Analytical study of Diabetic patient's data to answer the question what makes the US population susceptible to diabetes.

Shivanjali Khare,
Professor, Computer Science Department, University of New Haven, CT, USA
Email: skhare@newhaven.edu

Smit Kakadiya,
Master's student, Computer Science Department, University of New Haven, CT, USA
Email: smkakdiya201@gmail.com

Saurav Aich,
Master's student, Computer Science Department, University of New Haven, CT, USA
Email: paddyaich@gmail.com

Bhargav Patel,
Master's student, Computer Science Department, University of New Haven, CT, USA
Email: bonnypatel2286@gmail.com

Abstract:
Diabetes in 21st century has changed completely previously it used to affect elderly or obese people; now a days diabetes has become a common enemy regardless of age. Technology has also taken a leap forward to invent various methods of detecting diabetes or the chances of getting diabetes. The team has taken datasets with according activity labels to predict when the testing data will answer the question **"Am I diagnosed with diabetes"** 34.2 million or 10.5 percent of the US population is suffering from diabetes; and there is an estimation that 10.2 percent of the population has already been diagnosed with diabetes. This has made diabetes the seventh leading cause of death. The data has been collected from github where data has been divided into women with pregnancies issue it also includes data with glucose count skin thinness BMI (Body Mass Index) and age. This data will give us a clear idea of the growth and with age we can find out if the person (he/she) will be getting diabetes.

[1] Introduction:

The categorization of several sorts of datasets that can be used to assess if a person is diabetic is the main topic of this research. The cost of various types of datasets will be included in the answer to this problem. As a result, the purpose of this study is to use classifiers to correctly classify data so that a physician may pick the optimal datasets for diagnosing in a safe and expensive manner. SVM (Support vector Machine) is the prediction analysis that has been chosen for this project. SVM being a supervised learning which when associated with a learning algorithm which will fetch the classification and regression analysis. SVM Uses a subset of training points in the decision function (called support vectors), as a result it is memory efficient. The established dataset should help us find and predict whether the person will be diagnosed based on the activity labels in the dataset.

[2] Related Work:

The initial phase of the project was to go through similar research papers which would give an idea about how to go ahead with the project. Google scholar gave us numerous

literature by virtue of which we as a group was able to perceive the notion about how to go forward. Here, we are going to mention basic information about literature that we have reviewed and discussed.

[2.1] Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool, P.Yasodha, N.R. Ananthanarayanan Pachiyappa's college for women, Sri Chandrase kharendra Saraswathi Viswa Mahavidyalaya, March 2021, International Journal of Computer Applications Technology and Research Volume 3– Issue 9.

This research was carried out in India with the use of the Weka tool, which is our project's principal software. The purpose of this work is to appropriately classify datasets so that a doctor may select the finest datasets for disease diagnosis in a safe and cost effective manner. The symptoms will serve as the foundation for the doctor's theory as he examines the patient. The answer to such a project will aid doctors in making judgments and make the process more agile, as well as lower health-care expenses and patient wait times. For the data mining task, related attributes are picked. To train the dataset utilizing all of the qualifiers, 66% of the dataset was randomly selected. The training dataset was folded ten times for cross validation. The remaining data was used to create a test dataset, with 34% of the data chosen at random. The following characteristics were considered to create the best model that can accurately anticipate the research question's answer.

- Time: Referred to as the time required to complete training of the dataset.
- Relative absolute Error: the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.
- Mean Squared Error: This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value.
- Mean Absolute Error: The average of the difference between predicted and actual value in all test cases.
- Root Relative Squared Error: The total squared error made relative to what the error would have been if the prediction had been the average of the absolute value.
- Kappa Statistic: A measure of the degree of nonrandom agreement between observers.

The following is the methodology (classifiers that were utilized to train the training data).

- J48: A pruned or unpruned decision tree can be generated.
- Rep Tree: Using information gain/variance, it constructs a decision and regression tree and prunes it using reduced-error pruning (with back fitting). Values for numeric attributes are only sorted once.
- LAD Tree: Using the logistical technique, create a multiclass alternating decision tree. LAD Tree creates a LAD Tree with several classes. It can accept more than two different classes as inputs. Using the Logistics Strategy, it does additive logistic regression.

According to this study, the J48 classifier produced the greatest results, with an accuracy of 70.59 percent and a training time of 0.29 seconds.

[2.2] An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set, Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput, Department of Computer Science, Barkatullah University, December 2015, International Journal of Computer

Applications (0975 – 8887) Volume 131 – No. 2.

The purpose of this research study is comparable to the goal of the last research article. They employed a unique method in which they categorised the dataset and then compared several data mining strategies in Weka using the explorer, knowledge flow, and experimenter interfaces. This research employs three techniques: the first employs an explorer interface and relies on algorithms such as Naïve Bayes, SMO, J48, REP Tree, and RANDOM Tree, which are used in regions to represent, apply, and learn knowledge of statistics, yielding significant results. The Experimenter interface is used in the second technique. This research enables the creation of experiments for the application of algorithms like Nave Bayes, J48, REP Tree, and RANDOM Tree to datasets. These algorithms can be tested on an experimenter and the outcomes analyzed. It sets the test option to utilize 10 fold cross validation. Knowledge Flow is the third approach. We classified the accuracy of multiple algorithms Naïve Bayes, SMO, J48, REP Tree, and random Tree on different data sets in this experiment and compared with the results to see which method performed the best. The following is the methodology of classifiers that were used to train the data.

- Correctly Classified Accuracy: The percentage of correctly classified tests that are accurate.
- Erroneously Classified Accuracy: The percentage of tests that are classified incorrectly.
- Time: How long does it take to construct a model that can forecast disease?
- Mean Absolute Error: The number of defects used to evaluate algorithm classification accuracy.

The data was successfully trained using Naïve Bayes in this research study, yielding an accuracy of 76.3201 percent. Associated with blood cells, heart rate, personality traits, and medical conditions, the proposed future study can be applied to blood groups to establish the association between diabetic and detecting cancer patients.

[2.3] Accurate and rapid screening model for potential diabetes mellitus, Dongmei Pei (Department of Family Medicine, Shengjing Hospital, China Medical University), Yang Gong (University of Texas Health Science Center at Houston), Hong Kang (University of Texas Health Science Center at Houston), Chengpu Zhang (Department of Family Medicine, Shengjing Hospital, China Medical University), Qiyong Guo (Department of radiology, Shengjing Hospital, China Medical University), March 2019, MC Medical Informatics and Decision Making volume 19, Article number: 41.

Dongmei Pei, Yang Gong, Hong Kang, Chengpu Zhang, and Qiyong Guo conducted research on diabetes prediction and early diagnosis. J48, AdaboostM1, SMO, Bayes Net, and Naive Bayes are five classifiers used to identify people with diabetes based on nine non-invasive and easily accessible clinical parameters such as age, BMI, hypertension, history of cardiovascular disease, and eating habits. A total of 4205 data entries were used in the study. The decision tree classifier J48 has the best performance (accuracy = 0.9503, precision = 0.950, recall = 0.950, F-measure = 0.948, and AUC = 0.964), according to the results. Age is the most important factor, followed by family history of diabetes, work stress, BMI, salty food choice, physical activity, hypertension, gender, and a history of cardiovascular disease or stroke, according to the decision tree structure. The conclusion that can be derived from this study report is that decision tree analyses can be used to screen individuals for early diabetes risk without the requirement for invasive procedures.

[2.4] Diabetic Prediction with WEKA Tool, Dr. Pankaj Bhambri (Dept. of Information Technology, Guru Nanak Dev Engineering College, Ludhiana), Dr. Vijay Kumar Sinha (Dept. of Comp. Sci. & Engg., Chandigarh Engineering College, Mohali), Dr. Inderjit Singh (Dept. of Comp. Sci. & Engg., Guru Nanak Dev Engineering College, Ludhiana), September 2020, Journal of Critical Reviews, Volume 7, Issue 9.

Data mining is a well-established field of study that is continually expanding. This isn't your typical analysis. It aids in the prediction of the future by detecting and discovering many valuable patterns in big collections of data sets. It can also be characterized as a method for discovering previously undiscovered, valuable patterns and regularities in enormous amounts of industrial and business data. Classification methods such as decision trees are highly widespread and simple to utilize. WEKA supports standard tree algorithms such as J48, REP Tree, Random Tree, and LMT. The findings are tentative, but they are beneficial to the medical data mining community. It has been demonstrated that making a few key adjustments can increase the performance of several algorithms. As in this scenario, many algorithms, ensemble methods, re - sampling, and feature extraction all worked together to increase the greatest algorithms' actual quality.

[2.5] Diagnosis of diabetes type-II using hybrid machine learning based ensemble model, Abid Sarwar1 • Mehbob Ali2 • Jatinder Manhas1 • Vinod Sharma, Bharati Vidyapeeth's Institute of Computer Applications and Management, 12 December 2018.

This paper discusses type-II diabetes and uses a specialist methodology to diagnose it. The goal of this job is to examine several machine learning algorithms for classification models in the context of illness, i.e., to determine whether or not a subject is sick. The Ensemble technique improves performance by implementing the classifying abilities of multiple classifiers, and the risks of misclassifying a given instance are greatly reduced, resulting in improved overall classification results. Furthermore, this diagnostic tool is checked by validating denary cross attestation; additionally, the outcome is compared to the actual real-world interpretation of the cases.

[3] Data Exploration:

Data Exploration in simple words can be explained as the technique to understand various aspects of the data. Dataset just provides the data but what information we are trying to fetch out and how we can achieve it. Data Exploration also helps us to relate variables with each other and plot it in a map which gives us a clear view about the data. Every dataset has discrepancies like null values or redundancy which need to be cleaned. The cleaner the data the better view about the data can be gained. Data exploration is usually performed in 3 major steps:

- Understanding the data.
- Cleaning the data.
- Relationship between the variables.

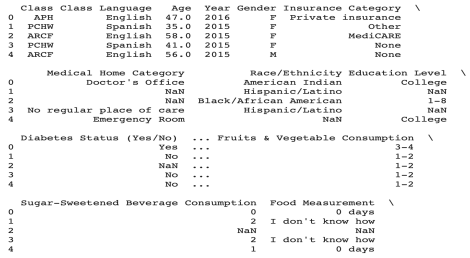
Jupyter Notebook a python framework has been used to carry out the data exploration. For data exploration certain libraries need to be called among them pandas, numpy and seaborn. These libraries help to break down the data and then visualize it in different forms. The first step was to upload the csv file to jupyter notebook.

```
In [2]: data = pd.read_csv("diabetes_dataset.csv")
```

The above command helped to upload the csv file to the python framework.

```
In [3]: print(data.head())
```

The above command prints the head of the table which takes the first 4 rows of the database and displays the information.



	Class	Class Language	Age	Year	Gender	Insurance Category	Medical Home Category	Race/Ethnicity	Education Level
0	APH	English	47.0	2016	F	Private insurance	Doctor's Office	American Indian	College
1	PCHW	Spanish	35.0	2015	F	Other	NaN	Hispanic/Latino	NaN
2	ARCF	English	58.0	2015	F	MediCARE	NaN	Black/African American	1-8
3	PCHW	Spanish	41.0	2015	F	None	No regular place of care	Hispanic/Latino	NaN
4	ARCF	English	56.0	2015	M	None	Emergency Room	NaN	College

	Diabetes Status (Yes/No)	...	Fruits & Vegetable Consumption	Sugar-Sweetened Beverage Consumption	Food Measurement
0	Yes	...	3-4	0	0 days
1	No	...	1-2	2	I don't know how
2	NaN	...	1-2	NaN	NaN
3	No	...	1-2	2	I don't know how
4	No	...	1-2	1	0 days

After it is seen that the data was able to be fetched out of the database. The next step was to know how many rows and columns are there in the dataset.

```
In [5]: print(data.shape)
```

```
(1688, 25)
```

The data.shape provided us the answer where it was seen that there are a total of 1688 rows and 25 columns. Every Dataset has unique values for example the Gender column contains 2 unique values which are male and female. Now every column in the dataset has their own unique values and therefore it's essential to figure it out and through python this can be achieved. The below command fetches all the unique values from all the columns.

```
In [8]: print(data.nunique())
```

```
Class                                        4
Class Language                               3
Age                                         75
Year                                         3
Gender                                       3
Insurance Category                           6
Medical Home Category                        9
Race/Ethnicity                               7
Education Level                              8
Diabetes Status (Yes/No)                     3
Heart Disease (Yes/No)                       2
High Blood Pressure (Yes/No)                 2
Tobacco Use (Yes/No)                         2
Previous Diabetes Education (Yes/No)         2
Diabetes Knowledge                           3
Fruits & Vegetable Consumption               5
Sugar-Sweetened Beverage Consumption         5
Food Measurement                             6
Carbohydrate Counting                        5
Exercise                                     7
Problem Area in Diabetes (PAID) Scale Score 80
ZIP code (address)                           0
ZIP code (city)                              0
ZIP code (state)                             0
ZIP code (zip)                              86
dtype: int64
```

And if we wanted to know the details of particular column that could also be performed via the command listed below:

```
In [9]: print(data['Gender'].unique())
```

```
['F' 'M' nan 'f']
```

```
In [10]: print(data['Race/Ethnicity'].unique())
```

```
['American Indian' 'Hispanic/Latino' 'Black/African American' nan 'Asian'
 'White' 'Other' 'Unknown']
```

These above commands helped us in understanding the data, how much data has been stored, what is the count, the unique values these helped us to gain more knowledge about the data. It gave us the picture and insightful information about the dataset. The second stage of data exploration is cleaning of the data. This step includes identifying null values across the dataset and if any column is not required can be dropped via the cleaning of data. When data is clean the analysis of the data becomes easier as without the discrepancies it becomes easier to visualize and create relationships which is the third phase of data exploration. First step was to calculate the null values.

```
In [11]: print(data.isnull().sum())
```

This above command gave a total number of null values present across the dataset. It calculates all the null values in every column and gives the result as to how many are there across the dataset.

```
Class                                        0
Class Language                               0
Age                                         32
Year                                         0
Gender                                      37
Insurance Category                         113
Medical Home Category                       91
Race/Ethnicity                              37
Education Level                            289
Diabetes Status (Yes/No)                    30
Heart Disease (Yes/No)                     100
High Blood Pressure (Yes/No)                93
Tobacco Use (Yes/No)                       124
Previous Diabetes Education (Yes/No)       125
Diabetes Knowledge                         154
Fruits & Vegetable Consumption              52
Sugar-Sweetened Beverage Consumption        53
Food Measurement                            63
Carbohydrate Counting                       66
Exercise                                    78
Problem Area in Diabetes (PAID) Scale Score 1056
ZIP code (address)                        1688
ZIP code (city)                           1688
ZIP code (state)                          1688
ZIP code (zip)                             175
dtype: int64
```

After studying the result it was observed that there are some columns which are not required and can be dropped from the dataset as it would not provide any insightful information on that topic.

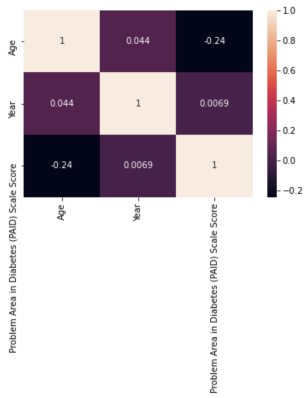
```
In [12]: student = data.drop(['ZIP code (state)'], axis=1)
         student = student.drop(['ZIP code (zip)'], axis=1)
         student = student.drop(['ZIP code (city)'], axis=1)
         student = student.drop(['ZIP code (address)'], axis=1)
```

After the completion of the cleaning the next step remains is analyzing and visualizing the variables which can be achieved building a relationship among them.

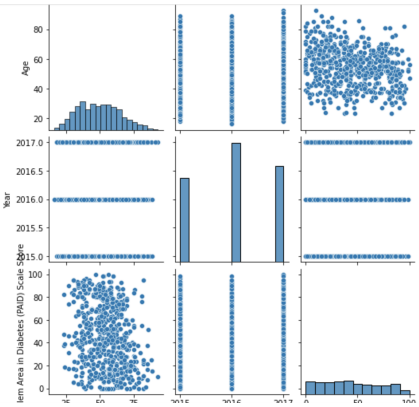
```
In [15]: sns.heatmap(corelation, xticklabels = corelation.columns, yticklabels = corel
```

This command helps to visualize the correlation between the variables.

```
Out[15]: <AxesSubplot:>
```

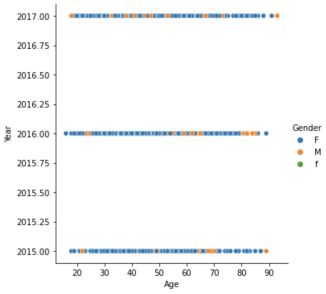


In the above image a heatmap is generated which is a correlation matrix which gives a wider perspective of the dataset for advance analysis. In this image it is observed that year, age, and problems in diabetes scale score are the variables which can be further considered for an advanced analysis. Heatmap is usually used for integer values as it does not take any categorical values into account which are the string values and for that there is another command which can get a better view of the categorical data. The pairplot on the other hand takes 2 variables into account and the variables can be continuous categorical or boolean as well and pairplot is a group of plots for the variables in the dataset.



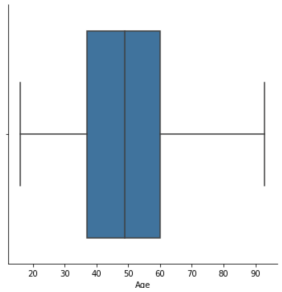
The next step is to plot data on the relationship between two numeric variables when compared with one categorical data variable. The scatter plot helps in achieving this goal.

```
In [18]: sns.relplot(x= 'Age', y= 'Year', hue= 'Gender', data= student)
Out[18]: <seaborn.axisgrid.FacetGrid at 0x7fe69b293f10>
```



The last plotting which was used to visualize the data was the categorical plot where the distribution of the variable across the dataset can be visualized.

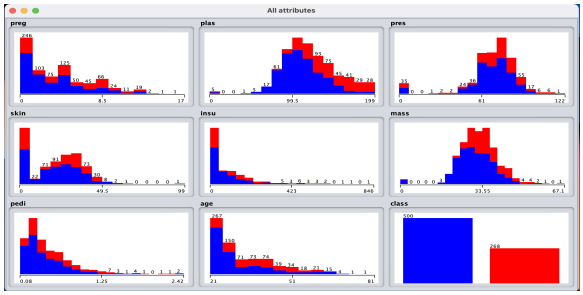
```
In [29]: sns.catplot(x='Age', kind= 'box', data= student)
Out[29]: <seaborn.axisgrid.FacetGrid at 0x7fe6b8f94c40>
```



[4] Data Modeling:

We used two techniques to do data modeling:

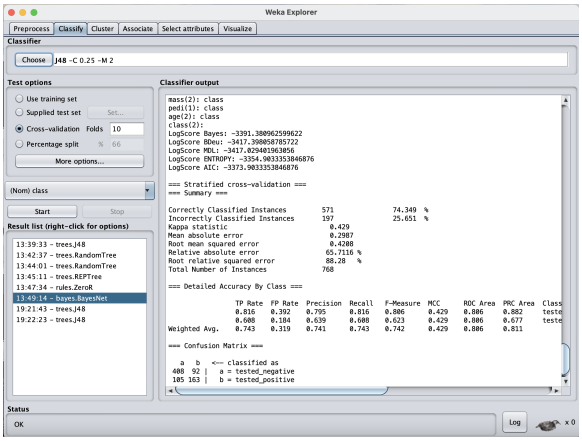
the J48 Pruned Tree and Bayes.Net. Weka Explorer, an open source machine learning tool, was used to classify data in order to assess the dataset's accuracy and come to a conclusion. Weka's main graphical user interface (GUI) displays a histogram of attribute distributions for a single attribute at a time. The histogram shows all of the ranges as well as the number of samples that fall into each range. Preg samples appear to have 17 distinct values and 2 unique values with a standard deviation of 3.845, while plasma appears to have 136 distinct values and 19 unique values, according to the dataset. These are the numerous patterns that data mining employs in order to accomplish categorization and other tasks.



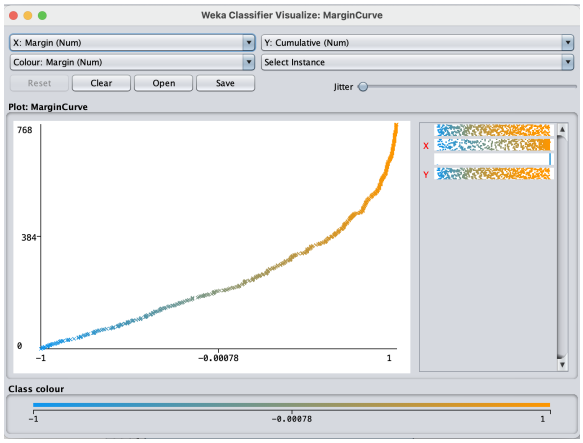
J48 Pruned Tree classification makes the data easier to interpret, and it also helps to reduce the danger of overfitting to the training data. Instead of focusing on the underlying notion, the tree concentrates on the intrinsic qualities that are unique to the training data. After completing the classification, 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%

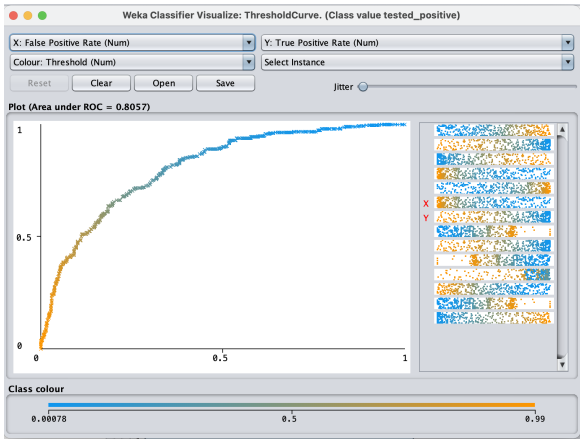
It has also been observed that the size of the tree also differs, the bigger the tree the accuracy and the prediction result deteriorates. The tree with the highest accuracy and prediction shows that the plasma count was the attribute from where the data has been split and according to the test result the weightage of the attribute has been presented. The Naive Bayes theorem is classified by Bayes.Net. The Bayes theorem is used to classify items. It requires high, or naive, independence between data point properties. The Bayes.Net algorithm follows a probability distribution across all classes, and the evidence that emerges after cross folding can be divided into independent portions.



Following the classification, it was discovered that Bayes.Net has 408 true instances, 105 false instances, and 163 true negatives. It also shows that the ROC curve is made up of charting the true positive rate (TPR) versus the false positive rate (FTR) created once the classification is finished.



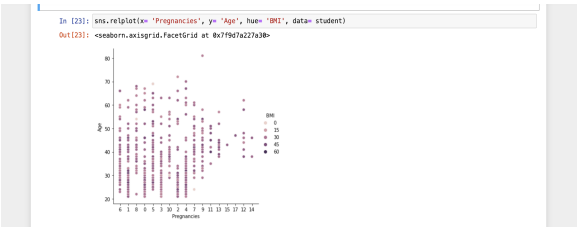
The Precision number is 0.795, which is a very good reading, according to the image above. It also demonstrates that while the accuracy of J48 may vary, the categorization performed is more constant. The user can alternatively calculate the threshold curve using Bayes' Theorem. It can be described as a method of determining classifier accuracy that is independent of the tradeoff that the user makes when performing the classification.



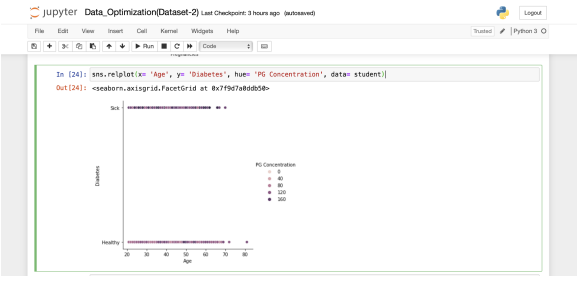
Both graphs indicate a ROC threshold value of 0.8, which is a more accurate model. The model's authentic factor is enhanced by the fact that both positive and negative test results gave the same outcome.

[5] Data Optimization:

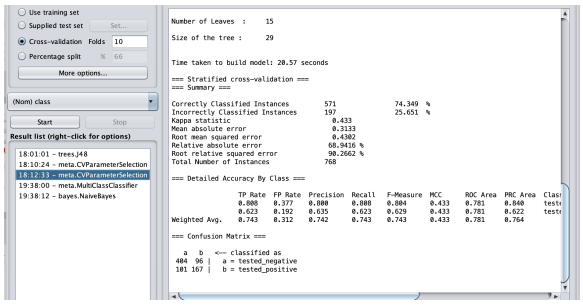
The initial dataset's Data Exploration had anomalies, resulting in a reduced accuracy rate. The number of True Positive Rates (TPR) and False Positive Rates (FTR) in the results never yielded the desired outcome. As a result, the decision was made to revisit the data exploration phase to see if the results provided an alternative perspective on the dataset. Following the mapping and plotting of the various features, the plotting of the relationship between pregnancies, age, and BMI yielded positive results, indicating that as age grows, so does BMI. The same consequences can be seen with pregnancies.



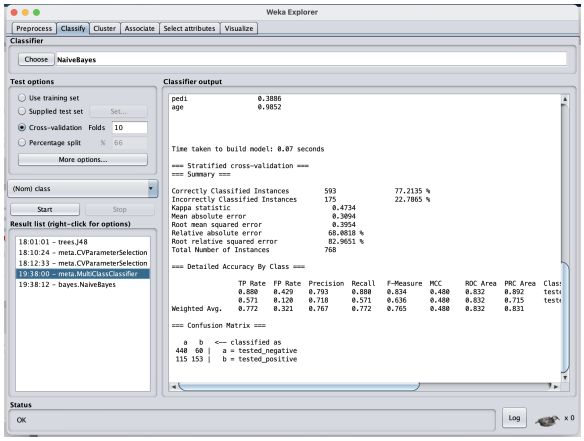
Following that, the relationship between Age, Diabetes, and PG Concentration, or the relationship between these three variables, was plotted. With increasing age, there is a larger risk of developing diabetes due to a higher PG concentration. Women who are pregnant have experienced the same problem. When a woman's age rises and her PG concentration rises, her risk of developing diabetes rises as well. The graphic below depicts healthy and unwell people based on their PG Concentration.



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.



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 it were improved to 77.2135%. The number of tested_negative rates was increased from 422 to 440.



[6] Conclusion:

The best results were shown by J48 pruned tree where the accuracy was 76.8229 percent and 23.1771 percent 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 percent and 25.651 percent 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. The above results were performed without optimization. Later when the optimization was performed there was a change in our findings and the result did change. The accuracy of Bayes.net was 77.2135 percent, which was higher than the J48 trimmed tree's 76.8229 percent. The tree was also unstable, so we had to optimize it. After analyzing the data, we can deduce that PG concentration is a crucial determinant in the development of diabetes, based on the work of both datasets. High-glucose diets have also been linked to high blood sugar levels.

[7] Future Work:

Datasets offered by doctors and health professionals can assist us in calculating alcohol consumption and calorie intake, allowing the general public to know how much alcohol and calories they should eat in a day or month to stay safe. Providing this information to the fast food sector may aid in the development or production of new products or items that promote a healthy lifestyle.

[8] Appendix for link to the GitHub repository:

Topic Name	Repository Link

Data Exploration	https://github.com/Saurav-Aich/Phase-4-Data_Exploration
Data Modeling	https://github.com/Saurav-Aich/Phase-5-Modeling-data
Data Optimization	https://github.com/Saurav-Aich/Phase-6-Optimization

[9] Reference:

1)https://en.wikipedia.org/wiki/Weka_(machine_learning)

2)https://www.kaggle.com/estefanytorres/data-exploration-in-numpy

3)https://github.com/matplotlib/matplotlib

4)https://towardsdatascience.com/speed-up-jupyter-notebooks-20716cbe2025

5) Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

6) Liu B., Hsu W., and Chen S., (1997) "Using general impressions to analyze discovered classification rules," Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

7)http://grb.mnsu.edu/grbts/doc/manual/ J48_Decision_T rees.html, accessed

8) T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning (1997) the McGraw- Hill Companies, Inc., pp. 52-78.

9) Stutz J., P. Cheeseman. (1996) Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

10) Srikant,R.,Vu,Q.andAgrawal,R.,(1997), "Mining association rules with item constraints," Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, USA, pp 67-73.

11) Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19(1), 1-9.

12) Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 13.

13) Sujni, P., & Latha, B. C. (2017). Prediction of diabetes using a classification model. Al Dar Research Journal For Sustainability. 2.

14) Nikhar, S., & Karandikar, A. M. (2016). Prediction of heart disease using machine learning algorithms. International Journal of Advanced Engineering, Management and Science, 2(6), 239484.

15) Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.