Diabetes prediction using Machine Learning

Submitted By

M S Sinchan

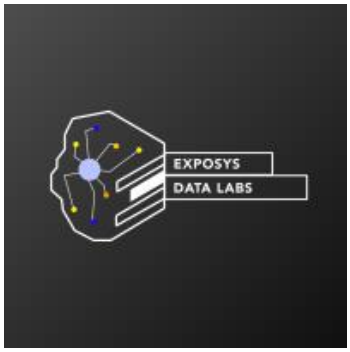
(4NM18EC085)

Department of Electronics and Communication Engineering

NMAM Institute of Technology, Nitte

Udupi, Karnataka 574110

June 2021



Exposys Data Labs

P.M R. Residency, Ground Floor, No-5/3 Sy. No.10/6-1
Doddaballapur Main Road, Singanayakanahalli, Yelahanka,
Bengaluru, Karnataka 560064

www.exposysdata.com

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Diabetes is a disease whereby blood sugar (glucose) is not metabolized in the body. In this work different methods of machine learning techniques are used to predict diabetes using dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset has eight parameters. Standard Machine learning procedures are used, the data is collected, visualized and prepared by cleaning. Two methods are used for cleaning,
Method 1- Column deletion and Method 2- Replacing by mean.
Following algorithms are applied to both methods
1) k-nearest neighbor algorithm, 2) decision tree, 3) logistic regression, 4) support-vector machine, 5) naïve bayes and 6) artificial neural network.
Confusion matrix for each algorithm is displayed and accuracy is calculated. By the observation made, all the algorithms resulted in more than 73% of accuracy.

1. INTRODUCTION

1) Diabetes

Diabetes is a disease whereby blood sugar (glucose) is not metabolized in the body. This increases the glucose in the blood to alarmingly high levels. This is known by the name hyperglycemia. In this condition, body is unable to produce sufficient insulin. The other possibility is that body cannot respond to the produced insulin. Diabetes is incurable; it has to be controlled. A diabetic person can develop severe complications like nerve damage, heart attack, kidney failure and stroke. According to statistics in 2017, an estimated 8.8% of global population has diabetes. This is likely to increase to 9.9% by year 2045.
There are a few different types of diabetes:

- Type 1 diabetes is an autoimmune disease. The immune system attacks and destroys cells in the pancreas, where insulin is made. It's unclear what causes this attack. About 10 percent of people with diabetes have this type.
- Type 2 diabetes occurs when your body becomes resistant to insulin, and sugar builds up in your blood.
- Prediabetes occurs when your blood sugar is higher than normal, but it's not high enough for a diagnosis of type 2 diabetes.
- Gestational diabetes is high blood sugar during pregnancy. Insulin-blocking hormones produced by the placenta cause this type of diabetes.

The general symptoms of diabetes include:

- increased hunger
- increased thirst
- weight loss
- frequent urination
- blurry vision
- extreme fatigue
- sores that don't heal

Symptoms in men
In addition to the general symptoms of diabetes, men with diabetes may have a decreased sex drive, erectile dysfunction (ED), and poor muscle strength.

Symptoms in women
Women with diabetes can also have symptoms such as urinary tract infections, yeast infections, and dry, itchy skin.

Diagnosis

Anyone who has symptoms of diabetes or is at risk for the disease should be tested. Women are routinely tested for gestational diabetes during their second or third trimesters of pregnancy.

Doctors use these blood tests to diagnose prediabetes and diabetes:

- The fasting plasma glucose (FPG) test measures blood sugar after fasted for 8 hours.
- The A1C test provides a snapshot of blood sugar levels over the previous 3 months.

To diagnose gestational diabetes, doctor will test blood sugar levels between the 24th and 28th weeks of pregnancy.

- During the glucose challenge test, blood sugar is checked an hour after the person drinks a sugary liquid.
- During the 3 hours glucose tolerance test, blood sugar is checked after fast overnight and then after drinking a sugary liquid.

Accurate screening and diagnosis of diabetes require more effective features and have a high demand on the judgment which can be closer to the nature of the disease. Some studies found that if we consider metabolic changes in diabetes from the perspective of body metabolism, doctors can better make a diagnosis of the type of diabetes and help patients with the more appropriate diabetic treatment. Metabolomics is a new discipline that has been developed in recent years to analyse all the low molecular weight metabolites of a certain organism or cell qualitatively and quantitatively. Through the change of endogenous metabolites and intermediates in diabetes and the evolution of coping rules, the metabolic status of the body can be further understood. On the basis of the study of early screening and diagnostic criteria for diabetes, diagnostic standards are increased from the initial clinical symptoms and signs to FPG, OGTT, HbA1c and other physiological parameters. Simultaneously clinical and demographic signs are also included in the diagnostic reference, such as sex, age, race/ethnicity, haemoglobin disease/anaemia, body mass index (BMI), cardiovascular disease, family history/ Genetic, medication records, etc. However, there is still no way to find out the pathogenesis of diabetes from the field of biology. It is urgent to clarify the pathology and diagnostic criteria of diabetes, it has a great significance in delaying the occurrence and development of diabetes, choosing drugs, reducing the incidence of diabetic complications and extending life expectancy. With the continuous development of artificial intelligence and data mining technology, researchers begin to consider using machine learning techniques to search for the characteristics of diabetes. Machine learning techniques can find implied pathogenic factors in virtue of analysing and using diabetic data, with a high stability and accuracy in diabetic diagnosis. Therefore, machine learning techniques which can find out the reasonable threshold of risky factors and physiological parameters provide new ideas for screening and diagnosis of diabetes. Earlier diagnosis of Diabetes increases the chances of preventing it from becoming severe.

2) Machine Learning

Machine Learning is a concept which allows the machine to learn from examples and experience, and that too without being explicitly programmed. So instead of you writing the code, what you do is you feed data to the generic algorithm, and the algorithm/ machine builds the logic based on the given data.

Types of Machine Learning:

Supervised learning

In supervised learning, we use known or labeled data for the training data. Since the data is known, the learning is, therefore, supervised, i.e., directed into successful execution. The input data goes through the Machine Learning algorithm and is used to train the model. Once the model is trained based on the known data, you can use unknown data into the model and get a new response.

Types of supervised learning algorithms:

- Polynomial regression
- Random forest
- Linear regression
- Logistic regression
- Decision trees
- K-nearest neighbors
- Naive Bayes

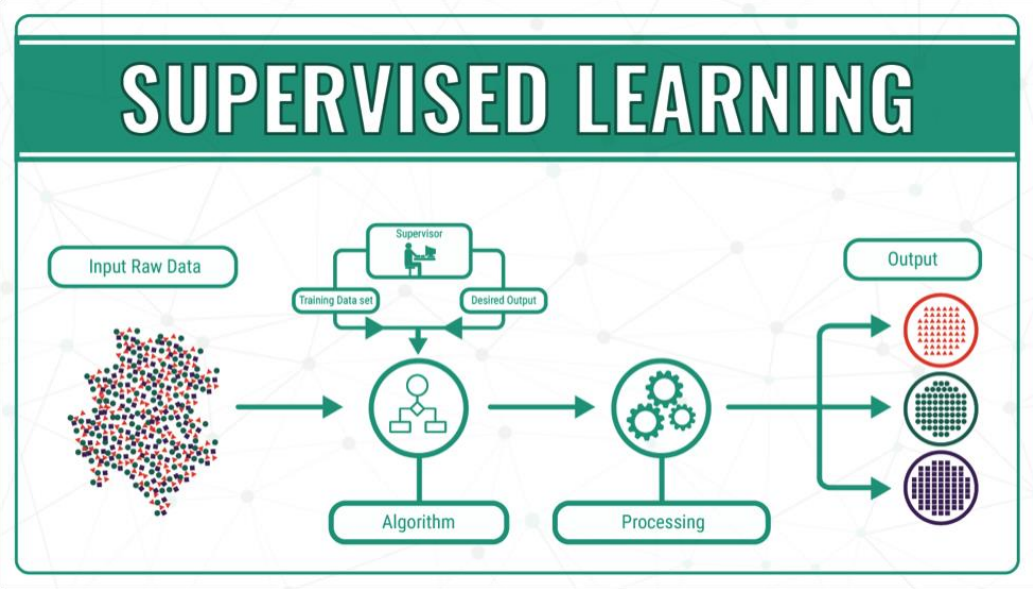


Figure 1-1: Supervised learning

Unsupervised learning

In unsupervised learning, the training data is unknown and unlabeled - meaning that no one has looked at the data before. Without the aspect of known data, the input cannot be guided to the algorithm, which is where the unsupervised term originates from. This data is fed to the Machine Learning algorithm and is used to train the model. The trained model tries to search for a pattern and give the desired response. In this case, it is often like the algorithm is trying to break code like the Enigma machine but without the human mind directly involved but rather a machine.

Types of Unsupervised learning algorithms:

- Partial least squares
- Fuzzy means
- Singular value decomposition
- K-means clustering
- Apriori
- Hierarchical clustering
- Principal component analysis

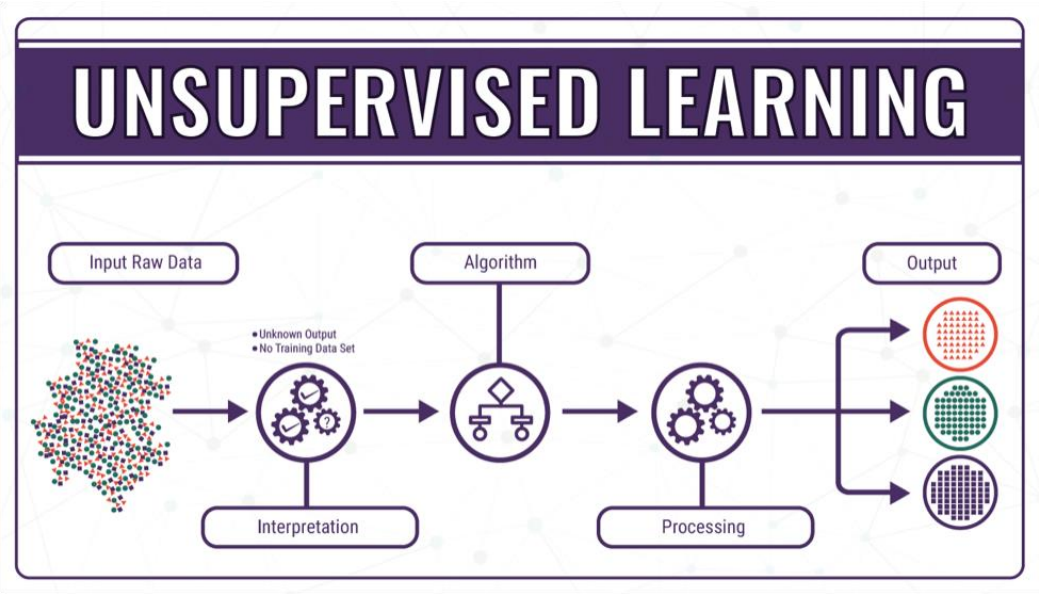


Figure 1-2: Unsupervised learning

Reinforcement learning

In Reinforcement learning, the algorithm discovers data through a process of trial and error and then decides what action results in higher rewards. Three major components make up reinforcement learning: the agent, the environment, and the actions. The agent is the learner or decision-maker, the environment includes everything that the agent interacts with, and the actions are what the agent does.

Reinforcement learning occurs when the agent chooses actions that maximize the expected reward over a given time. This is easiest to achieve when the agent is working within a sound policy framework.

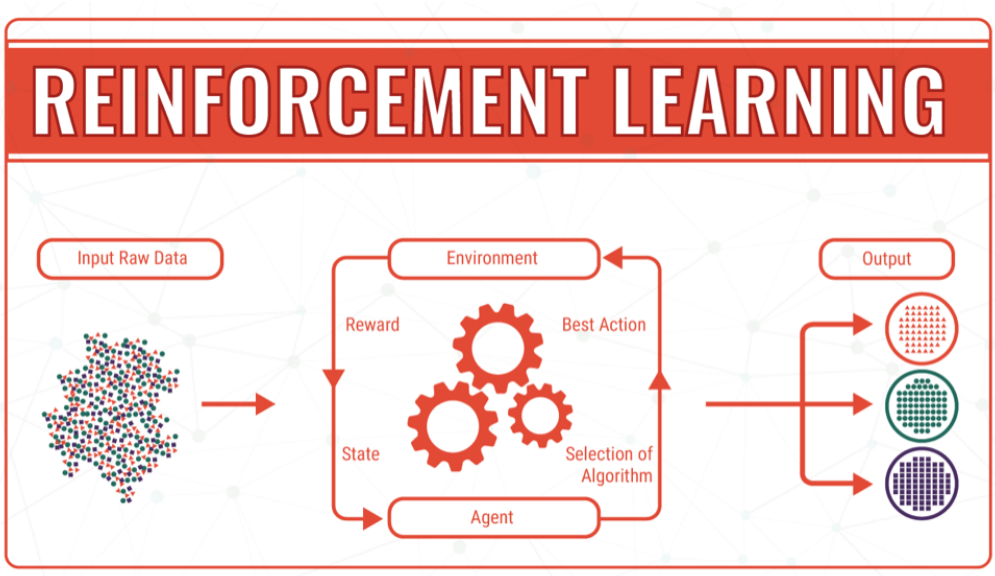


Figure 1-3: Reinforcement learning

We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyse data and deliver results faster, with the use of machine learning technology doctors can make good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare services. Healthcare is the most prime example of how machine learning is use in medical field.

2. METHODOLOGY

1) Data Collection

Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

- The quantity & quality of our data dictate how accurate our model is
- The outcome of this step is generally a representation of
- We are using pre-collected data, from Kaggle.

2) Data Preparation

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

3) Choosing Model

- Different algorithms are for different tasks; choose the right one

4) Training Model

- The goal of training is to answer a question or make a prediction correctly as often as possible
- Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)
- Each iteration of process is a training step

5) Evaluate the Model

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data

- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

6) Parameter Tuning

- This step refers to hyperparameter tuning, which is an "artform" as opposed to a science
- Tune model parameters for improved performance
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

7) Make Predictions

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

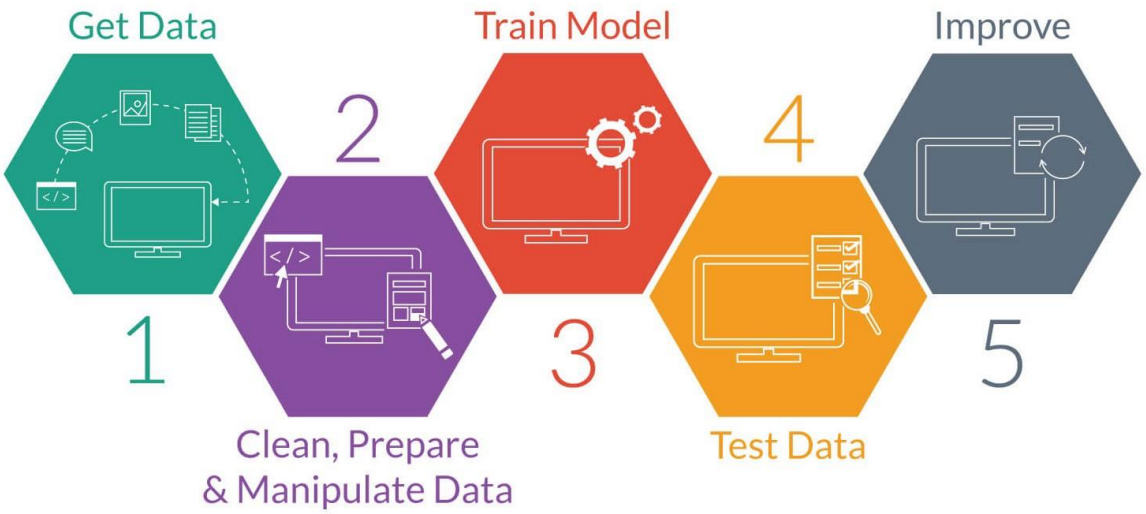


Figure 2-1: Machine learning flow

3. IMPLIMENTATION

1) Dataset

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable.

2) Data cleaning

Ways to handle missing data:

There are mainly two ways to handle missing data, which are:

By deleting the particular row:
The first way is used to commonly deal with null values. In this way, delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.

Not considered this as dataset size reduced by fifty percent.

By calculating the mean:
In this way, calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

By deleting negatively influencing parameter:
Parameter which decreased the accuracy of matrix is not considered.

3) Algorithms

K-Nearest Neighbors algorithm (KNN)
KNN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

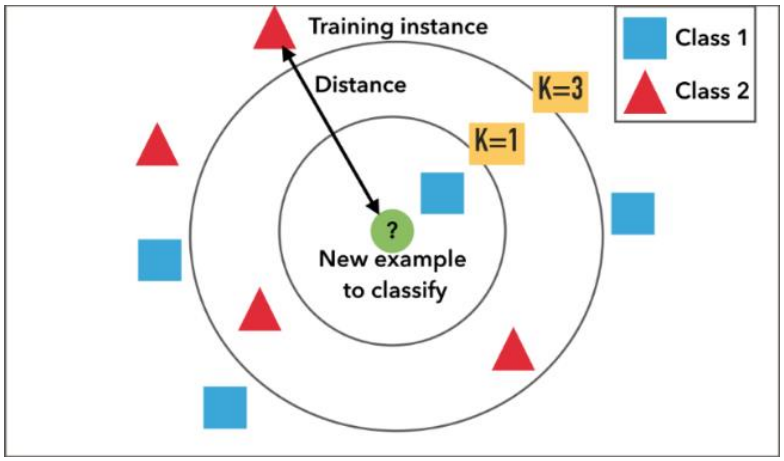


Figure 3-1: K-Nearest Neighbor

Decision tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

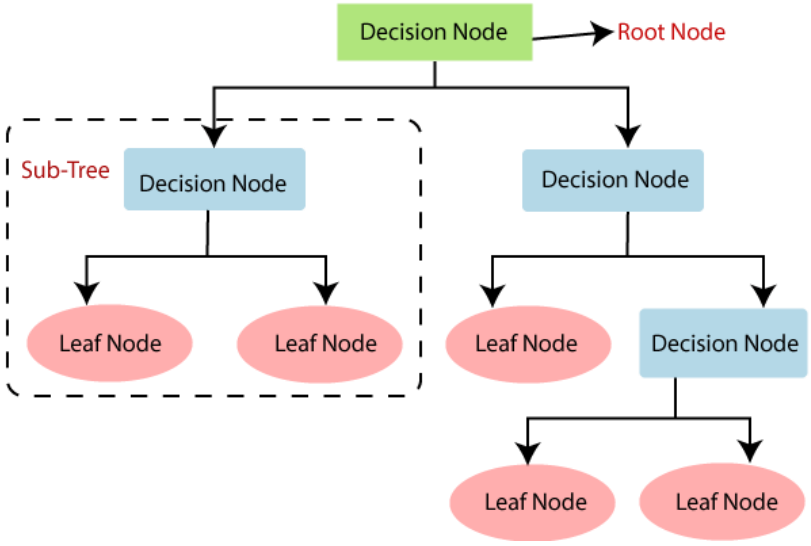


Figure 3-2: Decision tree

Logistic regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

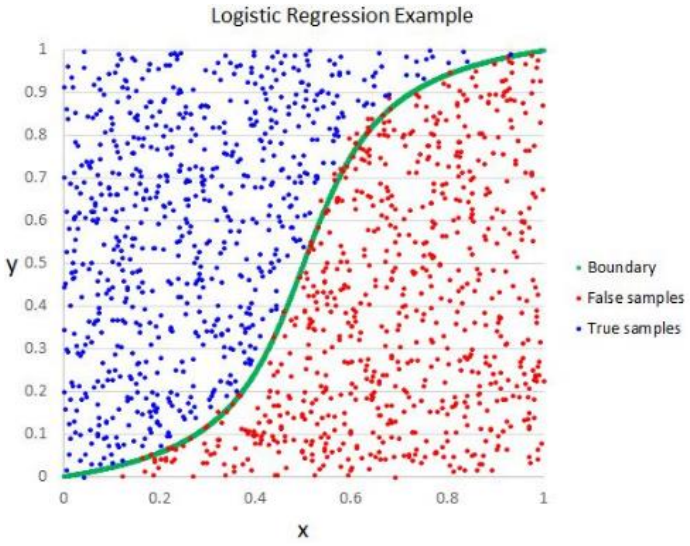


Figure 3-3: Logistic regression

Support-vector machine

In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

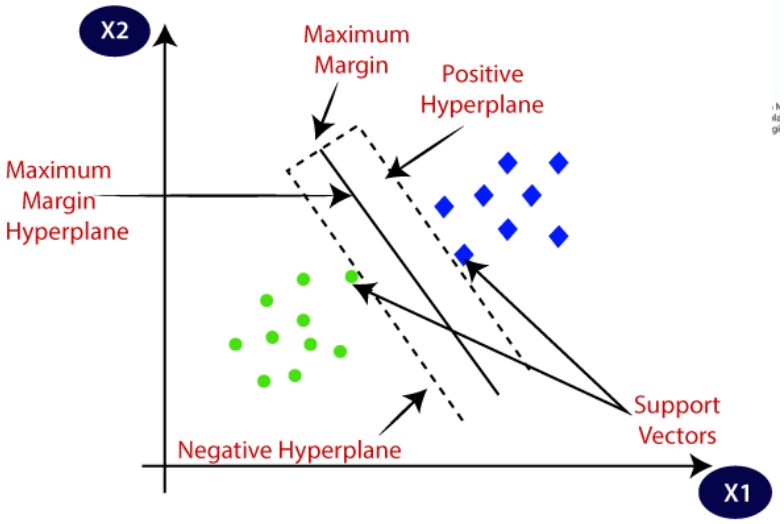


Figure 3-4: Support vector machine

Naïve Bayes

Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications. It performs well in multi-class predictions as compared to the other Algorithms. It is the most popular choice for text classification problems.

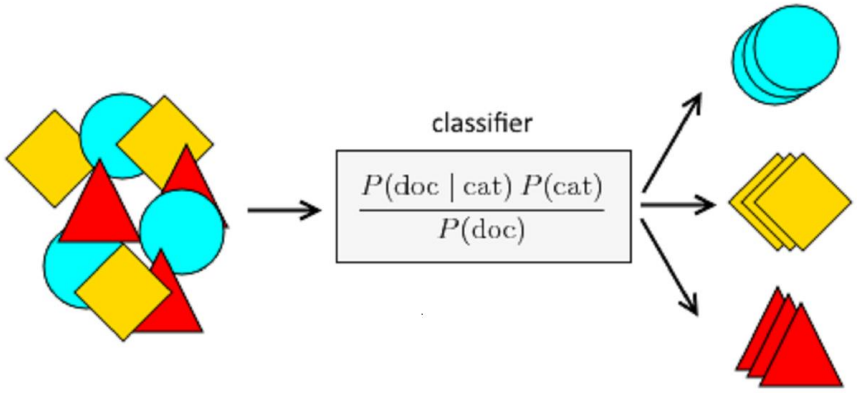


Figure 3-5: Naïve bayes

Artificial neural networks

Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

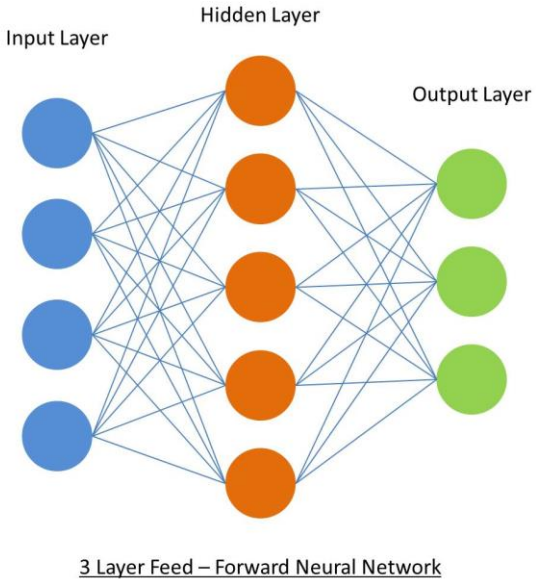


Figure 3-6: Artificial neural networks

Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

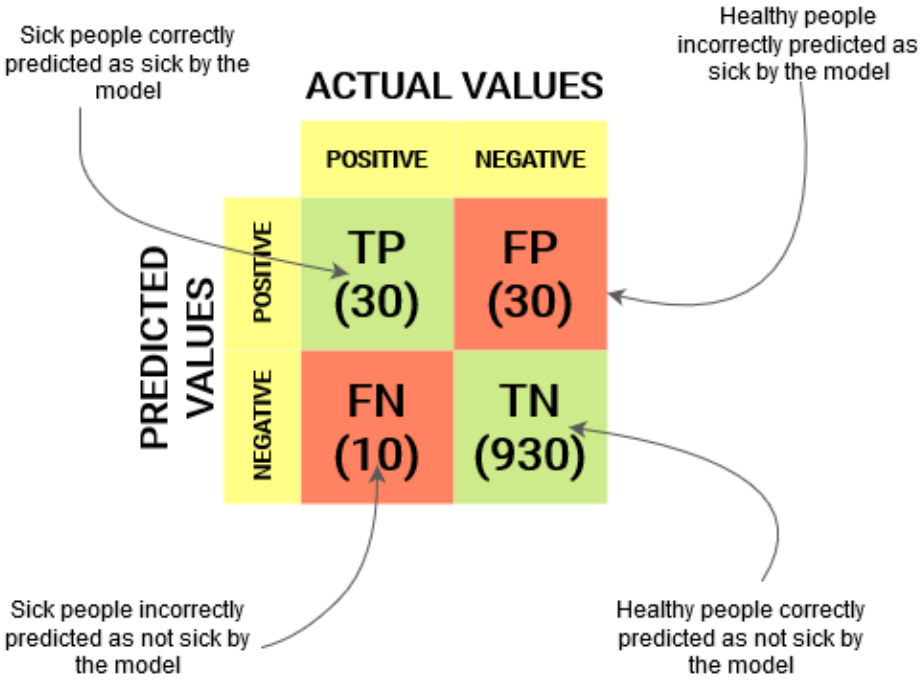


Figure 3-7: Confusion matrix

4. RESULTS

This chapter gives details of the results obtained in this project

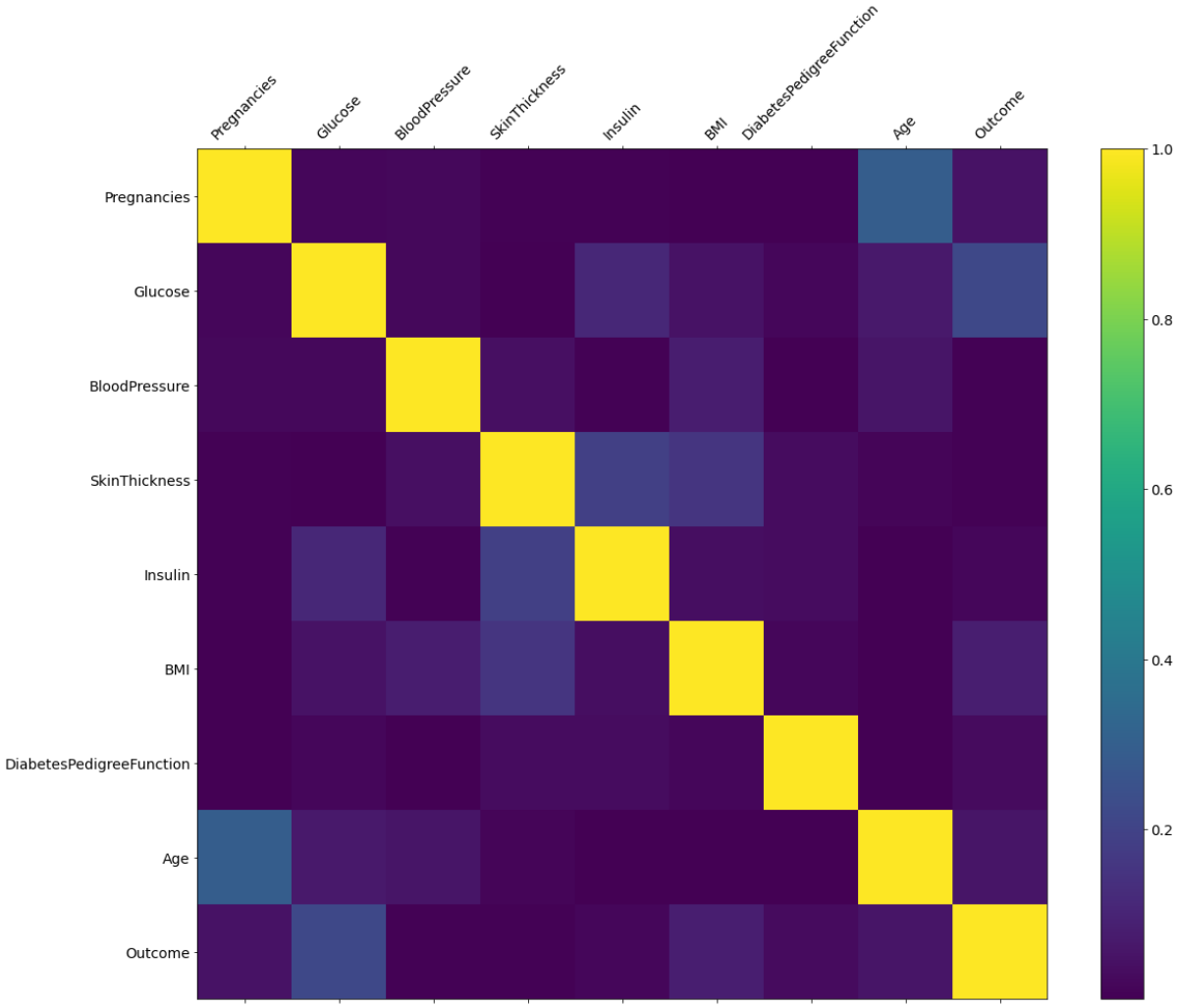


Figure 4-1: Correlation matrix

Figure 4-1 shows the visualization of correlation matrix of features of the diabetic dataset

Correlation is an indication about the changes between two variables. We can plot correlation matrix to show which variable is having a high or low correlation in respect to another variable.

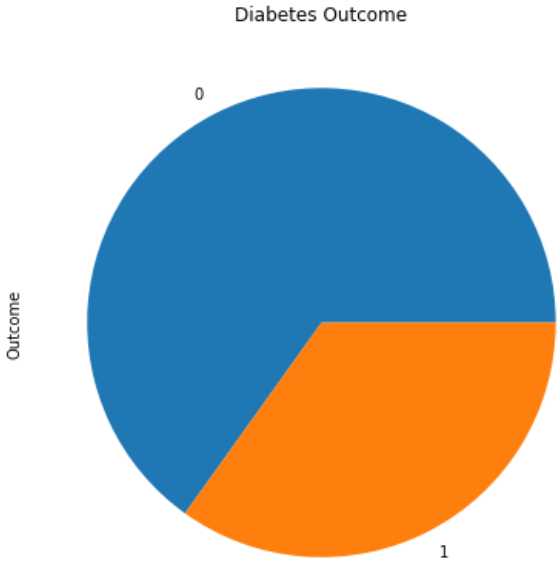


Figure 4-2: Percentage of positive and negative outcomes

Figure 4-2 shows the pie chart representing percentage of positive (represented by 1) and negative (represented by 0) outcomes

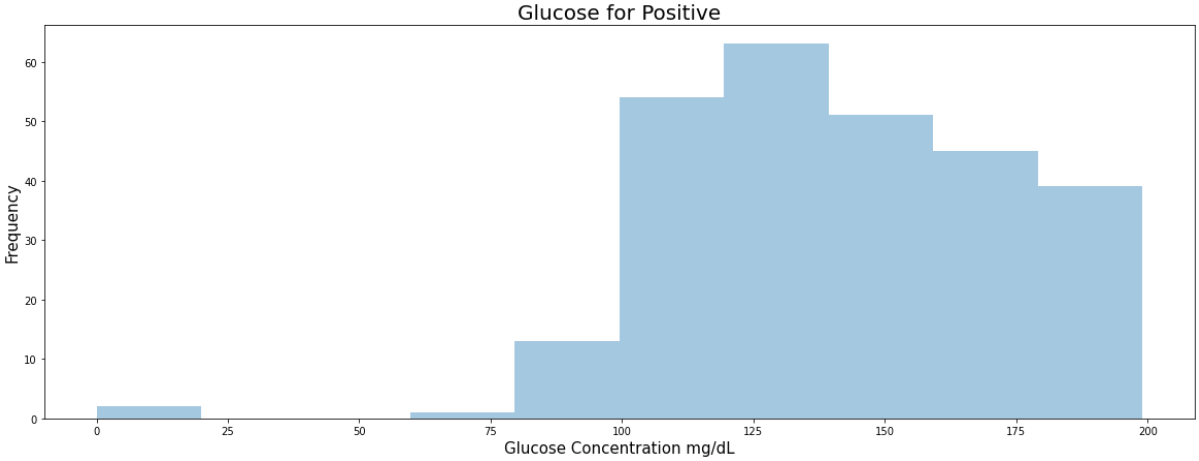


Figure 4-3: Glucose positive histogram

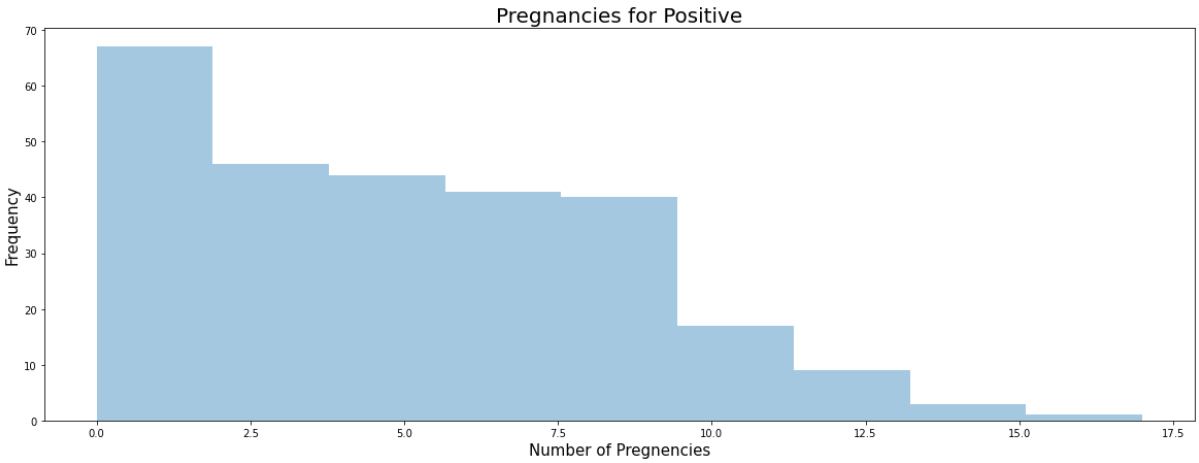


Figure 4-4: Pregnancies for positive histogram

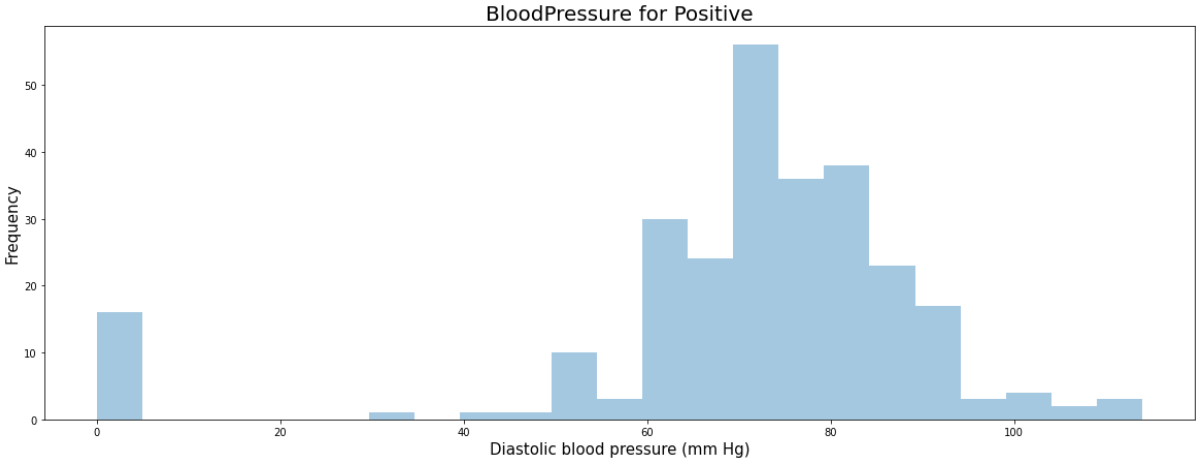


Figure 4-5: Blood Pressure positive histogram

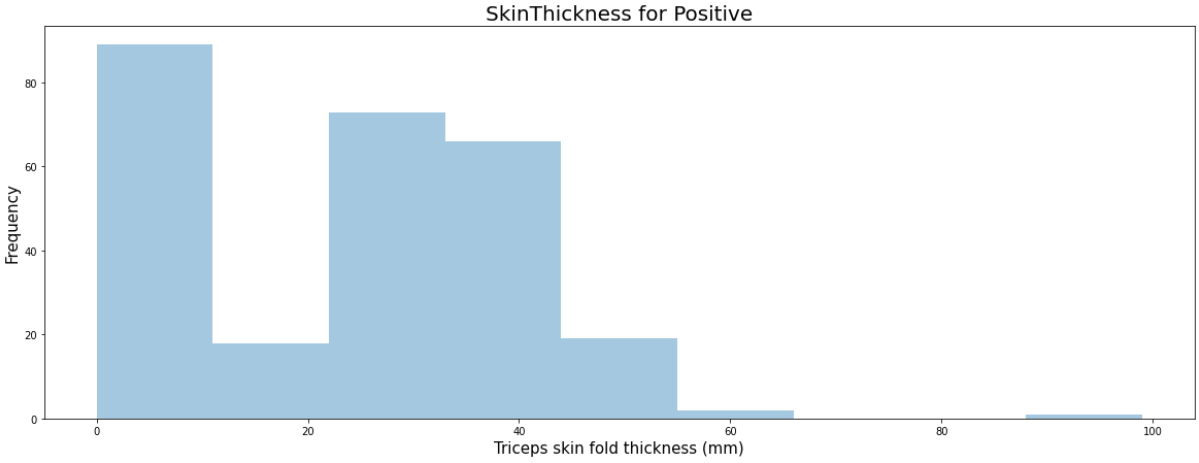


Figure 4-6: Skin Thickness positive histogram

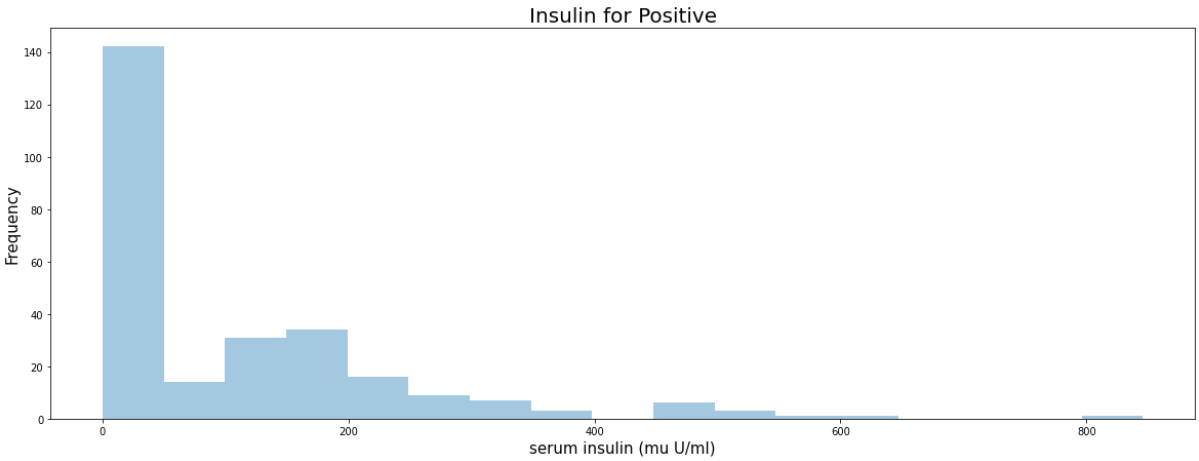


Figure 4-7: Insulin for positive histogram

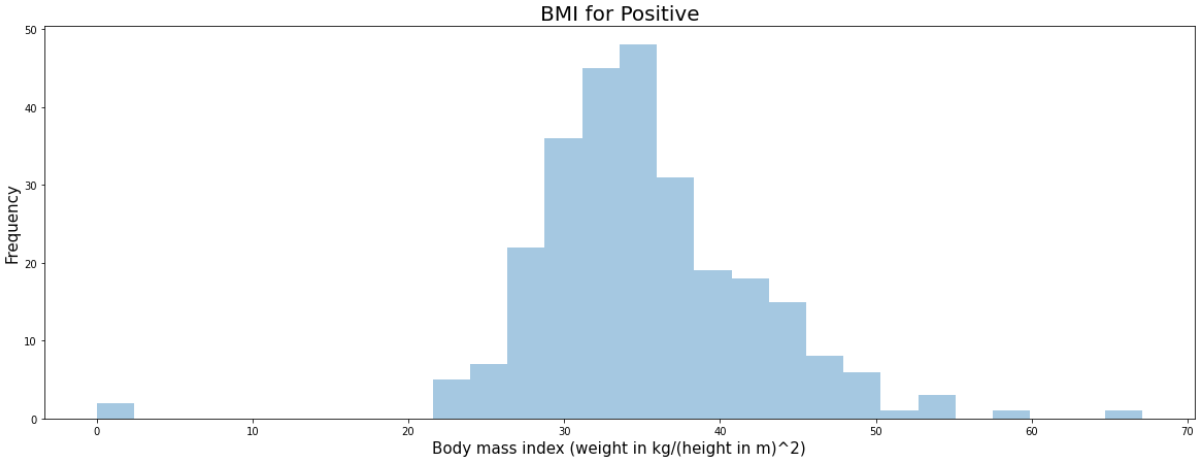


Figure 4-8: BMI for positive histogram

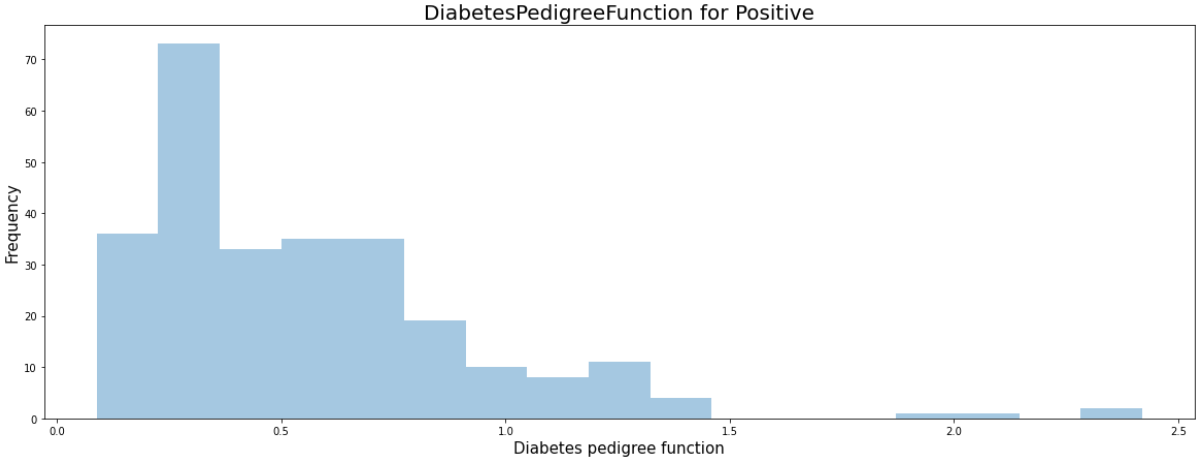


Figure 4-9: Diabetes pedigree function for positive histogram

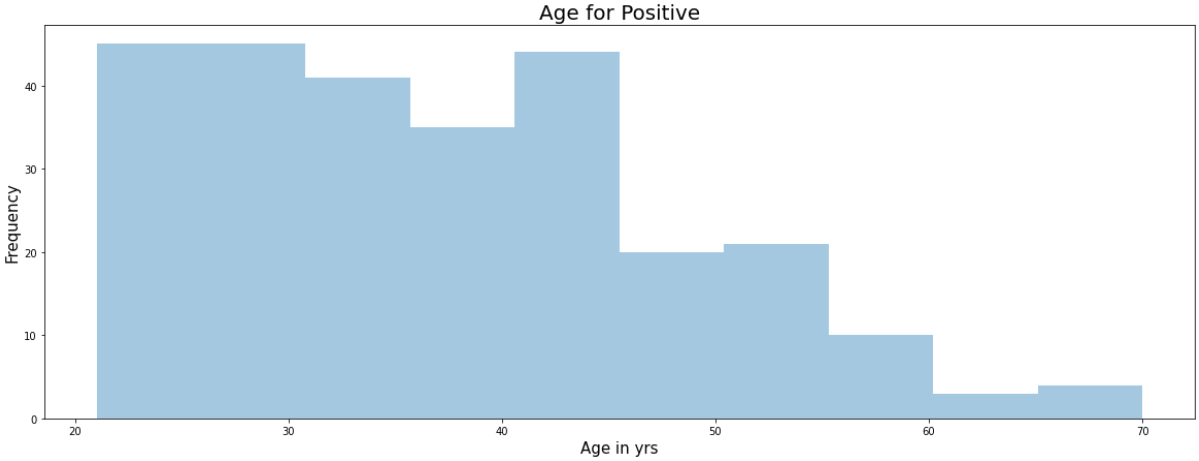


Figure 4-10: Age for positive histogram

Figure 4-3 to Figure 4-10 are the histograms of features contributing to diabetic condition.

Different algorithms were applied to the data set and their confusion matrix and accuracies are noted in Table 4-1 and Table 4-2 respectively.

Here Method 1: Column deletion method

Method 2: Replacing by mean method

Table 4-1: Confusion matrix for different algorithms

	KNN	Decision tree	Logistic regression	SVM	naive bayes
method 1	$\begin{bmatrix} 107 & 0 \\ 1 & 46 \end{bmatrix}$	$\begin{bmatrix} 107 & 0 \\ 0 & 47 \end{bmatrix}$	$\begin{bmatrix} 107 & 0 \\ 0 & 47 \end{bmatrix}$	$\begin{bmatrix} 107 & 0 \\ 0 & 47 \end{bmatrix}$	$\begin{bmatrix} 107 & 0 \\ 0 & 47 \end{bmatrix}$
method 2	$\begin{bmatrix} 96 & 11 \\ 17 & 30 \end{bmatrix}$	$\begin{bmatrix} 85 & 22 \\ 19 & 28 \end{bmatrix}$	$\begin{bmatrix} 94 & 13 \\ 18 & 29 \end{bmatrix}$	$\begin{bmatrix} 96 & 11 \\ 18 & 29 \end{bmatrix}$	$\begin{bmatrix} 87 & 16 \\ 20 & 31 \end{bmatrix}$

Table 4-2: Accuracies for different algorithms

	KNN	Decision tree	Logistic regression	SVM	naive bayes	ANN
method 1	0.9935	1.0	1.0	1.0	1.0	0.9610
method 2	0.8182	0.7338	0.7987	0.8117	0.7662	0.8117

5. CONCLUSION

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and predicts the probability of getting disease.

By the observation made, all the algorithms resulted in more than 73% of accuracy.

This model can be used with different datasets to predict different types of diseases.

REFERENCES

[1] https://www.sciencedirect.com/science/article/pii/S2405959518304624
[2] https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning
[3] https://www.healthline.com/health/diabetes
[4] https://technative.io/why-unsupervised-machine-learning-is-the-future-of-cybersecurity/
[5] https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html
[6] https://centricconsulting.com/blog/machine-learning-a-quick-introduction-and-five-core-steps/
[7] https://www.irjet.net/archives/V6/i12/IRJET-V6I12122.pdf
[8] https://medium.com/analytics-vidhya/an-end-to-end-data-science-project-on-diabetes-9a70c8368d2a
[9] https://www.kaggle.com/uciml/pima-indians-diabetes-database
[10]https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7
[11]https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_support_vector_machine.htm
[12] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
[13]https://insightimi.wordpress.com/2020/04/04/naive-bayes-classifier-from-scratch-with-hands-on-examples-in-r/
[14]https://www.the-diy-life.com/running-an-artificial-neural-network-on-an-arduino-uno/
[15] https://www.javatpoint.com/data-preprocessing-machine-learning
[16]https://www.datasciencecentral.com/profiles/blogs/why-logistic-regression-should-be-the-last-thing-you-learn-when-b
[17] https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/