The open source language that is selected for this assignment is **R language** and the tool that is to be used here is the **R Studio.R Markdown** file is being used here to do the combination of documentation and coding using the **knit option**.

Task 1

There are several open source machine learning packages available in R. From this, the package “**Rweka**” is taken to execute the task. Weka is a tool that is used to perform machine learning solutions using Java. This package RWeka helps in incorporating the functionalities present in weka and to blend it with R language. This is the major reason in choosing this package as it provides a variety of option while executing and performing different data mining algorithms.It consist of the interface code along with the tools for performing data pre processing, classification , regression , clustering,association rules and visualization thereby serving as a very useful package for machine learning purposes.At present, it requires Java 8 or higher and R 2.6.0 or higher version to run smoothly.It supports various Classification algorithms like C4.5 and PART etc.

Task2

The given data present in a text document is in such a way that the patients are in the columns and the attributes are in the rows. In total,there are 376 columns denoting the patients and 10 rows denoting the attributes. Inorder for the package RWeka to process the data, row names and column names are provided correspondingly. Then, the data is transposed because inorder for the package to process the data, the attributes are needed to be present in the columns with the target column which is needed for classification purposes.

```
library(RWeka)#loading the packing
autoimmune <- read.delim("C:/Users/sayei/Downloads/autoimmune.txt",header = F)#importing the given data and giving Header as False so that age row values are not taken as column names
colnames(autoimmune)<-c(1:376)#providing column name to denote each patient from 1 to 376
rownames(autoimmune)<-c('Age','Blood_Pressure','BMI','Plasma_level','Autoimmune_disease','Adeverse_events','Drug_in_serum','Liver_function','Activity_test','Secondary_text')#providing the details as given in the assignment question
x<-as.data.frame(t(autoimmune))#transposing the matrix inorder to get patients in the rows and attributes in columns
```

Task 3

From the given conditions, two classfication models are considered namely **C4.5** and **Naive Bayes Classification**.Firstly, C4.5 classification is done.C4.5 is an algorithm which is used to create decision tree. It is considered to be a betterment from that of its predecessor ID3.It makes use of the same information entropy concept as that of ID3. It makes use of the training data representing things which are classified before and generates a single tree with the use of it.It can be used for wide range of data like continuous and categorical.It also deals with incomplete data and uses single pass pruning to tackle over fitting.

```
C45 <- J48(Autoimmune_disease~., data=x)#C4.5 is referred to as J48 in RWeka package. In this, the target attribute is given as Autoimmune_disease as given in the question from the data x.
summary(C45)#Summary after C4.5

##
## === Summary ===
##
## Correctly Classified Instances        258               68.617  %
## Incorrectly Classified Instances       118               31.383  %
## Kappa statistic                          0     
## Mean absolute error                      0.4307
## Root mean squared error                  0.464 
## Relative absolute error                 99.9149 %
## Root relative squared error             99.9998 %
## Total Number of Instances              376     
## 
## === Confusion Matrix ===
## 
##    a   b   <-- classified as
##  258   0 |   a = negative
##  118   0 |   b = positive
```

From this it is seen that the target attribute Autoimmune_disease has 258 negative and 118 positive values with the correctly classified instance as 68.6%.

Now, the naive bayes classification model is considered. It is a probabilistic classifier which makes use of the bayesian theorem in order to perform classification.In this, the assumption that every attritbute that is being classified is independent to each other is considered. It helps in determining a class through a set of features by means of probability. It is preferred alot due to is simple mechanism and quickness in processing and ability to not care about non essential features.

```
library(e1071)
Naive_Bayes_Model=naiveBayes(Autoimmune_disease~., data=x)#importing the required package and then performing the naive bayes model by giving the target variable autoimmune to process the data x with it
summary(Naive_Bayes_Model)#providing the summary of the generated model

##         Length Class  Mode     
## apriori 2      table  numeric  
## tables  9      -none- list     
## levels  2      -none- character
## call    4      -none- call

print(Naive_Bayes_Model$apriori)#giving out the output

## Y
## negative positive 
##      258      118
```

Task 4

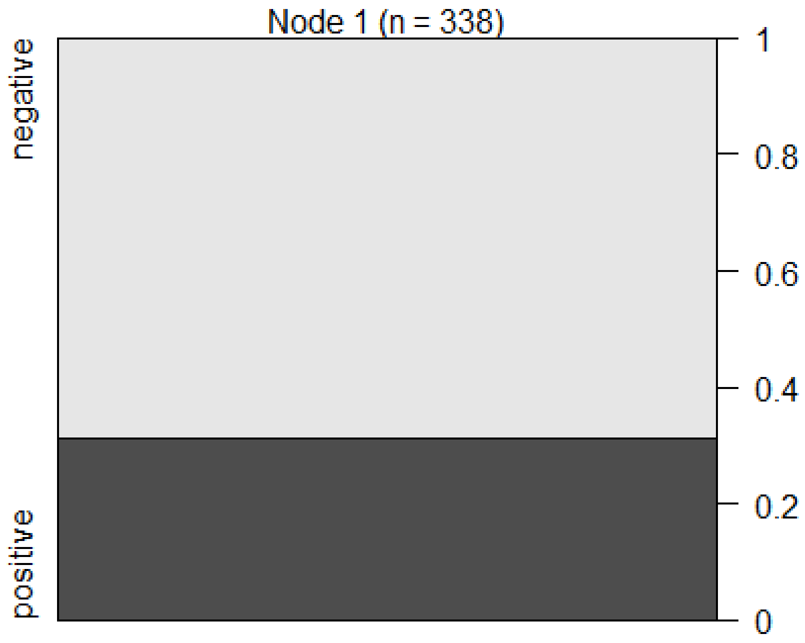
After this, 10 fold cross validation is done for both the algorithms to estimate the future occurences .This is at first applied to the data as follows,

```
#C45 10-fold process
x<-x[sample(nrow(x)),]#randomising the order of the row in the data x for better testing reasons
folds <- cut(seq(1,nrow(x)),breaks=10,labels=FALSE)#Creating 10 fold command
#Perform 10 fold cross validation
for(i in 1:10){#Segementing the data by fold using which() function
  testIndexes <- which(folds==i,arr.ind=TRUE)
  testData <- x[testIndexes, ]
```

```
 trainData <- x[-testIndexes, ]
}
C45 <- J48(Autoimmune_disease~., data=trainData)#performing the C4.5 classification in the trainData
segmented from x
summary(C45)#summary of c45 model on the train data

##
## === Summary ===
##
## Correctly Classified Instances         233               68.9349 %
## Incorrectly Classified Instances       105               31.0651 %
## Kappa statistic                          0
## Mean absolute error                      0.4283
## Root mean squared error                  0.4628
## Relative absolute error                 99.9016 %
## Root relative squared error             99.9997 %
## Total Number of Instances              338
##
## === Confusion Matrix ===
##
##    a   b   <-- classified as
##  233   0 |   a = negative
##  105   0 |   b = positive

plot(C45)#giving the graphical output of the train data
```



```
predictions2 <- predict(C45, testData[,1:10])#predicting the autoimmune_disease possibility with respect to the attributes present in the data
print(table(predictions2, testData$Autoimmune_disease))#giving out the final result

##
## predictions2 negative positive
##     negative       25       13
##     positive        0        0
```

From the above Summary, it is seen that the correctly classified instance stands at 68.9% ,with absolute mean error and root mean error at 0.4 and 0.6 respectively.When the same model is done for the test data ,the correctly classified instance stood at 65.8% and the summation of the values from the train data and test data gave the correct split as it is in the original case which indicates that the model prediction is done successfully.

```
#Naive bayes 10 fold cross by using the cross folded data done previously above
Naive_Bayes_Model1=naiveBayes(Autoimmune_disease~., data=trainData)#applying Naive bayes classifier model in traindata
NB_Predictions1=predict(Naive_Bayes_Model1,trainData)#predicting the values with respect to traindata
print(table(NB_Predictions1,trainData$Autoimmune_disease))#giving out the resulting confusion matrix

##
## NB_Predictions1 negative positive
##        negative      230        0
##        positive        3      105

NB_Predictions2=predict(Naive_Bayes_Model1,testData)#predicting the values for the test data
print(table(NB_Predictions2,testData$Autoimmune_disease))#giving out the confusion matrix for the test case

##
## NB_Predictions2 negative positive
##        negative       24        6
##        positive        1        7
```

From this, it is seen that in the train data there are 231 clear negative and 105 clear positive with 2 values that tend to favour negative more from the 338 ones.In the test data ,there are 21 clear negative and 4 clear postive with 13 mixed ones with 9 one of sligthly towards positive and 4 towards negative. So a cumulative of both of them gives 118 positive and 258 negative from the total of 376 in the given data.This indicates that the given model has been done to perform naive bayes classifier successfully.

Task 5

After obtaining all the results, it is seen that the models give a very similar output and does not vary significantly. This is because the accuracy levels of both the models doesn’t vary much corresponding to this data set and give out similar result with slight variations due to the factors corresponding to the respective models like mean error absolute error,root mean squared error and so on.The reason can also be due to the fact that they are done independantly and hence no possibility for the values to get changed drastically.

References

1.https://machinelearningmastery.com/non-linear-classification-in-r-with-decision-trees/

2.https://gist.github.com/duttashi/a51c71acb7388c535e30b57854598e77

3.https://cran.r-project.org/web/packages/RWeka/RWeka.pdf

4.https://hackerbits.com/data/c4-5-data-mining-algorithm/

5.http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/