AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Project Cover Page

Project Title:	Comparison of Cancer Prediction with Supervised Learning Models		
Project No:	1	Date of Submission:	07-08-2022
Course Title:	DATA WAREHOUSING AND DATA MINING		
Course Code:	CSE 4285	Section:	C
Semester:	Summer 2021-22	Course Teacher:	AKINUL ISLAM JONY

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	SAIDUL MURSALIN KHAN	19-41766-3	BSc [CSE]	
2	MD. SHAHRIAR ISLAM SHAIKAT	19-39972-1	BSc [CSE]	
3	MD. NAZIM HASAN	19-40118-1	BSc [CSE]	
4	TANHA REJA	19-40151-1	BSc [CSE]	

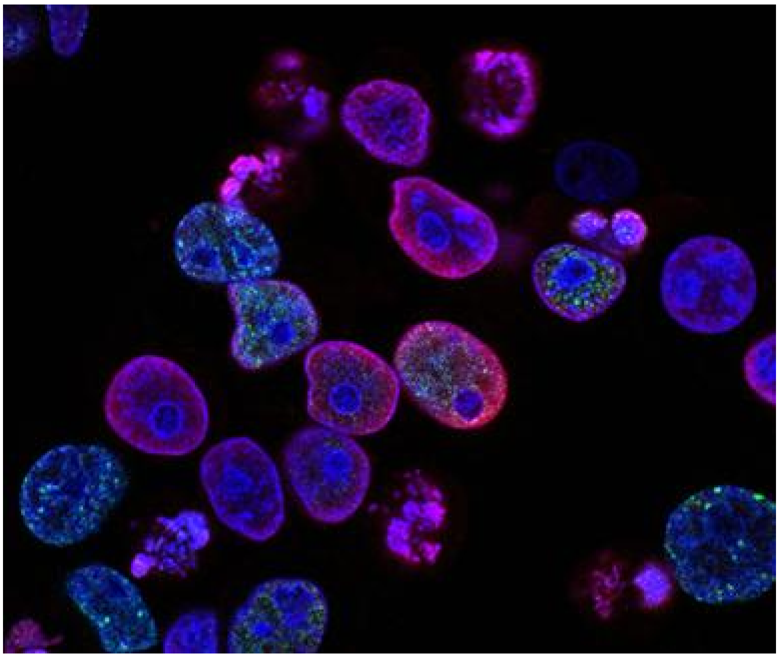
Faculty use only

FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Catalog

Section 1: Project Overview

Many people's lives are cut short due to cancer. However, due to the age of big data we are able to combat this malicious disease in a way. Big Data helps to analyse massive amount of information available and also helps to categorize data according to different attributes so that more detailed information is available, forming patterns which doctors can use to predict & treat cancer. We have collected a dataset from Kaggle. This dataset contains information about hundreds of cancer patients about their lifestyles. With this dataset we are going to train our machine learning model to identify patterns and relationships among the data in order to predict the possibility of having cancer.

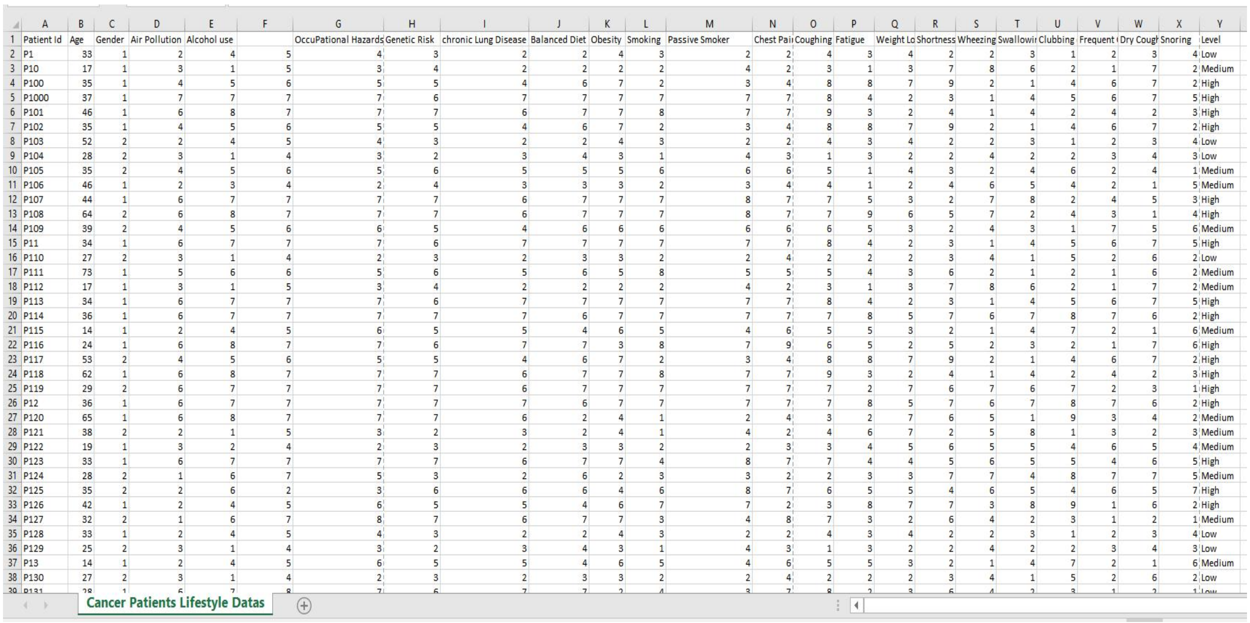


Section 2: Dataset Overview

Kaggle Link of Dataset:

https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

Snapshot of Dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Patient Id	Age	Gender	Air Pollution	Alcohol use		OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pai	Coughing	Fatigue	Weight Lo	Shortness	Wheezing	Swallowi	Clubbing	Frequent	Dry Cough	Snoring	Level
2	P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	Low
3	P10	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	Medium
4	P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
5	P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
6	P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	High
7	P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
8	P103	52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	Low
9	P104	28	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	Low
10	P105	35	2	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2	4	6	2	4	1	Medium
11	P106	46	1	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	6	5	4	2	1	5	Medium
12	P107	44	1	6	7	7	7	7	6	7	7	7	8	7	7	5	3	2	7	8	2	4	5	3	High
13	P108	64	2	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7	2	4	3	1	4	High
14	P109	39	2	4	5	6	6	5	4	6	6	6	6	6	6	5	3	2	4	3	1	7	5	6	Medium
15	P11	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
16	P110	27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	Low
17	P111	73	1	5	6	6	5	6	5	6	5	8	5	5	5	4	3	6	2	1	2	1	6	2	Medium
18	P112	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	Medium
19	P113	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
20	P114	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	High
21	P115	14	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	Medium
22	P116	24	1	6	8	7	7	6	7	7	3	8	7	9	6	5	2	5	2	3	2	1	7	6	High
23	P117	53	2	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
24	P118	62	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	High
25	P119	29	2	6	7	7	7	7	6	7	7	7	7	7	7	2	7	6	7	6	7	2	3	1	High
26	P12	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	High
27	P120	65	1	6	8	7	7	7	6	2	4	1	2	4	3	2	7	6	5	1	9	3	4	2	Medium
28	P121	38	2	2	1	5	3	2	3	2	4	1	4	2	4	6	7	2	5	8	1	3	2	3	Medium
29	P122	19	1	3	2	4	2	3	2	3	3	2	2	3	3	4	5	6	5	5	4	6	5	4	Medium
30	P123	33	1	6	7	7	7	7	6	7	7	4	8	7	7	4	4	5	6	5	5	4	6	5	High
31	P124	28	2	1	6	7	5	3	2	6	2	3	3	2	2	3	3	7	7	4	8	7	7	5	Medium
32	P125	35	2	2	6	2	3	6	6	6	4	6	8	7	6	5	5	4	6	5	4	6	5	7	High
33	P126	42	1	2	4	5	6	5	5	4	6	7	7	2	3	8	7	7	3	8	9	1	6	2	High
34	P127	32	2	1	6	7	8	7	6	7	7	3	4	8	7	3	2	6	4	2	3	1	2	1	Medium
35	P128	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	Low
36	P129	25	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	Low
37	P13	14	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	Medium
38	P130	27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	Low

Description About Dataset:

Our dataset contains 1000 instances and 25 attributes. This attributes contain **Patient Id, Age, Gender,** environment they live in or exposed to such as **Air Pollution** intensity level from 1 to 8, 1 being the lowest & 8 being the highest, their lifestyle such as **Diet**, their bad habits like **Alcohol Usage, Smoking** etc. And finally classifier attribute **Level** indicating their Low, Medium & High possibility of having cancer.

Here are the details of all attributes & values of our dataset:

Attribute	Values
Patient Id	P1 - P1000
Age	14 - 77
Gender	1(Male) - 2(Female)

Attribute	Intensity Level
Air Pollution	1(Lowest) - 8(Highest)
Alcohol Use	1(Lowest) - 8(Highest)
Dust Allergy	1(Lowest) - 8(Highest)
Occupational Hazards	1(Lowest) - 8(Highest)
Genetic Risk	1(Lowest) - 7(Highest)
Chronic Lung Disease	1(Lowest) - 7(Highest)
Balanced Diet	1(Lowest) - 7(Highest)
Obesity	1(Lowest) - 7(Highest)
Smoking	1(Lowest) - 8(Highest)
Passive Smoker	1(Lowest) - 8(Highest)
Chest Pain	1(Lowest) - 9(Highest)
Coughing of Blood	1(Lowest) - 9(Highest)
Fatigue	1(Lowest) - 9(Highest)
Weight Loss	1(Lowest) - 8(Highest)
Shortness of Breath	1(Lowest) - 9(Highest)
Wheezing	1(Lowest) - 8(Highest)
Swallowing Difficulty	1(Lowest) - 8(Highest)
Clubbing of Finger Nails	1(Lowest) - 9(Highest)
Frequent Cold	1(Lowest) - 7(Highest)
Dry Cough	1(Lowest) - 7(Highest)
Snoring	1(Lowest) - 7(Highest)
Classifier Attribute	**Values**
Level	Low - Medium - High

Figures:

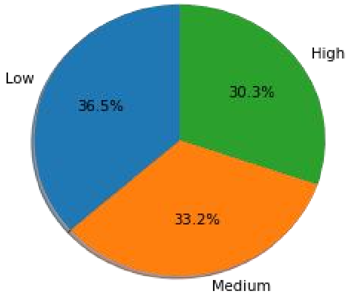


Figure 1 : Percentage of 3 Possibility Levels Of Cancer Patients

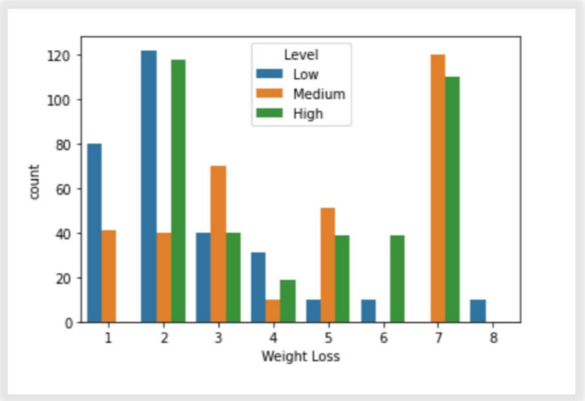


Figure 2

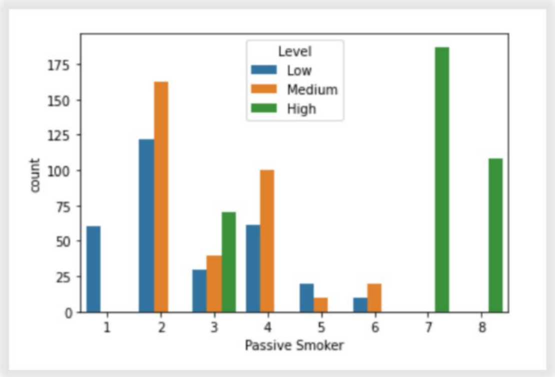


Figure 4

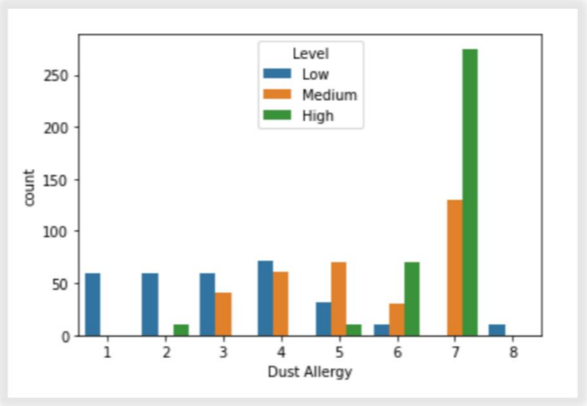


Figure 3

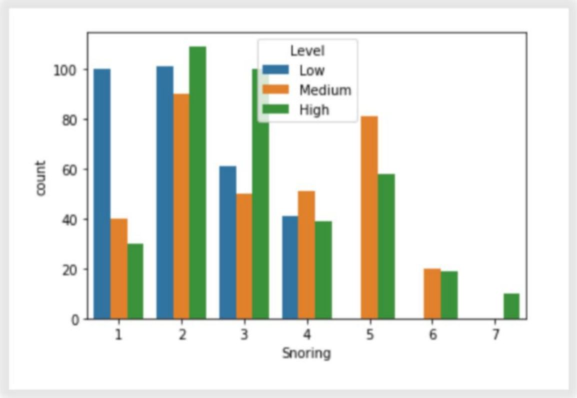


Figure 5

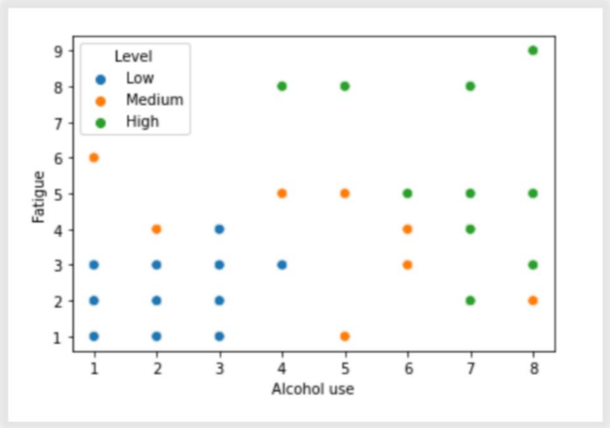
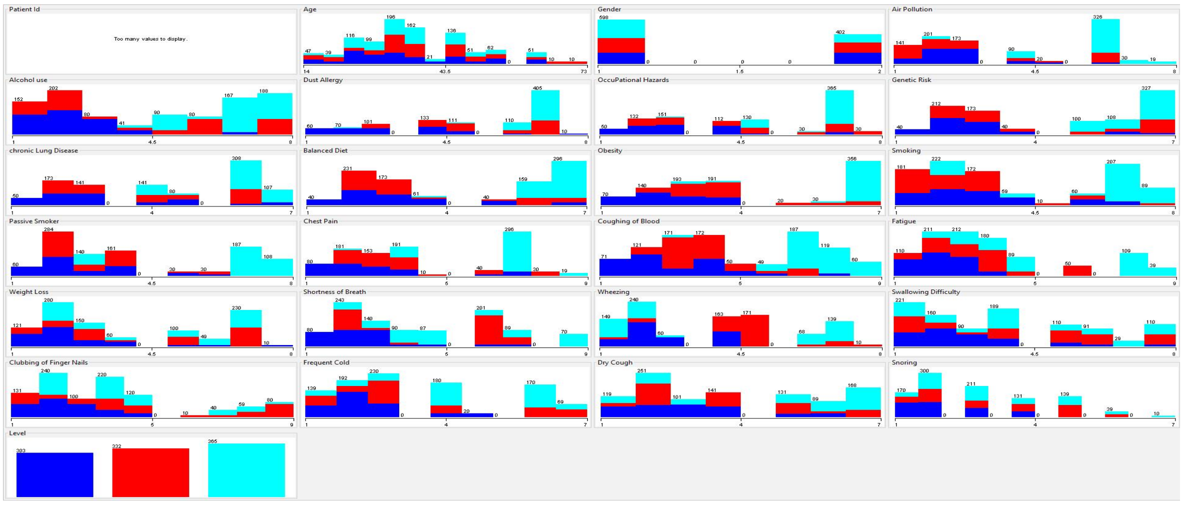


Figure 6

Section 3: Model Development

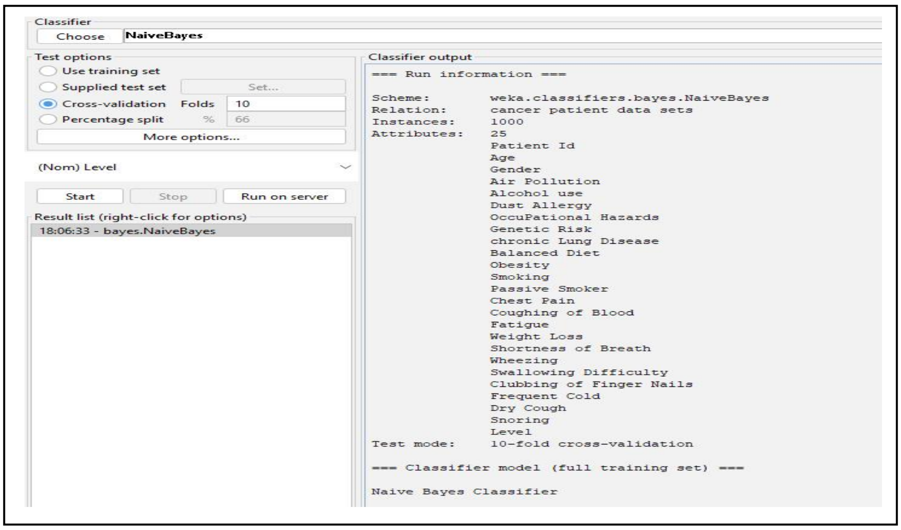
All Attributes Visualization:



Model Performance Test:

We will use K-Fold Cross Validation in order to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset.

Naïve Bayes Classifier Model:



Classifier: NaiveBayes

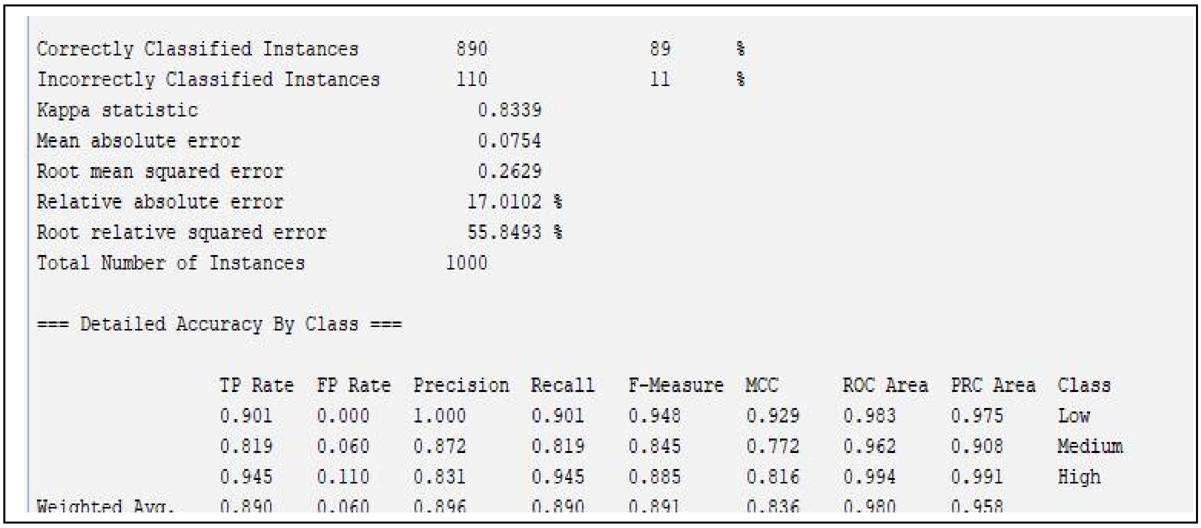
● Cross-validation Folds 10

○ Percentage split % 66

18:06:33 - bayes.NaiveBayes

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     cancer patient data sets
Instances:    1000
Attributes:   25
              Patient Id
              Age
              Gender
              Air Pollution
              Alcohol use
              Dust Allergy
              OccuPational Hazards
              Genetic Risk
              chronic Lung Disease
              Balanced Diet
              Obesity
              Smoking
              Passive Smoker
              Chest Pain
              Coughing of Blood
              Fatigue
              Weight Loss
              Shortness of Breath
              Wheezing
              Swallowing Difficulty
              Clubbing of Finger Nails
              Frequent Cold
              Dry Cough
              Snoring
              Level
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier
```

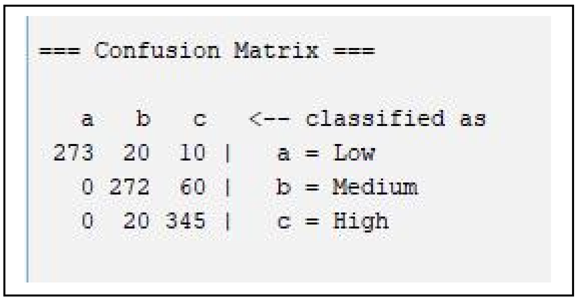
Predictive Accuracy:



```
Correctly Classified Instances         890               89      %
Incorrectly Classified Instances       110               11      %
Kappa statistic                          0.8339
Mean absolute error                      0.0754
Root mean squared error                  0.2629
Relative absolute error                 17.0102 %
Root relative squared error             55.8493 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.901    0.000    1.000      0.901    0.948      0.929    0.983     0.975     Low
                 0.819    0.060    0.872      0.819    0.845      0.772    0.962     0.908     Medium
                 0.945    0.110    0.831      0.945    0.885      0.816    0.994     0.991     High
Weighted Avg.    0.890    0.060    0.896      0.890    0.891      0.836    0.980     0.958
```

We can achieve 89% predictive accuracy by using Naïve Bayes Classifier Model. From 1000 instance, Naïve Bayes can classify 890 instances correctly and 110 instance incorrectly.

Confusion Matrix:

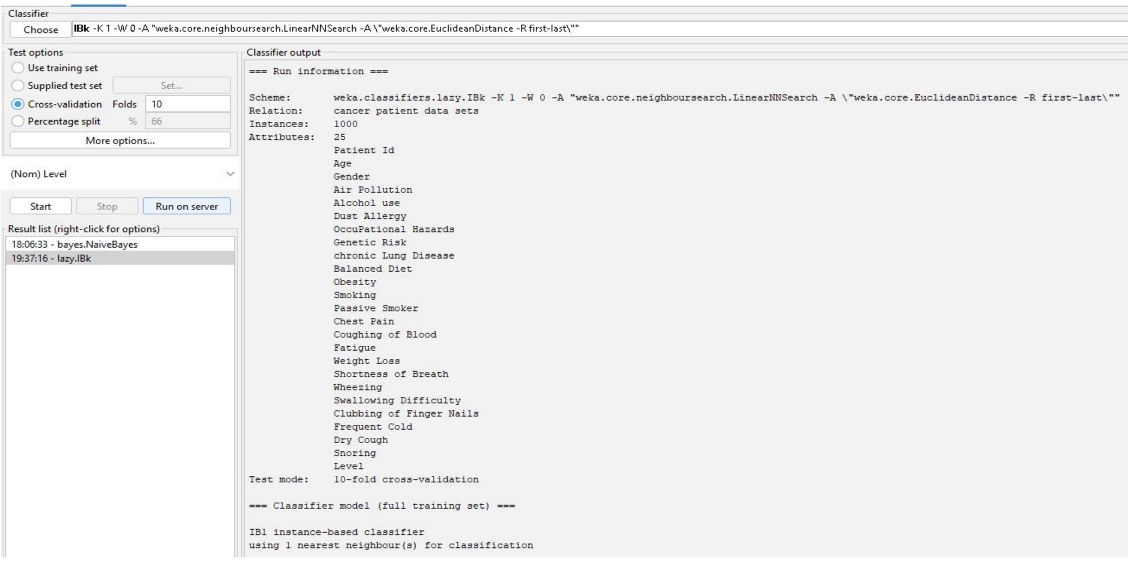


```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 273  20  10 |   a = Low
   0 272  60 |   b = Medium
   0  20 345 |   c = High
```

Low	**Medium**	**High**
True Negative = 273	False Positive = 20	True Negative = 10
False Negative = 0	True Positive = 272	False Negative = 60
True Negative = 0	False Negative = 20	True Negative = 345

The confusion matrix is listed at the bottom, and you can see that a wealth of classification statistics are also presented. Confusion matrices is visualizing important predictive analytics like recall, specificity, accuracy, and precision. It give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives. The confusion matrix assigns letters a and b to the class values and provides expected class values in rows and predicted class values (“classified as”) for each column.

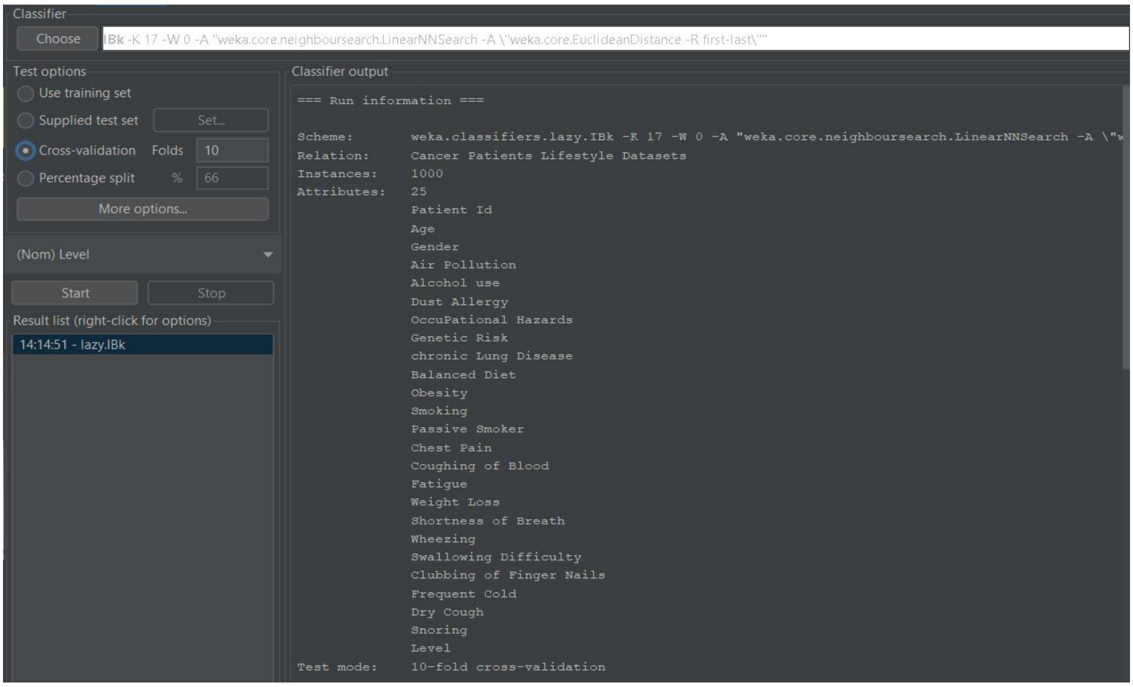
K-Nearest Neighbors (KNN) Classification Model:



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     cancer patient data sets
Instances:    1000
Attributes:   25
              Patient Id
              Age
              Gender
              Air Pollution
              Alcohol use
              Dust Allergy
              OccuPational Hazards
              Genetic Risk
              chronic Lung Disease
              Balanced Diet
              Obesity
              Smoking
              Passive Smoker
              Chest Pain
              Coughing of Blood
              Fatigue
              Weight Loss
              Shortness of Breath
              Wheezing
              Swallowing Difficulty
              Clubbing of Finger Nails
              Frequent Cold
              Dry Cough
              Snoring
              Level
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
```

1 FOLD

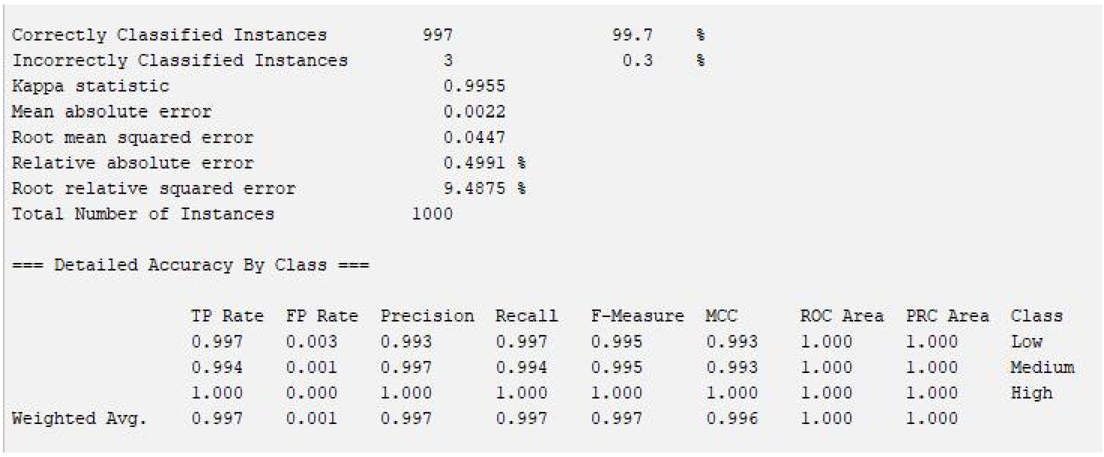


IBk -K 17 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 17 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"
Relation:     Cancer Patients Lifestyle Datasets
Instances:    1000
Attributes:   25
              Patient Id
              Age
              Gender
              Air Pollution
              Alcohol use
              Dust Allergy
              OccuPational Hazards
              Genetic Risk
              chronic Lung Disease
              Balanced Diet
              Obesity
              Smoking
              Passive Smoker
              Chest Pain
              Coughing of Blood
              Fatigue
              Weight Loss
              Shortness of Breath
              Wheezing
              Swallowing Difficulty
              Clubbing of Finger Nails
              Frequent Cold
              Dry Cough
              Snoring
              Level
Test mode:    10-fold cross-validation
```

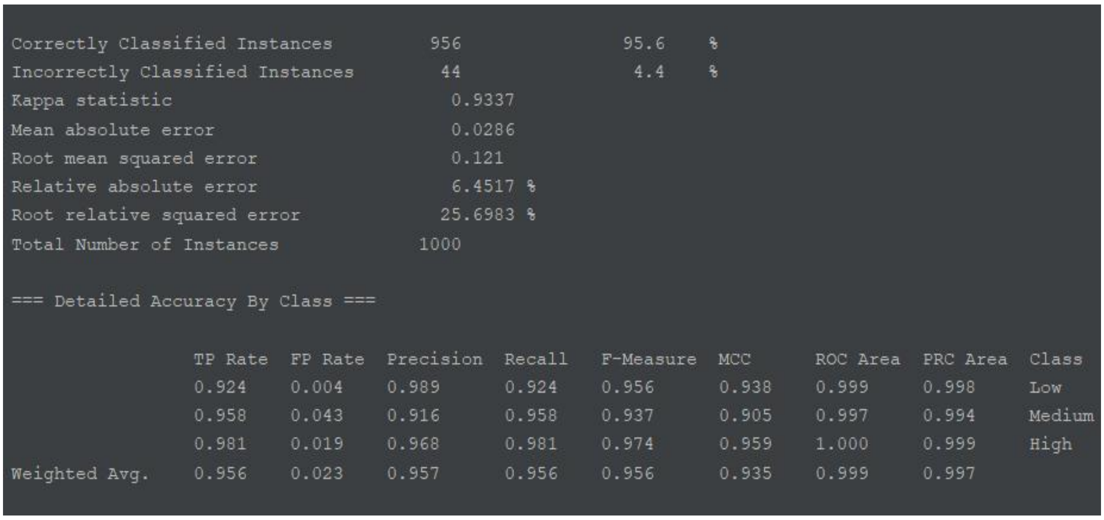
17 FOLD

Predictive Accuracy:



```
Correctly Classified Instances         997               99.7    %
Incorrectly Classified Instances         3                0.3    %
Kappa statistic                          0.9955
Mean absolute error                      0.0022
Root mean squared error                  0.0447
Relative absolute error                  0.4991 %
Root relative squared error              9.4875 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.003    0.993      0.997    0.995      0.993    1.000     1.000     Low
                 0.994    0.001    0.997      0.994    0.995      0.993    1.000     1.000     Medium
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     High
Weighted Avg.    0.997    0.001    0.997      0.997    0.997      0.996    1.000     1.000
```

1 FOLD



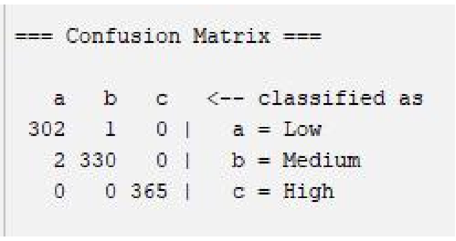
```
Correctly Classified Instances         956               95.6    %
Incorrectly Classified Instances        44                4.4    %
Kappa statistic                          0.9337
Mean absolute error                      0.0286
Root mean squared error                  0.121
Relative absolute error                  6.4517 %
Root relative squared error             25.6983 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.924    0.004    0.989      0.924    0.956      0.938    0.999     0.998     Low
                 0.958    0.043    0.916      0.958    0.937      0.905    0.997     0.994     Medium
                 0.981    0.019    0.968      0.981    0.974      0.959    1.000     0.999     High
Weighted Avg.    0.956    0.023    0.957      0.956    0.956      0.935    0.999     0.997
```

17 FOLD

For 1 Fold, we can achieve 99% predictive accuracy by using K-nearest neighbors (KNN) model. From 1000 instance, KNN can classify 997 instances correctly and 3 instance incorrectly.

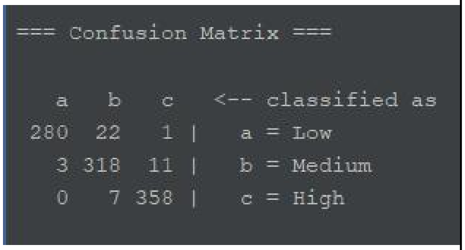
And For 17 Fold, we can achieve 95.6% predictive accuracy by using K-nearest neighbors (KNN) model. From 1000 instance, KNN can classify 956 instances correctly and 44 instance incorrectly.

Confusion Matrix:



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 302   1   0 |   a = Low
   2 330   0 |   b = Medium
   0   0 365 |   c = High
```

1 FOLD



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 280  22   1 |   a = Low
   3 318  11 |   b = Medium
   0   7 358 |   c = High
```

17 FOLD

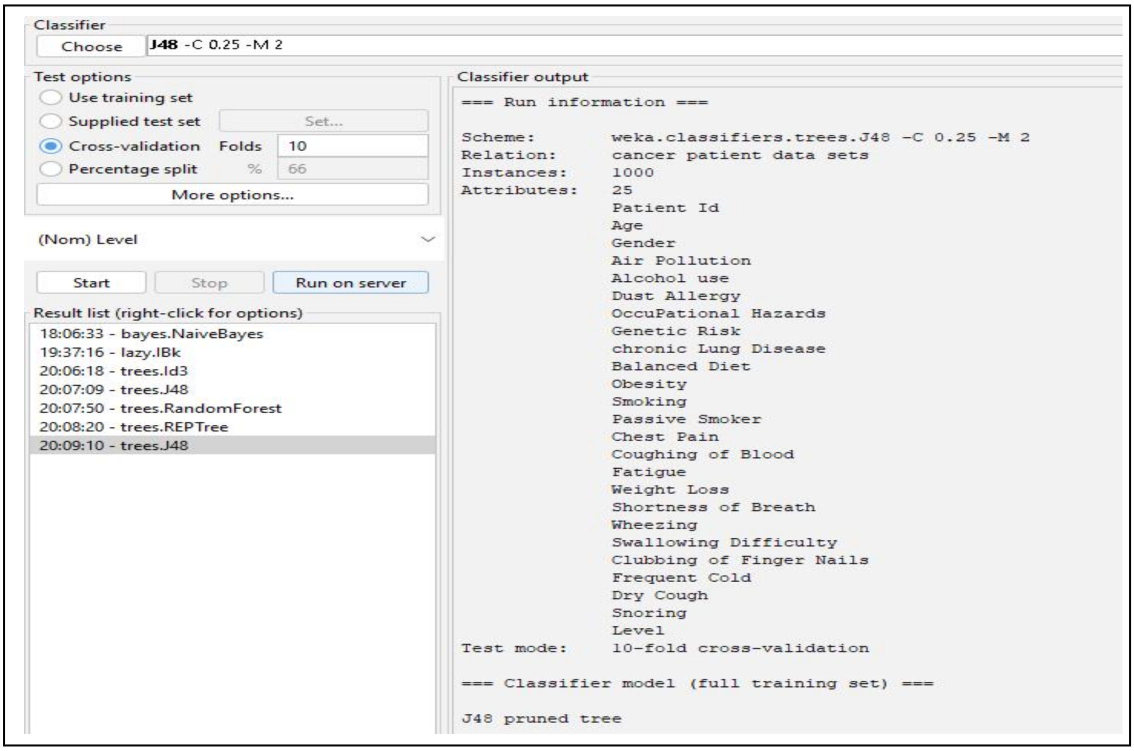
Low	Medium	High
True Negative = 302	False Positive = 1	True Negative = 0
False Negative = 2	True Positive = 330	False Negative = 0
True Negative = 0	False Negative = 0	True Negative = 365

Table 1: 1 FOLD

Low	Medium	High
True Negative = 280	False Positive = 22	True Negative = 1
False Negative = 3	True Positive = 318	False Negative = 11
True Negative = 0	False Negative = 7	True Negative = 358

Table 2: 17 FOLD

Decision Tree Classifier Model:



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

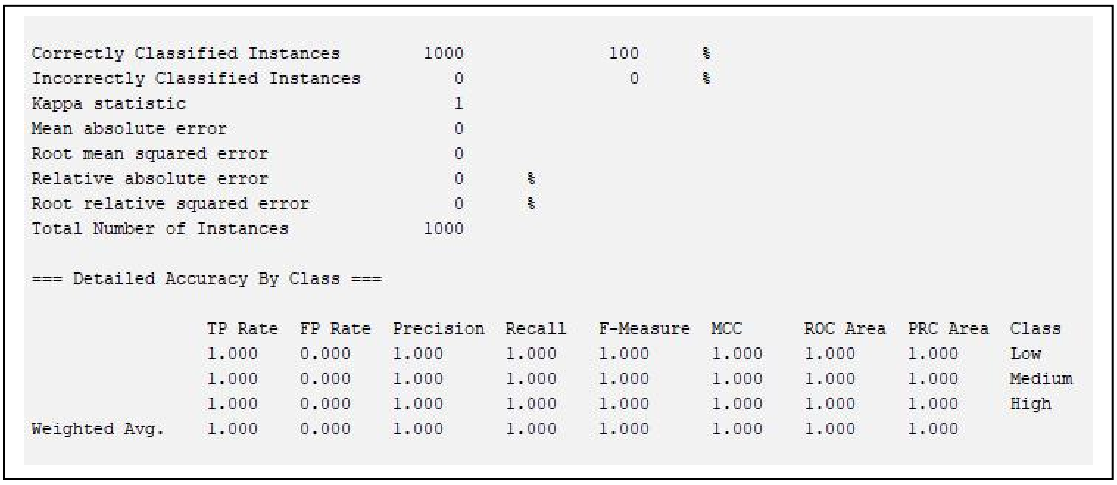
(Nom) Level

Result list (right-click for options)

```
18:06:33 - bayes.NaiveBayes
19:37:16 - lazy.IBk
20:06:18 - trees.Id3
20:07:09 - trees.J48
20:07:50 - trees.RandomForest
20:08:20 - trees.REPTree
20:09:10 - trees.J48
```

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     cancer patient data sets
Instances:    1000
Attributes:   25
              Patient Id
              Age
              Gender
              Air Pollution
              Alcohol use
              Dust Allergy
              OccuPational Hazards
              Genetic Risk
              chronic Lung Disease
              Balanced Diet
              Obesity
              Smoking
              Passive Smoker
              Chest Pain
              Coughing of Blood
              Fatigue
              Weight Loss
              Shortness of Breath
              Wheezing
              Swallowing Difficulty
              Clubbing of Finger Nails
              Frequent Cold
              Dry Cough
              Snoring
              Level
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
```

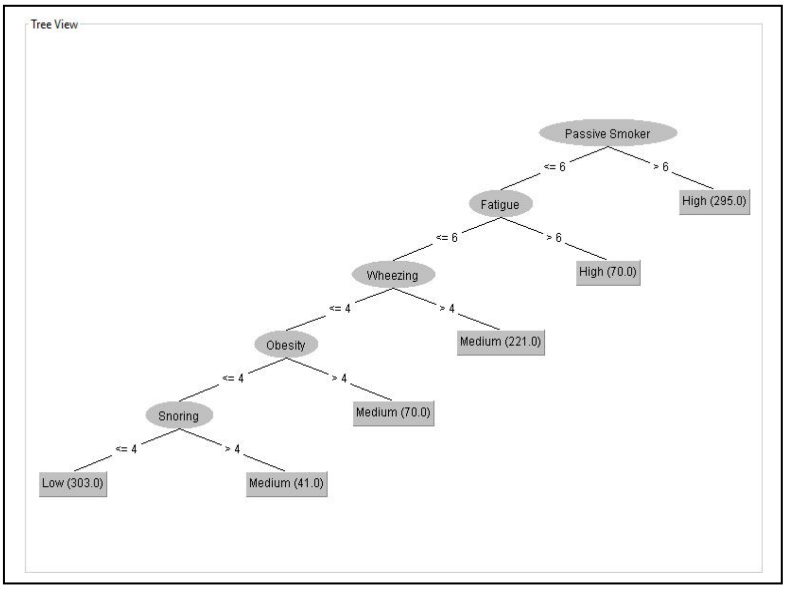
Predictive Accuracy:

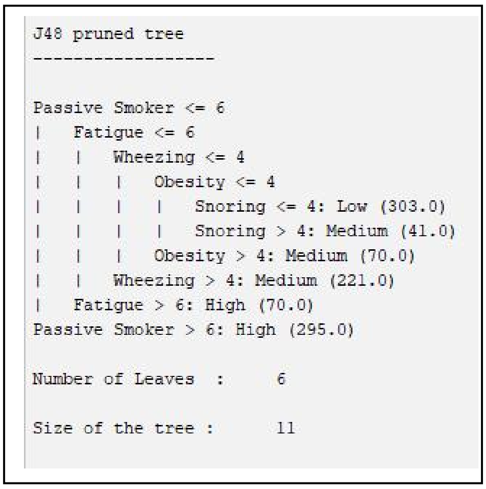


```
Correctly Classified Instances        1000              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Low
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Medium
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     High
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

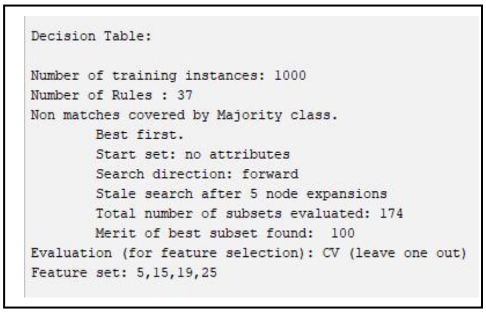
Here, We can achieve 100% predictive accuracy by using Decision Tree Classifier Model. From 1000 instance, Decision Tree can classify 1000 instances correctly and no instance incorrectly.

Tree:





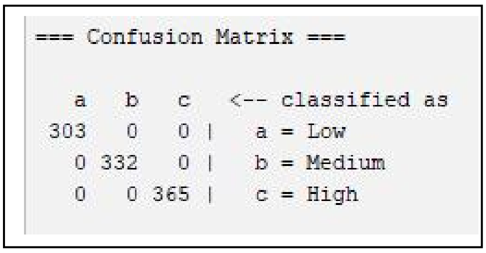
```
J48 pruned tree
------------------

Passive Smoker <= 6
|   Fatigue <= 6
|   |   Wheezing <= 4
|   |   |   Obesity <= 4
|   |   |   |   Snoring <= 4: Low (303.0)
|   |   |   |   Snoring > 4: Medium (41.0)
|   |   |   Obesity > 4: Medium (70.0)
|   |   Wheezing > 4: Medium (221.0)
|   Fatigue > 6: High (70.0)
Passive Smoker > 6: High (295.0)

Number of Leaves  :     6

Size of the tree :      11
```



```
Decision Table:

Number of training instances: 1000
Number of Rules : 37
Non matches covered by Majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 174
        Merit of best subset found:  100
Evaluation (for feature selection): CV (leave one out)
Feature set: 5,15,19,25
```

Here, We can see the Decision Tree and it's information. Here I have used J48 Model and Decision Table. The tree has 6 leaves, it's size is 11. This model provides us 37 rules in order to classify whole dataset. Total number of subset evaluated is 174. Merit of best subset found is 100. Feature set is 5, 15, 19, 25.

Confusion Matrix:



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 303   0   0 |   a = Low
   0 332   0 |   b = Medium
   0   0 365 |   c = High
```

Low	**Medium**	**High**
True Negative = 303	False Positive = 0	True Negative = 0
False Negative = 0	True Positive = 332	False Negative = 0
True Negative = 0	False Negative = 0	True Negative = 365

Section 4:Discussion & Conclusion

Comparison:

	Naïve Bayes	**KNN**	**Decision Tree**
Training Time	Slower than KNN, faster than Decision Tree	Fastest	Slower than Naïve Bayes and KNN
Accuracy	89%	95.6%	100%
Misclassification	110 instances	44 instances	0 instances
Works well for	Large Datasets	Small Datasets	Both Large and Small Datasets

Discussion:

In this project we have observed the results for three different Supervised Learning based Classification Model. We can see the step by step procedure of Naïve Bayes, KNN and Decision Tree Algorithm. The accuracy are accordingly 89%, 95.6% & 100%.

Naïve Bayes gives us lowest accuracy and Decision Tree gives us highest accuracy. By the confusion matrix, we have defined the performance for all three of the classification algorithm. Those visualizes and summarizes the performance of a classification algorithm. After conducting this experiment we can say that Decision Tree is the most effective algorithm for this dataset.