Data Analytics Assignment 2

Code ▾

Name: Lim Yu Jin

import necessary libraries:

Hide

```
library(dplyr)
library(tidyr)
install.packages("tree")
library(tree)
install.packages("e1071")
library(e1071)
install.packages(("ROCR"))
library(ROCR)
install.packages("randomForest")
library(randomForest)
install.packages("adabag")
library(adabag)
install.packages("rpart")
library(rpart)
install.packages("neuralnet")
library(neuralnet)
install.packages("pROC")
library(pROC)
install.packages("neuralnet")
library(neuralnet)
library(nnet)
install.packages('class')
library(class)
```

Question 1

lets have a quick look at the dataset.

Hide

```
humid <-  read.csv("HumidPredict2023D.csv")
head(humid)
```

	Y…	Location	MinT…	MaxT…	Rainfall	Evaporation	Sunshi…	WindGustDir	WindGustSp
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<
1	2018	36	2.8	19.7	0.0	*NA*	*NA*	*NA*	
2	2013	36	7.3	19.6	0.0	3.6	11.9	WSW	
3	2019	36	21.2	34.8	0.4	*NA*	*NA*	*NA*	
4	2019	*NA*	7.1	16.7	0.0	*NA*	*NA*	SW	
5	2019	8	13.2	22.7	4.8	1.2	8.6	ESE	

	Y…	Location	MinT…	MaxT…	Rainfall	Evaporation	Sunshi…	WindGustDir	WindGustSp
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<
6	2017	1	10.3	22.1	0.0	*NA*	*NA*	NW	

6 rows | 1-10 of 22 columns

Summary of the dataset containing the mean, min and max of each attribute.

Hide

```
summary(humid)
```

```
      Year         Location        MinTemp         MaxTemp         Rainfall        Evaporatio
n       Sunshine     WindGustDir       
 Min.   :2007   Min.   : 1.00   Min.   :-8.70   Min.   :-4.10   Min.   :  0.000   Min.   :  
0.0   Min.   : 0.00   Length:100000     
 1st Qu.:2011   1st Qu.:12.00   1st Qu.: 7.40   1st Qu.:17.90   1st Qu.:  0.000   1st Qu.:  
2.6   1st Qu.: 4.90   Class :character  
 Median :2014   Median :25.00   Median :11.80   Median :22.60   Median :  0.000   Median :  
4.8   Median : 8.50   Mode  :character  
 Mean   :2014   Mean   :24.85   Mean   :11.99   Mean   :23.21   Mean   :  2.215   Mean   :  
5.5   Mean   : 7.65                     
 3rd Qu.:2017   3rd Qu.:37.00   3rd Qu.:16.70   3rd Qu.:28.20   3rd Qu.:  0.600   3rd Qu.:  
7.4   3rd Qu.:10.60                     
 Max.   :2019   Max.   :49.00   Max.   :33.90   Max.   :48.20   Max.   :371.000   Max.   :10
3.6   Max.   :14.50                     
 NA's   :1031   NA's   :1014    NA's   :2221    NA's   :2046    NA's   :3367      NA's   :484
57   NA's   :52812                      
 WindGustSpeed     WindDir9am          WindDir3pm         WindSpeed9am    WindSpeed3pm    Pres
sure9am      Pressure3pm        Cloud9am    
 Min.   :  6.00   Length:100000      Length:100000      Min.   : 0.00   Min.   : 0.00   Min.   
: 979.1   Min.   : 978.9   Min.   :0.0    
 1st Qu.: 31.00   Class :character   Class :character   1st Qu.: 7.00   1st Qu.:13.00   1st Q
u.:1013.2   1st Qu.:1010.7   1st Qu.:1.0    
 Median : 39.00   Mode  :character   Mode  :character   Median :13.00   Median :19.00   Media
n :1017.8   Median :1015.4   Median :5.0    
 Mean   : 40.16                                         Mean   :14.03   Mean   :18.74   Mean   
:1017.9   Mean   :1015.4   Mean   :4.5    
 3rd Qu.: 48.00                                         3rd Qu.:19.00   3rd Qu.:24.00   3rd Q
u.:1022.7   3rd Qu.:1020.2   3rd Qu.:7.0    
 Max.   :135.00                                         Max.   :87.00   Max.   :87.00   Max.   
:1041.1   Max.   :1040.1   Max.   :9.0    
 NA's   :8304                                           NA's   :2598    NA's   :3885    NA's   
:11663    NA's   :11641    NA's   :41671  
    Cloud3pm        Temp9am         Temp3pm       RainToday            RISK_MM             MH
T         
 Min.   :0.00    Min.   :-6.00   Min.   :-5.10   Length:100000      Min.   :  0.000   Min.   
:0.00   
 1st Qu.:2.00    1st Qu.:12.20   1st Qu.:16.60   Class :character   1st Qu.:  0.000   1st Q
u.:0.00   
 Median :5.00    Median :16.60   Median :21.10   Mode  :character   Median :  0.000   Median 
:0.00   
 Mean   :4.52    Mean   :16.88   Mean   :21.68                      Mean   :  2.204   Mean   
:0.49   
 3rd Qu.:7.00    3rd Qu.:21.50   3rd Qu.:26.40                      3rd Qu.:  0.600   3rd Q
u.:1.00   
 Max.   :9.00    Max.   :40.20   Max.   :46.40                      Max.   :371.000   Max.   
:1.00   
 NA's   :44391   NA's   :2313    NA's   :3920                       NA's   :3430      NA's   
:5686   
```

Structure of the data set which tells us the class for each attribute.

Hide

```
str(humid)
```

```
'data.frame':	100000 obs. of  22 variables:
 $ Year         : int  2018 2013 2019 2019 2019 2017 2018 2009 2014 2010 ...
 $ Location     : int  36 36 36 NA 8 1 18 6 18 28 ...
 $ MinTemp      : num  2.8 7.3 21.2 7.1 13.2 10.3 8.8 2.4 11.7 13.9 ...
 $ MaxTemp      : num  19.7 19.6 34.8 16.7 22.7 22.1 28 11.8 22.3 18.9 ...
 $ Rainfall     : num  0 0 0.4 0 4.8 0 0 0.8 0 0.2 ...
 $ Evaporation  : num  NA 3.6 NA NA 1.2 NA NA NA NA 4.2 ...
 $ Sunshine     : num  NA 11.9 NA NA 8.6 NA NA NA NA 7 ...
 $ WindGustDir  : chr  NA "WSW" NA "SW" ...
 $ WindGustSpeed: int  NA 41 NA 28 31 63 31 39 22 43 ...
 $ WindDir9am   : chr  NA "W" NA "SSW" ...
 $ WindDir3pm   : chr  NA "WSW" NA "SSW" ...
 $ WindSpeed9am : int  NA NA NA 9 4 24 6 22 2 19 ...
 $ WindSpeed3pm : int  NA 26 NA 15 11 13 13 17 13 19 ...
 $ Pressure9am  : num  NA 1017 NA 1025 1030 ...
 $ Pressure3pm  : num  NA 1016 NA 1024 1028 ...
 $ Cloud9am     : int  NA 0 NA 8 7 NA NA 6 NA 4 ...
 $ Cloud3pm     : int  NA 1 NA 6 3 NA NA 8 NA 6 ...
 $ Temp9am      : num  11.4 12.8 25 10.5 17.2 20.1 16.6 8.1 14.1 17.5 ...
 $ Temp3pm      : num  16.1 18.7 23.3 16 20.2 15.3 26.5 10.1 21.2 17.1 ...
 $ RainToday    : chr  "No" "No" "No" "No" ...
 $ RISK_MM      : num  2.4 0 2 0 3.8 5.6 0 6.2 0 0 ...
 $ MHT          : int  1 0 0 0 1 1 1 0 0 1 ...
```

dimension of the dataset, which is 100000 rows, 22 columns

Hide

```
dim(humid)
```

```
[1] 100000     22
```

Hide

```
more_humid_days  <-  nrow(humid[humid$RainToday  ==  'Yes', ])
no_humid_days <- nrow(humid[humid$RainToday == 'No', ])

proportion  <-  more_humid_days/  no_humid_days
proportion
```

```
[1] 0.3036415
```

proportion of days is more humid compared to those where it is less humid is 0.3036415

Hide

```
humid1 <- select(humid, -Year)
str(humid1)
```

```
'data.frame':	100000 obs. of  21 variables:
 $ Location     : int  36 36 36 NA 8 1 18 6 18 28 ...
 $ MinTemp      : num  2.8 7.3 21.2 7.1 13.2 10.3 8.8 2.4 11.7 13.9 ...
 $ MaxTemp      : num  19.7 19.6 34.8 16.7 22.7 22.1 28 11.8 22.3 18.9 ...
 $ Rainfall     : num  0 0 0.4 0 4.8 0 0 0.8 0 0.2 ...
 $ Evaporation  : num  NA 3.6 NA NA 1.2 NA NA NA NA 4.2 ...
 $ Sunshine     : num  NA 11.9 NA NA 8.6 NA NA NA NA 7 ...
 $ WindGustDir  : chr  NA "WSW" NA "SW" ...
 $ WindGustSpeed: int  NA 41 NA 28 31 63 31 39 22 43 ...
 $ WindDir9am   : chr  NA "W" NA "SSW" ...
 $ WindDir3pm   : chr  NA "WSW" NA "SSW" ...
 $ WindSpeed9am : int  NA NA NA 9 4 24 6 22 2 19 ...
 $ WindSpeed3pm : int  NA 26 NA 15 11 13 13 17 13 19 ...
 $ Pressure9am  : num  NA 1017 NA 1025 1030 ...
 $ Pressure3pm  : num  NA 1016 NA 1024 1028 ...
 $ Cloud9am     : int  NA 0 NA 8 7 NA NA 6 NA 4 ...
 $ Cloud3pm     : int  NA 1 NA 6 3 NA NA 8 NA 6 ...
 $ Temp9am      : num  11.4 12.8 25 10.5 17.2 20.1 16.6 8.1 14.1 17.5 ...
 $ Temp3pm      : num  16.1 18.7 23.3 16 20.2 15.3 26.5 10.1 21.2 17.1 ...
 $ RainToday    : chr  "No" "No" "No" "No" ...
 $ RISK_MM      : num  2.4 0 2 0 3.8 5.6 0 6.2 0 0 ...
 $ MHT          : int  1 0 0 0 1 1 1 0 0 1 ...
```

Question 2

null values in rainfall and Evaporation attribute modify to 0

removes rows that containing null values

Attribute Raintoday yes = 1, no = 0

Hide

```
humid1$Rainfall[is.na(humid1$Rainfall)]  =  0
humid1$Evaporation[is.na(humid1$Evaporation)]  =  0
humid1 <- na.omit(humid1)

humid1$RainToday[humid1$RainToday  ==  'Yes'] <-  1
humid1$RainToday[humid1$RainToday == 'No'] <- 0
```

Wind Direction N = 0, NNE = 1, NE = 2, ENE = 3, E = 4, ESE = 5, SE = 6, SSE = 7, S = 8, SSW = 9, SW = 10, WSW = 11, W = 12, WNW = 13, NW = 14, NNW = 15

Hide

```
humid1$WindGustDir[humid1$WindGustDir == 'N'] <- 0
humid1$WindGustDir[humid1$WindGustDir == 'NNE'] <- 1
humid1$WindGustDir[humid1$WindGustDir == 'NE'] <- 2
humid1$WindGustDir[humid1$WindGustDir == 'ENE'] <- 3
humid1$WindGustDir[humid1$WindGustDir == 'E'] <- 4
humid1$WindGustDir[humid1$WindGustDir == 'ESE'] <- 5
humid1$WindGustDir[humid1$WindGustDir == 'SE'] <- 6
humid1$WindGustDir[humid1$WindGustDir == 'SSE'] <- 7
humid1$WindGustDir[humid1$WindGustDir == 'S'] <- 8
humid1$WindGustDir[humid1$WindGustDir == 'SSW'] <- 9
humid1$WindGustDir[humid1$WindGustDir == 'SW'] <- 10
humid1$WindGustDir[humid1$WindGustDir == 'WSW'] <- 11
humid1$WindGustDir[humid1$WindGustDir == 'W'] <- 12
humid1$WindGustDir[humid1$WindGustDir == 'WNW'] <- 13
humid1$WindGustDir[humid1$WindGustDir == 'NW'] <- 14
humid1$WindGustDir[humid1$WindGustDir == 'NNW'] <- 15

humid1$WindDir9am[humid1$WindDir9am == 'N'] <- 0
humid1$WindDir9am[humid1$WindDir9am == 'NNE'] <- 1
humid1$WindDir9am[humid1$WindDir9am == 'NE'] <- 2
humid1$WindDir9am[humid1$WindDir9am == 'ENE'] <- 3
humid1$WindDir9am[humid1$WindDir9am == 'E'] <- 4
humid1$WindDir9am[humid1$WindDir9am == 'ESE'] <- 5
humid1$WindDir9am[humid1$WindDir9am == 'SE'] <- 6
humid1$WindDir9am[humid1$WindDir9am == 'SSE'] <- 7
humid1$WindDir9am[humid1$WindDir9am == 'S'] <- 8
humid1$WindDir9am[humid1$WindDir9am == 'SSW'] <- 9
humid1$WindDir9am[humid1$WindDir9am == 'SW'] <- 10
humid1$WindDir9am[humid1$WindDir9am == 'WSW'] <- 11
humid1$WindDir9am[humid1$WindDir9am == 'W'] <- 12
humid1$WindDir9am[humid1$WindDir9am == 'WNW'] <- 13
humid1$WindDir9am[humid1$WindDir9am == 'NW'] <- 14
humid1$WindDir9am[humid1$WindDir9am == 'NNW'] <- 15

humid1$WindDir3pm[humid1$WindDir3pm == 'N'] <- 0
humid1$WindDir3pm[humid1$WindDir3pm == 'NNE'] <- 1
humid1$WindDir3pm[humid1$WindDir3pm == 'NE'] <- 2
humid1$WindDir3pm[humid1$WindDir3pm == 'ENE'] <- 3
humid1$WindDir3pm[humid1$WindDir3pm == 'E'] <- 4
humid1$WindDir3pm[humid1$WindDir3pm == 'ESE'] <- 5
humid1$WindDir3pm[humid1$WindDir3pm == 'SE'] <- 6
humid1$WindDir3pm[humid1$WindDir3pm == 'SSE'] <- 7
humid1$WindDir3pm[humid1$WindDir3pm == 'S'] <- 8
humid1$WindDir3pm[humid1$WindDir3pm == 'SSW'] <- 9
humid1$WindDir3pm[humid1$WindDir3pm == 'SW'] <- 10
humid1$WindDir3pm[humid1$WindDir3pm == 'WSW'] <- 11
humid1$WindDir3pm[humid1$WindDir3pm == 'W'] <- 12
humid1$WindDir3pm[humid1$WindDir3pm == 'WNW'] <- 13
humid1$WindDir3pm[humid1$WindDir3pm == 'NW'] <- 14
humid1$WindDir3pm[humid1$WindDir3pm == 'NNW'] <- 15

humid1$WindGustDir <- as.numeric(humid1$WindGustDir)
humid1$WindDir9am <- as.numeric(humid1$WindDir9am)
humid1$WindDir3pm <- as.numeric(humid1$WindDir3pm)
```

```
humid1$RainToday  <-  as.numeric(humid1$RainToday)
str(humid1)
```

```
'data.frame':	30414 obs. of  21 variables:
 $ Location     : int  8 28 33 36 46 23 38 32 21 33 ...
 $ MinTemp      : num  13.2 13.9 11.7 8.5 19.9 7.2 11.6 18.1 13.8 14.9 ...
 $ MaxTemp      : num  22.7 18.9 31.1 20 29.2 15 19.2 24.2 36.7 38.3 ...
 $ Rainfall     : num  4.8 0.2 0 0.2 7.6 0.8 0 0 0 0 ...
 $ Evaporation  : num  1.2 4.2 9.2 4.2 11.2 2 4.6 7.2 12 9 ...
 $ Sunshine     : num  8.6 7 12.7 5.9 10.8 6.4 7.7 6 12.7 11.8 ...
 $ WindGustDir  : num  5 4 3 12 0 11 9 9 10 11 ...
 $ WindGustSpeed: int  31 43 50 59 56 63 39 39 28 37 ...
 $ WindDir9am   : num  10 5 2 15 10 15 12 10 3 3 ...
 $ WindDir3pm   : num  5 5 4 12 6 12 8 10 9 12 ...
 $ WindSpeed9am : int  4 19 31 9 15 15 15 19 7 15 ...
 $ WindSpeed3pm : int  11 19 19 31 26 31 19 24 11 26 ...
 $ Pressure9am  : num  1030 1026 1020 1010 1009 ...
 $ Pressure3pm  : num  1028 1025 1016 1009 1007 ...
 $ Cloud9am     : int  7 4 0 6 6 1 6 6 0 1 ...
 $ Cloud3pm     : int  3 6 0 7 6 6 2 7 0 1 ...
 $ Temp9am      : num  17.2 17.5 24.3 15 25.4 12.2 14.3 19.2 22.7 25.2 ...
 $ Temp3pm      : num  20.2 17.1 30.2 12.7 25.6 11.7 18.2 21.7 34.2 36.4 ...
 $ RainToday    : num  1 0 0 0 1 0 0 0 0 0 ...
 $ RISK_MM      : num  3.8 0 0 1.8 0 12.4 0 0 0 0 ...
 $ MHT          : int  1 1 0 0 0 0 1 0 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:69586] 1 2 3 4 6 7 8 9 11 12 ...
  ..- attr(*, "names")= chr [1:69586] "1" "2" "3" "4" ...
```

Hide

```
summary(humid1)
```

```
    Location        MinTemp         MaxTemp         Rainfall       Evaporation        Sunshin
e       WindGustDir     WindGustSpeed
 Min.   : 4.00   Min.   :-6.70   Min.   : 7.20   Min.   :  0.00   Min.   : 0.000   Min.   : 
0.000   Min.   : 0.000   Min.   :  9.00  
 1st Qu.:16.00   1st Qu.: 8.50   1st Qu.:18.60   1st Qu.:  0.00   1st Qu.: 2.600   1st Qu.: 
5.000   1st Qu.: 4.000   1st Qu.: 31.00  
 Median :28.00   Median :13.00   Median :23.80   Median :  0.00   Median : 4.800   Median : 
8.600   Median : 7.000   Median : 39.00  
 Mean   :26.46   Mean   :13.33   Mean   :24.09   Mean   :  2.38   Mean   : 5.249   Mean   : 
7.715   Mean   : 7.297   Mean   : 41.03  
 3rd Qu.:38.00   3rd Qu.:18.10   3rd Qu.:29.40   3rd Qu.:  0.60   3rd Qu.: 7.200   3rd Qu.:1
0.700   3rd Qu.:11.000   3rd Qu.: 48.00  
 Max.   :49.00   Max.   :30.20   Max.   :48.10   Max.   :367.60   Max.   :72.200   Max.   :1
4.500   Max.   :15.000   Max.   :126.00  
   WindDir9am       WindDir3pm      WindSpeed9am    WindSpeed3pm    Pressure9am      Pressure
3pm         Cloud9am        Cloud3pm    
 Min.   : 0.000   Min.   : 0.000   Min.   : 2.00   Min.   : 2.00   Min.   : 979.1   Min.   : 
978.9   Min.   :0.000   Min.   :0.000  
 1st Qu.: 3.000   1st Qu.: 4.000   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.:1012.8   1st Qu.:1
010.3   1st Qu.:1.000   1st Qu.:2.000  
 Median : 7.000   Median : 8.000   Median :15.00   Median :19.00   Median :1017.3   Median :1
014.8   Median :5.000   Median :5.000  
 Mean   : 7.012   Mean   : 7.484   Mean   :15.55   Mean   :19.75   Mean   :1017.3   Mean   :1
014.9   Mean   :4.238   Mean   :4.297  
 3rd Qu.:11.000   3rd Qu.:11.000   3rd Qu.:20.00   3rd Qu.:26.00   3rd Qu.:1022.0   3rd Qu.:1
019.6   3rd Qu.:7.000   3rd Qu.:7.000  
 Max.   :15.000   Max.   :15.000   Max.   :81.00   Max.   :72.00   Max.   :1041.1   Max.   :1
040.1   Max.   :8.000   Max.   :9.000  
    Temp9am         Temp3pm        RainToday         RISK_MM            MHT        
 Min.   :-0.70   Min.   : 4.80   Min.   :0.0000   Min.   :  0.000   Min.   :0.0000  
 1st Qu.:13.00   1st Qu.:17.30   1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.:0.0000  
 Median :17.60   Median :22.20   Median :0.0000   Median :  0.000   Median :0.0000  
 Mean   :18.08   Mean   :22.57   Mean   :0.2217   Mean   :  2.409   Mean   :0.4882  
 3rd Qu.:23.10   3rd Qu.:27.70   3rd Qu.:0.0000   3rd Qu.:  0.600   3rd Qu.:1.0000  
 Max.   :39.10   Max.   :46.10   Max.   :1.0000   Max.   :371.000   Max.   :1.0000  
```

Hide

```
dim(humid1)
```

```
[1] 30414    21
```

Question 3

Split 70% of data to training, 30% for testing

Hide

```
set.seed(32637888)

train.row  =  sample(1:nrow(humid1),  0.7*nrow(humid1))
humid.train = humid1[train.row,]
humid.test  =  humid1[-train.row,]
humid.train$MHT  =  as.factor(humid.train$MHT)
humid.test$MHT = as.factor(humid.test$MHT)
```

Question 4

Decision Tree Model

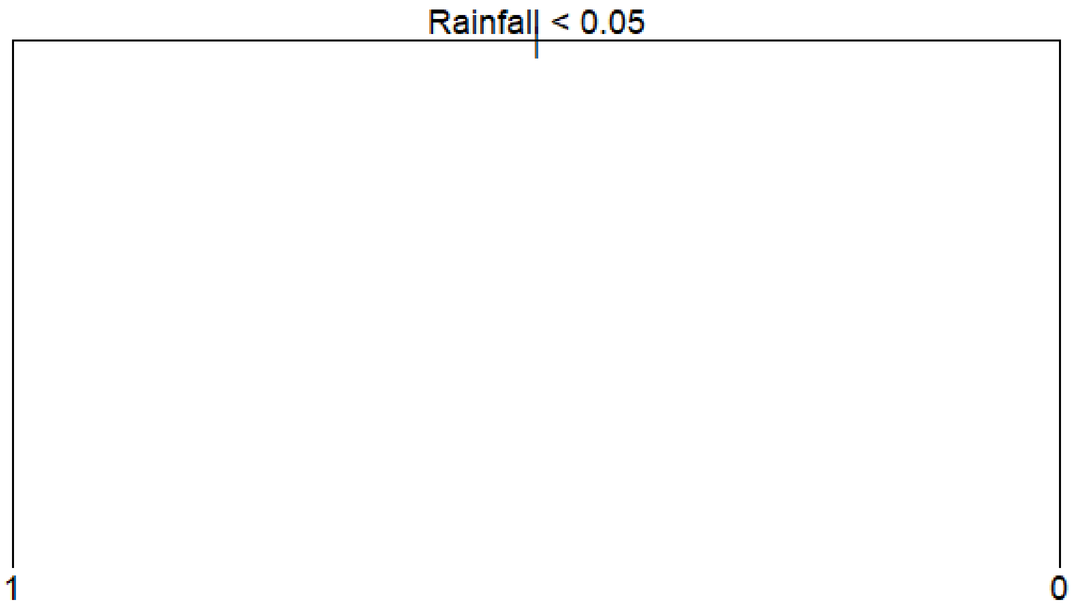
Hide

```
humid.tree=tree(MHT ~., data = humid.train)
summary(humid.tree)
```

```


Classification  tree:
tree(formula = MHT ~ ., data = humid.train)
Variables actually used in tree construction:
[1] "Rainfall"
Number of terminal nodes:  2
Residual mean deviance:  1.37 = 29160 / 21290
Misclassification error rate: 0.4432 = 9436 / 21289
```

Hide

```
plot(humid.tree)
text(humid.tree, pretty = 0)
```



Naive Bayes Model

Hide

```
humid.bayes = naiveBayes(MHT ~. , data = humid.train)
summary(humid.bayes)
```

```
          Length Class  Mode
apriori    2     table  numeric
tables    20     -none- list
levels     2     -none-  character
isnumeric 20     -none- logical
call       4     -none- call
```

Bagging Model

Hide

```
humid.bag = bagging(MHT ~., data = humid.train)
summary(humid.bag)
```

```
           Length  Class   Mode
formula          3 formula call
trees          100 -none-  list
votes        42578 -none-  numeric
prob         42578 -none-  numeric
class        21289 -none-  character
samples    2128900 -none-  numeric
importance      20 -none-  numeric
terms            3 terms   call
call             3 -none-  call
```

Boosting Model

Hide

```
humid.boost <- boosting(MHT ~ ., data=humid.train, mfinal=3)
summary(humid.boost)
```

```
           Length Class   Mode
formula        3  formula call
trees          3  -none-  list
weights        3  -none-  numeric
votes      42578  -none-  numeric
prob       42578  -none-  numeric
class      21289  -none-  character
importance    20  -none-  numeric
terms          3  terms   call
call           4  -none-  call
```

Random Forest Model

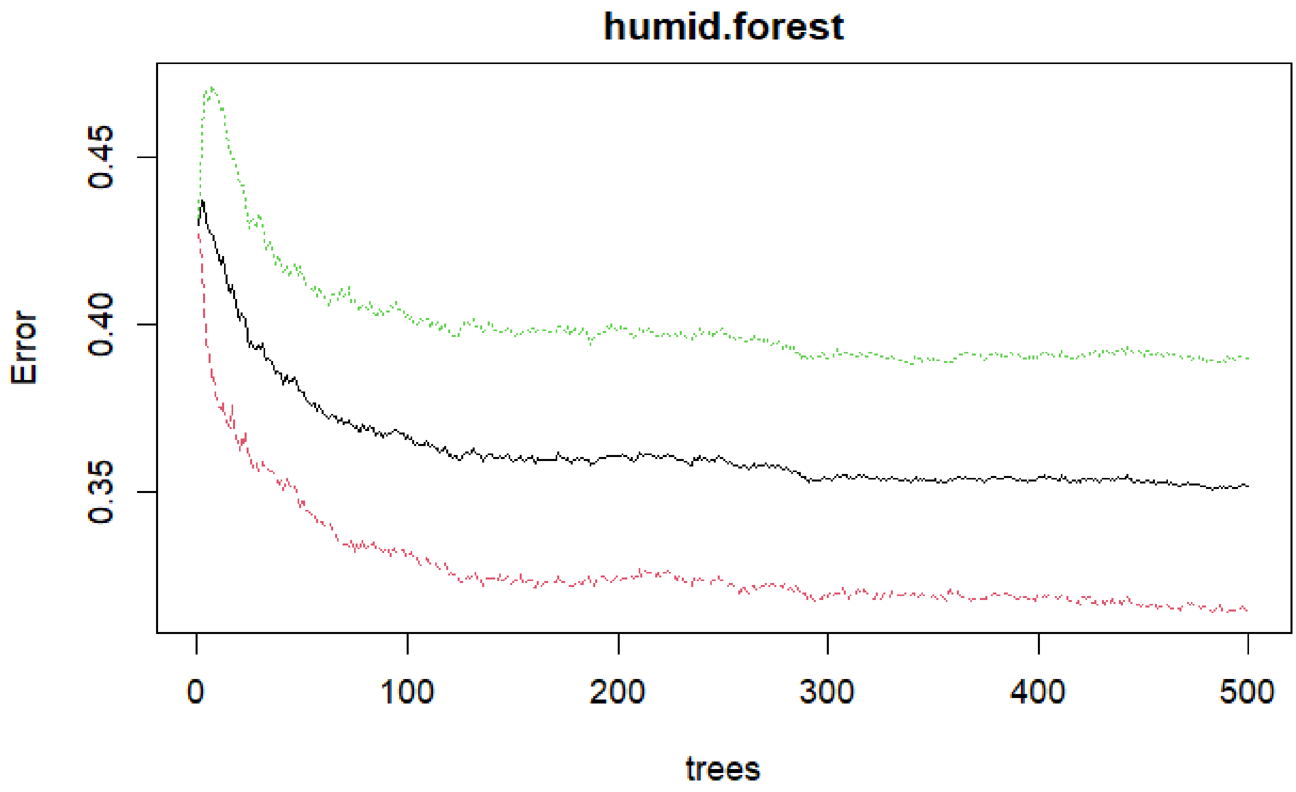
Hide

```
humid.forest <- randomForest(MHT~., data=humid.train)
summary(humid.forest)
```

```
                Length Class  Mode
call                3  -none- call
type                1  -none- character
predicted       21289  factor numeric
err.rate         1500  -none- numeric
confusion           6  -none- numeric
votes           42578  matrix numeric
oob.times       21289  -none- numeric
classes             2  -none- character
importance         20  -none- numeric
importanceSD        0  -none- NULL
localImportance     0  -none- NULL
proximity           0  -none- NULL
ntree               1  -none- numeric
mtry                1  -none- numeric
forest             14  -none- list
y               21289  factor numeric
test                0  -none- NULL
inbag               0  -none- NULL
terms               3  terms  call
```

Hide

```
plot(humid.forest)
```



Question 5

Confusion Matrix :

Decision Tree

Hide

```
humid1.tree.predict = predict(humid.tree, humid.test, type = "class")
tree.matrix <- table(actual = humid.test$MHT, predicted = humid1.tree.predict)
confusionMatrix(tree.matrix)
```

```
Confusion Matrix and Statistics

      predicted
actual    0    1
     0 1929 2763
     1 1214 3219

               Accuracy : 0.5642
                 95% CI : (0.5539, 0.5744)
    No Information Rate : 0.6556
    P-Value [Acc > NIR] : 1

                  Kappa : 0.136

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6137
            Specificity : 0.5381
         Pos Pred Value : 0.4111
         Neg Pred Value : 0.7261
             Prevalence : 0.3444
         Detection Rate : 0.2114
   Detection Prevalence : 0.5142
      Balanced Accuracy : 0.5759

       'Positive' Class : 0

```

Naive Bayes

Hide

```
humid1.bayes.predict = predict(humid.bayes, humid.test, type = "class")
bayes.matrix <- table(actual = humid.test$MHT, predicted = humid1.bayes.predict)
confusionMatrix(bayes.matrix)
```

```
Confusion Matrix and Statistics

      predicted
actual    0    1
     0 2400 2292
     1 1603 2830
                                         
               Accuracy : 0.5732          
                 95% CI : (0.5629, 0.5833)
    No Information Rate : 0.5613          
    P-Value [Acc > NIR] : 0.01161         
                                          
                  Kappa : 0.1493          
                                          
 Mcnemar's Test P-Value : < 2e-16         
                                          
            Sensitivity : 0.5996          
            Specificity : 0.5525          
         Pos Pred Value : 0.5115          
         Neg Pred Value : 0.6384          
             Prevalence : 0.4387          
         Detection Rate : 0.2630          
   Detection Prevalence : 0.5142          
      Balanced Accuracy : 0.5760          
                                          
       'Positive' Class : 0               
```

Bagging

Hide

```
humid1.bag.predict = predict(humid.bag, humid.test, type = "class")
bag.matrix  <-  humid1.bag.predict$confusion
confusionMatrix(bag.matrix)
```

```
Confusion Matrix and Statistics

               Observed Class
Predicted Class    0    1
              0 3397 2428
              1 1295 2005
                                          
               Accuracy : 0.592           
                 95% CI : (0.5818, 0.6021)
    No Information Rate : 0.5142          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.1775          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.7240          
            Specificity : 0.4523          
         Pos Pred Value : 0.5832          
         Neg Pred Value : 0.6076          
             Prevalence : 0.5142          
         Detection Rate : 0.3723          
   Detection Prevalence : 0.6384          
      Balanced Accuracy : 0.5881          
                                          
       'Positive' Class : 0               
```

Boosting

Hide

```
humid1.boost.predict = predict(humid.boost, humid.test, type = "class")
boost.matrix <- humid1.boost.predict$confusion
confusionMatrix(boost.matrix)
```

```
Confusion Matrix and Statistics

               Observed Class
Predicted Class    0    1
              0 3119 2147
              1 1573 2286
                                          
               Accuracy : 0.5923          
                 95% CI : (0.5822, 0.6024)
    No Information Rate : 0.5142          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.1811          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.6647          
            Specificity : 0.5157          
         Pos Pred Value : 0.5923          
         Neg Pred Value : 0.5924          
             Prevalence : 0.5142          
         Detection Rate : 0.3418          
   Detection Prevalence : 0.5771          
      Balanced Accuracy : 0.5902          
                                          
       'Positive' Class : 0               
```

Random Forest

Hide

```
humid1.forest.predict = predict(humid.forest, humid.test, type = "class")
forest.matrix <- table(actual = humid.test$MHT, predicted = humid1.forest.predict)
confusionMatrix(forest.matrix)
```

```
Confusion Matrix and Statistics

      predicted
actual    0    1
     0 3255 1437
     1 1703 2730

               Accuracy : 0.6559
                 95% CI : (0.646, 0.6656)
    No Information Rate : 0.5433
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3101

 Mcnemar's Test P-Value : 2.255e-06

            Sensitivity : 0.6565
            Specificity : 0.6551
         Pos Pred Value : 0.6937
         Neg Pred Value : 0.6158
             Prevalence : 0.5433
         Detection Rate : 0.3567
   Detection Prevalence : 0.5142
      Balanced Accuracy : 0.6558

       'Positive' Class : 0

```

Question 6

Hide

```
roc(humid.test$MHT,as.numeric(humid1.tree.predict))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.tree.predict))

Data: as.numeric(humid1.tree.predict) in 4692 controls (humid.test$MHT 0) < 4433 cases (humi
d.test$MHT 1).
Area under the curve: 0.5686
```

Hide

```
ROC.bayes  <-  roc(humid.test$MHT,as.numeric(humid1.bayes.predict))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

```
ROC.bayes
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.bayes.predict))

Data: as.numeric(humid1.bayes.predict) in 4692 controls (humid.test$MHT 0) < 4433 cases (humi
d.test$MHT 1).
Area under the curve: 0.575
```

Hide

```
ROC.bag <- roc(humid.test$MHT,as.numeric(humid1.bag.predict$class))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

```
ROC.bag
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.bag.predict$class))

Data: as.numeric(humid1.bag.predict$class) in 4692 controls (humid.test$MHT 0) < 4433 cases
(humid.test$MHT 1).
Area under the curve: 0.5881
```

Hide

```
ROC.boost <- roc(humid.test$MHT,as.numeric(humid1.boost.predict$class))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

```
ROC.boost
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.boost.predict$class))

Data: as.numeric(humid1.boost.predict$class) in 4692 controls (humid.test$MHT 0) < 4433 cases
(humid.test$MHT 1).
Area under the curve: 0.5902
```

Hide

```
ROC.forest  <-  roc(humid.test$MHT,as.numeric(humid1.forest.predict))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

```
ROC.forest
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.forest.predict))

Data: as.numeric(humid1.forest.predict) in 4692 controls (humid.test$MHT 0) < 4433 cases (hum
id.test$MHT 1).
Area under the curve: 0.6548
```

Hide

```
plot(roc(humid.test$MHT,as.numeric(humid1.tree.predict)))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

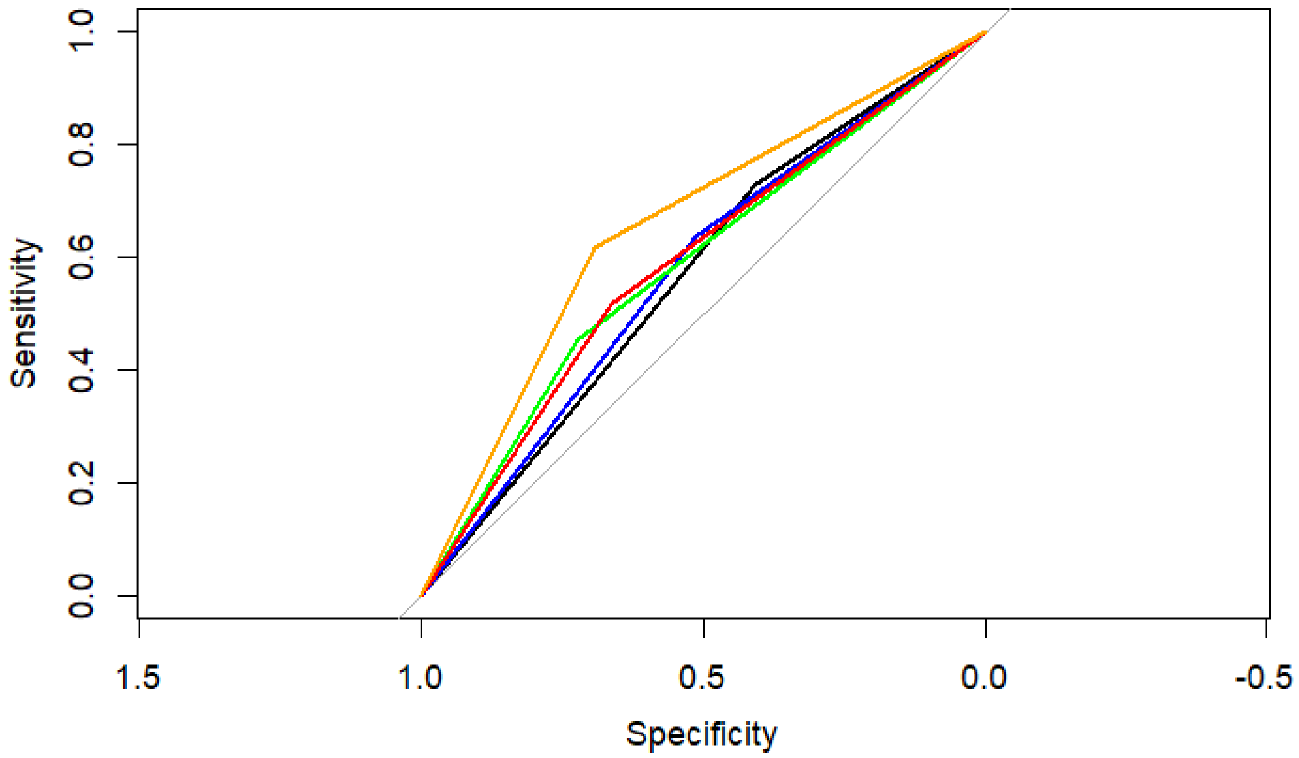
```
lines.roc(ROC.bayes, col= "blue" )
```

Hide

```
lines.roc(ROC.bag, col= "green" )
lines.roc(ROC.boost,  col=  "red"  )
```

Hide

```
lines.roc(ROC.forest,  col=  "orange"  )
```



black line: Decision Tree AOC

blue line: Naive Bayes AOC

green line: Bagging AOC

red line: Boosting AOC

orange line: Random Forest AOC

Question 7

Hide

```
Accuracy <- c(confusionMatrix(tree.matrix)$overall[1],confusionMatrix(bayes.matrix)$overall
[1],confusionMatrix(bag.matrix)$overall[1],confusionMatrix(boost.matrix)$overall[1],confusion
Matrix(forest.matrix)$overall[1])
AOC  <-  c(roc(humid.test$MHT,as.numeric(humid1.tree.predict))$auc[1],ROC.bayes$auc[1],ROC.bag
$auc[1],ROC.boost$auc[1],ROC.forest$auc[1])
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

Hide

```
Model <- c("Decision Tree","Naive Bayes", "Bagging", "Boostng",  "Random Forest" )
data.frame(Model,Accuracy,AOC)
```

Model <chr>	Accuracy <dbl>	AOC <dbl>
Decision Tree	0.5641644	0.5686351

Model	Accuracy	AOC
<chr>	<dbl>	<dbl>
Naive Bayes	0.5731507	0.5749514
Bagging	0.5920000	0.5881440
Boostng	0.5923288	0.5902132
Random Forest	0.6558904	0.6547849

5 rows

Best model is Random Forest, because highest accuracy and AOC

Question 8

Hide

```
summary(humid.tree)
```

```


Classification tree:
tree(formula = MHT ~ ., data = humid.train)
Variables actually used in tree construction:
[1] "Rainfall"
Number of terminal nodes:  2
Residual mean deviance:  1.37 = 29160 / 21290
Misclassification error rate: 0.4432 = 9436 / 21289
```

Decision Tree model most significant variable : Rainfall

Hide

```
sort(humid.bag$importance,decreasing = TRUE)
```

```
     Rainfall       RISK_MM      Cloud9am       Temp9am    WindDir9am    WindDir3pm        Max
Temp       Temp3pm      Location     RainToday
  33.94569345   24.24494564   14.48104318    9.38848002    3.95284368    3.73231377    3.6469
4787    2.14519258    1.72423133    0.83009249
WindGustSpeed   Pressure3pm   Pressure9am      Sunshine   WindGustDir   Evaporation        Min
Temp  WindSpeed3pm  WindSpeed9am      Cloud3pm
   0.55004611    0.25779811    0.25637351    0.25597865    0.23793093    0.12509903    0.0822
8387    0.07537639    0.06732938    0.00000000
```

Bagging model most significant variable : Rainfall, RISK_MM

Hide

```
sort(humid.boost$importance,decreasing = TRUE)
```

```
     Rainfall       RISK_MM      Cloud9am       Temp9am    WindDir3pm WindGustSpeed         Max
Temp    WindDir9am       Cloud3pm   Evaporation 
    38.681537     28.417553     12.346927      7.851302      5.144715      3.539174      2.20
3104      1.815688      0.000000      0.000000 
     Location       MinTemp   Pressure3pm   Pressure9am     RainToday      Sunshine         Tem
p3pm   WindGustDir   WindSpeed3pm  WindSpeed9am 
     0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.00
0000      0.000000      0.000000      0.000000 
```

Boosting model most significant variable : Rainfall, RISK_MM

Hide

```
sort(humid.forest$importance[,1],decreasing = TRUE)
```

```
     Sunshine        MinTemp        Temp3pm        Temp9am        MaxTemp    Pressure9am    Pressur
e3pm    Evaporation  WindGustSpeed    WindSpeed3pm 
    762.40575      744.78676      737.28138      727.91400      712.66200      702.33716      701.4
0374      632.39301      561.21465      507.76871 
 WindSpeed9am       Location     WindDir9am     WindDir3pm     WindGustDir       Cloud9am         RIS
K_MM       Cloud3pm        Rainfall      RainToday 
    499.78226      472.51735      467.40352      456.08507      435.48354      392.15904      371.8
2992      346.81709      323.12890       82.09918 
```

Random Forest model most significant variable : Sunshine, MinTemp, Temp3pm, Temp9am, MaxTemp, Pressure9am, Pressure3pm

Overall most significant variable: Rainfall

Overall not significant variable: Evaporation, WindSpeed3pm, WindSpeed9am, Cloud3pm, WindDir9am, WindDir3pm, Location, RainToday

Hence the not significant variables above could be ommited because they have very little effect on performance.

Question 9

According to the model created in Question 4, i know that the most significant variable is Rainfall. Hence, we can use rainfall to make a prediction to predict is tomorrow raining or not.

If value of rainfall is larger than 0.05 ,tomorrow will not be raining, else it will be raining tomorrow.

Hide

```
head(humid.test[,c("Rainfall","MHT")],10)
```

	Rainfall <dbl>	MHT <fctr>
34	0.8	0
37	0.0	0
57	0.0	0
67	1.0	1

	Rainfall	MHT
	<dbl>	<fctr>
76	0.0	0
85	7.2	0
100	0.0	0
137	0.0	0
164	0.0	0
169	0.0	1

1-10 of 10 rows

So, by using this we found out that the accuracy of this model is 5/10 = 0.5

Question 10

Decision Tree Pruning

Hide

```
cv.tree(humid.tree, FUN = prune.misclass)
```

```
$size
[1] 2 1

$dev
[1]  9549 10416

$k
[1] -Inf  980

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Hide

```
prunedtree = prune.misclass(humid.tree, best = 4)
```

```
Warning: best is bigger than tree size
```

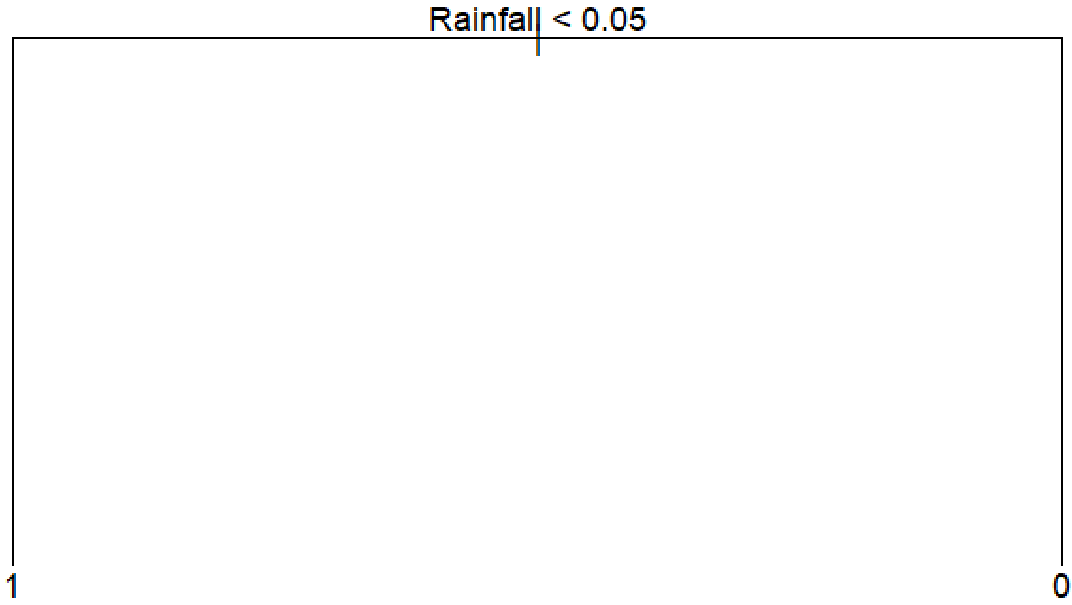
Hide

```
summary(prunedtree)
```

```
Classification tree:
tree(formula = MHT ~ ., data = humid.train)
Variables actually used in tree construction:
[1] "Rainfall"
Number of terminal nodes:  2
Residual mean deviance:  1.37 = 29160 / 21290
Misclassification error rate: 0.4432 = 9436 / 21289
```

Hide

```
plot(prunedtree)
text(prunedtree, pretty = 0)
```



After pruning is the same model as the Decision Tree in Question 4, so the Decision Tree in Question 4 can be considered as a good tree already.

Hence its accuracy and Area Under Curve Values will be the same.

Important factors: Rainfall

Using attribute Rainfall giving us an accuracy of 0.564 is better than other attributes.

Question 11

Artificial neural network

Hide

```
humid.neural <- nnet(MHT~.-MHT, data = humid.train,size = 4, decay = 0.0001, maxit = 500)
```

```
# weights:  89
initial  value 15821.538914 
iter  10 value 14662.206443
iter  20 value 14498.962434
iter  30 value 14413.344186
iter  40 value 14362.119823
iter  50 value 14305.665291
iter  60 value 14275.371804
iter  70 value 14269.156926
iter  80 value 14268.841492
iter  90 value 14268.631485
iter 100 value 14268.230177
iter 110 value 14267.411362
final  value 14266.968358 
converged
```

Confusion Matrix of ANN:

Hide

```
humid1.neural.predict <- predict(humid.neural, humid.test, type = 'class')
confusionMatrix(as.factor(humid1.neural.predict),  humid.test$MHT)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2618 1632
         1 2074 2801
                                          
               Accuracy : 0.5939          
                 95% CI : (0.5837, 0.604) 
    No Information Rate : 0.5142          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.1893          
                                          
 Mcnemar's Test P-Value : 4.352e-13       
                                          
            Sensitivity : 0.5580          
            Specificity : 0.6319          
         Pos Pred Value : 0.6160          
         Neg Pred Value : 0.5746          
             Prevalence : 0.5142          
         Detection Rate : 0.2869          
   Detection Prevalence : 0.4658          
      Balanced Accuracy : 0.5949          
                                          
       'Positive' Class : 0               
                                          
```

AOC of ANN:

Hide

```
roc(humid.test$MHT,as.numeric(humid1.neural.predict))
```

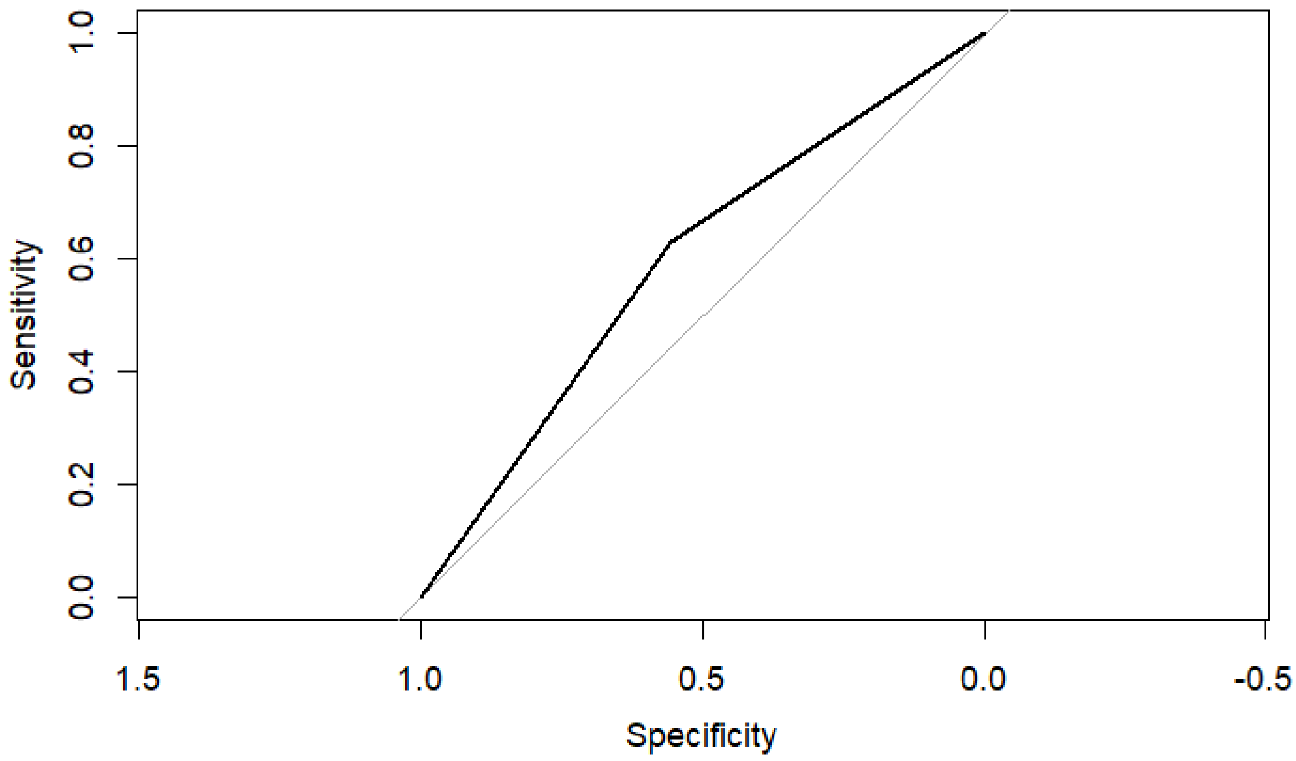
```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```


Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.neural.predict))

Data: as.numeric(humid1.neural.predict) in 4692 controls (humid.test$MHT 0) < 4433 cases (hum
id.test$MHT 1).
Area under the curve: 0.5949
```

Hide

```
plot(roc(humid.test$MHT,as.numeric(humid1.neural.predict)))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```



Accuracy: 0.5939, AOC: 0.5949. So the Artificial neural network model is better than the other 4 models because its accuracy and AOC is lower than the other models.

Question 12

K-th Nearest Neighbors Model

package used: class

package link: class: Functions for Classification (r-project.org) (https://cran.r-project.org/web/packages/class/class.pdf)

This model classifies through the nearest points on the graph, it groups according to the distance between the points. Nearer points will be form a group.

Hide

```
trainControl <- trainControl(method="repeatedcv", number=10, repeats=3)
humid.knn <- train(MHT~., data=humid.train, method="knn",
                  metric="Accuracy" ,trControl=trainControl)
humid1.knn.predict <- predict(humid.knn, newdata = humid.test)
confusionMatrix(humid1.knn.predict, humid.test$MHT)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2923 1922
         1 1769 2511
                                          
               Accuracy : 0.5955          
                 95% CI : (0.5854, 0.6056)
    No Information Rate : 0.5142          
    P-Value [Acc > NIR] : < 2e-16         
                                          
                  Kappa : 0.1896          
                                          
 Mcnemar's Test P-Value : 0.01235         
                                          
            Sensitivity : 0.6230          
            Specificity : 0.5664          
         Pos Pred Value : 0.6033          
         Neg Pred Value : 0.5867          
             Prevalence : 0.5142          
         Detection Rate : 0.3203          
   Detection Prevalence : 0.5310          
      Balanced Accuracy : 0.5947          
                                          
       'Positive' Class : 0               
```

Hide

```
roc(humid.test$MHT,as.numeric(humid1.knn.predict))
```

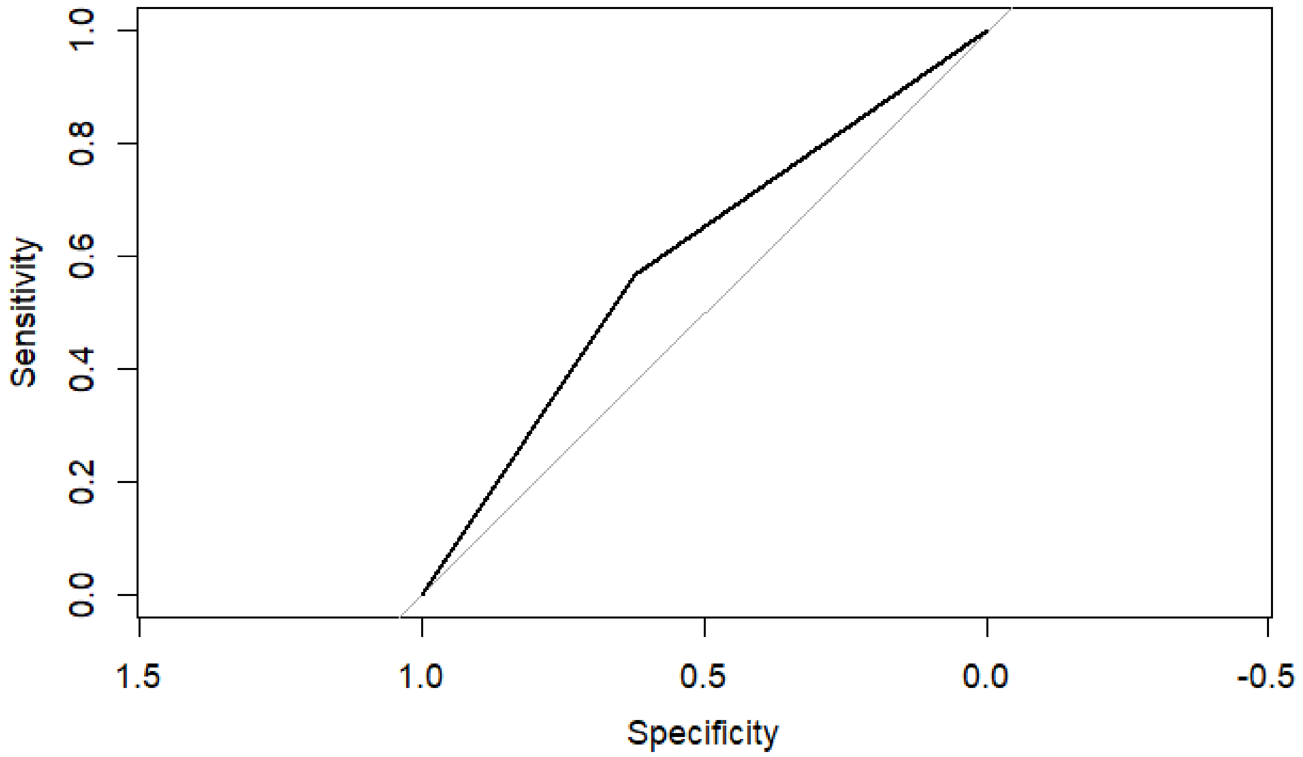
```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```

Call:
roc.default(response = humid.test$MHT, predictor = as.numeric(humid1.knn.predict))

Data: as.numeric(humid1.knn.predict) in 4692 controls (humid.test$MHT 0) < 4433 cases (humid.
test$MHT 1).
Area under the curve: 0.5947
```

Hide

```
plot(roc(humid.test$MHT,as.numeric(humid1.knn.predict)))
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```



Accuracy: 0.5955, AOC: 0.5947

This model is better than the Decision Tree model and the Naive Bayes model because its accuracy and AOC is larger than them.