Phase 6: Optimization

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

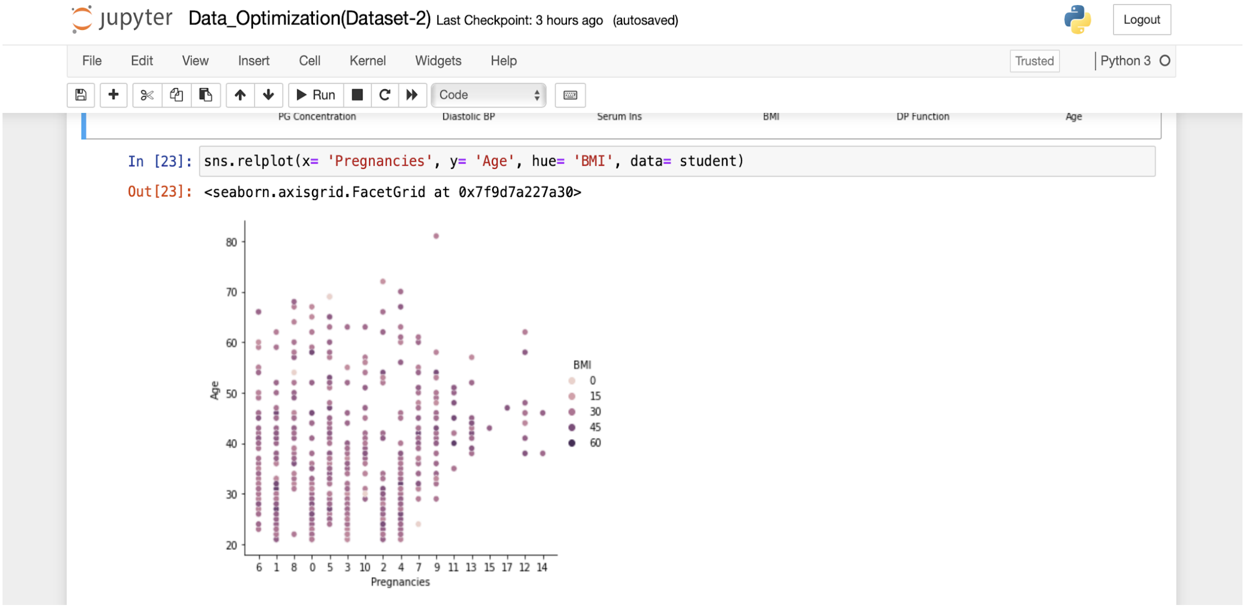
Selected Dataset and Research Question: -

Our dataset has been taken from WebMD which contains data from the year 1990 to 2012 where people across different age groups along with their food habits and vital information. These data have been collected to answer the question which is What makes the US population susceptible to diabetes?

Data Exploration : -

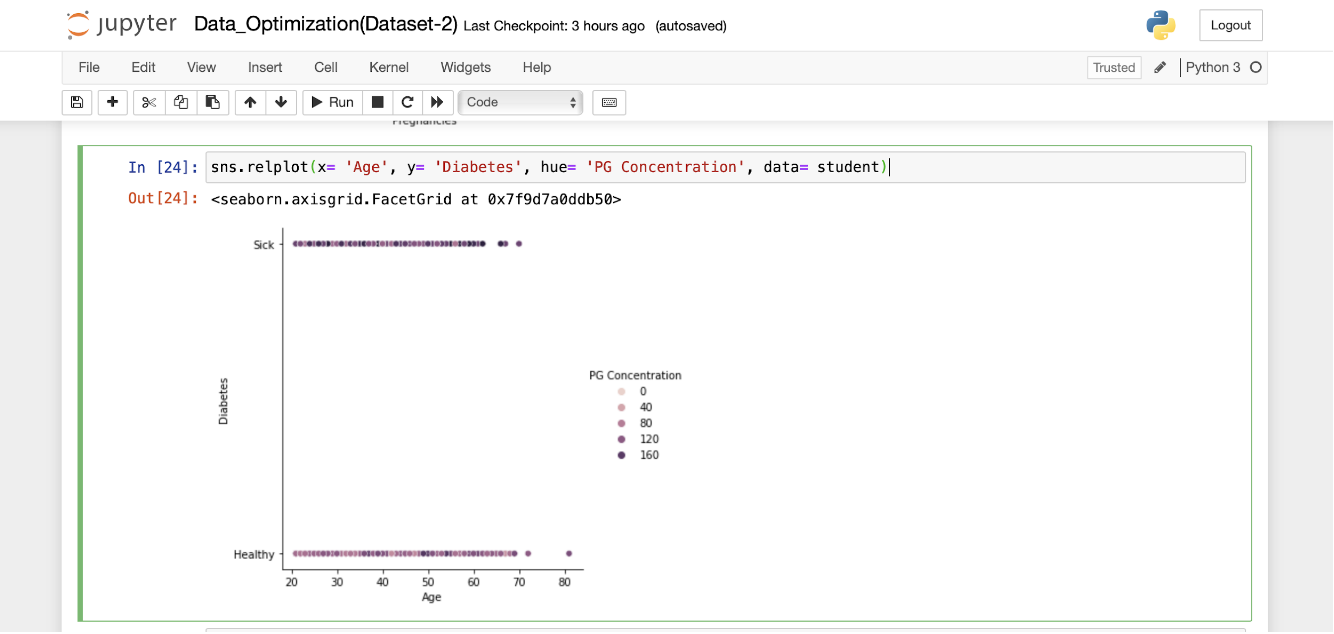
The Data Exploration with the first dataset had abnormalities which led to a lower accuracy rate. The number of True Positive Rate(TPR) and the number of FTR(False Positive Rate) the results never gave the desired output. This led to the decision to revisit the data exploration phase to figure out if the results gave a different outlook for the given dataset.

After mapping and plotting the various attributes the plotting of the relation between Pregnancies, age and BMI showed promising results it also showed as the age increases the BMI also increases. With pregnancies the same results can be seen.



```
In [23]: sns.relplot(x= 'Pregnancies', y= 'Age', hue= 'BMI', data= student)
Out[23]: <seaborn.axisgrid.FacetGrid at 0x7f9d7a227a30>
```

After this the next plotting that was considered was the relation between Age, Diabetes and PG Concentration, the relation between these three attributes. With increasing age, the risks of having diabetes with a higher concentration of PG has been observed. The same condition has been observed with women who are pregnant. As the age of women increases and if the PG concentration is on the high then, the risk of getting diabetes also gets increased. The below image of healthy and sick people according to the PG Concentration.



```
In [24]: sns.relplot(x= 'Age', y= 'Diabetes', hue= 'PG Concentration', data= student)
Out[24]: <seaborn.axisgrid.FacetGrid at 0x7f9d7a0ddb50>
```

Data Mining Technique Used :-

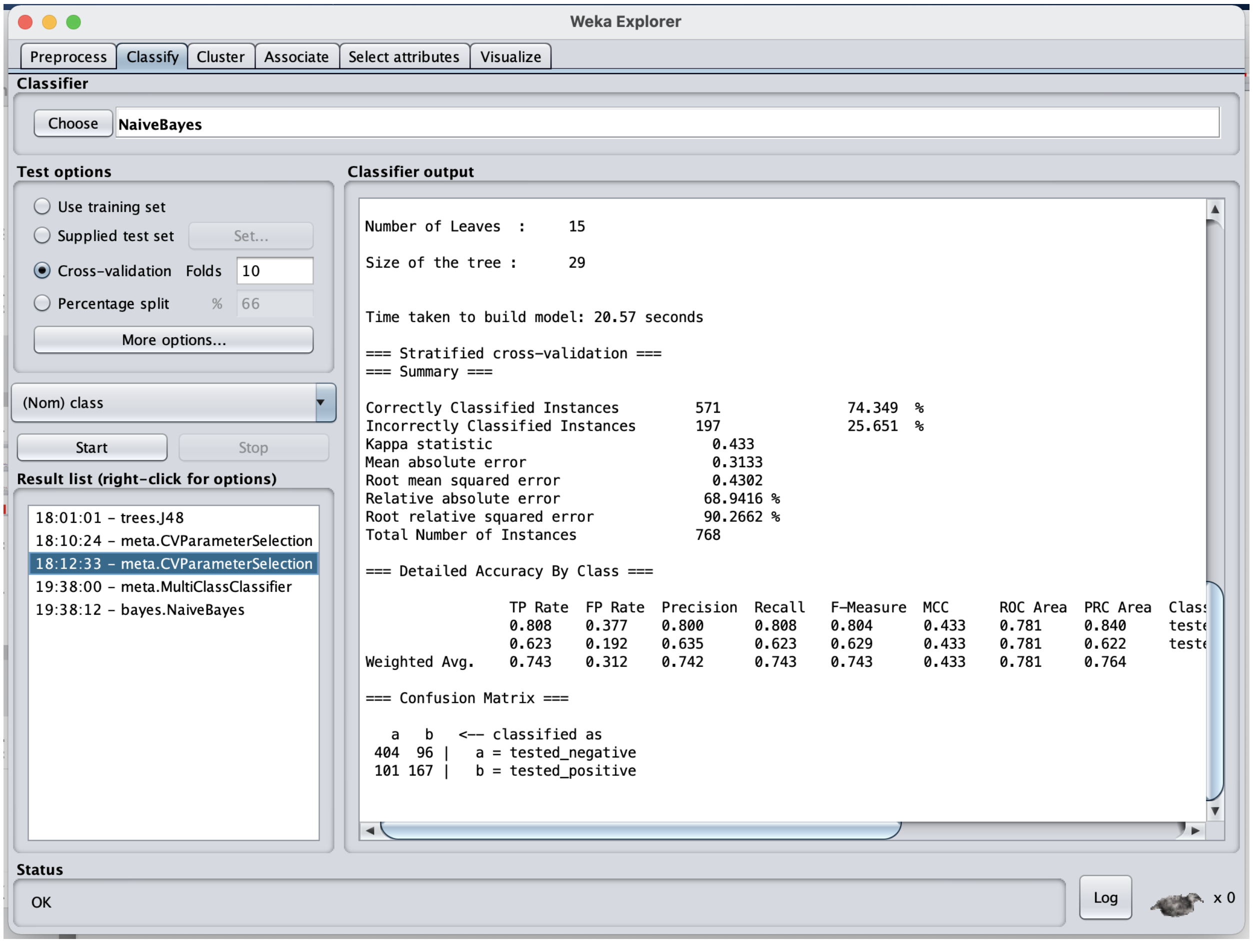
J48 Pruned Tree

Primarily J48 Pruned Tree was which had 768 instances with 9 attributes. After conducting 10 folds of cross validation. The accuracy of uprunning set to true was 73.8281%

When using the J48 Pruned Tree the C V parameters are the key parameters upon which the cross validation works. With the default values it gives the accuracy of 73.8 %. When meta learners "The C V parameter" can be adjusted by changing the confidence factor from 0.1to 1.0 in 10 folds. Surprisingly, the accuracy rate was improved to 74.349%. The number of tested_negative has increased to 404.

Optimizing Technique Used : -

C V Parameter Selection



Classifier: NaiveBayes

Cross-validation Folds: 10

Percentage split %: 66

(Nom) class

Result list:

18:01:01 – trees.J48
18:10:24 – meta.CVParameterSelection
18:12:33 – meta.CVParameterSelection
19:38:00 – meta.MultiClassClassifier
19:38:12 – bayes.NaiveBayes

```
Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 20.57 seconds

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

Correctly Classified Instances         571               74.349  %
Incorrectly Classified Instances       197               25.651  %
Kappa statistic                          0.433
Mean absolute error                      0.3133
Root mean squared error                  0.4302
Relative absolute error                 68.9416 %
Root relative squared error             90.2662 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.808    0.377    0.800      0.808    0.804      0.433    0.781     0.840     teste
                 0.623    0.192    0.635      0.623    0.629      0.433    0.781     0.622     teste
Weighted Avg.    0.743    0.312    0.742      0.743    0.743      0.433    0.781     0.764

=== Confusion Matrix ===

   a   b   <-- classified as
 404  96 |   a = tested_negative
 101 167 |   b = tested_positive
```

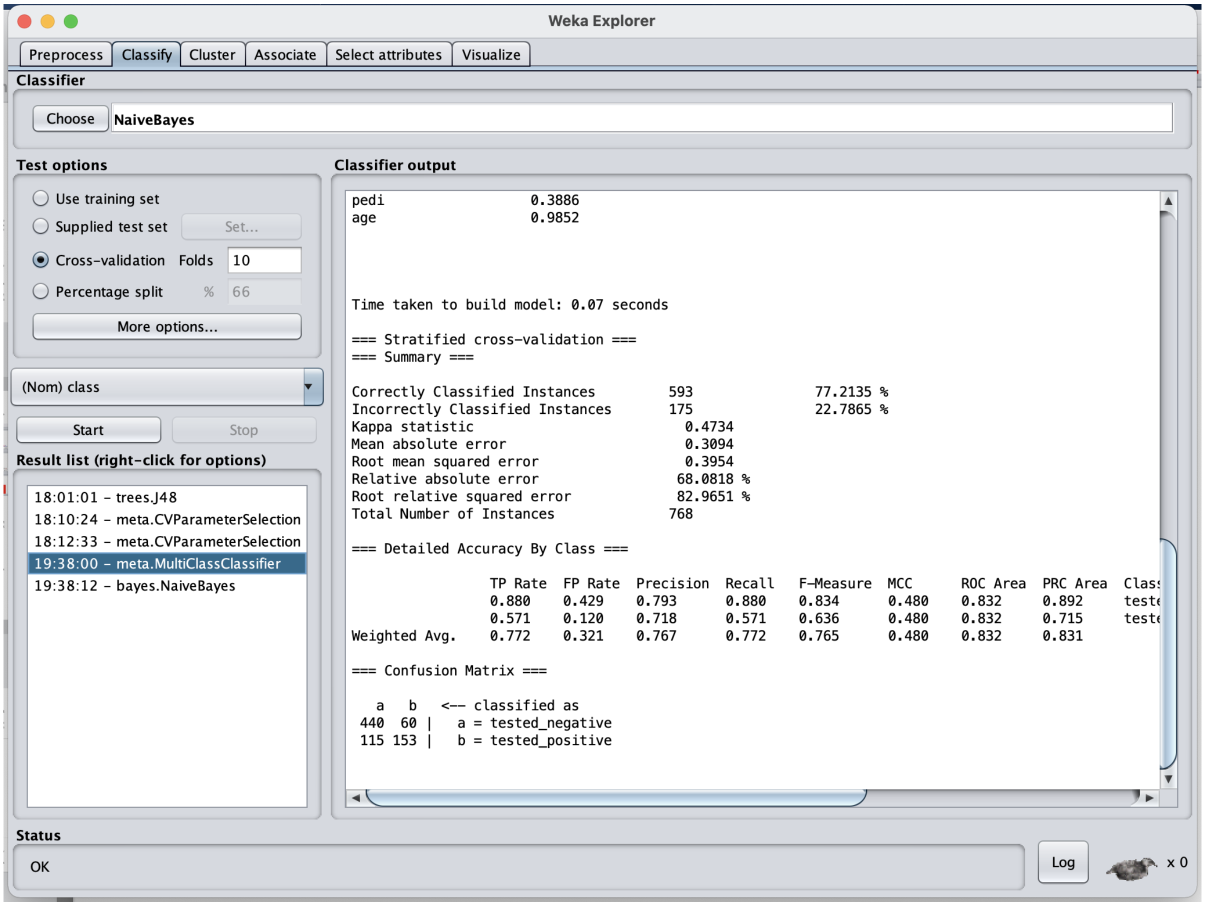
Status: OK

Naïve Bayes : -

Initially a probabilistic algorithm like algorithm gave us a very promising result with an accuracy rate of 76.3021% which is very promising than J48. There are several meta classifiers with Naïve Bayes like Class Classifier where the sum of weights across all instances in the data is the same. The results after using was improved to 77.2135%. The number of tested_negative rate was increased from 422 to 440.

Optimizing Technique Used : -

Meta Class Classifier.



Classifier: NaiveBayes

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

(Nom) class

Result list (right-click for options)

- 18:01:01 - trees.J48
- 18:10:24 - meta.CVParameterSelection
- 18:12:33 - meta.CVParameterSelection
- 19:38:00 - meta.MultiClassClassifier
- 19:38:12 - bayes.NaiveBayes

Classifier output

```
pedi                  0.3886
age                   0.9852



Time taken to build model: 0.07 seconds

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

Correctly Classified Instances         593               77.2135 %
Incorrectly Classified Instances       175               22.7865 %
Kappa statistic                          0.4734
Mean absolute error                      0.3094
Root mean squared error                  0.3954
Relative absolute error                 68.0818 %
Root relative squared error             82.9651 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.880    0.429    0.793      0.880    0.834      0.480    0.832     0.892     teste
                 0.571    0.120    0.718      0.571    0.636      0.480    0.832     0.715     teste
Weighted Avg.    0.772    0.321    0.767      0.772    0.765      0.480    0.832     0.831

=== Confusion Matrix ===

   a   b   <-- classified as
 440  60 |   a = tested_negative
 115 153 |   b = tested_positive
```

Status

OK

Conclusion : -

Optimization has made us understand that preliminary results cannot be considered as results. There are several meta classifiers which can be added to the primary classifiers where factors such as the confidence coefficient and class variables can be set which gives a better output and better accuracy. Data exploration should be performed on a regular basis to notice the other changes that are taking place in the dataset. Weka and other tools help us in visualizing the data. Better tweaking the set variables can give better accuracy.

Github Repository Link : -

https://github.com/Saurav-Aich/Phase-6-Optimization