Early Diabetes Prediction Based on Machine Learning Approach

Submitted by

Rabiul Islam

ID: 171-35-188

Department of Software Engineering
Daffodil International University

Supervised By

Ms. Marzia Ahmed
Lecturer
Department of Software Engineering
Daffodil International University

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Software Engineering.

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Thesis entitled on “Early Diabetes Prediction Based on Machine Learning Approach”, submitted by Rabiul Islam, ID No: 171-35-188 to the Department of Software Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc in Department of Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

------------------------------ External Examiner

Abu Shamim Aminur Razzaque
Director
Computer Ease Limited

------------------------------ Internal Examiner 1

SK. Fazlee Rabby
Lecturer
Department of Software Engineering
Daffodil International University

------------------------------ Internal Examiner 2

Md. Rajib Mia
Lecturer
Department of Software Engineering
Daffodil International University

DECLARATION

I hereby declare that, this thesis has been done by me under the supervision **Ms. Marzia Ahmed, Lecturer, Department of Software Engineering,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:

Marzia

Ms. Marzia Ahmed
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by:

Rabiul

Rabiul Islam
ID: 171-35-188
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for Her divine blessing makes me possible to complete the final year thesis successfully. I really grateful and wish my indebtedness to Ms. Marzia Ahmed, Lecturer, Department of Software Engineering, Daffodil International University. Deep Knowledge and keen interest of my supervisor in the field of “Data Mining and Machine Learning “to carry out this Thesis. Her endless patience, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to Dr. Imran Mahmud, Associate Professor and Head In-Charge, Department of Software Engineering, Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The number of diabetic patients around the world is increasing alarmingly day by day. It has now become a threat to our human society. This disease is usually caused by eating heavy sugary foods and not following a proper diet. However, nowadays machine learning algorithms can be used to easily and accurately predict for diabetics by checking and sorting out different types of symptoms. This can greatly reduce our mortality rate and make us more aware of diabetes. The purpose of my work is to make patients aware of and predict diabetes in advance using machine learning algorithms. I have used three algorithms for this task- Logistic Regression, Gaussian Naive Bayes, Random Forest. The overall performance of the three algorithms is evaluated in exceptional steps which includes accuracy, precision, F1 score, ROC accuracy, Recall, Standard deviation and K-Fold mean accuracy. Analyzing the all algorithms, it is seen that the random forest algorithm gave the best result of 86%.

CHAPTER 1
INTRODUCTION

1.1 Introduction

A complete or relative deficiency of a hormone called insulin in our body causes metabolic disorders that increase the amount of glucose in the blood and at one point it is excreted in the urine. This overall condition is called diabetes. In this disease, the amount of glucose in the blood increases chronically. The importance of insulin must be understood before understanding the disease of diabetes. Insulin is one such hormone. Which regulates the metabolism of carbohydrates and fats in the body. Glucose cannot enter the body without insulin. It accumulates in the blood vessels. In such a situation the person does not get his energy. It makes a person suffer from diabetes. In a healthy person, the amount of glucose in the blood plasma is less than 5.7 ml when fasting and 6.6 ml less than two hours after a meal. Diabetes is defined as the amount of glucose in the blood plasma when fasting is more than 8.1 ml mole or if the amount of glucose in the blood plasma is more than 11.1 ml mole two hours after eating 75 grams of glucose.

There are basically two types of diabetes- Type 1 diabetes and Type 2 diabetes.

Type 1: Insulin is not produced in the body of such patients at all. This type of diabetes is usually seen in people under 30 years of age. These patients have to take insulin injections to control diabetes. Otherwise, the blood sugar rises very fast and in a short time, he becomes unconscious due to acid poisoning in the blood and dies.

Type 2: Most of the patients in this category are above 30 years of age. Nowadays, the number of such patients under the age of thirty is increasing day by day. Their body makes insulin. However, if insulin is not injected if necessary, they do not get poisoned like type-1 patients. That means they are not insulin dependent. In many cases they can be treated with changes in diet and regular exercise.

There can be many causes of diabetes like- Eating too much junk food increases the amount of calories and fat in the body. Due to which the level of sugar in insulin in the body increases. Genetic diseases can also cause diabetes. Diabetes can be caused by excess body weight and weight gain. Again, not exercising daily physical activities can lead to diabetes.

Excessive stress can lead to diabetes. If a person smokes excessively, it can result in diabetes. Again, taking the wrong medication without a doctor's advice can cause a person to develop diabetes. Diabetes is caused by a person consuming tea, cold drinks and sugary foods. I am working on type 2 of diabetics in this work. I have used machine learning algorithms in my work to predict diabetics.

1.2 Motivation

At the current rate at which the number of diabetic patients is increasing, a study has found that by 2035, about 595 million people worldwide will have diabetes, which about 90% of people will have affected type 2 diabetes. Recent studies have shown that it depends not only on age but also on many other factors such as insulin, blood sugar, BMI level, age etc. If we can maintain daily proper nutrition diet and exercise we are less likely to get diabetes. Diabetes is a very complex disease so we can easily get infected with other diseases. Nowadays, medical science has improved a lot using machine learning methods. The use of these methods can provide an early prediction for diseases such as diabetes. As a result, I can easily become aware and protect ourselves from disease.

If I can get good results using machine learning methods for early detection of diseases then I will have unprecedented success in medicine. If physicians can easily get an idea of patient’s illnesses, it will benefit both the patient and the physician. If we can easily diagnose diseases using medical datasets in machine learning method then our medical cost will be reduced and patients will be able to make themselves aware easily. By using this method, people can easily become more aware of themselves and keep their life disease free. It is very easy to prevent a disease if it is predicted before it occurs.

The main reason for the growing number of diabetics around the world is the lack of awareness about our own bodies. There are very few people who take the necessary treatment before diabetes and follow the proper advice of doctors. Most patients seek medical attention after developing diabetes, but if they had been aware in advance, the disease would not have settled in their body. This method will help us to easily give a preliminary idea about whether a person is likely to be diabetic, thus greatly reducing the number of people with diabetes.

1.3 Rational of the study

Working with medical science is really very difficult and important. Because countless people around the world are dying from a variety of diseases, the number of diabetic patients among them is alarming. Currently, countless people around the world are affected by this disease. If this condition continues, many people will affected diabetes in the future. However, if we can be aware of this before I get diabetes and change our quality of life eating habits, then the chances of getting this disease will be greatly reduced. As a result, I have developed a model using machine learning methods to determine the likelihood of developing diabetes, using data from a variety of diseases to provide a preliminary idea of the likelihood of developing diabetes. This will make it easier for us to be aware and our medical costs will be greatly reduced, as well as the use of the intelligent device will bring about a massive change in our medical science.

1.4 Research Questions

- What is diabetes?
- How can we conscious people about diabetes?
- How to collect data?
- What are the benefits of diabetes prediction?
- How to analysis and pre-process data?
- How will this work help other researchers?
- What are the future works of diabetes prediction?
- Is the research relevant or not?
- How diabetes prediction model works?

1.5 Expected Output

My project is related about research based. So, my main goal is to publish research papers on related projects. The main purpose of this diabetes predication work is to make people aware of diabetes and to use advanced technology to diagnose diabetes. Through this work, I want to easily identify diabetes using machine learning methods, as well as raise awareness among patients with diabetes. I want to get better accuracy using the algorithms we apply in this work. I also want people to have an idea about the prognosis of diabetes and know the causes of diabetes.

1.6 Report Layout

The following report is arranged in the subsequent manner. This report has 5 chapters. Each chapter has different subparts which are described in detail.

- In chapter 1, I discussed Introduction, Motivation, Rational of the Study, Research Questions, Expected Output and Report Layout.
- In chapter 2, I discussed about Related Works, Research Summary, Scope of the Problem and Challenges.
- In chapter 3, I discuss about Introduction, Research Title and Apparatus, Dataset Description, Dataset Visualization, Correlation between features, Data Pre-processing, Classification Techniques, Model Evaluation Techniques, Receiver Operating Characteristic (ROC) and Area Under Curve (AUC).
- In chapter 4, I explained about Introduction, Model Performance and Analysis and Summary.
- In chapter 5, Discussion about the Work flow of the Study, Conclusions, Limitation and Future Work.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Diabetes is a commonplace lengthen disease and then humans did now not realize this disease how they affected it and after that they do not apprehend what is going to take place subsequent with them. Many forms of illnesses are created via diabetes including complexity of crucial organs and different organs of our frame. If we can be aware of the symptoms and causes of diabetes, it can be greatly reduced. Type 1 diabetes happens while pancreas will now not able to produce insulin. The insulin hormone balances our blood glucose type 1 diabetes generally occurs due to abnormally blood sugar degree. Loss of insulin within the blood then lack of insulin-producing beta cells inside the pancreas are the number one motive of type 1 diabetes. It's also called insulin subordinate diabetes mellitus. It can be you get type 1 diabetes through ancestral, in case your dad and mom has it and it is most determined in children. We are able to see the signs of type 1 diabetes like thirst, tiredness, weight reduction, common urination and growth in urge for food in a diabetic man or woman.

Type 2 diabetes in the main arrives while our frame’s cells and tissues ineffectively respond to insulin and it is the maximum common diabetes in humans. It is found that 90% human beings suffering from type 2 diabetes and 10% with the aid of type 1 diabetes and gestational diabetes. The body can't use and make insulin because of excessive blood sugar. Human beings who have type 2 diabetes they take remedy to improve the body’s insulin and attempt to lower the blood sugar of level that is produced by using liver. Human beings with at any age, type 2 diabetes may be arrived. But it's far most usually located in middle age or older human beings. Type 2 diabetes can be extend sickness with different fitness disorder like coronary heart disease, stroke, nerve damage, blindness, kidney harm and different part of human body if blood sugar level is not adequately controlled through treatment.

Many types of hormones secrete at some point of women being pregnant. The ones hormones grow blood sugar stage in the body and that’s why gestational diabetes occurs.

There is a possibility occurs type 2 diabetes and obesity later, who has gestation diabetes. Baby might die before or after beginning if gestation diabetes is untreated. There is no clear sample of inheritance of Type 2 diabetes. Therefore, the awareness and drug can improve the fitness of people as well as there may be no permanent treatment for diabetes.

2.2 Related Work

Research on diabetics pedicure has played an important role in medicine through the use of machine learning algorithms. A number of classification algorithms have been used in most of the research work on the prediction of diabetes, the work that has been done so far on this diabetes is discussed below-

Author Deepti Sisodia et al. [2] worked about diabetes prediction using classification algorithms for PIMA Indians diabetes database. They used three types of algorithms-decision Tree, Support Vector Machines and Naive Bayes. Overall the best accuracy was 76.30 percent. They validated the results using Receiver Operating Characteristic curves to accurately measure the results. The Author Muhammad Azeem Sarwar et al. [5] discussed diabetes using six different type of Machine Learning classification algorithms. In this work the maximum accuracy was 77%.

Author Md. Maniruzzaman et al. [3] worked about diabetes prediction using National Health and Nutrition Examination Survey (NHANES) (2009–2012) in US. They applied ML classification algorithms and used K2, K5, and K10 partition protocols. In this work the combination of Logistic Regression and Random Forest based classifier gives highest accuracy 94.10%. Author Dr. D. Asir Antony Gnana Singh et al. [4] presented a diabetes prediction system to diagnosis diabetes using medical data. They applied Naive Bayes, Multilayer Perceptron and Random Forest machine learning algorithm and k-fold cross validation, percentage split and use training dataset with preprocessing technique and without preprocessing technique. The pre-processing technique conducts better average accuracy for Naive Bayes.

Md. Aminul Islam et al. [7] discussed few ML algorithms to classification the dataset and compared those effects. Their remark turned into the expecting diabetes in early degree plays an essential role for a patient’s appropriate remedy approach. Hasan Temurtas et al. [6] analyzed the data to predict the diabetes in primary stage. They applied Levenberg–Marquardt method and a probable neural network architurer used.

Henock M. Deberneh et al. [11] predicted of type 2 diabetes using ML classification algorithms. In this work they collected medical record from (2013-2018) at a private medical institute called Hanaro Medical foundation in Seoul, South Korea. The model's performance proved reasonable and good in predicting the occurrence of T2D in the Korean population. Neha Prerna Tigga et al. [1] have utilized six types of ML methods on PIMA Indian dataset and their own dataset for predicted type 2 diabetes. After that they've as compared both datasets every different and got 94.10 percent accuracy from Random forest Classifier.

Han Wu et al. [8] labored to improve accuracy of prediction and make a version that might be able to adaptive to a couple of dataset. They used total three datasets and used WEKA toolkit for pre-processing, classifying, clustering, associating algorithms, and the visible interface. K-means cluster algorithm and logistic regression had been used on statistics. They attained ninety five percent accuracy that's 3.04% higher than others.

Author Huma Naz et al. [10] used deep learning to make a model for the risk measurement of diabetes in early stage. They have utilized total four diverse classifiers Artificial Neural Network, Naive Bayes, Decision tree and Deep learning. For data preprocessing they have used sampling technique (linear sampling, shuffled sampling, stratified sampling, and automatic sampling) on the dataset. Deep learning gave the highest accuracy rate of 98.07%. M. M. Faniqul Islam et al. [9] utilizes three machine learning algorithm. After that they applied tenfold cross validation and percentage split evaluation technique and best result achieved by Random Forest which is 97.4% for 10-fold cross validation and 99% for percentage split method.

Author Vinaytosh Mishraused et al. [12] used Logistic Regression to predecting diabetes. For this dataset they used Age, Smoking, Parental Diabetes Mellitus, hypertension and Waist Circumference, sex, BMI, HBA1C statistics as the characteristic. The statistics analysis became carried out the use of the software program device IBM SPSS 20.0. They discovered the probability 78.56 percent, Cox and Snell R square Nagelkerke rectangular 0.628, and Nagelkerke R square 0.839.

In this work I have to use three different types of classification ML algorithms to predicted diabetes and compared results.

2.3 Research Summary

I have worked in this project to predict diabetes. For this project I used the PIMA dataset obtained from online and add some new data collect from Bangladesh hospital on this dataset. And run three machine learning algorithms on the dataset. The three models are Logistic Regression, Gaussian Naive Bayes and Random Forest. And finally we have earned good result from the work. Though we have observed there is no real implementation as people are comfortable with doctor for consultation. But people would rely on this computerized diabetes prediction as they consult with doctor if it works appropriately. Many researchers using a single model or another multiple model for their work.

2.4 Scope of the problem

Separate machine learning algorithms have been devised for information on the prediction of diabetes, the study looked at different architectures that had already been applied to diagnose diabetes. In this study, researchers will use machine learning architectures to generate diagnostic results to find more advanced ways of diagnosing and treating diseases, but the results can be used at an early stage. The results of our study will help physicians in the field of primary care and will be a relevant role model for other diseases. Diabetics can use our research work to get a clear idea about diabetes and to be aware, it can reduce the risk of this disease. As our research work is about prediction diabetes and it is one of the most significant and unique work of medical area. The aim of our hypothesis will give a better prediction about diabetes.

2.5 Challenges

Working with any kind of medical prediction is a matter of many challenges. This is because the prognosis of a disease without any kind of examination is not considered acceptable by many at present. But our project will have to work hard to bring acceptance to medical science and face many tests initially. But the hope is that the models we use are giving very good results. The biggest challenge in doing that is data collection. Usually no one wants to provide data in medical science, which makes it very difficult to do this related work. If more of this data can be collected and the machine can be operated, the quality of this work will be further enhanced.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this chapter I will discuss what algorithms, methods I used and how I worked in my project. I must have knowledge about when I work with any algorithm. It is important to have a prior idea of how an algorithm works on a data set and its potential output. I will also discuss in this chapter what technologies and instruments I have used in my work. From data collection to data pre-processing, leveling implementation, everything will be discussed. How to predict diabetes through data mining and its processes are discussed in this chapter.

The range of diabetes patient is growing at an alarming rate everywhere in the global. On this research work I tried to predict diabetes sickness at a completely early degree. I use supervised studying procedure on this thesis. So that at first, we split our data set into two parts. Training data for trained the model and testing data for measure how accurately the model can predict the disease. I practice several ML methods and various data preprocessing technique for predict the diabetes more accurately. The full work process is given below which will serve as a summary of the entire research work-

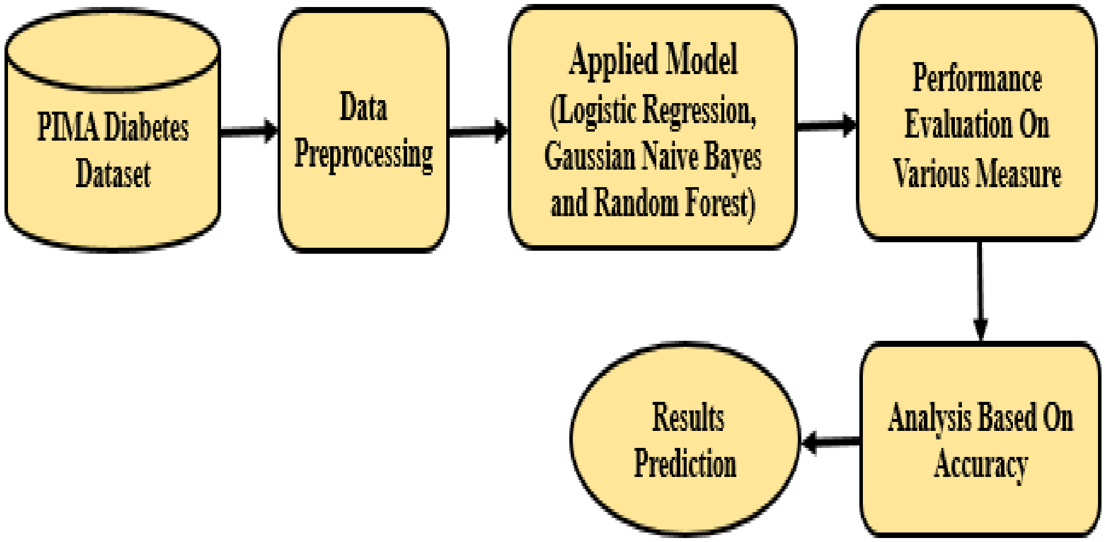

PIMA Diabetes Dataset
Data Preprocessing
Applied Model (Logistic Regression, Gaussian Naive Bayes and Random Forest)
Performance Evaluation On Various Measure
Analysis Based On Accuracy
Results Prediction

Figure 3.1: Workflow of depression detection system.

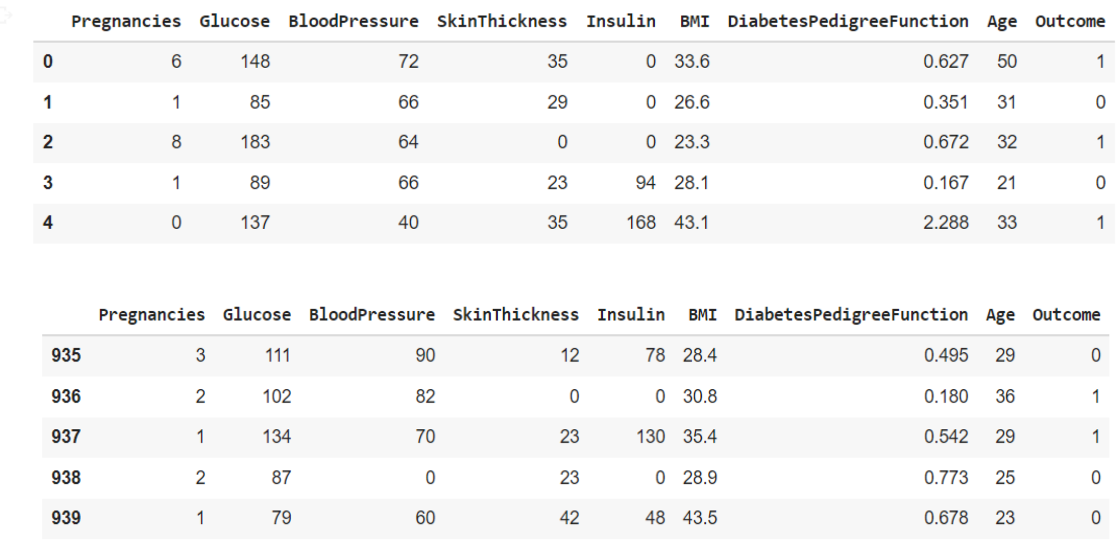
3.2 Research Title and Apparatus

Diabetes diseases a major problem in today's society. I tried to predict diabetes by using PIMA dataset and some new data collect from Bangladesh hospital. I used Machine Learning algorithms for my work. My main focus of research work was this diseases prediction. Topic of my research is “Early Diabetes Prediction Based on Machine Learning Approach”. It consists field of Machine Learning system. We used some software and hardware instruments. Names of the instruments are below:

- Software and Hardware:
  - Intel Core i5 7th generation with 8GB RAM
  - 1TB HDD
  - Google Colab with GPU
  - Google Chrome Browse
  - Key-board, mouse
- Development Tools:
  - Windows 10
  - Python 3.7
  - Numpy
  - Matplotlib
  - Seaborn
  - Pandas.

3.3 Dataset Description

In this work I used Pima Indians Dataset obtained from online and add some new data collect from Bangladesh hospital on this dataset. Which is predict the onset of diabetes based on diagnostic measures. In this dataset there are 940 data and 9 columns. The columns attributes are Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index, Diabetes pedigree function, Age (years) and Outcome. Here outcome 0 means not affected by diabetes and 1 means affected by diabetes. The data set is based on female data. Below is a short picture of dataset-



	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

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
935	3	111	90	12	78	28.4	0.495	29	0
936	2	102	82	0	0	30.8	0.180	36	1
937	1	134	70	23	130	35.4	0.542	29	1
938	2	87	0	23	0	28.9	0.773	25	0
939	1	79	60	42	48	43.5	0.678	23	0

Figure 3.3: Sample of dataset.

3.4 Dataset Visualization

Data visualization is a graphical presentation for a dataset like – pie chart, bar chart, graphs, histogram, count plot, pair plot, maps, scatter plot etc. It maintain the relation between data and image. For data mining related work it is very important to understand the full work easily. Data visualization help us to understanding data clearly, make the best decision among several decision to solve problem and comparative analysis.

3.4.1 Count Plot

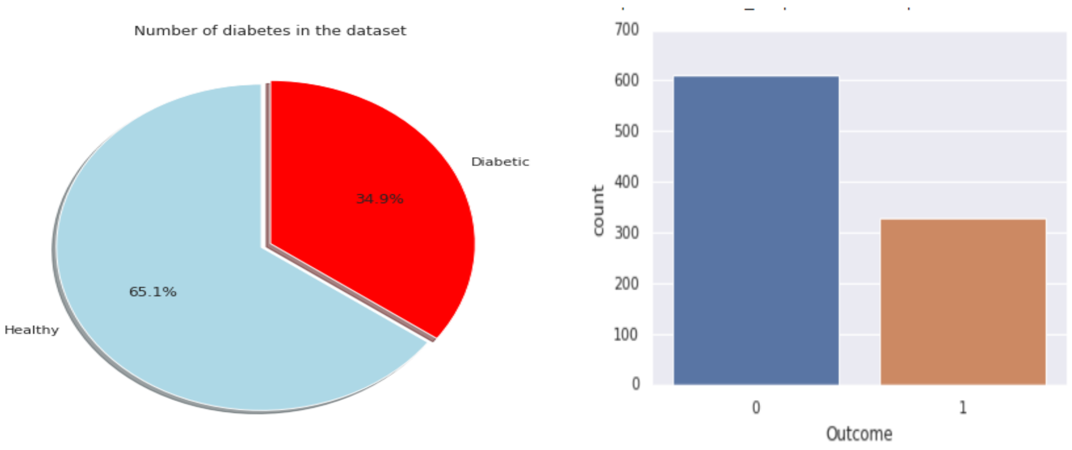


Figure 3.4.1: Count Plot for dataset Visualization

The dataset visualization from the pie chart (Figure 3.4.1), we can say that around 65.1% of the people are Healthy and 34.9% are Diabetic. Again the outcome of True case (1) is 328 and False case (0) is 612. We can see that this dataset is imbalance dataset.

3.4.2 Pair Plot

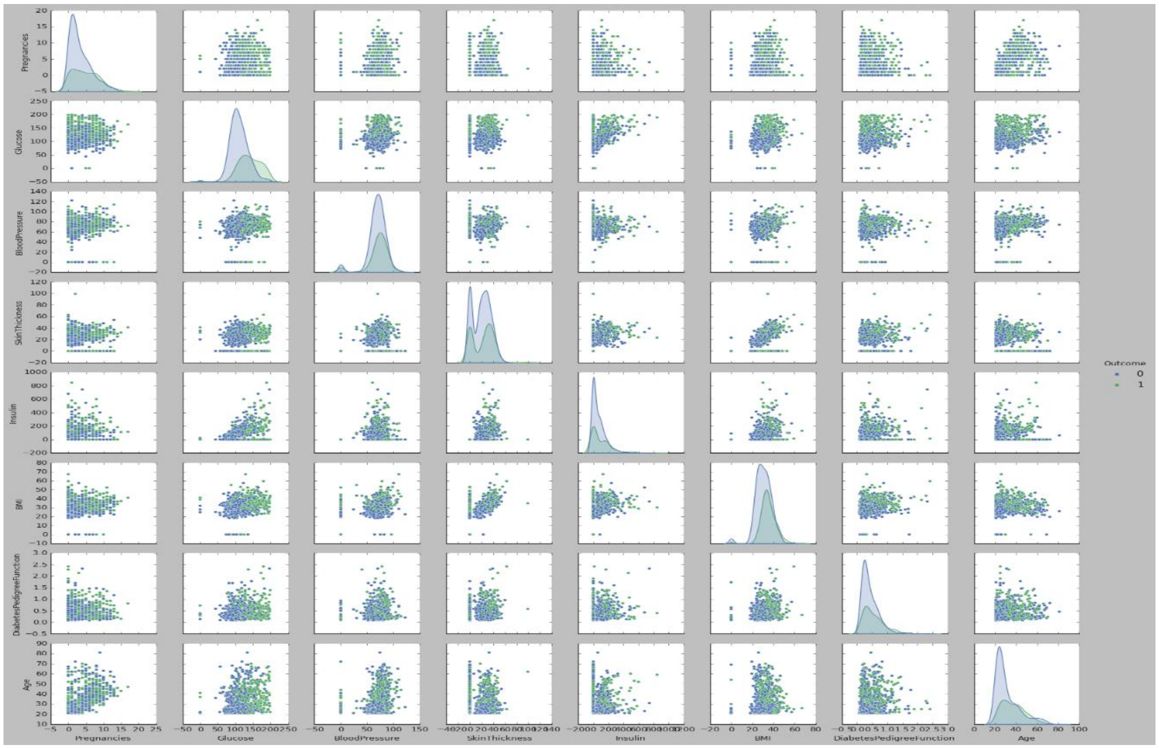


Figure 3.4.2: Pair Plot for dataset Visualization

In (Figure 3.4.2) pair plot visualization means the relationship between the dataset. It work for the data distribution.

3.4.3 Distribution Plot

Distribution is a statistical method, where I can understand the observation number within a given interval. Here I visualize each feature by histogram. Histogram is a graphical display of distribution numerical data. From visualization of frequency distribution anyone can easily understand the majority occurrence range in the dataset.

3.4.3.a Healthy vs Diabetic by Pregnancy

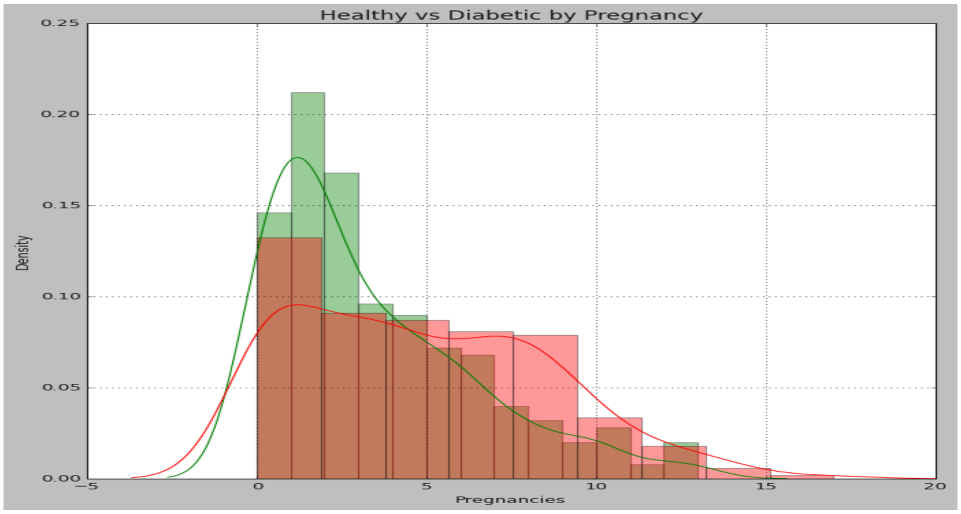


Figure 3.4.3.a: Distribution Plot for Healthy vs Diabetic by Pregnancy

From the graph (Figure 3.4.3.a), we can say that the Pregnancy isn't likely cause for diabetes as the distribution between the Healthy and Diabetic is almost same.

3.4.3.b Healthy vs Diabetic by Glucose

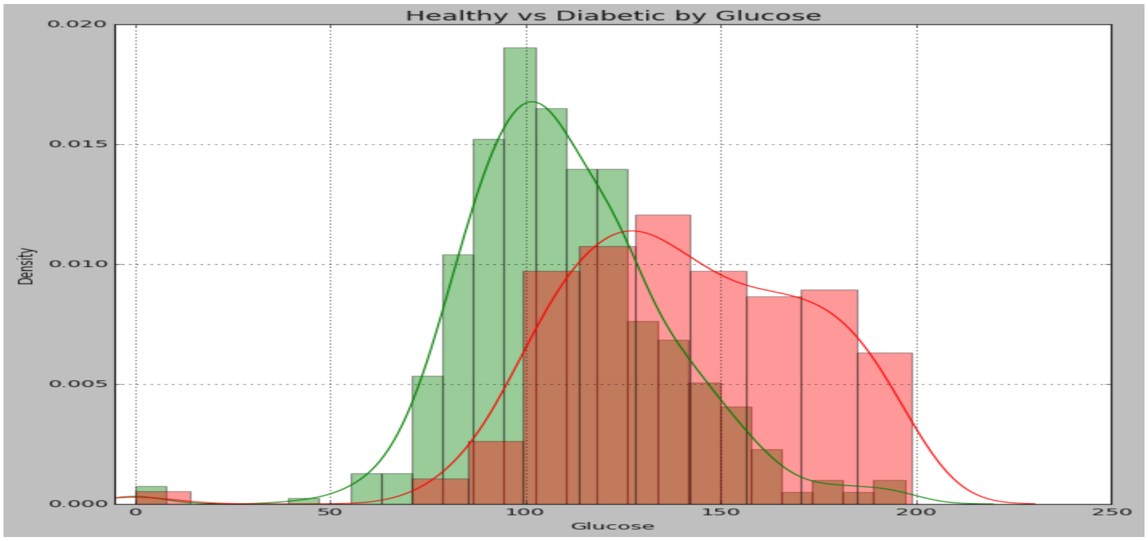


Figure 3.4.3.b: Distribution Plot for Healthy vs Diabetic by Glucose

Diabetes is a disease that happens in your blood glucose, which is also known as high blood sugar. Blood glucose comes from your principal supply of strength and the meals you consume. The Glucose level for a Normal Adult is around 120-130mg/dl anything above it means that the person is likely suffering from pre-diabetes and diabetes.

From the graph (Figure 3.4.3.b), I can see the Healthy person are more around 120mg/dl but it then gradually drops, and for diabetic person it is vice versa.

3.4.3.c Healthy vs Diabetic by Blood Pressure

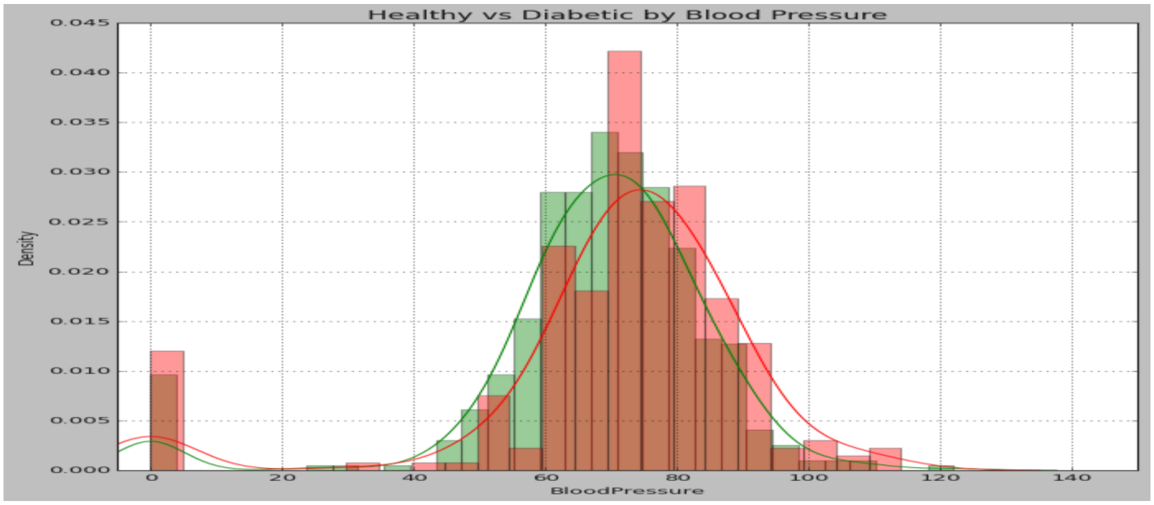


Figure 3.4.3.c: Distribution Plot for Healthy vs Diabetic by Blood Pressure

High blood pressure (hypertension) may be very not unusual in human beings with diabetes. Diabetes damages the arteries and goals them to harden, called atherosclerosis. It is able to reason excessive blood strain, if left untreated it can reason problems including vascular damage, coronary heart attack and kidney failure. For a Normal person BP should be below 120/80 mm Hg, the person with hypertension can be above 139/89 mm Hg.

From above graph (Figure 3.4.3.b), I can say that, diabetic and healthy people are evenly distributed with low and normal BP but, there are less healthy people who have high BP.

3.4.3.d Healthy vs Diabetic by Skin Thickness

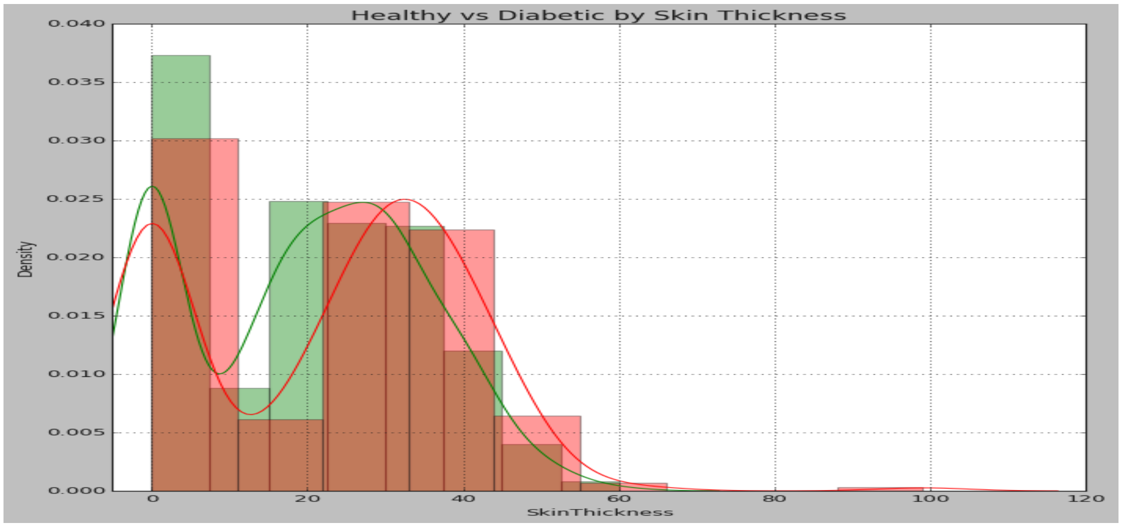


Figure 3.4.3.d: Distribution Plot for Healthy vs Diabetic by Skin Thickness

Changes in blood vessels due to diabetes can cause a skin condition known as diabetic dermatopathy. The Dermatopathy seems as scale patches which might be light brown or pink, frequently at the front of the foot. The patches do now not hurt, blister or itch and remedy is commonly now not important. From the (Figure 3.4.3.d) above, the distribution between healthy and diabetic individuals is sort of equal to the thickness of the skin.

3.4.3.e Healthy vs Diabetic by Insulin

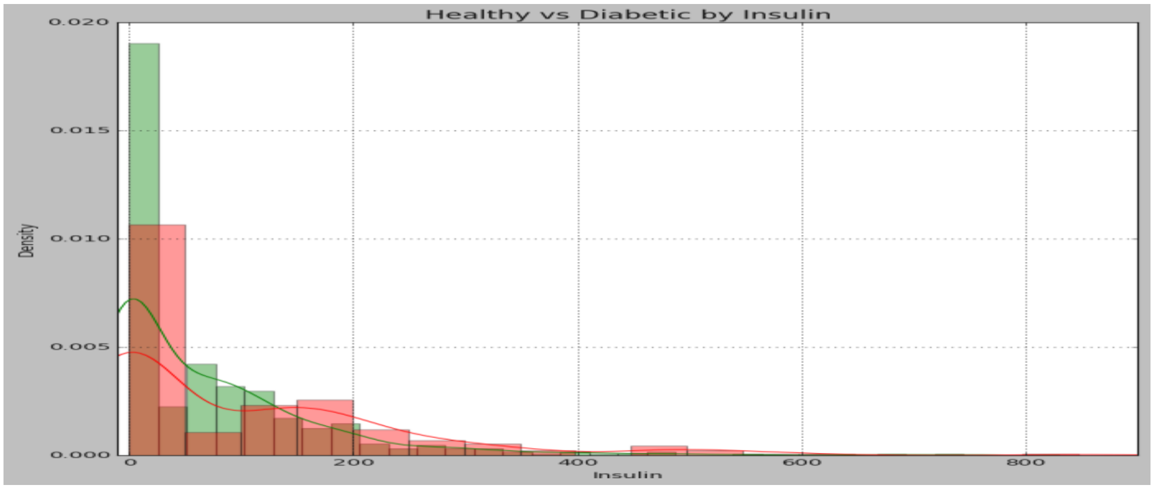


Figure 3.4.3.e: Distribution Plot for Healthy vs Diabetic by Insulin

Insulin is a hormone that lets in your pancreatic cells to use glucose. Whilst your body isn't making or the usage of insulin nicely, you could take guy-made insulin to govern your blood sugar. Insulin helps manipulate blood glucose levels by using signaling the liver and muscle groups and fats cells to receive glucose from the blood. Insulin allows cells get hold of glucose to use for strength. From the above (Figure 3.4.3.e) I can see that diabetic patients have an increase in insulin stages as they step by step increase.

3.4.3.f Healthy vs Diabetic by BMI

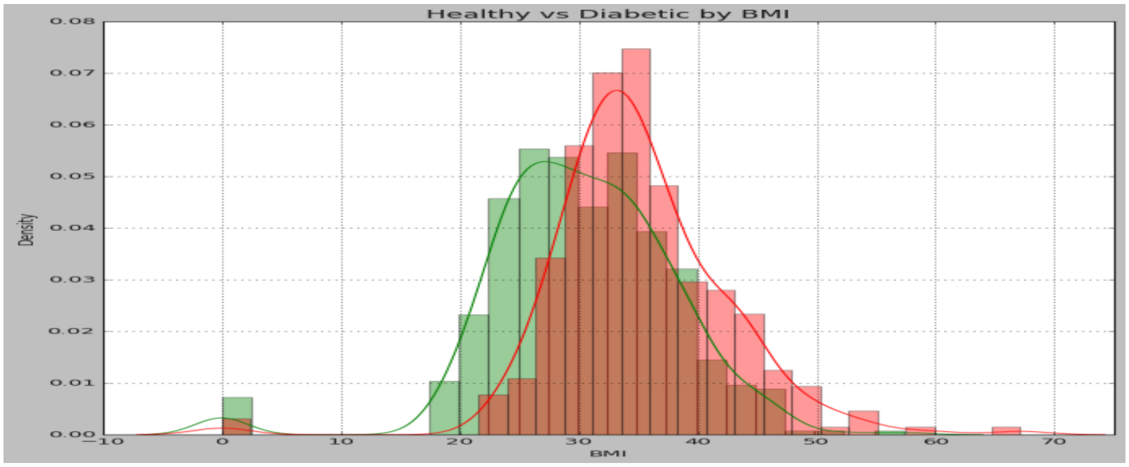


Figure 3.4.3.f: Distribution Plot for Healthy vs Diabetic by BMI

Being overweight or obese increases the risk of developing type 2 diabetes. I can determine this from the (Figure 3.4.3.f) above, as an increase in BMI reduces a person's chances of staying healthy and increases their chances of developing diabetes.

3.4.3.g Healthy vs Diabetic by Diabetes Pedigree Function

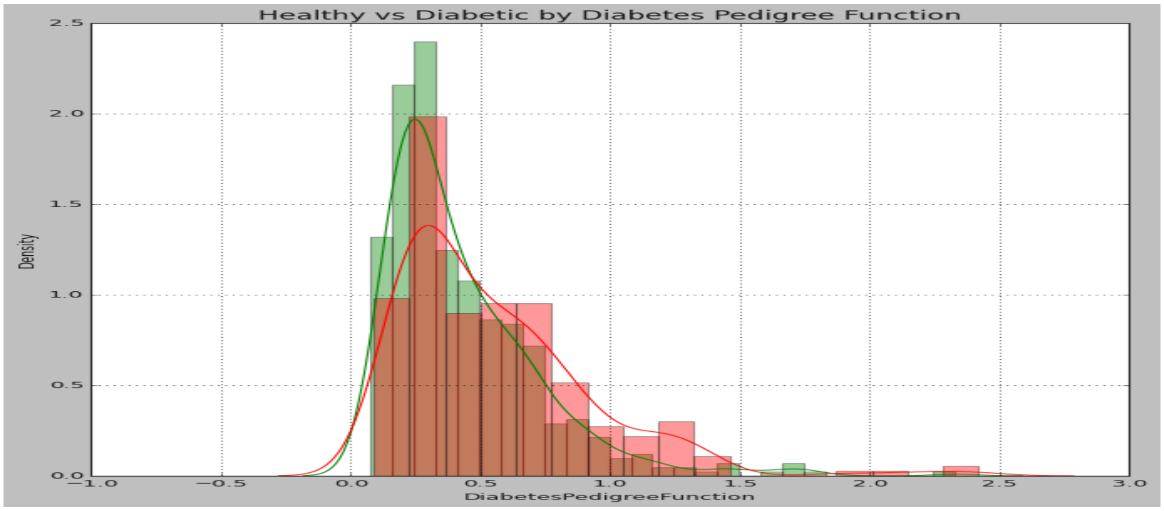


Figure 3.4.3.g: Distribution Plot for Healthy vs Diabetic by Diabetes Pedigree Function

Diabetes pedigree characteristic is a function that reduces the chances of diabetes primarily based on own family records. It affords patients with a records of diabetes mellitus in close household and a few information approximately the genetic relationships of those loved ones. From the (Figure 3.4.3.g), the function increases with the onset of diabetes that diabetes may be hereditary in that character.

3.4.3.h Healthy vs Diabetic by Age

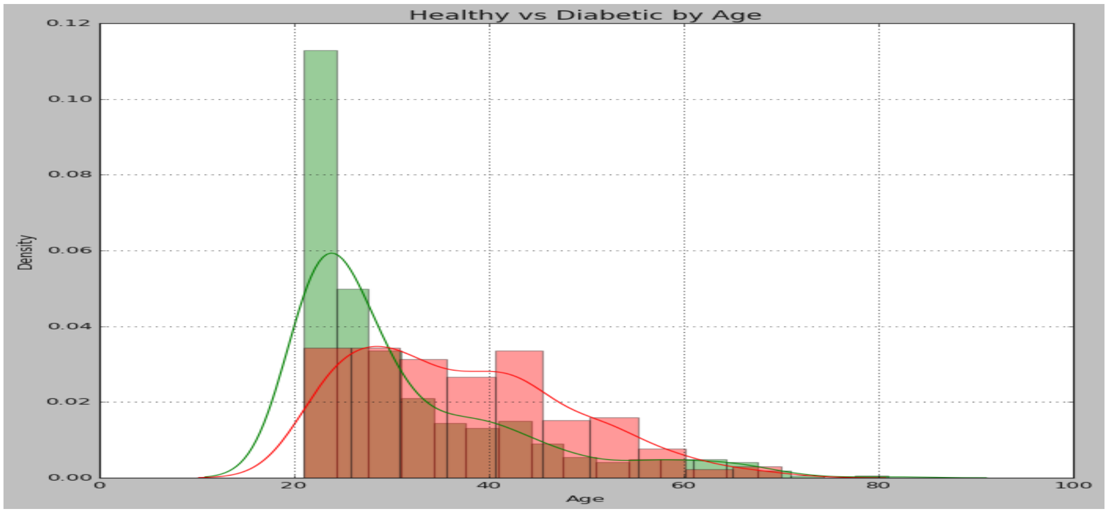


Figure 3.4.3.h: Distribution Plot for Healthy vs Diabetic by Age

As people age, they are at better chance for growing type 2 diabetes due to the combined impact of insulin resistance and impaired pancreatic islet characteristic with age. From the (Figure 3.4.3.h), I am able to see that there are more wholesome human beings across the age of 20-25, however with age, there is additionally diabetes, which shows that age and diabetes cross hand in hand.

3.5 Correlation between features

Correlation analysis is used to quantify the degree to which two variables are related. It's provide with a linear relationship between two variables. When I correlate feature variables with target variables get to know that how much dependency is there between particular variables and target variables. Here the right-hand side of the figure is shown the value of correlation that has marked two different color. The lighter side means less correlation and the darker side means strong correlation between them. Here, if I considered the dependent column "Outcome" I can find out the correlation of "Outcome" column with all other columns. From “Outcome” column I see that “Outcome” is less dependent on “Blood Pressure” and “Glucose” is very important column for predicting diabetes disease. If I want to drop any column, I can drop the “Blood Pressure” Column form our dataset.

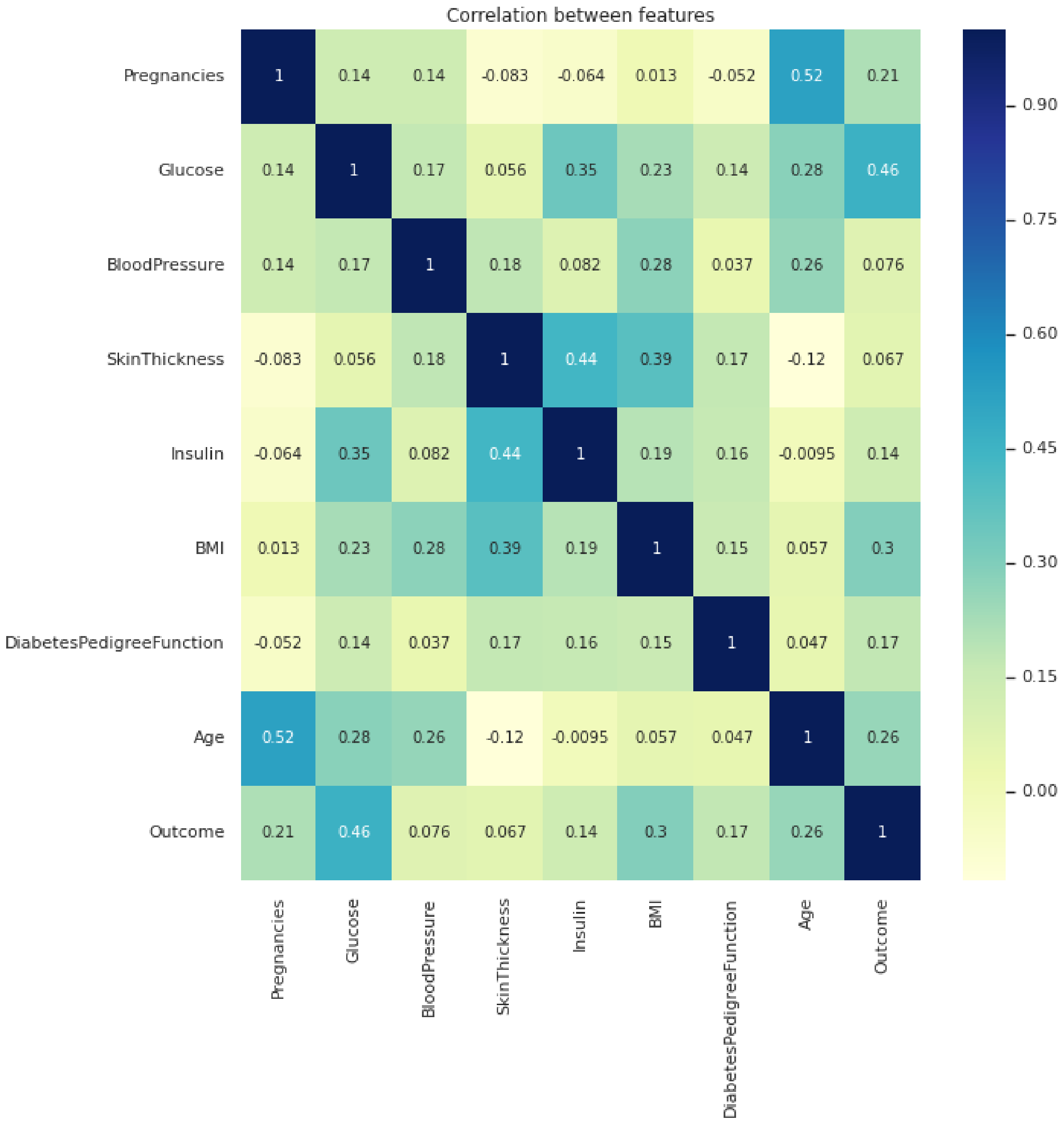


Figure 3.5: Heat map of Correlation

3.6 Data Pre-Processing

It is not possible to get an organized records from actual world. In actual world most of the time the dataset is incompatible, loss of specific vital behavior and incorporate many mistakes. Data preprocessing is a method to reduce this problem. Data Preprocessing is very important for records set. It could put together the uncooked facts for making it appropriate to construct distinct system studying version. We notice there is a lot of missing fee in different columns in our information set. In this data set, I drop the duplicates values, then check null values and check the number of zero values. Then I split the dataset and divided into train-test.

3.7 Classification Techniques

I have used three methods of machine learning algorithms to complete this thesis, through which I can get an idea about the prognosis of diabetes. I applied classification based ML approach which are Logistic Regression, Gaussian Naive Bayes and Random Forest. Now the algorithms are described below-

3.7.1 Logistic Regression

Logistic regression is comparable to multivariate regression and it creates a model to explain the effects of multiple predictors of response variables. The variable of a continuous result can be converted to a classified variable to be used for logistical regression. However, breaking uninterrupted variables in this manner is largely discouraged because it reduces accuracy. Logistic Regression is a class algorithm of supervised gaining knowledge of which deals with possibility to expect an outcome for dependent variable that's binary nature with the aid of using logistic function. It works with non-stop and discrete value and its purpose to find the best fitting version for independent and dependent variable relationship. In a graph we determined it like S form. Which means there is no chance that the value would be fraction, so the value would be both 0 and 1. And it in no way crosses the restrict. It keeps any relationship. The formula of Logistic Regression is,

$$\boldsymbol{f(x)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e^{-x}}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathbf{(1)}$$

3.7.2 Gaussian Naive Bayes

A Gaussian Naïve Bayes algorithm is a special type of Naïve Bayes algorithm that works with continuous data. Especially when features are in constant use. All its features follow Gaussian distribution i.e. general distribution. This model can only be found by looking for average and standard deviations of points within each label. It is a part of the supervised machine learning classification algorithm based on Naive Bayes. Although it is a common classification method, it has high efficiency. Gaussian Naïve Bayes is a form of Naïve Bayes and Probabilistic approach algorithm. It is utilized for continuous values. Firstly, the algorithm starts classification after that total block come. It is based on Gaussian distribution where the data is lied between a center points that means they will not very right or left. The formula is given below,

$$\boldsymbol{P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)} \textbf{.............................. (2)}$$

This function is called probability density function. Here, we can search out mean, variance and standard deviation together.

3.7.3 Random Forest

Random forest is kind of a powerful and popular ensemble classifier which is using decision tree algorithm in a randomize way. It is combined of multiple decision tree. To build every single decision tree it uses bagging technique. When we classify a new object, we got classification from each tree as tree vote. And the major vote for classification is accepted. That is why it rather provides more accurate result than single decision tree. And in the case of regression takes the average of the outputs by different trees. Random Forest algorithm of rules is a popular approach of rules of ML. Random Forest has many medical and biological troubles it can remedy both classification and regression troubles in healthcare offerings.

3.8 Model Evaluation Techniques

By applying different performance parameter, I can evaluate the performance of Machine learning algorithm which we applied in this thesis work. Confusion matrix is a two-by-two matrix that regulates the performance of a classification model. There can be 4 cases from where I know about the number of positive and negative occurrences were classified correctly or incorrectly.

Table 1: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	TN	FP
Actual Positive	FN	TP

True Positive (TP): True Positive means both actual class and predicted class is true (1) so, patient has complexity in reality and also classified true by the model.

True Negative (TN): True Negative means both actual class and predicted class is false (0) so, patient has not complexity in reality and also classified false by the model.

False Positive (FP): False Positive means actual class is false (0) but predicted class is true (1) so, patient has not complexity in reality but classified true by the model.

False Negative (FN): False Negative means actual class is true (1) but predicted class is false (0) so, patient has complexity in reality but classified false by the model.

The computation method of the measurement equations are as follows below-

Accuracy is measured by total number of correct classifications divided by total number of classifications. The formula is presented by,

$$\boldsymbol{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + FP + TN + FN}} \text{ ....} \textbf{ (3)}$$

According to the Confusion Matrix, Precision is the ratio between true-positive samples and predicted yes samples.

$$\boldsymbol{Precision} = \frac{\mathbf{TP}}{\mathbf{TP + FP}} \text{ ....} \textbf{ (4)}$$

Recall is also known as Sensitivity. According to the Confusion Matrix, Recall is the ratio true-positive samples and actual yes samples.

$$\boldsymbol{Recall} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \text{ ....} \textbf{ (5)}$$

Specificity is measured by true negative divided by total number of actual negative. The formula is presented by,

$$\mathbf{S}\boldsymbol{pecificity} = \frac{\mathbf{TN}}{\mathbf{TN + FP}} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ \mathbf{(6)}$$

F1 Score is measured by multiplication of precision and recall divided by addition of precision and recall and multiply by 2. The formula is presented by,

$$\mathbf{F1 = 2 *} \ \frac{\mathbf{precision * recall}}{\mathbf{precision + recall}} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ \mathbf{(7)}$$

3.9 Receiver Operating Characteristic (ROC)

ROC curve is one of the important evaluation metrics that should be used to check the performance of a classification model. It is also called relative operating characteristic curve. Because it is comparison of between two main characteristic True Positive Rate and False Positive Rate. It measures different thresholds value for determine the best threshold point of the model. Form ROC curve we understand how well the model for predicting positive and negative class.

3.10 Area Under Curve (AUC)

AUC is the area under ROC curve. This is use for determine the performance of the classifier. Its range is 0 to 1. The higher value of AUC means better performance 0.5 means the model has no separation capacity. For balanced dataset, AUC are useful for measure the performance. Higher value defines better performance. AUC curve helps us to choose the best model amongst the models for which we have plotted the ROC curve.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this work, I conducted different analysis to evaluate the three type's machine learning classification algorithms for diagnosis and prediction of diabetes disease. For this work I used 80% data for training and 20% data for testing. I tried to de predict diabetes using Logistic Regression, Gaussian Naive Bayes, Random Forest algorithm. It is very complicated to work with the prognosis of any kind of disease. Because if these predictions do not work properly, patients can suffer a lot. So I compared the results between the three models in this work and tried to give the best prediction.

At first, I trained our training dataset by using different machine learning algorithm and build a model then we tested on test dataset in this model. I compare the result with different classifier performance. The algorithms helped to make a good prediction based on various data and results like- K-Fold Mean Accuracy, Standard Deviation, ROC AUC accuracy, Precision, Recall, F1 score. Now the result discussion in tabular from for three models in below-

Table 2: Classification performance measurements

Model	**Accuracy**	**K-Fold Mean Accuracy**	**Standard Deviation**	**ROC AUC**	**Precision**	**Recall**	**F1 Score**
Random Forest	86.170	82.042	3.462	0.852	0.836	0.785	0.805
Logistic Regression	80.320	76.331	6.323	0.759	0.769	0.615	0.683
Gaussian NB	74.468	74.724	5.592	0.710	0.639	0.600	0.619

4.2 Model Performance and Analysis

In this section I discussed about my project working model and result analysis. The expected outcome of this research has been found by applying different machine learning algorithm. I use both cross validation and percentage split technique then we compare the algorithm for find out the best algorithm from which we can obtain the highest accuracy. For measure the performance of each model we built confusion matrix for both k fold cross validation and percentage split method.

For this work I used ROC-AUC curve Measures performance for classification problems in different threshold settings. ROC represents a probability curve and the degree or measure of AUC isolation. It shows how many models are able to distinguish between classes. By comparison, higher the AUC, better the model is at distinguishing between patients with the disease and no disease.

I used Precision-Recall Curves, It’s like the ROC curve and the accuracy-reconstitution curve is used to evaluate the effectiveness of the binary classification algorithm. It is often used in situations where classes are unbalanced. Like the ROC curve, the precision-recall curve provides a graphical representation of the classification function across multiple margins rather than a single value.

For each algorithm Performance Parameter, confusion matrix, precision – recall curve and ROC AUC curve is shown below-

4.2.1 Logistic Regression Algorithm Analysis

After fitting Logistic Regression model, my Accuracy Score is 0.803. In simplest terms, this means there is 80% chance that the model will be able to correctly predict all healthy and diabetic patients from our dataset.

I found the confusion matrix for both percentage split and k fold cross validation method.

For this model the Performance parameters (table 3) is below-

Table 3: Performance Parameters Report of Logistic Regression

Architecture	**Precision**	**Recall**	**F1-score**
Healthy (0)	0.82	0.90	0.86
Diabetes (1)	0.77	0.62	0.68
Macro Average	0.79	0.76	0.77

Now the confusion matrix (Figure 4.2.1.1) for this algorithm is below-

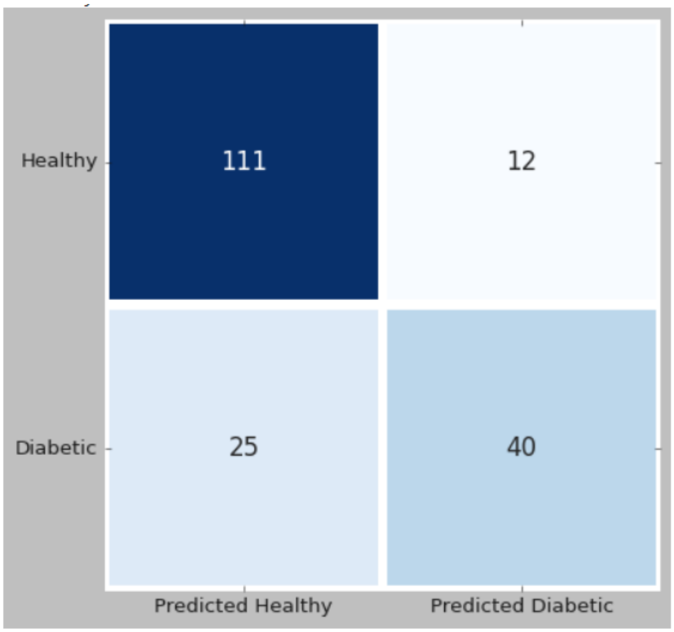


Figure 4.2.1.1: Confusion Matrix for Logistic Regression

Here the ROC AUC curve (Figure 4.2.1.2) and ROC AUC score is 0.85 for Logistic Regression algorithm. In simplest terms, this means there is 85% chance that the model will be able to separate the diabetes and healthy patients.

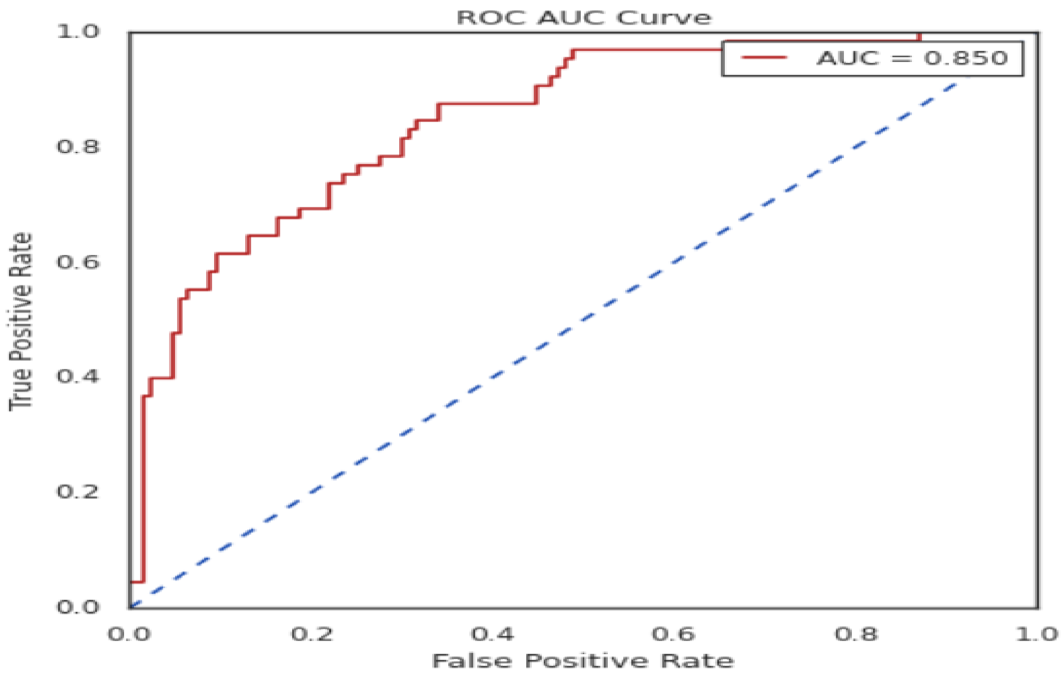


Figure 4.2.1.2: ROC AUC Curve for Logistic Regression

And Precision – Recall curve (Figure 4.2.1.3) and Precision-Recall score is 0.75 for Logistic Regression algorithm. In simplest terms, this means there is 75% chance that the model will be able to correctly predict all the healthy patient.

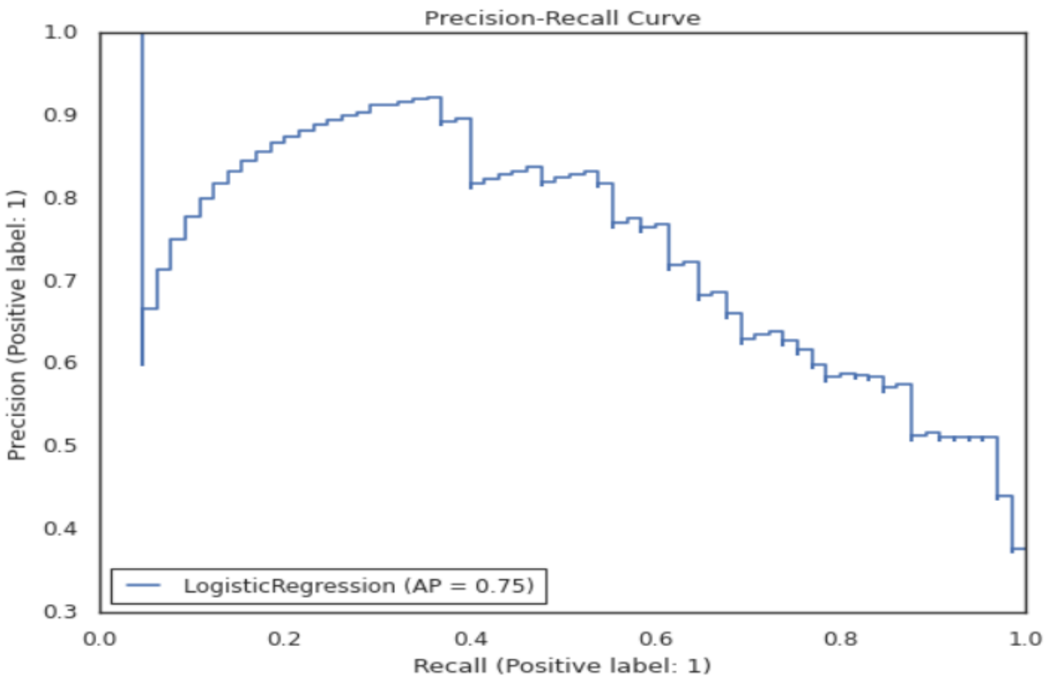


Figure 4.2.1.3: Precision – Recall Curve for Logistic Regression

4.2.2 Gaussian Naive Bayes Algorithm Analysis

After fitting Gaussian Naive Bayes model, my Accuracy Score is 0.74. In simplest terms, this means there is 74% chance that the model will be able to correctly predict all healthy and diabetic patients from our dataset.

I found the confusion matrix for both percentage split and k fold cross validation method. For this model the Performance parameters (table 4) is below-

Table 4: Performance Parameters Report of Gaussian Naive Bayes

Architecture	Precision	Recall	F1-score
Healthy (0)	0.80	0.82	0.81
Diabetes (1)	0.64	0.60	0.62
Macro Average	0.72	0.71	0.71

Now the confusion matrix (Figure 4.2.2.1) for this algorithm is below-

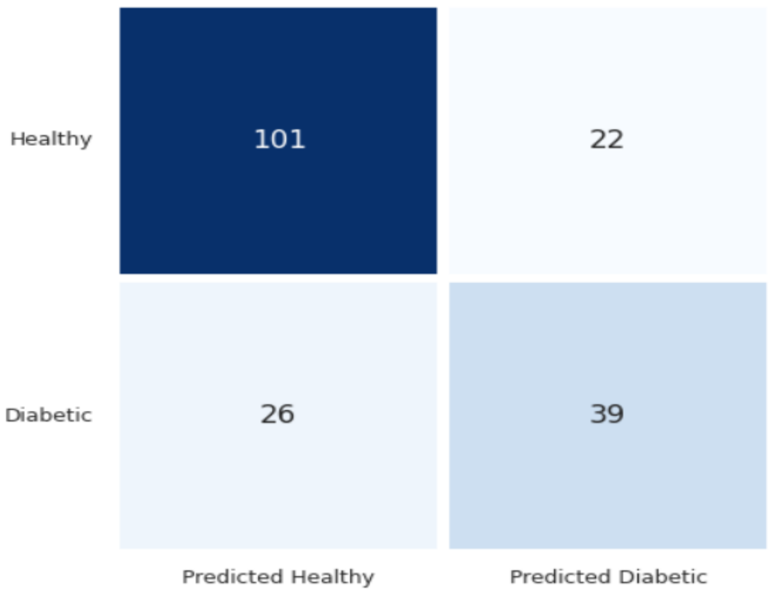


Figure 4.2.2.1: Confusion Matrix for Gaussian Naive Bayes

Here the ROC AUC curve (Figure 4.2.2.2) and ROC AUC score is 0.83 for Gaussian Naive Bayes algorithm. In simplest terms, this means there is 83% chance that the model will be able to separate the diabetes and healthy patients.

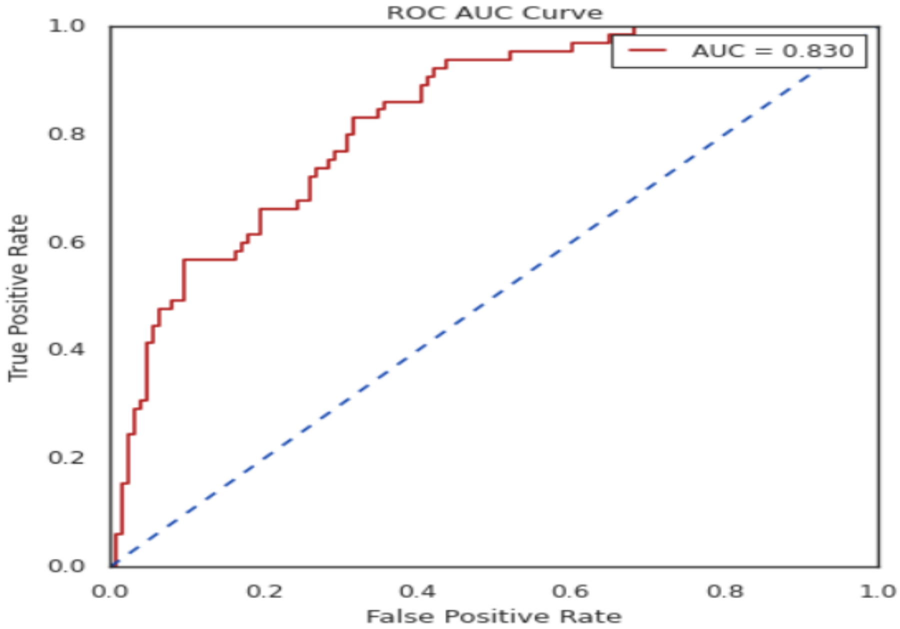


Figure 4.2.2.2: ROC AUC Curve for Gaussian Naive Bayes

And Precision – Recall curve (Figure 4.2.2.3) and Precision-Recall score is 0.69 for Gaussian Naive Bayes algorithm. In simplest terms, this means there is 69% chance that the model will be able to correctly predict all the healthy patient.

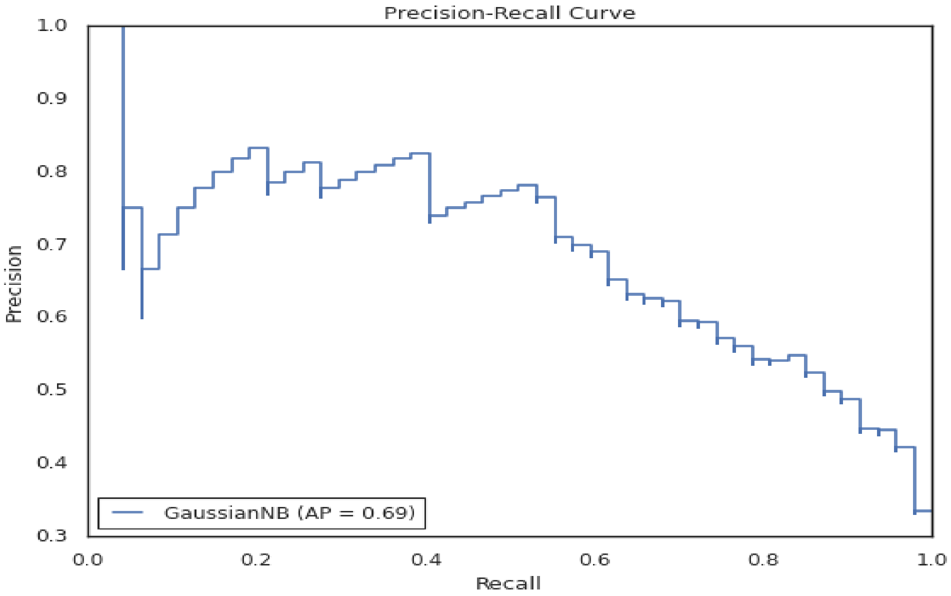


Figure 4.2.2.3: Precision – Recall Curve for Gaussian Naive Bayes

4.2.3 Random Forest Algorithm Analysis

After fitting Random Forest model, my Accuracy Score is 0.862. In simplest terms, this means there is 86% chance that the model will be able to correctly predict all healthy and diabetic patients from our dataset.

I found the confusion matrix for both percentage split and k fold cross validation method. For this model the Performance parameters (table 5) is below

Table 5: Performance Parameters Report of Random Forest

Architecture	Precision	Recall	F1-score
Healthy (0)	0.87	0.93	0.90
Diabetes (1)	0.84	0.74	0.79
Macro Average	0.86	0.83	0.84

Now the confusion matrix (Figure 4.2.3.1) for this algorithm is below-

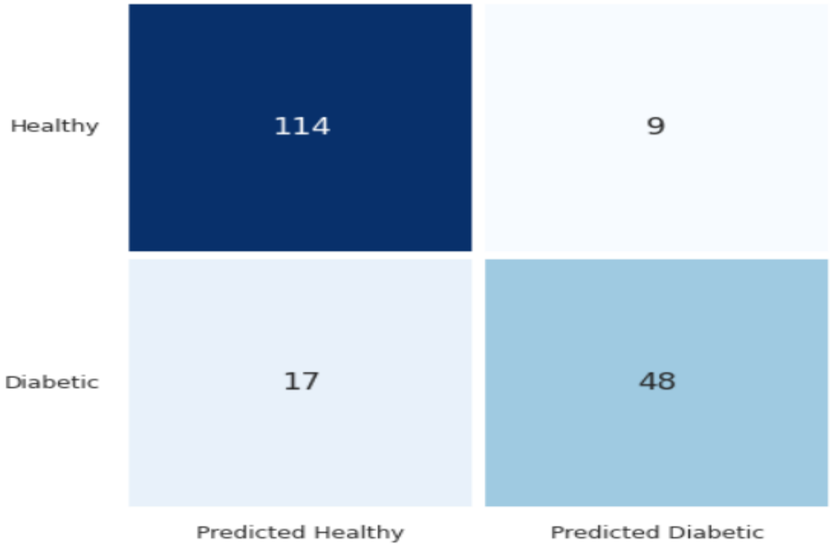


Figure 4.2.3.1: Confusion Matrix for Random Forest

Here the ROC AUC curve (Figure 4.2.3.2) and ROC AUC score is 0.94 for Random Forest Classifier algorithm. In simplest terms, this means there is 94% chance that the model will be able to separate the diabetes and healthy patients.

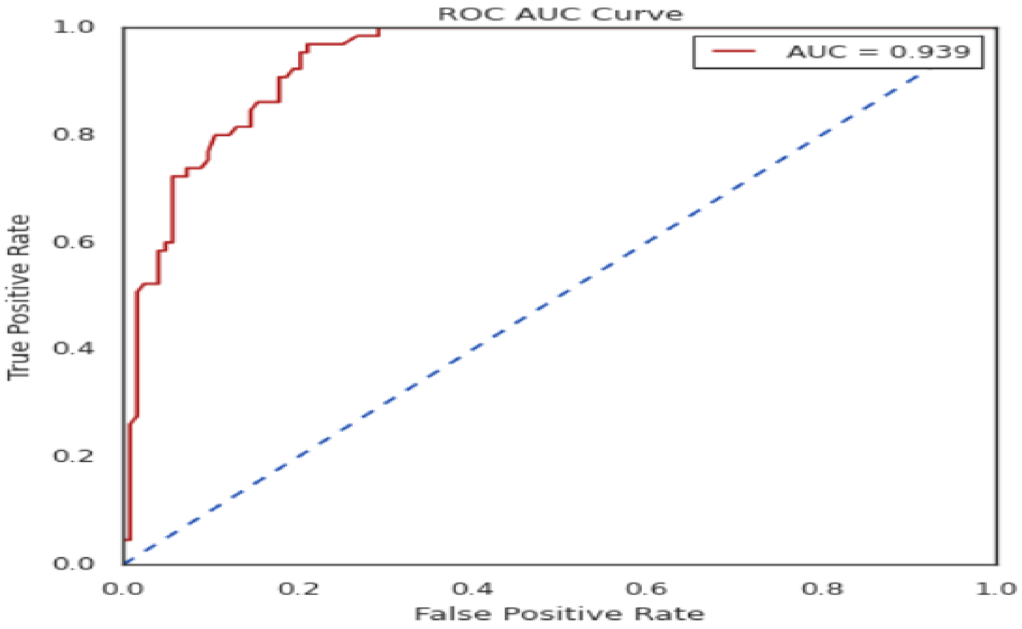


Figure 4.2.3.2: ROC AUC Curve for Random Forest

And Precision – Recall curve (Figure 4.2.3.3) and Precision-Recall score is 0.86 for Random Forest Classifier algorithm. In simplest terms, this means there is 86% chance that the model will be able to correctly predict all the healthy patient.

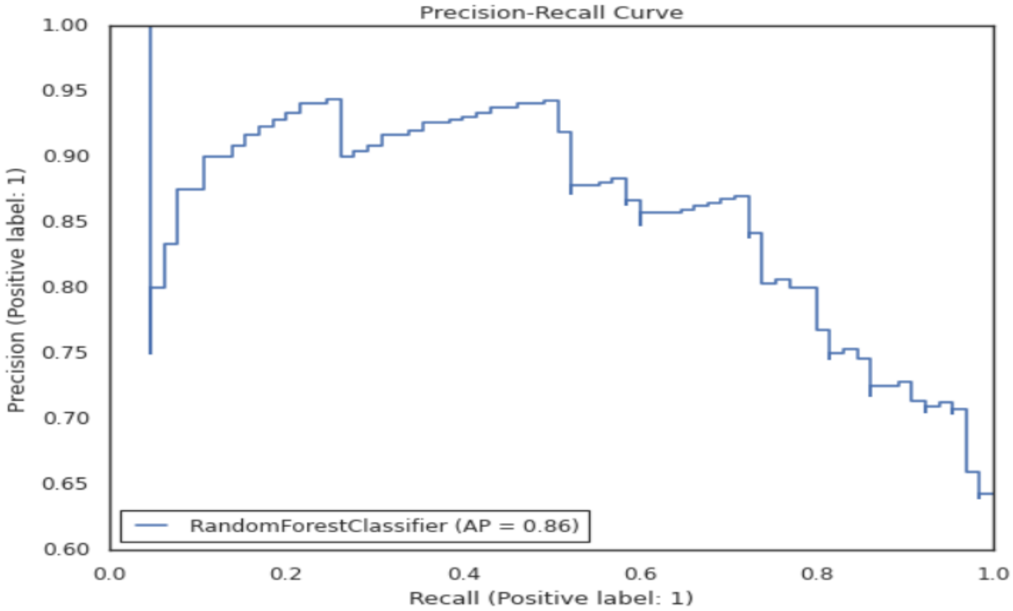


Figure 4.2.3.3: Precision – Recall Curve for Random Forest

4.3 Summary

In this project is designed to diagnose diabetes patients. Here Random Forest show the highest performance. But there are no such things that RF always gives the best performance. Again Gaussian Naive Bayes, Logistic Regression Classifier also can give best result in different perspective. The outcome of any work depends on how the work is done step-by-step. From data collection to data preparation, data processing, applying algorithms to dataset, after doing all these things properly I can see the best output or results. The expected model is found by comparing difference performance parameter in machine learning algorithm such as confusion matrix, precision, recall and F1-score for percentage split.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Work Flow of the Study

In this thesis I tried to find out the best model which can be able to predict diabetes more accurately at a very early stage. So that I applied different machine learning classifier. After applying and comparing this classifier I found that Random Forest performed very well than all other classifier. The work flow of this thesis is described below step by step.

Step 1: Data collection from online and offline sources.
Step 2: Data preprocessing.
Step 3: Divided training and testing data.
Step 4: Model trained by different algorithm.
Step 5: Predicting Test result.
Step 6: Applied different statistical matrix for measure model performance.
Step 7: Model compared.
Step 8: Find the best prediction model.

It’s very important to predict diabetes at an early stage. This model will help us to predict diabetes at a very early stage. If diabetes predicts at an early stage it can be reduced the effect of diabetes significantly.

5.2 Conclusion

In this study, I described three supervised learning-based machine learning approach. Therefore, I provide a dataset in the machine learning based model for early detection of diabetes monitoring. I divided my dataset into two parts training and testing. In percentage split method we used 80% as a training dataset and 20% for test the model for finding expected model. Next, I compared the performance of the three classifications used in the prediction of diabetes and the Confusion Matrix used to evaluate their effectiveness. The three classification models are Logistic Regression, Gaussian Naive Bayes and Random Forest Classifier. I found that Random Forest performed very well than all other classifier. By using small amount of attribute, I find this accuracy if I use more attribute may we will find better accuracy in this system.

5.3 Limitations

Every studies has its obstacles and my study is not any exception from others. It would be incredible if I will cowl the restrictions. In my thesis I used dataset that is very small and old. On this thesis I paintings with only woman dataset. This study will provide extra accurate accuracy via the use of extra recent and real-existence records which I can collect from different hospital in Bangladesh.

5.4 Future Work

In my experiments, associated with maximum of the observe work, each class set of rules become skilled. And examine a schooling set that includes each treasured and negative samples. Moreover amassing facts from a diffusion of responsibilities may be useful for the analysis and detection of chronic illnesses. Can distribute gadgets and sensors related to fitness, scientific and clinical facilities and greater accurately diagnosis and analysis results. From my studies perspective, there are distinct aspects. For this work inside the case of future studies. I've most effective investigated some popular supervised machines To study algorithms, it can choose greater algorithms to create the proper model of them. Chronic sickness diagnosis and overall performance can be in addition advanced. For better performance improved data preprocessing technique need to be used. By applying Advanced and combined algorithm we can develop this model.

REFERENCES

[1] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods.”

Procedia Comput. Sci., vol. 167, no. 2019, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[2] Deepti Sisodia, Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithms.” Procedia Computer Science, Volume 132, 2018, ISSN 1877-0509,https://doi.org/10.1016/j.procs.2018.05.122.

[3] Maniruzzaman, M., Rahman, M.J., Ahammed, B., “Classification and prediction of diabetes disease using machine learning paradigm.” *Health Inf Sci Syst* **8,** 7 (2020). https://doi.org/10.1007/s13755-019-0095-z.

[4] A. Gnana, E. Leavline, and B. Baig, “Diabetes Prediction Using Medical Data,” J. Comput. Intell. Bioinforma., vol. 10, no. January, pp. 1–8, 2017.

[5] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," *2018 24th International Conference on Automation and Computing (ICAC)*, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.

[6] Hasan Temurtas, Nejat Yumusak, Feyzullah Temurtas, “A comparative study on diabetes disease diagnosis using neural networks,” Expert Systems with Applications, Volume 36, Issue 4, 2009, Pages 8610-8615, ISSN 0957-4174,
https://doi.org/10.1016/j.eswa.2008.10.032.

[7] Md. Aminul Islam and Nusrat Jahan, “Prediction of Onset Diabetes using Machine Learning Techniques,” December 2017, International Journal of Computer Applications 180(5):7-11, DOI: 10.5120/ijca2017916020.

[8] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, no. December 2017, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006

[9] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, “Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques,” Computer Vision and Machine Intelligence in Medical Image Analysis, Advances in Intelligent Systems and Computing 992, https://doi.org/10.1007/978-981-13-8798-2_12.

[10] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” J. Diabetes Metab. Disord., vol. 19, no. 1, pp. 391–403, 2020, doi: 10.1007/s40200-020-00520-5.

[11] Henock M. Deberneh and Intaek Kim, “Prediction of Type 2 Diabetes Based on Machine Learning Algorithm,” March 2021, International Journal of Environmental Research and Public Health 18(6):3317, DOI: 10.3390/ijerph18063317.

[12] Mishra, V., Samuel, C., Sharma, S.K., “Use of machine learning to predict the onset of diabetes,” Int. J. Recent Adv. Mech. Eng. (IJMECH) 4(2) (2015), DOI: 10.14810/ijmech.2015.4202.

[13] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[14] Sen, S.K. and Dash, S., Application of Meta Learning Algorithms for the Prediction of Diabetes Disease. International Journal of Advance Research in Computer Science and Management Studies, 2014.