Phase 5: Modeling Data

Team Name: La Casa de Papel

Team Members:- Smit Kakadiya, Bhargav Patel, Saurav Aich

UNH Emails:-

Saurav Aich:- saich1@unh.newhaven.edu

Smit Mansukhbhai Kakadiya:- skaka3@unh.newhaven.edu

Bhargav Prakashchandra Patel:- bpate21@unh.newhaven.edu

Student ID:-

Saurav Aich:- 00718242

Smit Mansukhbhai Kakadiya:- 00703186

Bhargav Prakashchandra Patel:- 00711864

Team Head:-

Saurav Aich

Research Question : -

Our dataset has been taken from Webmd which contains data from the year 1990 to 2012 where people across different age groups and females who are pregnant. It also has values such as plasma, insulin . These data have been collected to answer the question which is What makes the US population susceptible to diabetes?

List of Data Mining Techniques used : -

- J48 Pruned Tree
- Bayes.Net

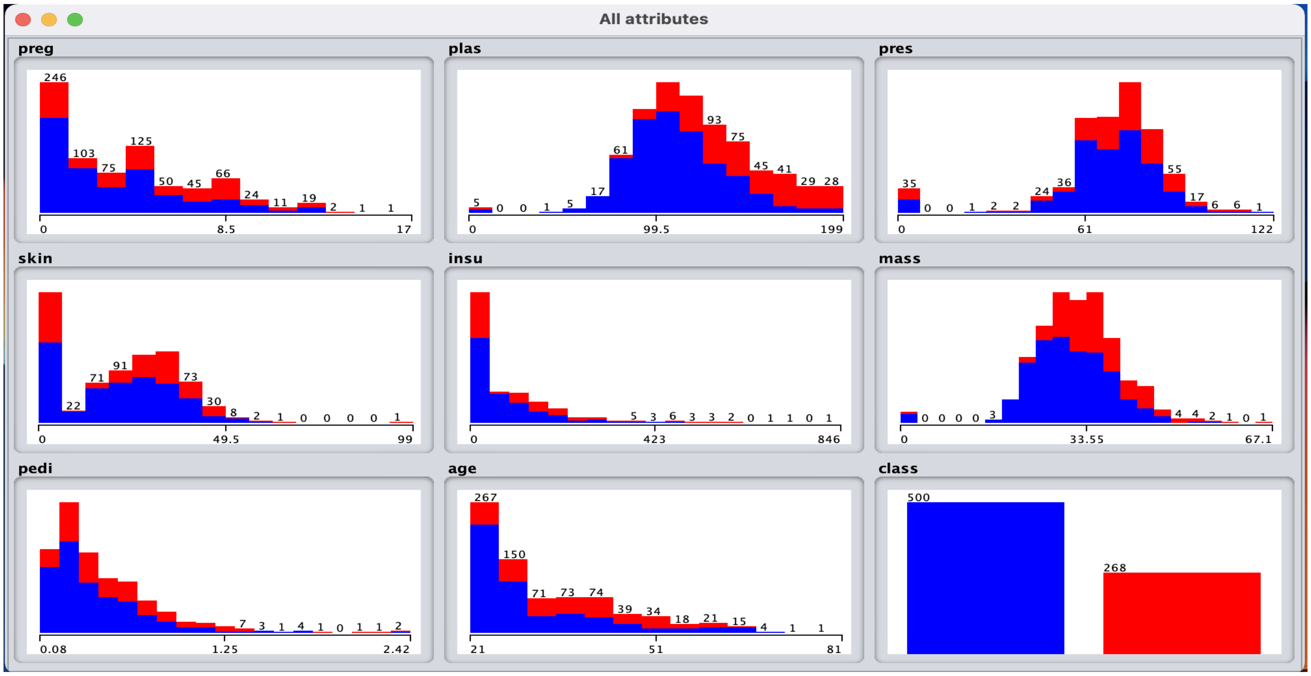
Hardware Used : -

Weka Explorer, an open source machine learning tool has been used to classify data to measure the accuracy of the dataset to come up with a conclusion.

Histogram : -

The main GUI of weka shows a histogram for attribute distributions for a single attribute at a time. The histogram displays all the ranges and how many samples fall in each range.

From the dataset it has been observed that preg samples appear to have 17 distinct values and 2 unique values with standard deviation of 3.845 and plasma appears to have 136 distinct values and 19 unique values. These can be explained as the various patterns that data mining uses to perform classification and other functions.



Outcome and Visualization : -

J48 Pruned Tree

This classification makes it easier to understand the data, the classification also helps in reducing the risk of overfitting to the training data. The tree instead of understanding the underlying concept focuses on the intrinsic properties and which are specific to the training data.

After completing the classification, the report details : -

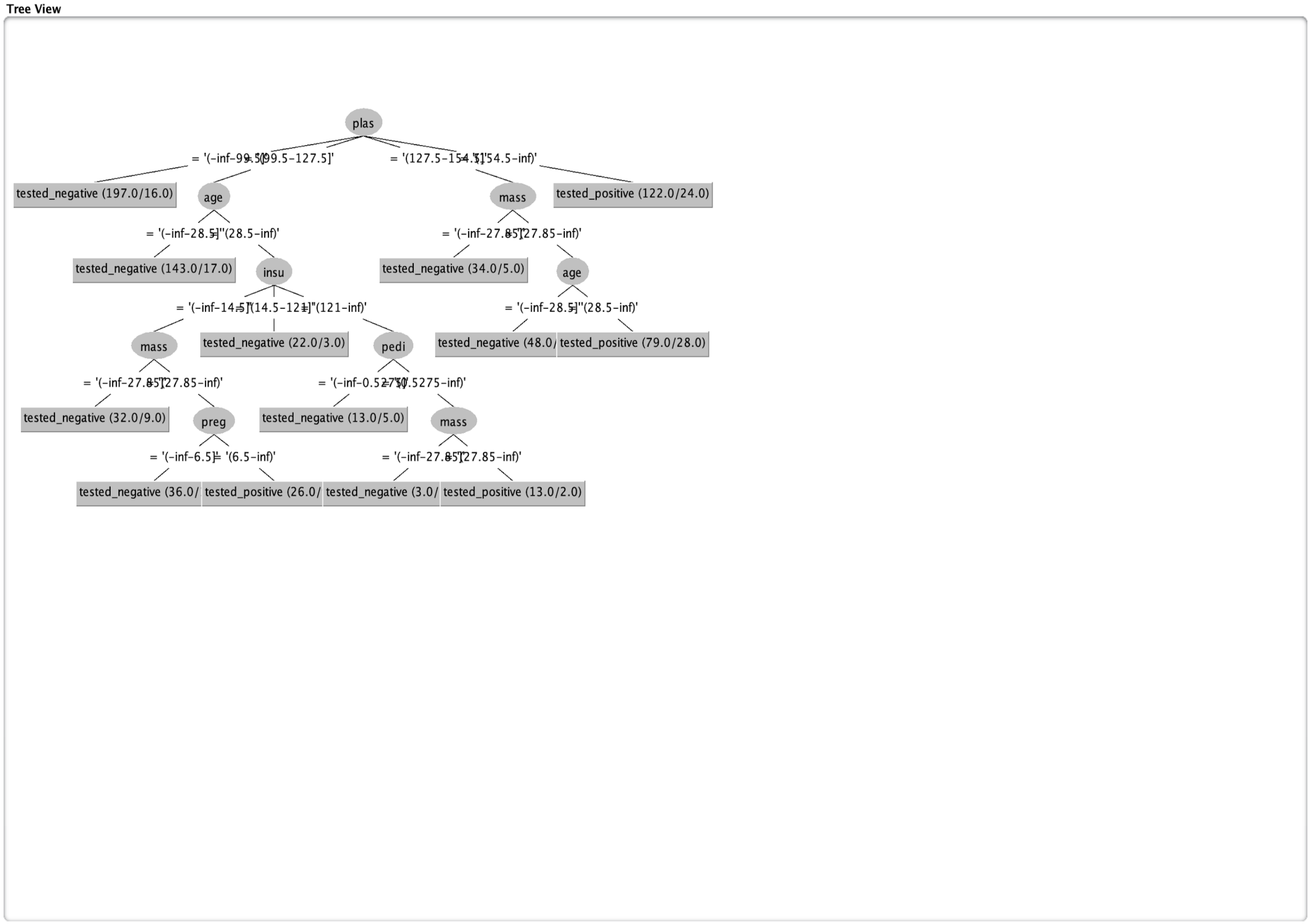
- No. of Instances : - 768
- No. of Attributes : - 9
- No. of Cross- Folds Validation conducted : - 10
- Size of the tree - 20 leaves
- No. Of Nodes :- 39 (Therefore 19 leaf nodes and 20 interior nodes)

The classification was conducted by two methods. The first one the option for the uprunning was set to false and the second time it was set to true which changes the accuracy rate.

- Accuracy with uprunning set to false : - 76.8229%
- Accuracy with uprunning set to true : - 78.2552%

It has also been observed that the size of the tree also differs, the bigger the tree the accuracy and the prediction result deteriorates.

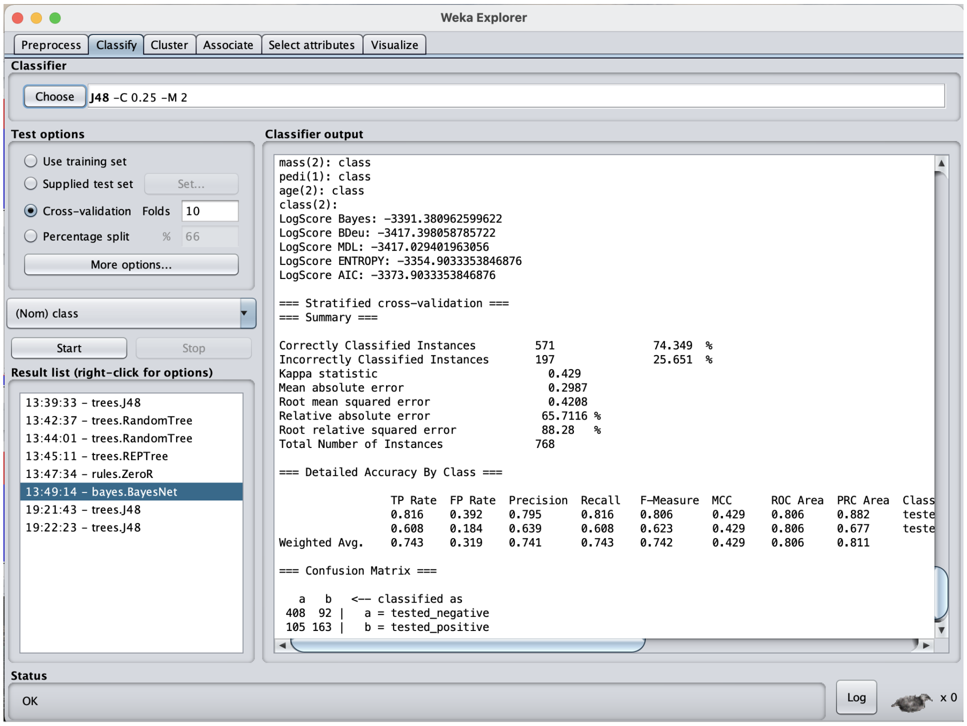
The tree with the highest accuracy and prediction shows that the plasma count was the attribute from where the data has been split and according to the test result the weightage of the attribute has been presented.



Bayes.Net: -

Bayes.Net is a classification of the Naïve Bayes theorem. It uses the Bayes theorem to classify objects. It presuppose strong, or naïve, independence, between attributes of data points.

The Bayes.Net follows the probabilistic distribution over all the classes and the evidence which comes out after the completion of cross folding can be split into parts that are independent.



J48 -C 0.25 -M 2

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

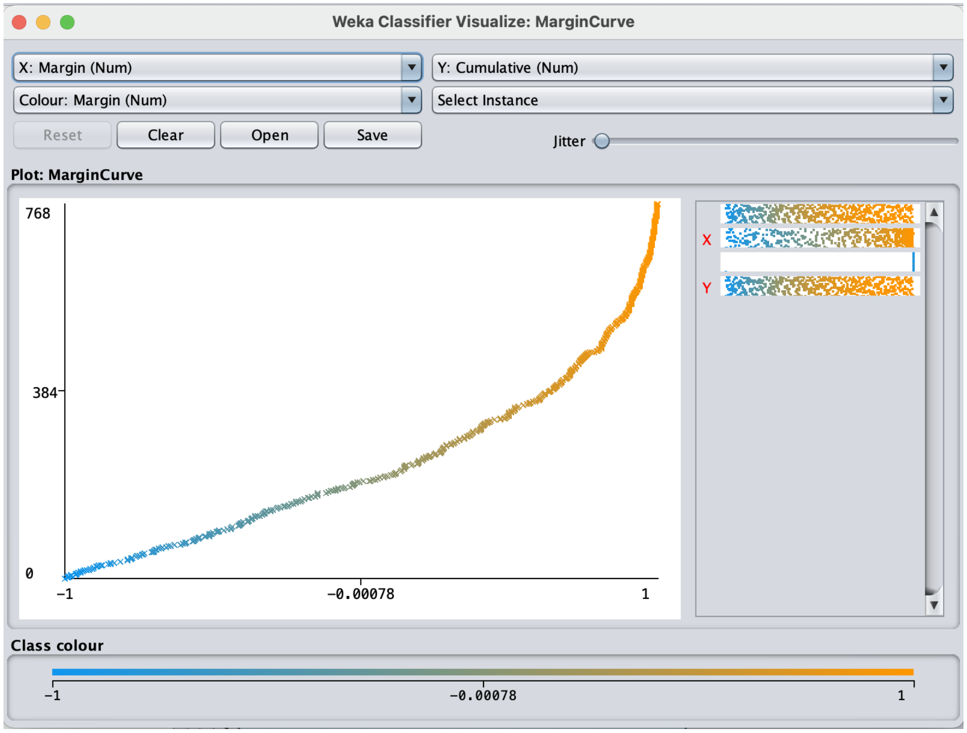
(Nom) class

```
13:39:33 - trees.J48
13:42:37 - trees.RandomTree
13:44:01 - trees.RandomTree
13:45:11 - trees.REPTree
13:47:34 - rules.ZeroR
13:49:14 - bayes.BayesNet
19:21:43 - trees.J48
19:22:23 - trees.J48
```

```
mass(2): class
pedi(1): class
age(2): class
class(2):
LogScore Bayes: -3391.380962599622
LogScore BDeu: -3417.398058785722
LogScore MDL: -3417.029401963056
LogScore ENTROPY: -3354.9033353846876
LogScore AIC: -3373.9033353846876

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         571               74.349  %
Incorrectly Classified Instances       197               25.651  %
Kappa statistic                          0.429
Mean absolute error                      0.2987
Root mean squared error                  0.4208
Relative absolute error                 65.7116 %
Root relative squared error             88.28   %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.816    0.392    0.795      0.816    0.806      0.429    0.806     0.882     teste
                 0.608    0.184    0.639      0.608    0.623      0.429    0.806     0.677     teste
Weighted Avg.    0.743    0.319    0.741      0.743    0.742      0.429    0.806     0.811

=== Confusion Matrix ===

   a   b   <-- classified as
 408  92 |   a = tested_negative
 105 163 |   b = tested_positive
```

OK

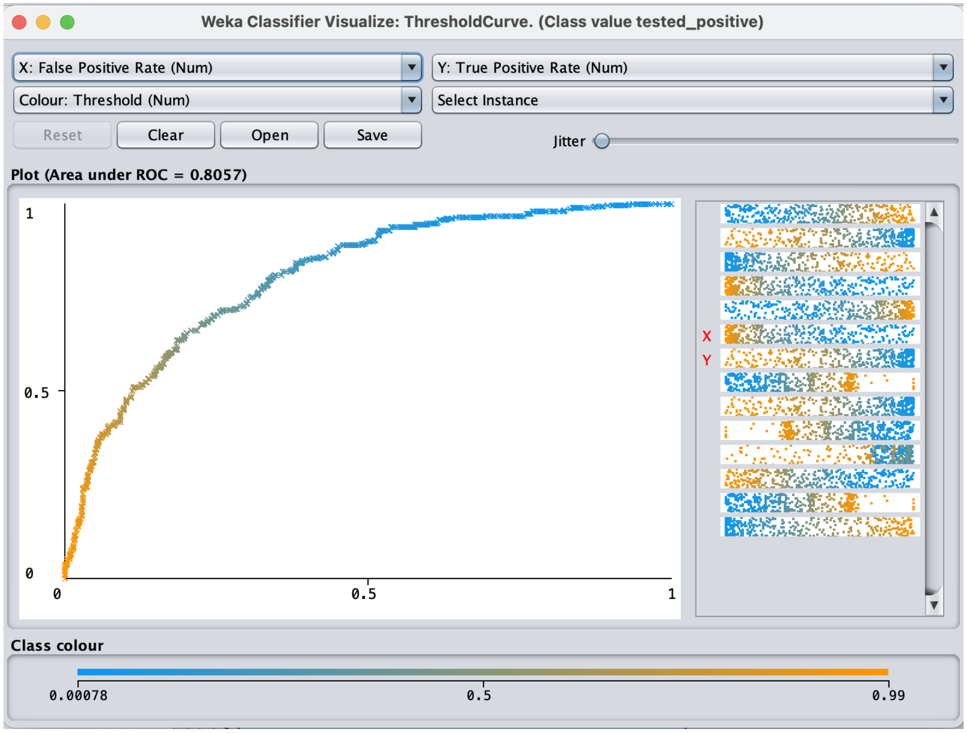
After conducting the classification it is observed that there are 408 true instances, 105 false instances and 163 true negatives for Bayes.Net. It also proves that the ROC curve is composed by plotting the true positive rate(TPR) against the false positive rate (FTR) generated after the classification is complete.



X: Margin (Num)
Y: Cumulative (Num)
Colour: Margin (Num)
Select Instance

From this above graph the visualization shows the Precision value is 0.795 which is quite an excellent reading. It also proves that with J48 the accuracy might differ but the classification which has been performed is more stable.

Bayes Theorem also lets the user calculate the threshold curve. It can be explained as the way of measuring classifier accuracy independent of the tradeoff that the user choose to perform the classification.



Weka Classifier Visualize: ThresholdCurve. (Class value tested_positive)

X: False Positive Rate (Num)

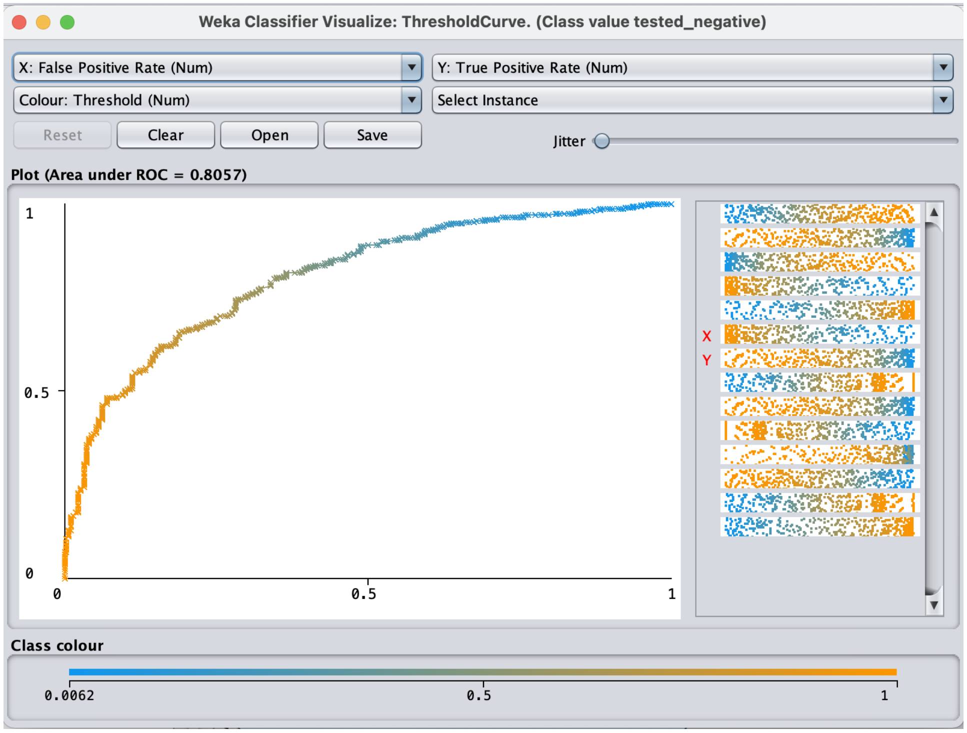
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.8057)

Class colour


Weka Classifier Visualize: ThresholdCurve. (Class value tested_negative)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.8057)

Class colour

Both the graphs predict a threshold value ROC of 0.8 which proves to be a better model. Both the positive and negative test result produced the same result which adds to the genuine factor of the model.

Conclusion : -

All the data in the datasets are trained and tested using 10 cross validation folds with Bayes.Net and J48 pruned tree and then the performance was evaluated, measured and compared with each other using weka. The best results were shown by J48 pruned tree where the accuracy was 76.8229% and 23.1771% correctly classified, but the No. Of true positive instances were 431 which is a big tree and the model cannot be considered as the best one as there was no stability. The Bayes.Net produced an accuracy of 74.349% and 25.651% incorrectly classified and there were 408 were TPR (true instances) and 105 were FTR (false instances), though the ROC came to a 0.8 which gives a more stable model to hold on.

Github Repository Link : -

https://github.com/Saurav-Aich/Phase-5-Modeling-data