting a Decision Tree Model and plotting the tree

```
d.tree <- ctree(quality.num ~
                  fixed.acidity
                + volatile.acidity
                + residual.sugar
                + chlorides
                + total.sulfur.dioxide
                + density
                + pH
                + sulphates
                + alcohol
                , data = train.data)


plot(d.tree, type = "simple")
```

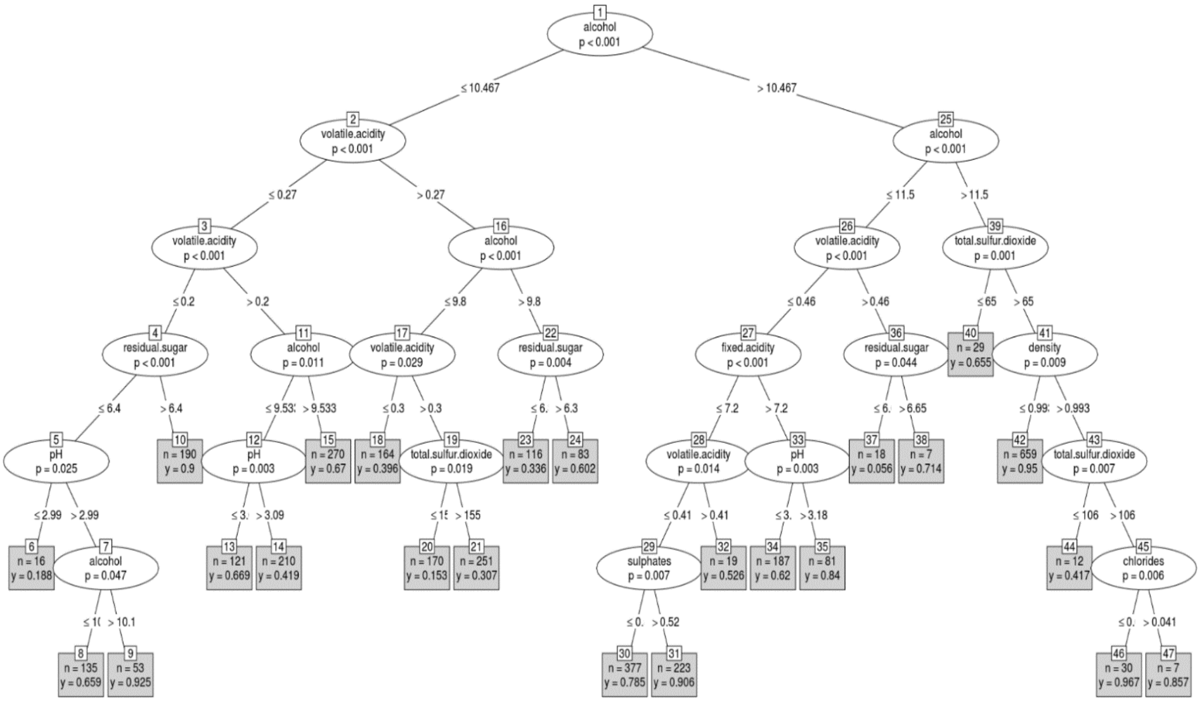


Fig 3: Decision Tree

Predictions for the Test Data Set

```
prob <- predict(d.tree, newdata = test.data, type = "response")

classes <- c()
for (i in prob) {
  if(i >= 0.5) {
    classes <- c(classes,1)
  } else {
    classes <- c(classes,0)
  }
}
```

Plotting Confusion Matrix and finding the Accuracy

```
conf.matrix <- table(Actual = test.data[,16], Predicted = classes)
print(conf.matrix)

##       Predicted
## Actual   0   1
##   Bad  276 238
##   Good 148 808

acc <- (conf.matrix[1,1]+conf.matrix[2,2])/length(test.data[,16])
acc <- round(acc, digits = 4)
print(paste("Accuracy = ",acc))

## [1] "Accuracy =  0.7374"
```

BIBLIOGRAPHY

Data set is obtained from:

https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub

Links referred:

https://www.r-bloggers.com/

https://cran.r-project.org/web/packages/