# Exploiting Machine Learning in diagnosis of Diabetes in Healthcare

Luksh

lukshkumar97@gmail.com

A Thesis in the Field of Machine Learning

for the Degree of Master

# Abstract

Diabetes is a disease in human body in which the blood sugar (glucose) exceeds its normal range. The body functions by generating glucose from food that goes in to our blood, then a particular hormone named as insulin takes the glucose from blood and supply it to the cells. Diabetes creates deficiency of insulin whose intensity is determined by classifying diabetes into two major categories known as Type 1 and Type 2 diabetes which is also recognized as *Mellitus*. In Type 1 diabetes the body does not make insulin because the immune system destroys the cells in the pancreas that make insulin. It is mostly diagnosed in younger aged people. In Type 2 diabetes the body does not make insulin or does not use it well, it is mostly diagnosed in elders and over-aged. The literature of diabetes empathizes the fact that there is no proper cure of diabetes therefore it is an open research problem which needs different methodologies and approaches for the precautions of diabetes. This creates the need of an intelligent machine learning algorithm that can robustly predict the potential candidate having diabetes. Therefore, this paper focus on conducting the research using different data science techniques and methodologies to analyze the data, extract insights out of it, build machine learning algorithms for it, and finally we evaluate the accuracy of the model. This would help us to predict whether the patient has diabetes or not given certain number of input features such as patient blood group, sugar level and so on. This paper focus on exploiting the use of machine learning in this domain and utilizing the best possible benefit from it into the field of healthcare and medicine. After analyzing the models, we have been able to classify whether the patient has diabetes or not with the accuracy of 79 per cent.

# Table of Content

# Chapter 1 - Introduction

Diabetes is a disease in human body that is cause by the malfunctioning of pancreas. The body functions by taking the sugar from food into the blood and a hormone called insulin is responsible to take that sugar from blood to cells in order to operate normally. Incase of diabetes, it is classified into various categories such as Type 1, Type 2, Gestational diabetes but the most frequent and important are Type 1 and Type 2 diabetes. In Type 1 diabetes, this insulin does not work [1] leading the malfunctioning in pancreas. It is mainly occurred in children or younger-aged group people. The Type 2 diabetes is the one in which the body does not use the insulin well [2] and it is mainly occurred in over-aged people. Gestational is caused in women while pregnancy which also tends to Type 2 diabetes if proper measure are not taken.

There have been many researches in the precaution and cure of diabetes but it is still an open problem whose discrete and obvious solution is still not available. Therefore, it rises a need for the bioinformatically technology that assist doctors and researches in this regard. After the advent of recent advancement in artificial intelligence and bioinformatics, the world has changed gears from a manual hand engineered research exploitation to an automated and statistical approach exploiting the technology. This paper focus to build robust model which can predict that whether the patient has a Type 2 diabetes or not. The methodology of this research will be discussed in detail in the upcoming chapters. The research is conducted on an authentic and authorized dataset of patient that have the Type 2 diabetes as well as the patients that do not have Type 2 diabetes in order to keep the data unbiased it was important to obtain the dataset of each class as equal as possible. This is not always possible to have equal observations in each observational experiment because the number of odds observations are always fewer than the rest. For example, the number of diabetes patients in a hospital is likely to be very less than all the remaining patients in the same hospital with some different or no disease. This is technically known as *class imbalance problem.* The paper caters this problem using the advance evaluation techniques known as ROC Curves which basically

generates the area under the curve and precision and recall which is another way to evaluate the class imbalance problem. This would be covered in detail in the upcoming chapter of research finding.

Healthcare has always been an important domain in the field of research because there are still many research gaps which needs to be resolved by inventing new evaluation methods and medical solutions for precaution and cure of various outlying diseases. This research was stagnant in the late nighty but after the boost from technology in the current century, the advent of internet and the availability of high computing has given new birth and motivation to solve such problems. Therefore, we specifically target the exploitation of diabetes using machine learning in order to explore what insights and performance it can give with regards to limited dataset. These findings would help us to better analyze how much data would generalize the results and whether technology can drastically help to resolve such issues or not.

Machine learning is a science of making the computer capable of learning without being explicitly programmed which means that now there is no need of a programmer sitting over the machine and updating the records to analyze the outcome. In computer programming, we put if and else statements to see whether which condition is true and which is false, but the problem is we can not put thousands and thousands of conditions without actually knowing what will these conditions be to program the outcome. Therefore, machine learning is a science that does this job for us by learning from the past experiences, updating its parameters (learned knowledge), and then improve its outcome. This cycle goes on and on until the system is in use. Therefore, we can prove that the more the data, the better it learns and improve itself. Moreover, we use the art of statistics in terms of transforming tabular data into visualizations so to interpret the aggregation like minimum blood level for children, average blood pressure for adults, outlying atmospheric pressure value and so on. This is done using descriptive and inferential statistics. We have run and tested our models using Python as a programming language which supports a lot of robust and well written packages enabling the engineers to exploit the most efficient algorithms at their ease. The visualizations are plotted using Matplotlib and Plotly opensource packages. These analysis would be discussed in the later chapter of discussion.

# Chapter 2 - Data and Methods Discussion

This chapter illustrate all the technical methodologies and procedures carried out for this research. It is fully occupied with data science primitives but we have tried to map each and every proof of concept with healthcare industry. Each section in this chapter explore different statistical findings for diabetes that is explained in non-technical terms in order to make it understandable for healthcare industry.

## 2.1 Dataset

This section explains the nature of dataset that is used to carry out the analysis and predictions. The snapshot of the few observations from the dataset is shown in the below attached Figure-1.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

Figure-1

It can be observed that there are eight features (inputs) in the dataset in which each one of the them has it own significance. The *Outcome* is the target variable whose value can be either 1 indicating that the patient has a diabetes and 0 indicating that the patient does not have diabetes, therefore we can say that the problem is *binary classification* since there are two possible outcomes for any given observation.

The datatypes of each feature is shown in the below attached Figure-2 which explains that what kind of value does each feature can have. It can be noticed that only BMI and DiabetesPedigreeFunction can have a value with decimal places allowed. There is no such feature expecting a string (textual) value such as low or high because all of these classes are better computed using the integer notations.

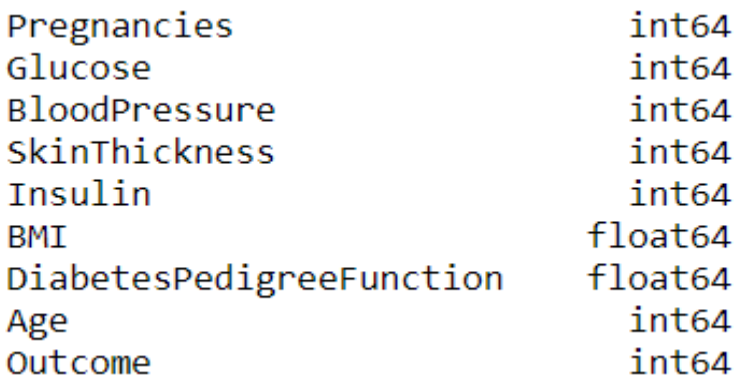


Figure-2

It is always very important to verify that the data contains no missing, incorrect, duplicate or inconsistent value. This lies under the process of *data cleaning* in data science whose objective is to make the data structured containing no such error at all. This dataset does not contain any missing values which can be seen in the below attached Figure-3.

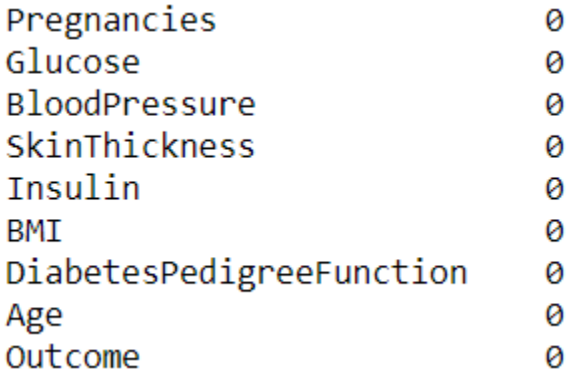


Figure-3

## 2.2 Descriptive Statistics and Analysis

### *Pregnancies Relation with Diabetes*

It is very necessary to explore each variable with context of its real nature. For example, the tendency of getting diabetes of Type 2 increases with age therefore it becomes important to analyze the Age variable in the dataset. Similarly, we are more interested in the exploration of those variables whose straight forward relationship does not exist so that our research can bring more novel results that help researchers to further work on it. Therefore, we have analyzed the *Pregnancies* variable here to see that how many unique values are there and what are they. By looking at the Figure-4, we can notice that the mostly the number of pregnancies is between 0-2, and then this frequency decreases as the number of pregnancies increases.

| | |
|---|---|
| 1 | 135 |
| 0 | 111 |
| 2 | 103 |
| 3 | 75 |
| 4 | 68 |
| 5 | 57 |
| 6 | 50 |
| 7 | 45 |
| 8 | 38 |
| 9 | 28 |
| 10 | 24 |
| 11 | 11 |
| 13 | 10 |
| 12 | 9 |
| 14 | 2 |
| 15 | 1 |
| 17 | 1 |

Figure-4

### *Basic Statistics*

The below attached Figure-5 demonstrates the aggregations of each variable. It can be noticed that these descriptive values are very hard to interpret that is why we will explore and interpret these values visually in the next section in detail.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 | 0.348958 |
| **std** | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| **25%** | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| **50%** | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| **75%** | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| **max** | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

Figure-5

## 2.3 Univariate Analysis and Visualizations

It is always very hard to interpret the descriptive statistics looking at the charts and numeric values. Therefore, we will use different visualizations to analyze and interpret the relationship between variables.

***Glucose Distribution***

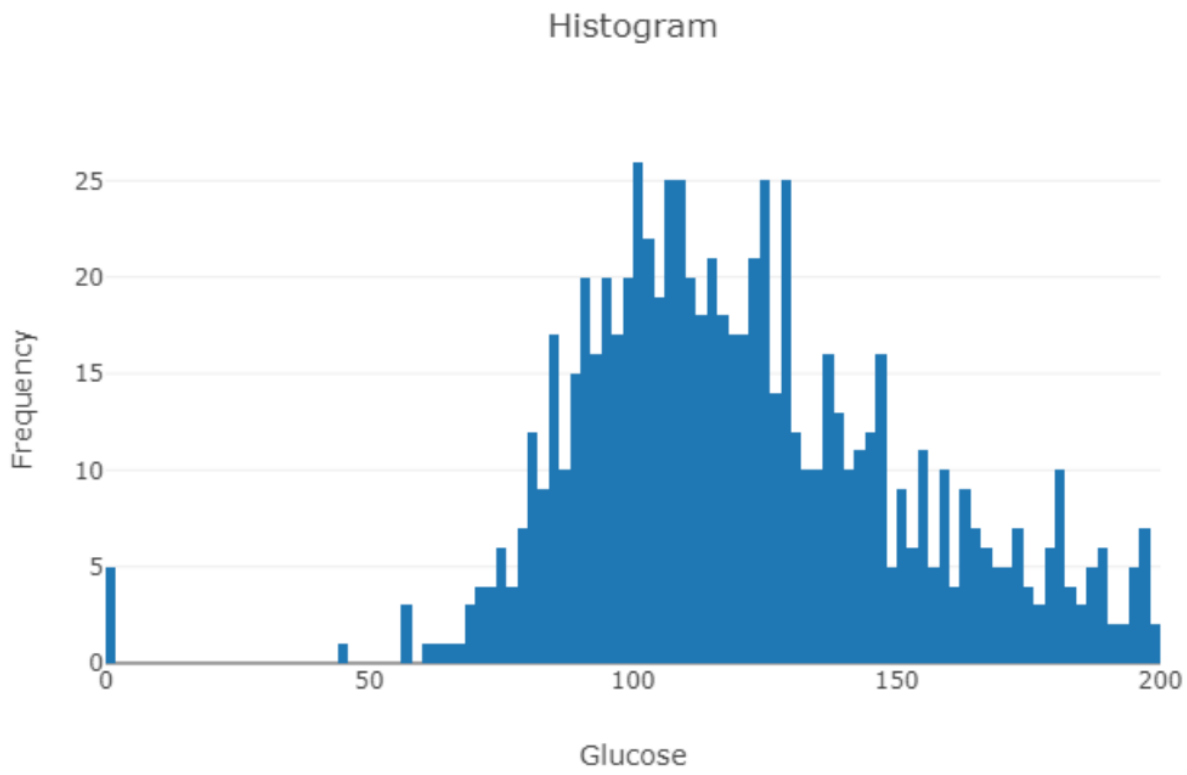


Figure-6

It can be analyzed from the above attached Figure-6 that the average glucose level in human blood is about 100 – 110 which can be seen on the x-axis whereas its frequency (how many people out of total sample) can be analyzed at y-axis. The graph is leaning towards the right which means that people whose blood sugar level has been increased are also captured in this image and those are the patients who have diabetes because as we know that diabetes occurs when the sugar level in the blood increase since the body is not using insulin to take that sugar and supply to the

cells. Therefore, this visual gives the interesting insight about the distribution of diabetes patients.

***Blood Pressure Distribution***

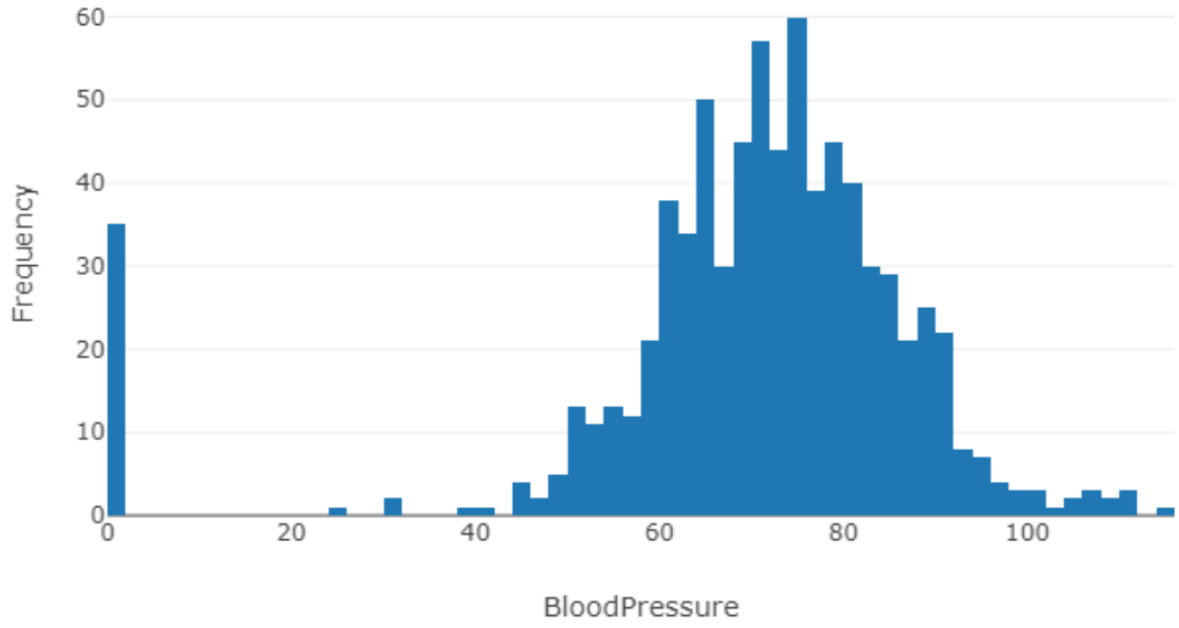


Figure-7

We can interpret the blood pressure in the same way we analyze the blood sugar (glucose) in the above visualization. The difference here is that it is more normal distribution which means that the number of patients have the same values throughout and the average blood pressure is about 70 – 75 which can be seen on the x-axis.

***Blood Pressure Boxplot***

This is another depiction of the average values, quartiles and mainly to analyze the outliers in the distribution of the blood pressure. By outlier, we mean that many unusual values in the dataset that can be either because of an error or it can be real value is tends to occur less frequently. This can be seen in the below attached Figure-8 of box plot. As it can be observed that there are few outliers whose value of blood pressure is between 30-35 and above 120.

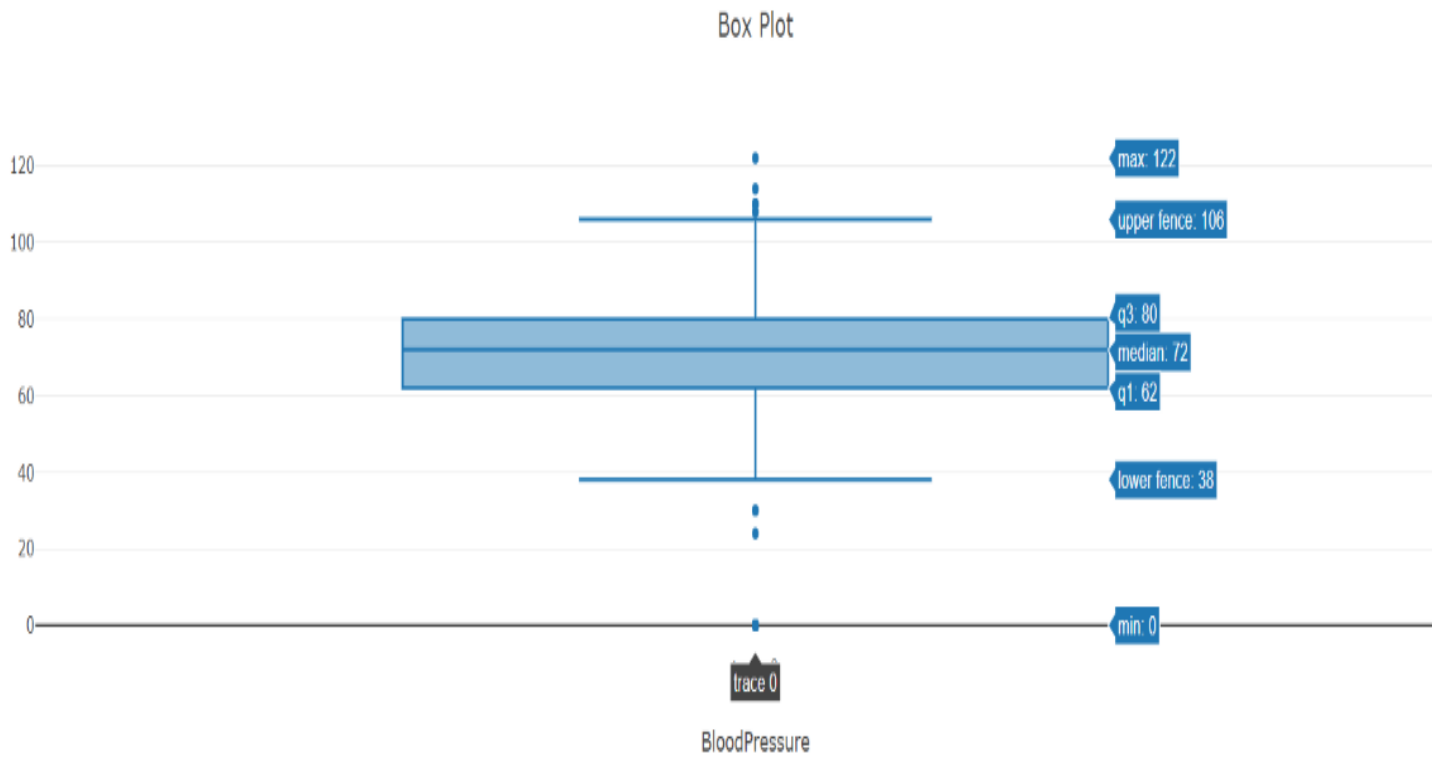


Figure-8

***Pregnancies Bar chart***

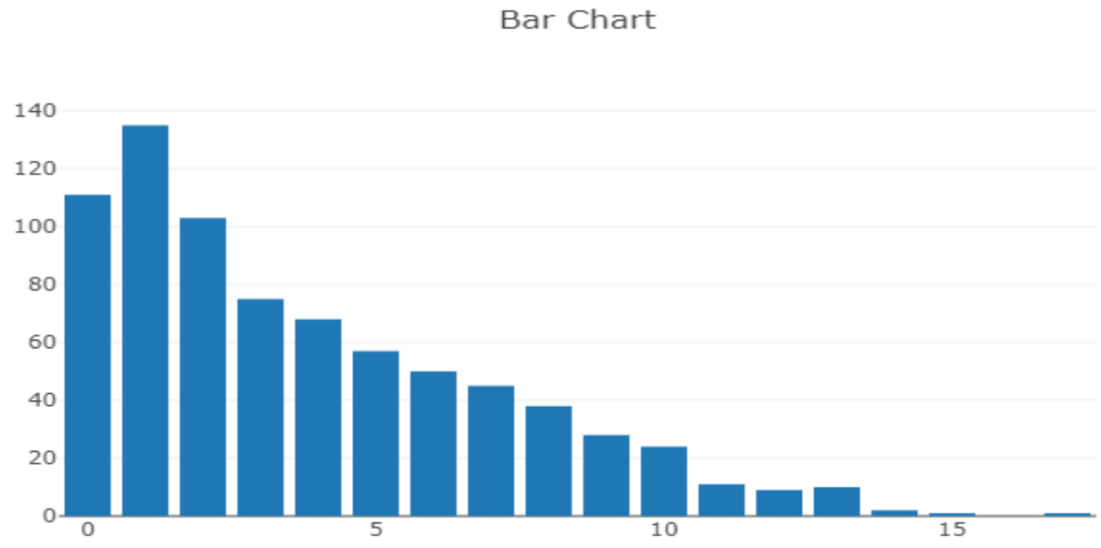


Figure-9

Since the Pregnancies is a categorical variable, we would not be using histogram for it rather we would use Bar plot. This bar plot shows that the number of 0 and 1 Pregnancies are dominating while it is clearly interpretable that as the number of pregnancies increase, the frequency of women giving birth to that many people decreases.

***Outcome Bar Chart***

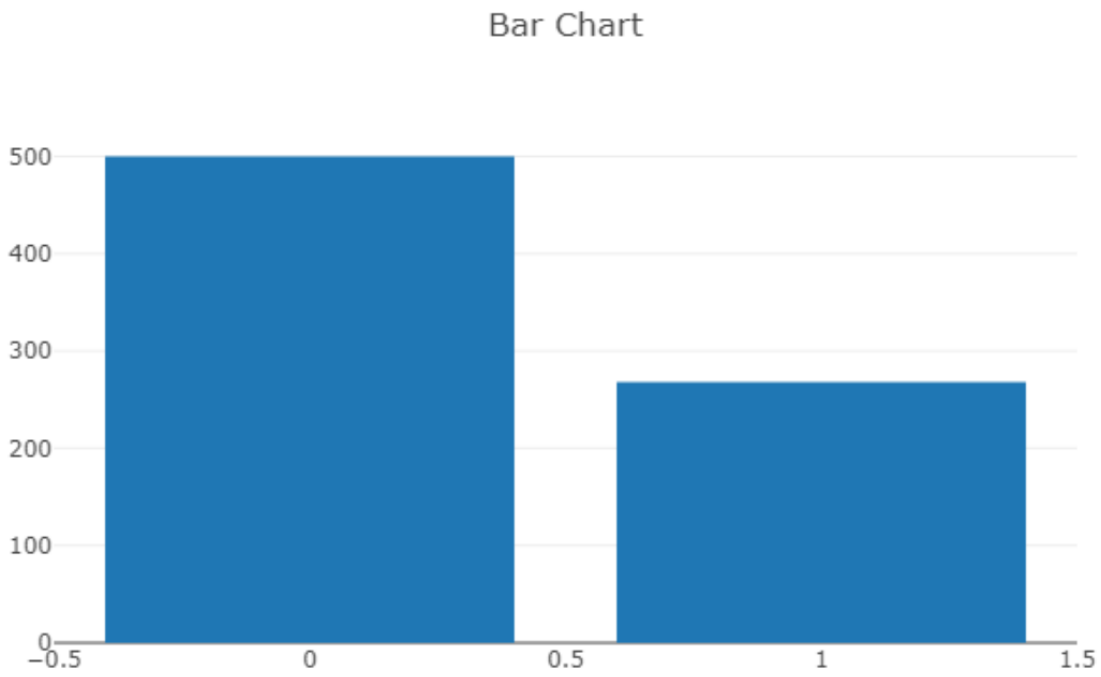


Figure-10

This is one of the interesting and vital visual because it clearly shows the class imbalance problem as discussed above. The number of 0 that is patients that does not have diabetes are 500 whereas the 1 that is patient having the diabetes are only around 250 which is not equal distribution at all. Therefore, this has to be catered well while evaluating the models for machine learning exploitation. From the healthcare point of view, this make sense because most of the patient in any hospital does not have diabetes compared to all those the remaining patients who have diabetes. Therefore, we can also observe the same nature in our dataset as well.

## 2.4 Multivariate Analysis and Visualizations

In multivariate visualizations we aim to extract the insights and relationship between more than one variable. This helps in understanding the correlation between variable and helps to find the better features for our model.

### *Age and Blood Pressure Relationship*

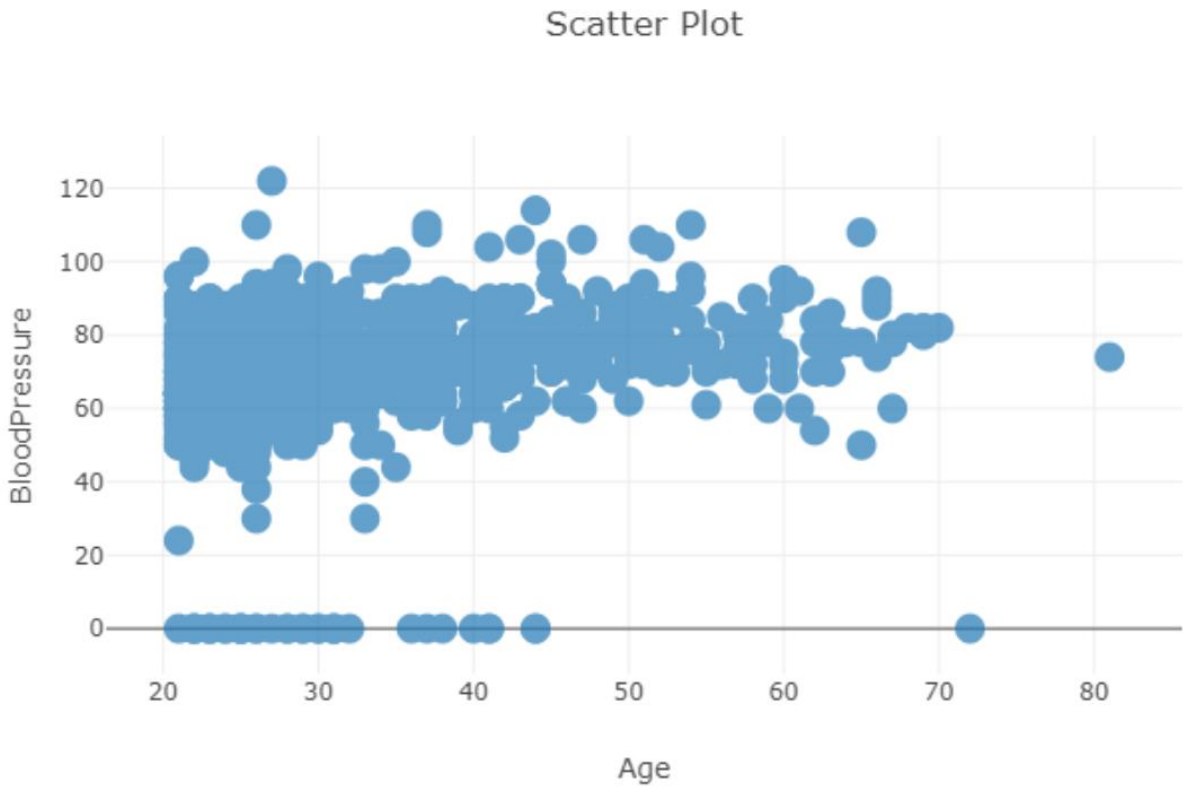


Figure-11

We can analyze that as the age increase the blood pressure has no significant impact on it which means that it is no truer that the blood pressure increases with age. With the recent generation, this has been a problem that they are unlikely to maintain a better and consistent blood pressure level. The graph can be interpreted in a way that from age 20-40 the area is more darker explaining that the observation in the dataset are mostly of younger people.

***Pregnancies and Outcome Relationship***

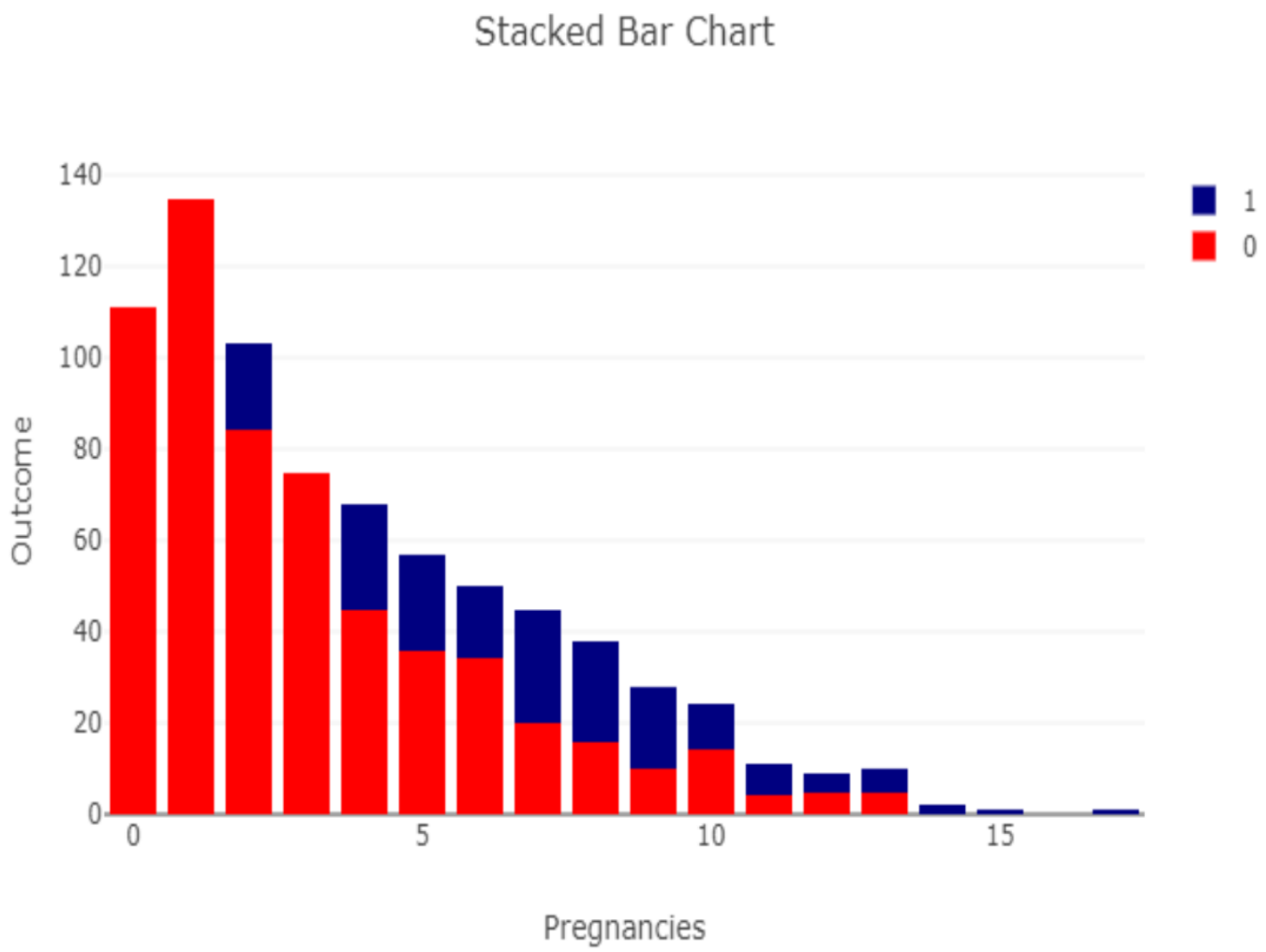


Figure-12

This stacked bar plot shows the relationship between two categorical variables the number of pregnancies and the outcome. This is very important visual which demonstrates that as the number of pregnancies increases the outcome tends to be true.

***Correlation Matrix***

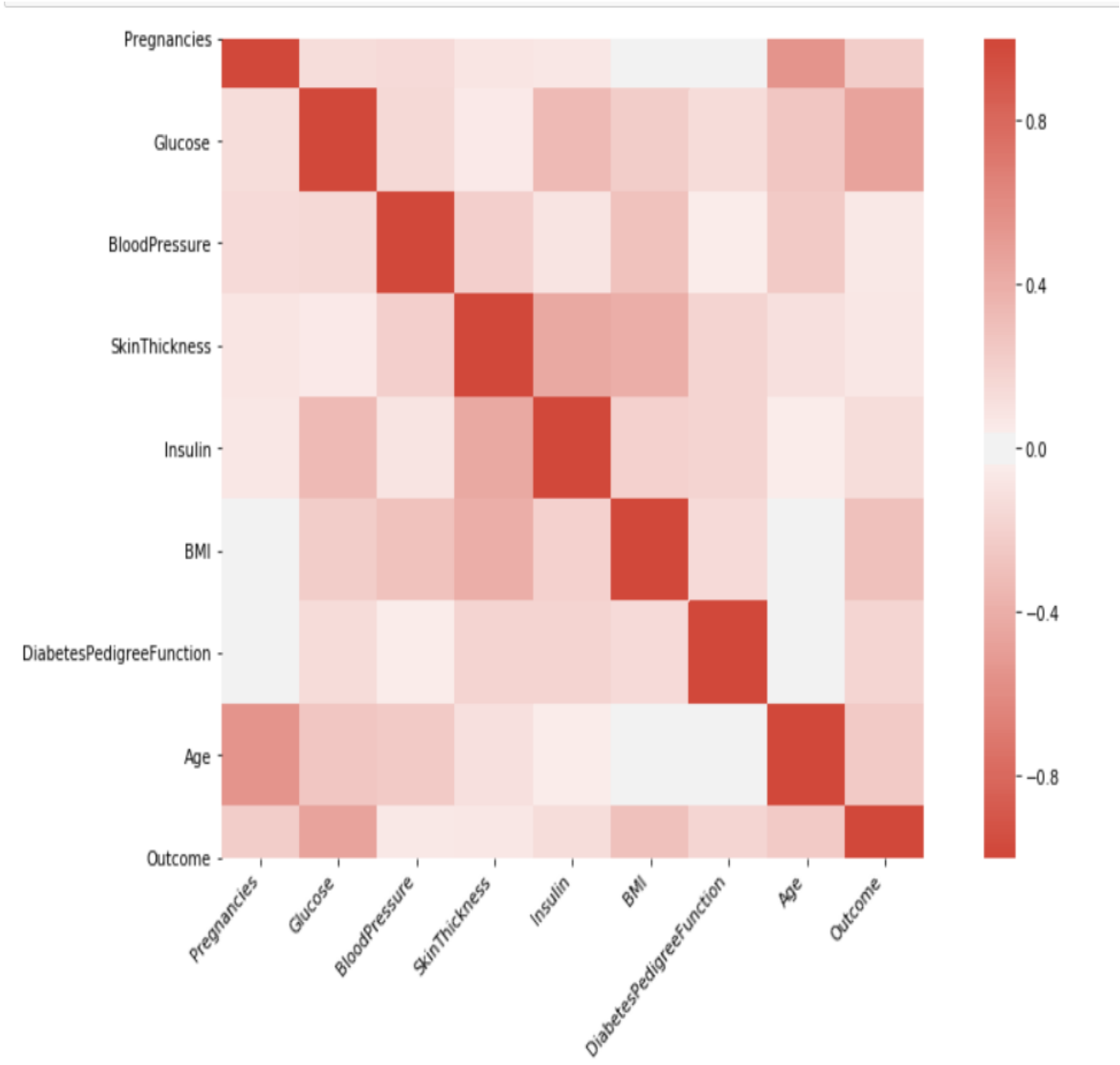


Figure-13

This visualization sums up all the relationship using the Pearson correlation. The darker the color, the positive relation it is between the two variables and vice versa for lighter colors.

# Chapter 3 – Research Findings

## 3.1 Algorithms

In this section we discuss about all the algorithms we have used in this research to classify the patient as having the diabetes or not. This includes variety of algorithms based on the nature of the dataset explored above.

### *Feature Elimination*

It's an algorithm based on either top-down or bottom-up approach, we take one variable and build the model on to it and then keep adding more variables to increase the accuracy in the latter, while we take all variables and keep removing variables in the latter. These are the most conventional techniques used when the dataset is not that big because what happens is that they make the calculations for each iteration until it converges (the same result is obtained every time you run the program). Therefore, they need heavy computational power and time to compute the weights for each model since this is a recursive algorithm. But, the best part about it is that it converges with the best possible accuracy on those variables, therefore since our dataset was reasonable, we could afford to run feature elimination on our dataset and get the desired results. This process can be analyzed from the below attached Figure-14.

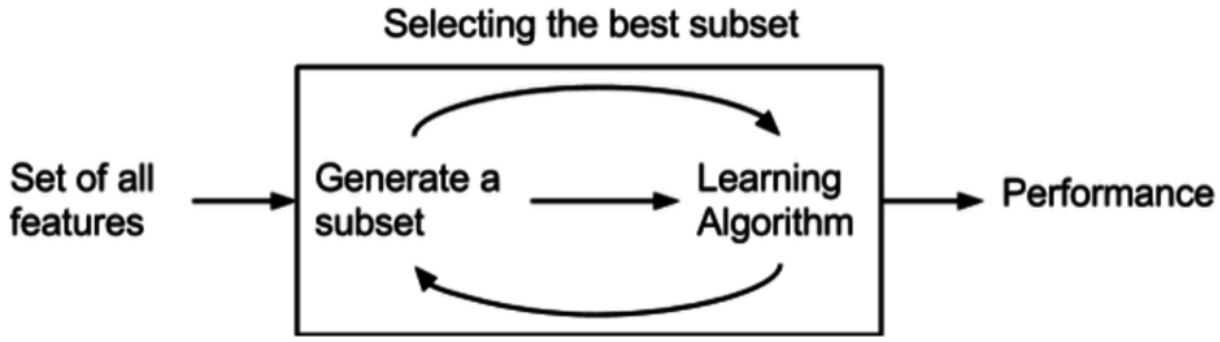


Figure-14

### *Logistic Regression*

It's a classification algorithm based on the idea of linear regression but the cost function (function which calculates the cost of the model by testing the results and improves itself using this function) is different in this case because it relies on logarithm or sigmoid function in order to generate the outcome as either 1 or 0 i.e. the binary output rather than a floating point number ranging from 0-1. Therefore this sigmoid function converts the probability into a fixed number either 0 or 1 which classify the output. The graph of sigmoid function can be seen in Figure-15.

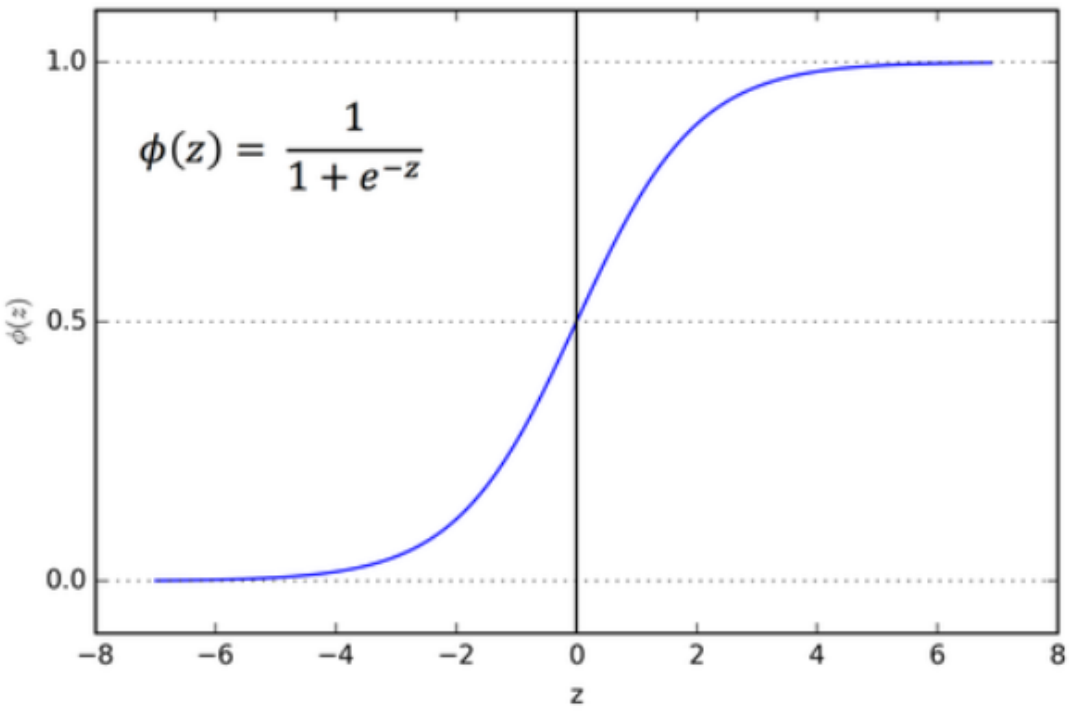


Figure-15

### *Results on Logistic Regression*

The results on logistic regression are quite impressive because the data did not have any outliers therefore the logistic regression was not disturbed by the unusual values. We managed to get 79 per cent accuracy on logistic regression classifier on test set. Which means that 79 per cent of the patients were classified correctly either as having the diabetes or either not having diabetes while 21 per cent of the patient were not classified correctly.

***Support Vector Machines***

It's another classification algorithm based on the idea of logistic regression but the cost function is different in a way that it divides the data of multiple classes in a more robust and sophisticated manner, it means that it creates the boundary decision line that divides the data of two classes properly with no nearby or overlapping values. This is why it is also known as *Large Margin Classifier*. This large margin or robust boundary can be visualized by looking into Figure-16 attached below.

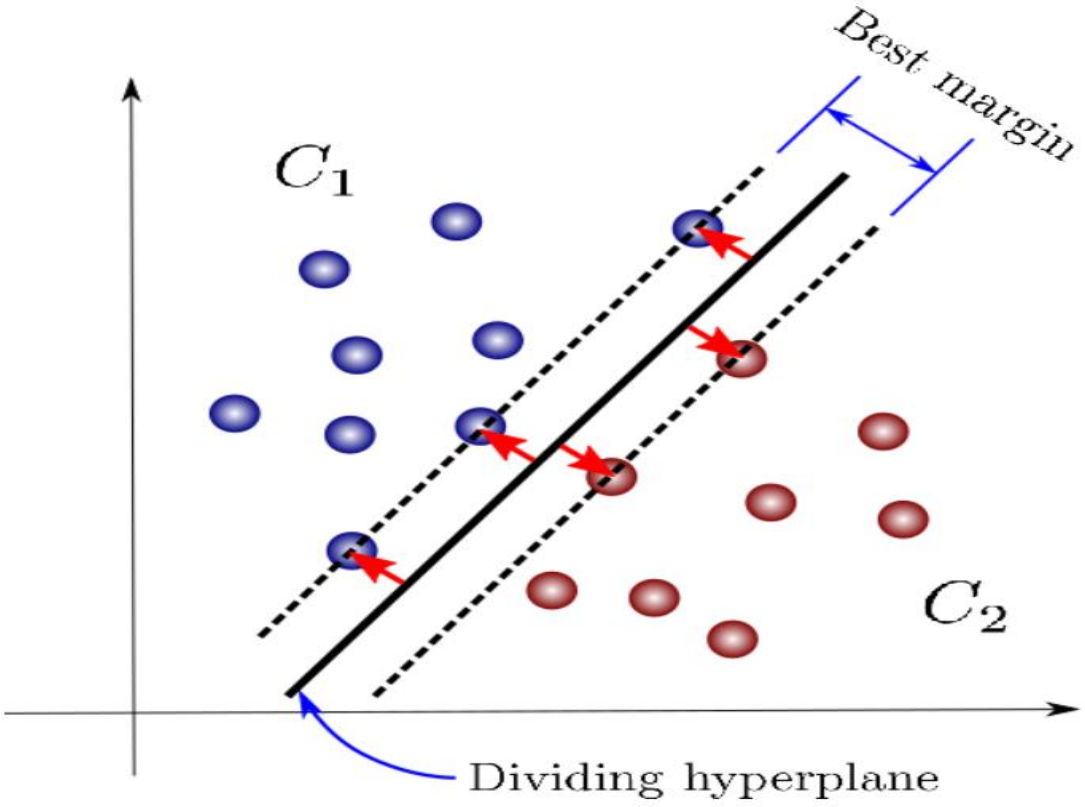


Figure-16

***Results on Support Vector Machine***

There are many different kernels (techniques to generate more data) available for SVM which are widely used such as Gaussian Kernel or Radial basis function kernel. They help to generalize the model well. We have managed to get 68 percent of accuracy of radial basis kernel on test set. It is fairly less than the logistic regression because the data generated by it was having intensive class imbalance problem therefore the accuracy falls down.

### 3.2 Evaluation of Methods

These algorithms and different methods must be evaluated before using them in industrial or practical implication, specially in healthcare department because it is very sensitive information and the results could have drastic effects, therefore the algorithms are better to reduce the human effort and can help out where humans can't help but they should not be trusted blindly.

There are several techniques to evaluate the models based on the nature of the dataset and as we have shown above our dataset is *imbalanced (unequal observations of positive and negative classes)*

***Accuracy***

It is the most common and trusted method to use but the condition is the dataset should not be imbalanced at all. Suppose we build a poor model which always says that no the patient does not have the diabetes and considering the class imbalance problem we say that 99 per cent observations in the test set are of patients that does not have the diabetes then the model would have 99 per cent accuracy because it would give the correct results 99 times out of 100 times. But this is wrong and the most useless model. Therefore, we need to think of another evaluation techniques.

***Precision and Recall***

It is one of the most favorable matric to evaluate models incase of imbalance data because here we does not care about the total dataset as a whole rather we care about the two segments (classes) of data. The one having the diabetes and the one not having the diabetes therefore the model has to give correct outcome in both cases. The formula for precision and recall are attached below in Figure-17.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Figure-17

Precision is basically how many correct outcomes the model has given out of total outcomes that model has given.

Recall is how many correct outcomes model has given out of total actual correct outcomes.

*F1-Score* is basically the average of both the Precision and Recall.

### *ROC Curves*

ROC stands for receiver operator characteristic which is another method which basically tells the AUC (Area under the curve) which means that the higher the area, the higher the model predicts 0 as 0 and 1 as 1. Therefore, we can say that the higher the area or ROC the better it predicts whether the patient has diabetes or not.

### *Results based on ROC and Precision and Recall*

The ROC for the logistic regression is 83 per cent on the test data which can be better analyzed from the below Figure-18.

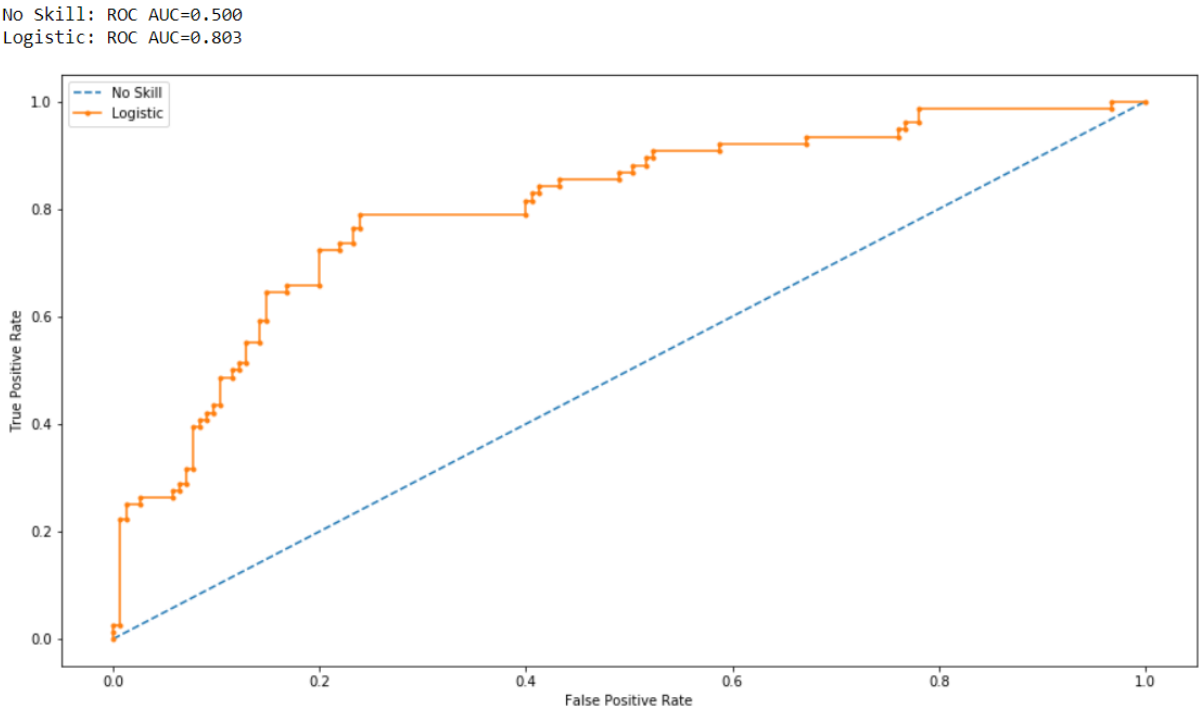


Figure-18

The ROC for the Support Vector Machine is 67 per cent as we discussed that the data generated was imbalance.

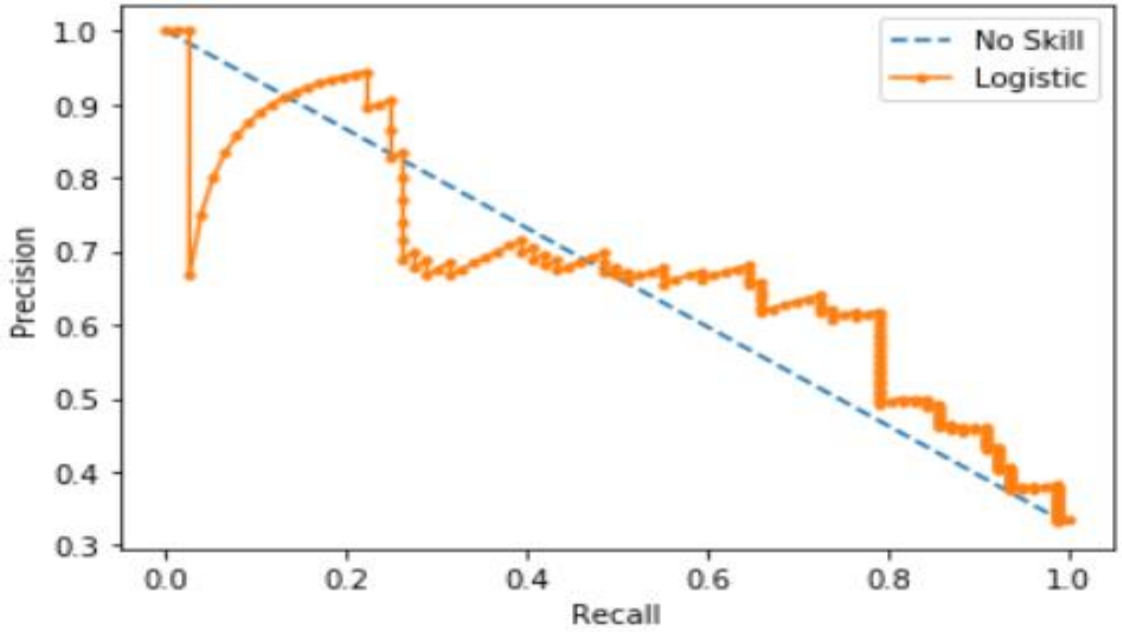


Figure-19

The Precision and Recall can be together analyzed in a confusion matrix (a 2x2 matrix incorporating all the information).

***Confusion Matrix***

| | |
|---|---|
| 118 | 12 |
| 25 | 37 |

True Label on Y-axis and Predicted Label on X-axis.

This tells us that the F1-score calculated from the above fromual described is around 70 per cent which means that we have successfully obtained the correct results 70 times out of 100 times.

# Chapter 4 – Conclusion

## Research Conclusion

The research was based on the idea of exploitation of machine learning in the area of diagnosis and prognosis of Type 2 diabetes. We collected the data from the authentic source which was imbalance. The statistical models were applied to find the relationship and to analyze the data. Later on, the machine learning algorithms were used to build the robust model which can predict whether the patient have diabetes or not, it turned out that we got good precision and recall whose average(F1-Score) was around 70 per cent which means that we can now use machine learning to explore the patient healthcare analytics specially to predict the Type 2 diabetes whose prove is given in this paper that it would turn out to be correct 70 times out of 100 times. Nevertheless, in serious cases, it is also suggested to investigate manual for confirmation and better cure.

# References

[1] Ahmad, A., Mustapha, A., Zahadi, E., Masah, N. and Yahaya, N. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. Digital Information Processing and Communications, pp.537-545.

[2] Alva, M., Hoerger, T., Zhang, P. and Gregg, E. (2017). Identifying risk for type 2 diabetes in different age cohorts: does one size fit all?. *BMJ Open Diabetes Research & Care*, 5(1), p.e000447.

# Appendix