VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

FINAL REPORT

SWE1011- SOFT COMPUTING

SLOT: D2+TD2

Lung Cancer Prediction Using Machine Learning Techniques

GROUP MEMBERS:

T MAHAMMAD SUHEL	16MIS0141
P KANAKA NAGA AKHIL	16MIS0188
R PRATAP KUMAR	16MIS0346
V GOWTHAM	16MIS0369

CERTIFICATE

This is to certify that the Project work entitled “Lung Cancer Prediction Using Machine Learning Techniques” that is being submitted by “T MAHAMMAD SUHEL, P KANAKA NAGA AKHIL, R PRATAP KUMAR, V GOWTHAM ” in M. Tech (S.E) for SWE1011: SOFT COMPUTING is a record of bonafide work done under my supervision. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted for any other course.

Signature of faculty

(CHIRANJI LAL CHOWDHARY)

AKNOWLEDGEMENT

We are thankful to the Department because of whom, we have gained confidence in Innovative Thinking and it also enhanced our professional skills as to become competent in this field.

In performing our project, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. The completion of this project gives us much Pleasure. We would like to show our gratitude to Prof. CHIRANJI LAL CHOWDHARY, SITE VIT University for giving us a good guideline for project throughout numerous consultations. We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in this project.

Thank you,

T MAHAMMAD SUHEL	16MIS0141
P KANAKA NAGA AKHIL	16MIS0188
R PRATAP KUMAR	16MIS0346
V GOWTHAM	16MIS0369

TABLE OF CONTENTS

Topics **Page Numbers**

ABSTRACT:

Today the number of diseases are gradually increasing day-by-day. Cancer is one of the most dangerous and deadliest disease in today's world. Based on the deaths from 2011-2015, there are about 163.5 death per 100000 men and women. The cancer like lung, prostrate, and colorectal cancers contribute up to 45% of cancer deaths. So it is very important to detect or predict before it reaches to serious stages. If cancer predicted in its early stages, then it helps to save the lives. Statistical methods are generally used for classification of risks of cancer i.e. high risk or low risk. Sometime it becomes difficult to handle the complex interactions of high-dimensional data. Machine learning techniques can be used to overcome these drawbacks which are cause due to the high dimensions of the data. So in this project we are going to use machine learning algorithms to predict the chances of getting cancer. We are going to use algorithms like Naive Bayes, decision tree, and svm etc. we will also compare the accuracy among these algorithms.

INTRODUCTION:

It is a challenging task to predict the results of cancer disease. The data analysis related to medical information data is very difficult because it contains a number of variables and they have hidden values. The machine learning algorithms are very powerful techniques to recognize patterns and to find relationships among a large number of variable of the medical data. This helps to predict results of cancer disease by making use of the datasets. The machine learning algorithms are used instead of statistical analysis techniques because it has the drawbacks in handling the high dimensions of the data. Some valid studies clear that machine learning methods improves the accuracy of predicting cancer susceptibility, recurrence and mortality up to 15–25%.

Cancer is caused due to uncontrolled cellular growth and reproduction. There are two types of tumors such as benign and malignant. Benign is a not harmful as it is localized and doesn’t spreads to the parts of the body. Unlike benign, malignant is harmful and spreads to the parts of the body. The other word for malignant tumor is cancer. It is important to distinguish between these tumors which helps in predicting cancer. There are various type of the cancer like lung, breast, prostrate, carnival etc. Each type of cancer has specific symptoms. Based on the symptoms the type of the cancer is predicted. In this project we are mainly focusing on the lung cancer prediction as 1 in 4 cancer deaths are from lung cancer.

Machine learning is branch of artificial intelligence that deals with various techniques such as statistical, optimization and probabilistic. These techniques helps the computer to predict results from past datasets which has large, complex data. We collect the standard dataset then it is pre-processed using the tool called rapid minier. The noisy, irrelevant, missing data is eliminated using this tool. Then we are going to use classification algorithms like decision tree, Naïve Bayes, and Random Forest algorithms to build a cancer risk prediction system is proposed here which predicts cancers and is also user friendly, time and cost saving. These are further compared for their accuracies.

LITERATURE REVIEW

S.NO	Paper Name	Year	Author	Techniques
1	Empirical Analysis on Cancer Dataset with Machine Learning Algorithms	2018	T. PanduRanga Vital, M. Murali Krishna	decision tree, Naïve Bayes, K-Star, and Random Forest
2	Lung cancer prediction using machine learning and advanced imaging techniques:	2018.	Timor Kadir, Fergus Gleeson	Convolutional neural networks ,deep learning
3	Machine Learning Approaches in Cancer Detection and Diagnosis: Mini Review	2017	Majid Murtaza Noor and Vinay Narwal	(SCILM) method, Knowledge base system learning method, Gene expression learning method, Convolution neural network
4	Machine learning applications in cancer prognosis and detection:	2014	Konstantino Kourou, Themis P.Exarchos	Artificial neural network, Bayesian network, Support vector machine and Decision tree
5	Predicting Breast Cancer Survivability using Data Mining Techniques:	2015	Abdelghani Bellaachia, Erhan Guven	Back propagated neural network, The Naïve Bayes and decision tree algorithms

T. PanduRanga Vital, M. Murali Krishna In this paper, research is based on the data collected from different districts of Andhra Pradesh with 1008 instances and 46 attributes which has both cancer and non-cancer data. They have used supervised machine learning algorithms like decision tree, Naïve Bayes, K-Star, and Random Forest because the dataset has the class label. For their dataset all applied algorithms show above 96% accuracy and k-star model is performed with 100% accuracy in predicting cancer.

Timor Kadir, Fergus Gleeson have proposed many Machine learning(ML) techniques to predict lung cancer and assist clinical managing incidental or screen detected indeterminate pulmonary modules. This technique helps us to reduce variability in nodule classification, improve decision making. It is also important to distinguish between benign and malignant tumor which helps us to get the overview on the type of cancer and their effects. In addition to ML lung cancer prediction approaches, they also proposed strengths and weakness of their approaches

Majid Murtaza Noor and Vinay Narwal They mainly concentrated on new research directions of machine learning in cancer prediction like Sparse compact incremental learning machine (SCILM) method, Knowledge base system learning method and Gene expression learning method and Convolution neural network learning method. To predict cancer accurately of which type it is, Machine learning has come up with the efficient technique that promised to give the best results. To classify between high risk and low risk of cancer many statistical methods have been used. But it is not as accurate as Machine learning

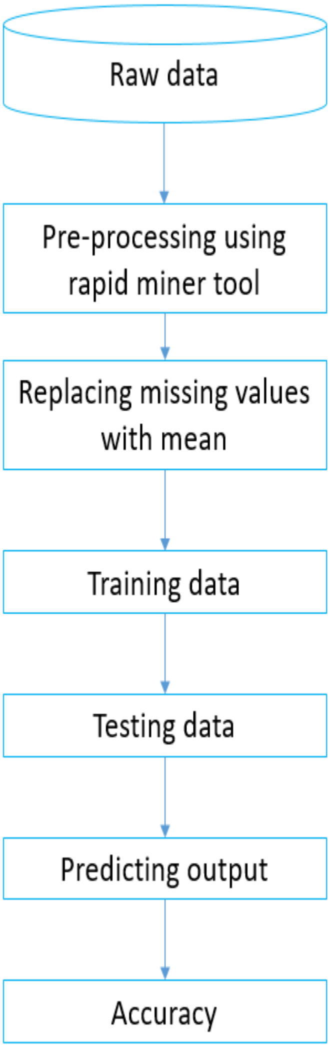
Konstantino Kourou, Themis P.Exarchos They has taken various data samples and input features and applied various recent ML approaches to those samples. ML techniques such as Artificial neural network, Bayesian network, Support vector machine and Decision tree have been used and accuracy is calculated and compared. Different data samples like prediction of cancer susceptibility, recurrence and survival were used. The early diagnosis and prognosis of cancer type is playing an important role in research center which can facilitate the management of patients. From the bio-medical and bio-informatics field many groups have been formed to classify the patients into high or low risk. There are varieties of Proper use of these ML techniques helps us to understand the cancer progression.

Abdelghani Bellaachia, Erhan Guven They made analysis on the prediction of survivability rate of breast cancer patients using data mining techniques. In this paper mainly three data mining techniques has been used such as Back propagated neural network, The Naïve Bayes and the C4.5 decision tree algorithms on SEER dataset. The results from three techniques will be calculated and compared. However, decision tree algorithm has been giving accurate results than other two data mining techniques.

METHODOLOGY:

- We collect the standard dataset.(dataset source: data.world)
- The dataset may contain some irrelevant, noisy data.
- To eliminate this we are using a tool called rapid miner.
- Dataset is pre-processed using this tool.
- It includes the elimination of noisy, irrelevant, missing data.
- Then we are going to use classification algorithms like decision tree, Naïve Bayes, and K Star algorithm to build a cancer risk prediction system and is also user friendly, time and cost saving.
- These algorithms are further compared for their accuracies.

DESIGN FRAMEWORK:


Raw data
Pre-processing using rapid miner tool
Replacing missing values with mean
Training data
Testing data
Predicting output
Accuracy

The lung dataset is collected from data world

The dataset have error values or missing values so Pre-processing is done using rapid miner tool

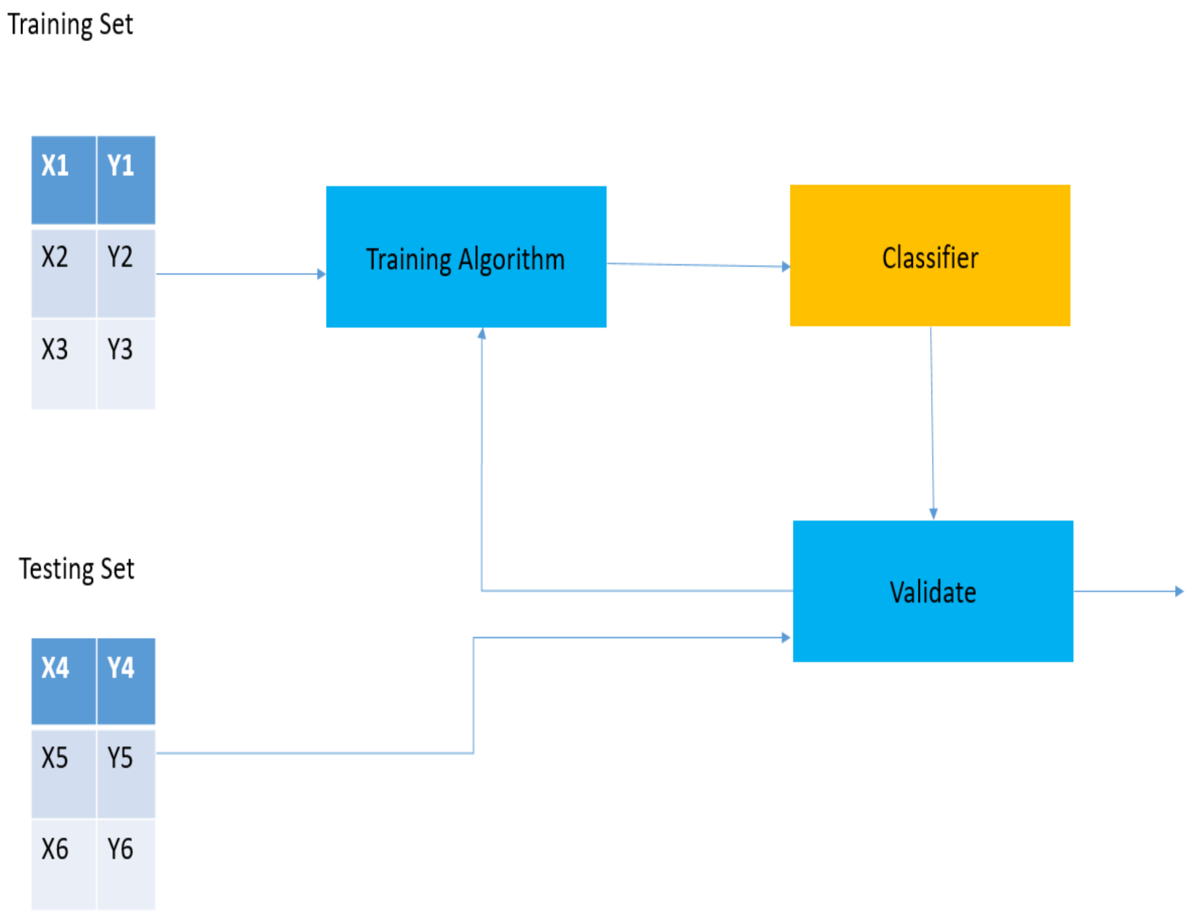
To avoid the missing of instance we replaced missing values with mean instead of removing them

The target dataset is splitted into two one for training(60%) and testing(40%)

The testing data is tested with the help of trained data

Then the output is predicted

The predicted values are compared with the actual values to predict accuracy



X1	Y1
X2	Y2
X3	Y3

X4	Y4
X5	Y5
X6	Y6

LANGUAGES/PLATEFORMS/TOOLS

Tools

Rapidminer

Rstudio

Language

R

ALGORITHMS:

Naive Bayes:

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms.

Naïve Bayes Algorithm:

- Statistical method for classification.
- Supervised learning method.
- Assumes an underlying probabilistic model, the Bayes theorem.
- Can solve problems involving both categorical and continuous valued attributes.
- Named after Thomas bayes, who proposed the bayes theorem.

It uses Bayesian Theorem : $P(H|X) = p(X|H) P(H) / p(X)$

K-Star (K*) Classification:

K* is an instance-based classifier that it is very simple and similar to k-nearest neighbors (KNN) algorithm. New data or information instances Xi where i = 1, 2, 3, …, k are allotted to the class that happens most often along with the k-nearest information or data instances or items Yj where j = 1, 2, 3, …, k. The most related items from the dataset are retrieved from entropic distance. By the method for entropic separation as a metric have various advantages, such as handling of missing values and real-esteemed features or attributes. The K* function can be analysed as Eq. 3.

$$K^*(I,x) = -\ln P^*(y_{i,},x) \quad - (3)$$

Where P is the transformational path probability from data point or instance x to instance y. It is very helpful to understand the probability that instance x will reach the destination y through a random walk.

Decision Tree Algortihm:

Decision tree (DT) algorithm is handled in the way that the element vector acts for a count of instances by training data samples. The classes for the newly generated instances are being found in this algorithm. This algorithm generates the rules for the prediction of the target variable. With the assistance of tree classification, the critical distribution of the data is easily understandable C4.5 is an expansion of ID3.

DATA PRE- PROCESSING

Dataset:

This dataset contains Attributes of lung cancer prediction.

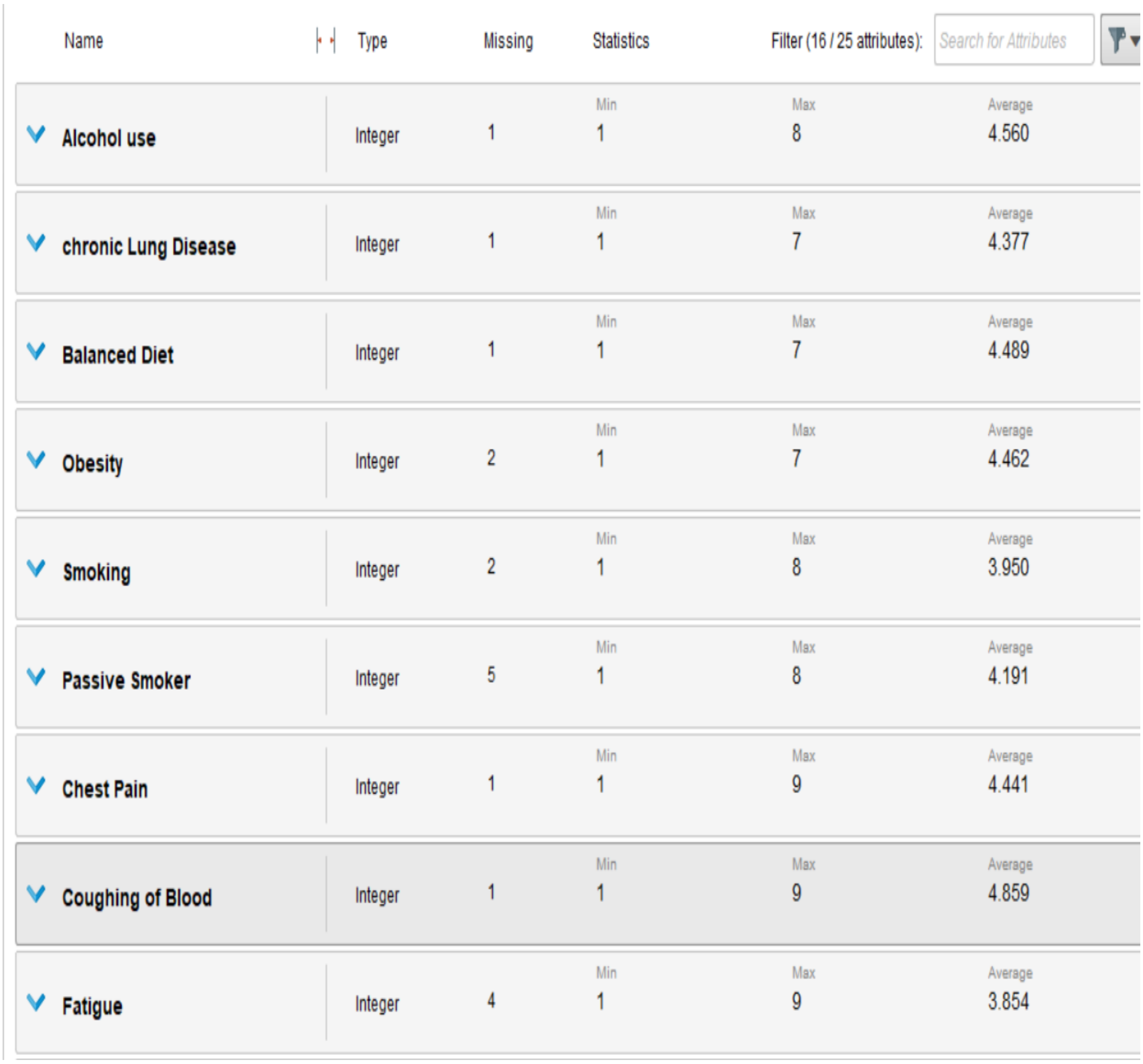
Dataset Information:

Number of attributes: 25

Number of instances: 1000

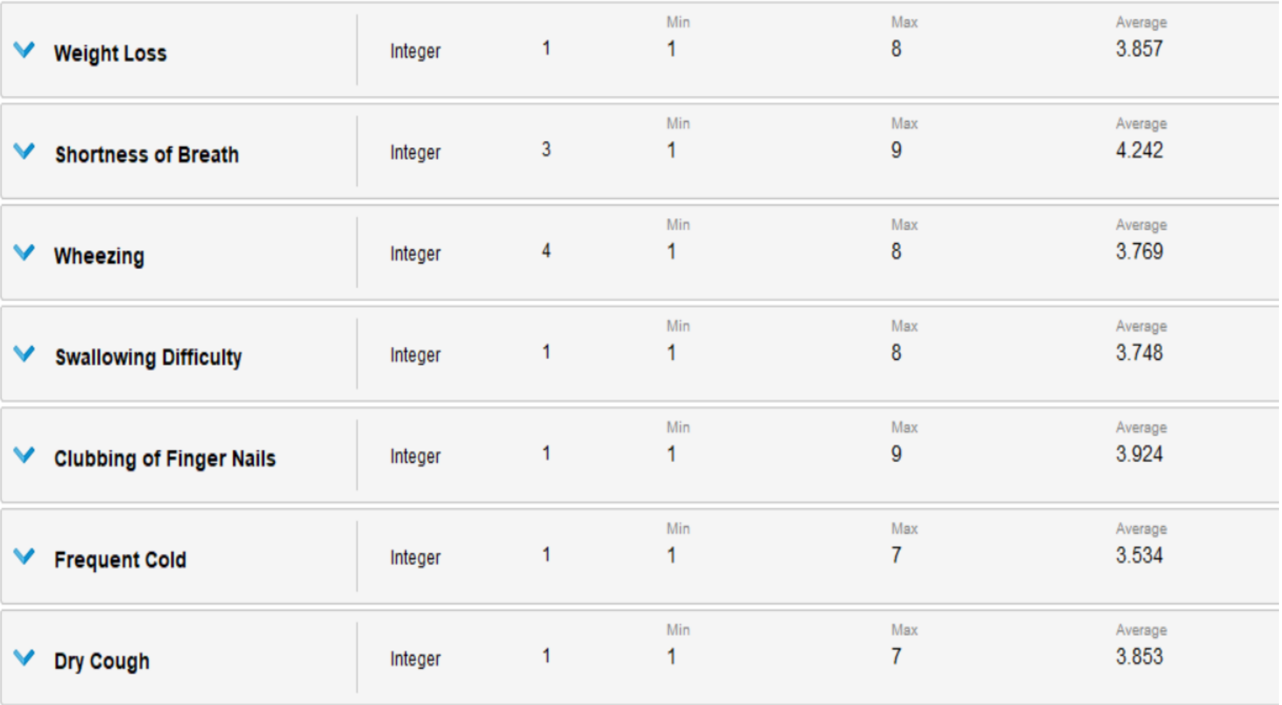
Number of attributes with missing values: 16

Missing attributes information is as follows:



Filter (16 / 25 attributes): Search for Attributes

Name	Type	Missing	Statistics		
			Min	Max	Average
Alcohol use	Integer	1	1	8	4.560
chronic Lung Disease	Integer	1	1	7	4.377
Balanced Diet	Integer	1	1	7	4.489
Obesity	Integer	2	1	7	4.462
Smoking	Integer	2	1	8	3.950
Passive Smoker	Integer	5	1	8	4.191
Chest Pain	Integer	1	1	9	4.441
Coughing of Blood	Integer	1	1	9	4.859
Fatigue	Integer	4	1	9	3.854



Weight Loss	Integer	1	Min 1	Max 8	Average 3.857
Shortness of Breath	Integer	3	Min 1	Max 9	Average 4.242
Wheezing	Integer	4	Min 1	Max 8	Average 3.769
Swallowing Difficulty	Integer	1	Min 1	Max 8	Average 3.748
Clubbing of Finger Nails	Integer	1	Min 1	Max 9	Average 3.924
Frequent Cold	Integer	1	Min 1	Max 7	Average 3.534
Dry Cough	Integer	1	Min 1	Max 7	Average 3.853

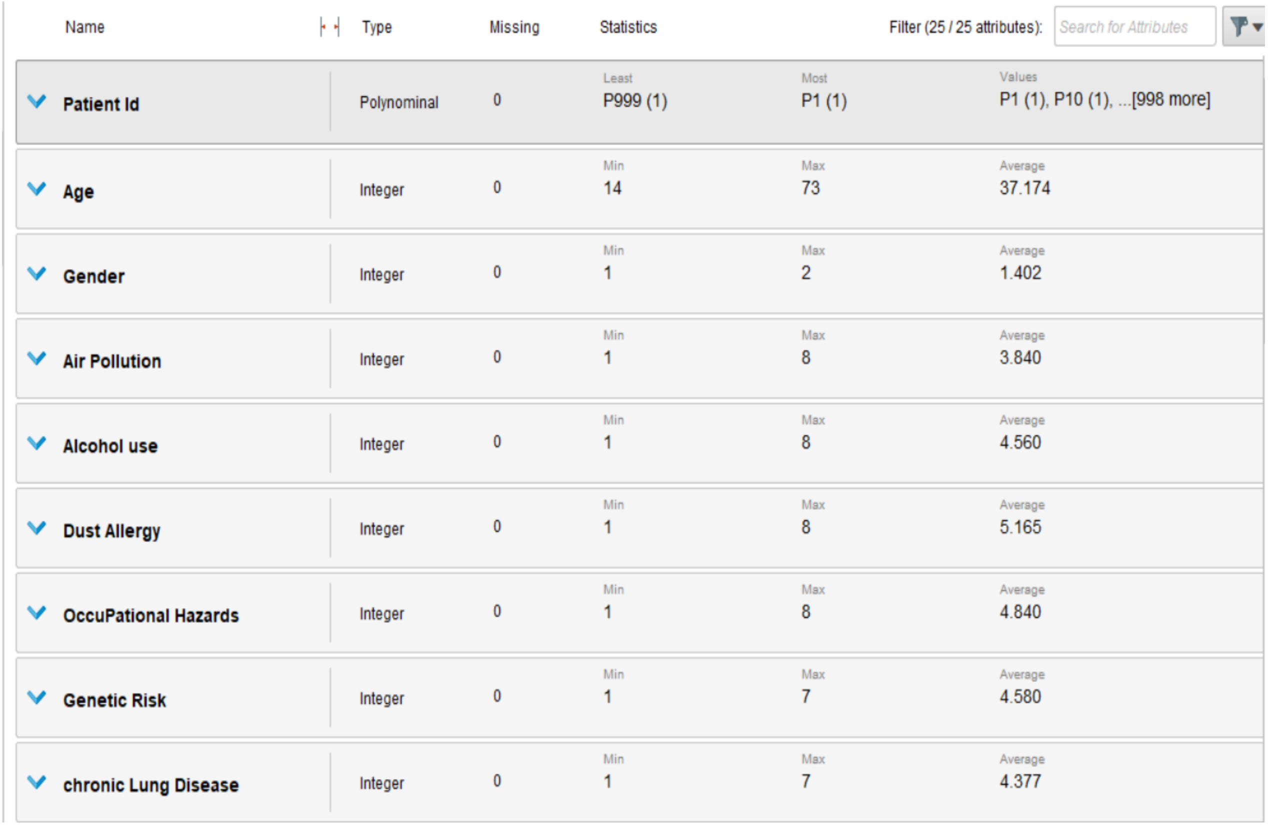
On Filtering (removing) the missing values:

Number of attributes: 25

Number of instances: 973

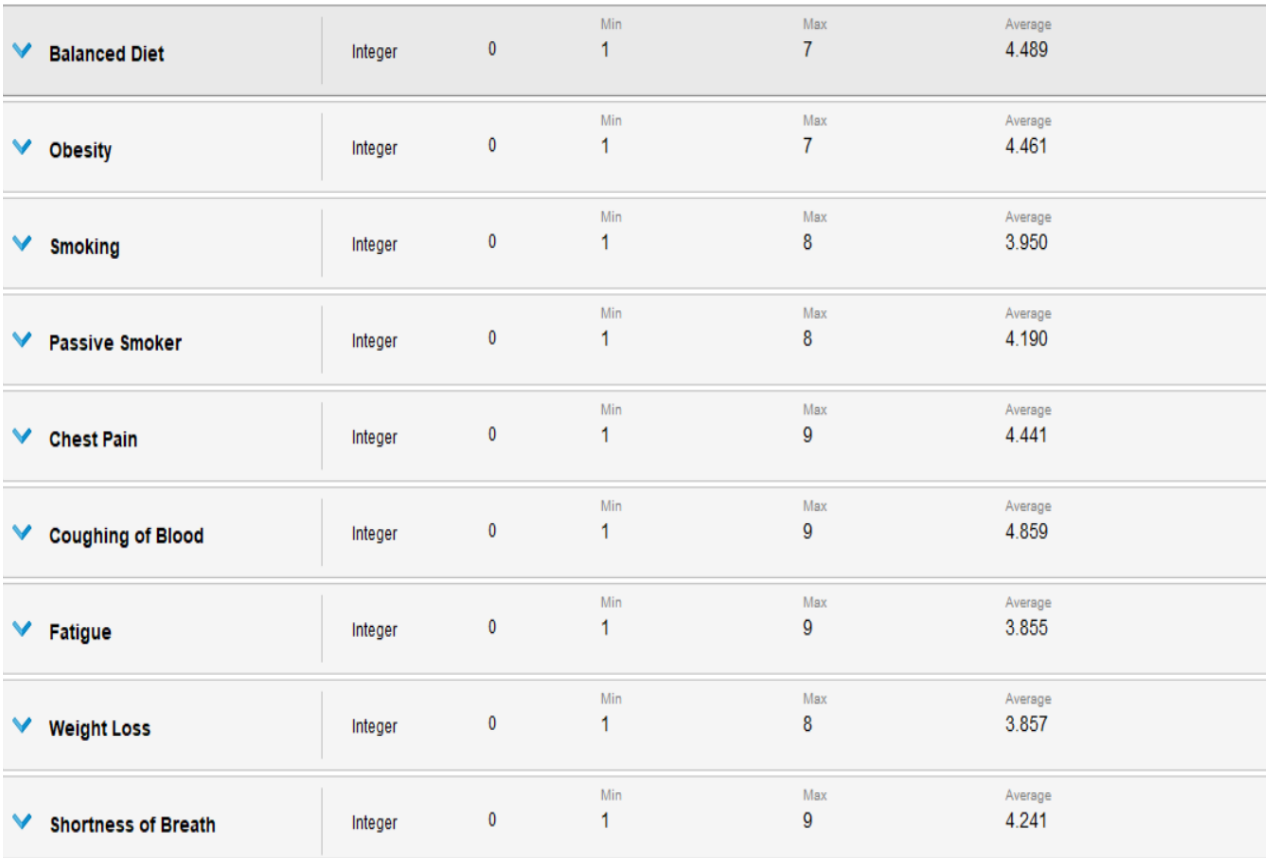
To avoid the loss of instances we replace missing values with **average.**

Then the statistical analysis is as follows:

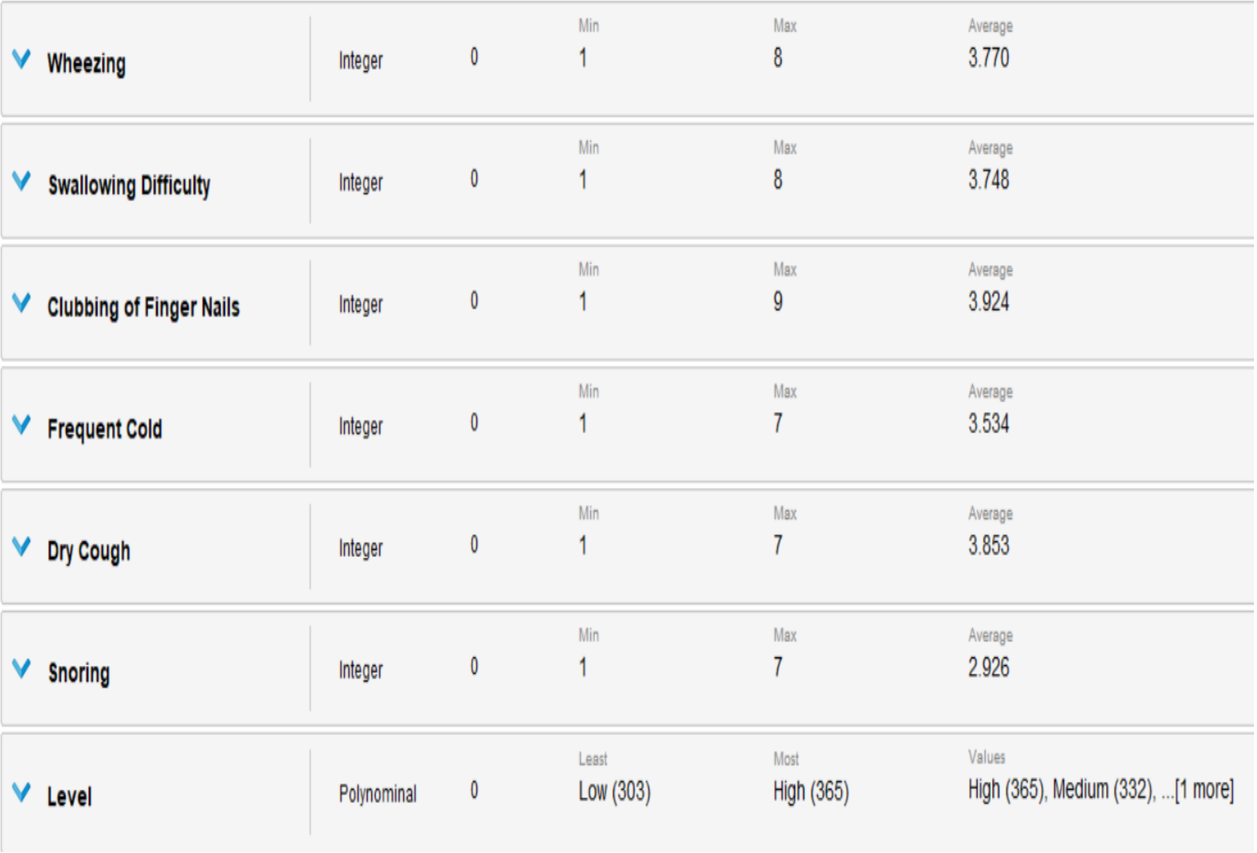


Filter (25 / 25 attributes): Search for Attributes

Name	Type	Missing	Statistics		
Patient Id	Polynominal	0	Least P999 (1)	Most P1 (1)	Values P1 (1), P10 (1), ...[998 more]
Age	Integer	0	Min 14	Max 73	Average 37.174
Gender	Integer	0	Min 1	Max 2	Average 1.402
Air Pollution	Integer	0	Min 1	Max 8	Average 3.840
Alcohol use	Integer	0	Min 1	Max 8	Average 4.560
Dust Allergy	Integer	0	Min 1	Max 8	Average 5.165
OccuPational Hazards	Integer	0	Min 1	Max 8	Average 4.840
Genetic Risk	Integer	0	Min 1	Max 7	Average 4.580
chronic Lung Disease	Integer	0	Min 1	Max 7	Average 4.377



			Min	Max	Average
Balanced Diet	Integer	0	1	7	4.489
Obesity	Integer	0	1	7	4.461
Smoking	Integer	0	1	8	3.950
Passive Smoker	Integer	0	1	8	4.190
Chest Pain	Integer	0	1	9	4.441
Coughing of Blood	Integer	0	1	9	4.859
Fatigue	Integer	0	1	9	3.855
Weight Loss	Integer	0	1	8	3.857
Shortness of Breath	Integer	0	1	9	4.241



			Min	Max	Average
Wheezing	Integer	0	1	8	3.770
Swallowing Difficulty	Integer	0	1	8	3.748
Clubbing of Finger Nails	Integer	0	1	9	3.924
Frequent Cold	Integer	0	1	7	3.534
Dry Cough	Integer	0	1	7	3.853
Snoring	Integer	0	1	7	2.926
Level	Polynominal	0	Least Low (303)	Most High (365)	Values High (365), Medium (332), ...[1 more]

SAMPLE CODE

Naive Bayes:

```
setwd("C://Users//PRATAP KUMAR//Documents//SWE2009 datamining project//data set")
cancer <- read.csv("lung1.csv")
summary(cancer)
test <- read.csv(file="C://Users//PRATAP KUMAR//Documents//SWE2009 datamining project//data set//test1.csv", header=TRUE, sep=",")
testdata<-data.frame(test)
#checking the distribution of the target variable
table(cancer$Level)
#Patitioning the dataset into training and testing sets
library(caret)
#pseudo-random number generator
set.seed(2)
# This will help to divide the package into training and testing sets.
inTrain1 <- createDataPartition(cancer$Level, p = 0.6, list = F)
datTrain1 <- cancer[inTrain1,]
datTest1 <- cancer[-inTrain1,]
#Check the rows and porportion of target variable for both training
nrow(datTrain1)
nrow(datTest1)
prop.table(table(datTrain1$Level))
prop.table(table(datTest1$Level))
#NaiveBayes in e1071
library(e1071)
#model building
# e1071model <- naiveBayes(CLASSLABEL ~ smoking + age +  cough, data=datTrain1)
e1071model <- naiveBayes(Level ~ ., data=datTrain1)
#prediction on test dataset
#Run the model again and predict classes by using the training set
```

```
e1071predictions <- predict(e1071model, datTest1)
#check prediction for the first top 5 rows in the testing data
head(e1071predictions, n=10)
head(datTest1,n=10)
#e1071predictions <- predict(e1071model, head(datTest1,n=1))
per<-predict(e1071model,testdata)
per
#print the confusion matrix
xtab <- table(e1071predictions, datTest1$Level)
library(caret)
#It is used to calculate the accuracy, precision, recall and F-Measure.
library(rminer)
confusionMatrix(xtab)
per<-predict(e1071model,testdata)
per
```

K-Star (K*) Classification:

```
require("class")
setwd("C://Users//PRATAP KUMAR//Documents//SWE2009 datamining project//data set")
cancer <- read.csv("lung1.csv")
# load cancer Dataset
str(cancer)
summary(cancer)
head(cancer)
set.seed(99) # required to reproduce the results
rnum<- sample(rep(1:1000)) # randomly generate numbers from 1 to 150
rnum
cancer<- cancer[rnum,] #randomize "cancer" dataset
head(cancer)
```

```
normalize <- function(x){
  return ((x-min(x))/(max(x)-min(x)))
}
cancer.new<-as.data.frame(lapply(cancer[,c(2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24)],normalize))
head(cancer.new)
cancer.train<- cancer.new[1:600,]
cancer.train.target<- cancer[1:600,25]
cancer.test<- cancer.new[601:1000,]
cancer.test.target<- cancer[601:1000,25]
summary(cancer.new)
anyNA(cancer.new)
model1<- knn(train=cancer.train, test=cancer.test, cl=cancer.train.target, k=31,prob=TRUE)
xtab<-table(cancer.test.target, model1)
library(caret)
#It is used to calculate the accuracy, precision, recall and F-Measure.
library(rminer)
confusionMatrix(xtab)
```

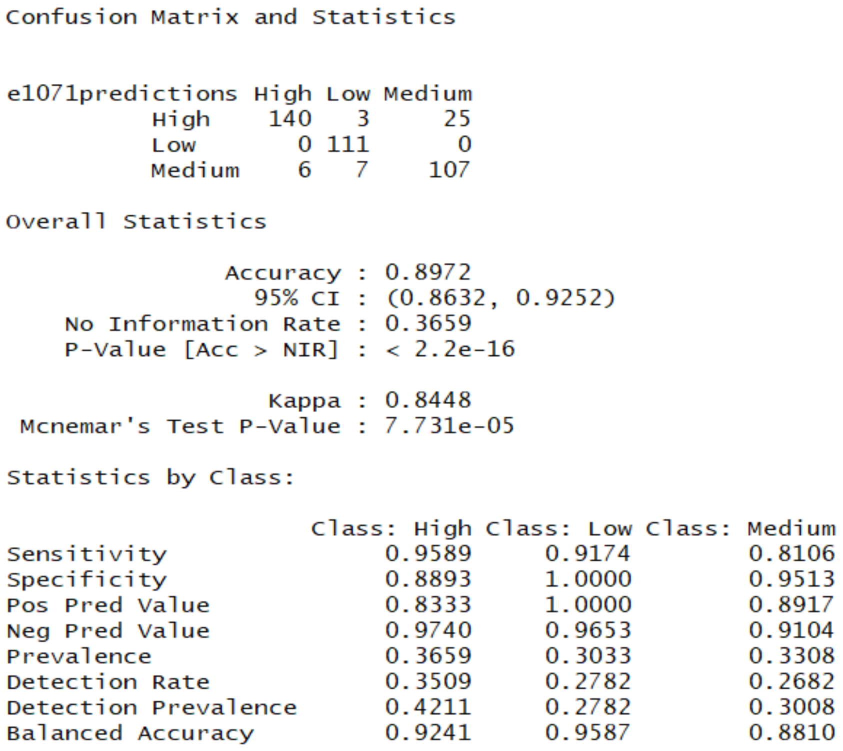
Decision Tree Algortihm:

```
setwd("C://Users//PRATAP KUMAR//Documents//SWE2009 datamining project//data set")
cancer <- read.csv("lung1.csv")
summary(cancer)
library(rpart)
library(rpart.plot)
library(caret)
set.seed(2)
inTrain1 <- createDataPartition(cancer$Level, p = 0.6, list = F)
cancer_train <- cancer[inTrain1,]
cancer_test <- cancer[-inTrain1,]
```

```
dtm<-
rpart(Level~Age+Gender+AirPollution+DustAllergy+OccuPationalHazards+GeneticRisk+ch
ronicLungDisease+BalancedDiet+Obesity+Alcoholuse+Smoking+PassiveSmoker+ChestPain
+CoughingofBlood+   Fatigue
        +WeightLoss+ShortnessofBreath+Wheezing+SwallowingDifficulty+ClubbingofFing
erNails +FrequentCold          +DryCough+Snoring,cancer_train,method="class")

#plot(dtm)

#text(dtm)

#rpart.plot(dtm)

rpart.plot(dtm,type=4,extra=101)

p<-predict(dtm,cancer_test,type="class")

xtab <-table(cancer_test$Level,p)

library(caret)

#It is used to calculate the accuracy, precision, recall and F-Measure.

library(rminer)

confusionMatrix(xtab)
```

RESULTS AND DISCUSSIONS

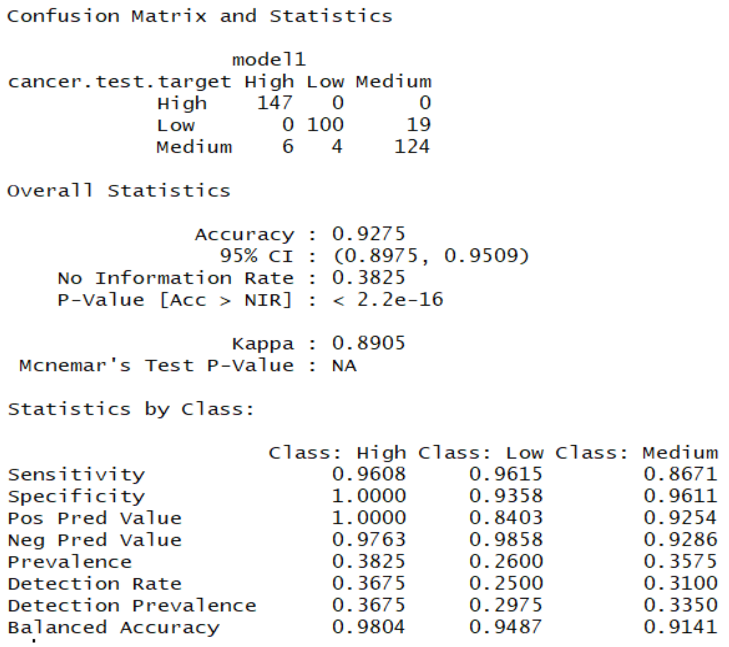
Navie Bayas



```
Confusion Matrix and Statistics


e1071predictions High Low Medium
          High    140   3     25
          Low       0 111      0
          Medium    6   7    107

Overall Statistics

               Accuracy : 0.8972
                 95% CI : (0.8632, 0.9252)
    No Information Rate : 0.3659
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8448
 Mcnemar's Test P-Value : 7.731e-05

Statistics by Class:

                     Class: High Class: Low Class: Medium
Sensitivity               0.9589     0.9174        0.8106
Specificity               0.8893     1.0000        0.9513
Pos Pred Value            0.8333     1.0000        0.8917
Neg Pred Value            0.9740     0.9653        0.9104
Prevalence                0.3659     0.3033        0.3308
Detection Rate            0.3509     0.2782        0.2682
Detection Prevalence      0.4211     0.2782        0.3008
Balanced Accuracy         0.9241     0.9587        0.8810
```

From the above results, using naïve bayes the accuracy is 89.2%

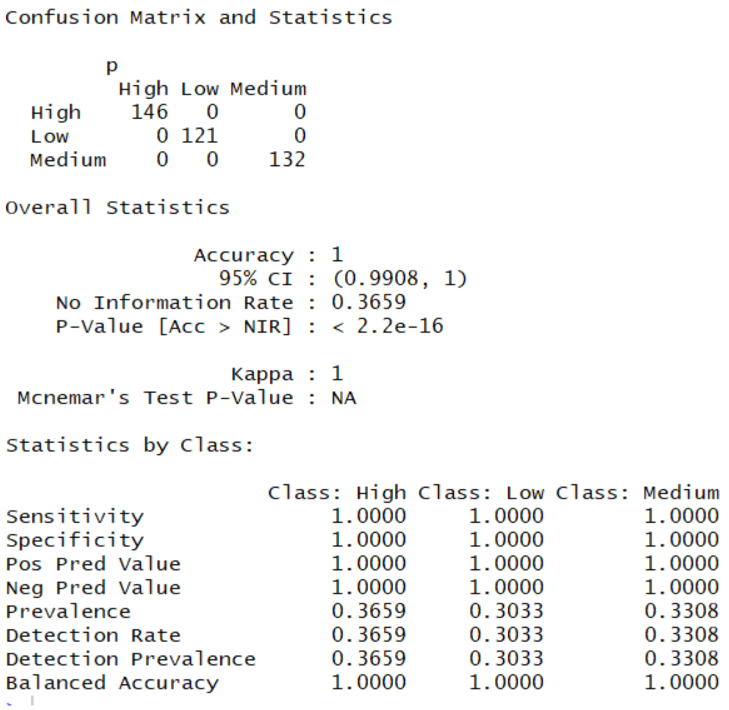
K-Star



```
Confusion Matrix and Statistics

                  model1
cancer.test.target High Low Medium
            High    147   0      0
            Low       0 100     19
            Medium    6   4    124

Overall Statistics

               Accuracy : 0.9275
                 95% CI : (0.8975, 0.9509)
    No Information Rate : 0.3825
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8905
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: High Class: Low Class: Medium
Sensitivity               0.9608     0.9615        0.8671
Specificity               1.0000     0.9358        0.9611
Pos Pred Value            1.0000     0.8403        0.9254
Neg Pred Value            0.9763     0.9858        0.9286
Prevalence                0.3825     0.2600        0.3575
Detection Rate            0.3675     0.2500        0.3100
Detection Prevalence      0.3675     0.2975        0.3350
Balanced Accuracy         0.9804     0.9487        0.9141
```

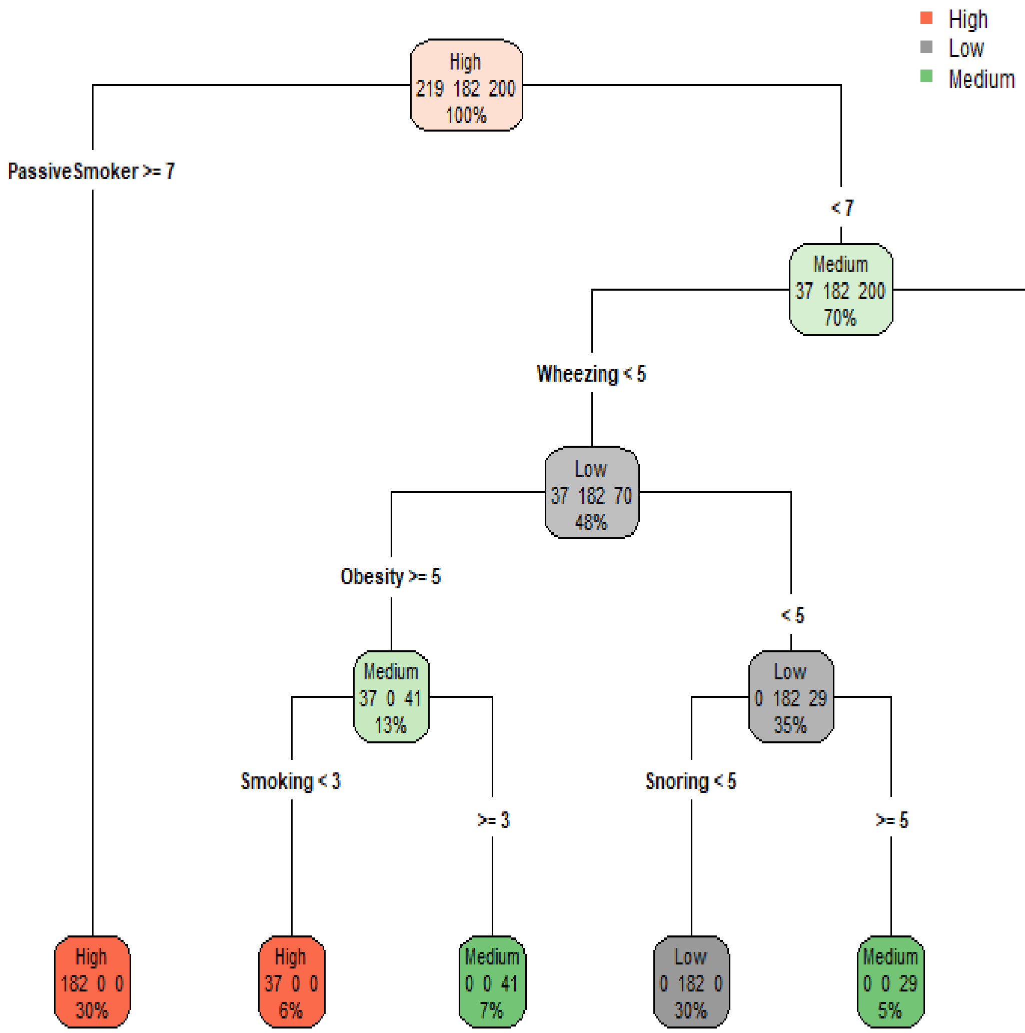
From the above results, using k-star the accuracy is 92.75%

Decision Tree



```
Confusion Matrix and Statistics

        p
         High Low Medium
  High    146   0      0
  Low       0 121      0
  Medium    0   0    132

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9908, 1)
    No Information Rate : 0.3659
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: High Class: Low Class: Medium
Sensitivity               1.0000     1.0000        1.0000
Specificity               1.0000     1.0000        1.0000
Pos Pred Value            1.0000     1.0000        1.0000
Neg Pred Value            1.0000     1.0000        1.0000
Prevalence                0.3659     0.3033        0.3308
Detection Rate            0.3659     0.3033        0.3308
Detection Prevalence      0.3659     0.3033        0.3308
Balanced Accuracy         1.0000     1.0000        1.0000
```

From the above results, using decision tree the accuracy is 100%



CONCLUSION

From the results

Algoritham	**Accuracy**
Navie Bayas	89.72
Decision Tree	100
K-Star	92.75

The data provides more number of lung cancer instances .In this, the most of classification algorithms like decision tree, Naïve Bayes, K-Star algorithms give the good results for prediction of lung cancer. The best algoritham for this dataset is Decision Tree as it qives 100% accuracy.

REFERENCE:

[1] Department of Biochemistry, Maharshi Dayanand University, Rohtak, India

Received: 01.10.2017, Accepted: 20.10.2017, Published: 30.10.2017

[2] Department of Radiology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK Submitted Apr 07, 2018. Accepted for publication May 22, 2018.

[3] Dept.of biological research, Ioannina, Greece Accepted for publication 15 November, 2014

[4] Department of Computer Science the George Washington University 2015

[5] Springer Nature Singapore Pte Ltd. 2019 J. Nayak et al. (eds.), Soft Computing in Data Analytics, Advances in Intelligent Systems and Computing 758

[6] https://data.world/cancerdatahp/lung-cancer-data

[7] Dubey, A.K., Umesh, G., Sonal, J.: Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pac. J. Cancer Prev. 16(10), 4237–4245 (2015)

[8] Kourou, K., et al.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13,8 –17 (2015)

[9] Lynch, C., et al.: Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int. J. Med. Inform.