**PREDICTION OF DIABETES**

# A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**Computer Science & Engineering**

# by

**K.Dhanush (2203A51680)**

**T.Ajay kumar (2203A51689)**

**B.Jeevan (2203A51722)**

# Under the guidance of

**Dr. E.L.N. Kiran Kumar**

**Associate Professor, CS & AI**



**Department** **of** **Computer Science & Engineering**

# 



**Department** **of** **Computer Science & Engineering**

# CERTIFICATE

This is to certify that the **AIML - Course Project** Report entitled **“PREDICTION OF DIABETES**" is a record of bonafied work carried out by the students K.Dhanush, bearing Roll No (2203A51680) , T.Ajay kumar , bearing Roll No (2203A517)89 , B.Jeevan , bearing Roll No (2203A51722) , during the academic year 2023-2024 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science & Engineering** by the SR University, Warangal

**DR.E.L.N. Kiran Kumar Dr.M.Sheshikala**

Assoc. Prof. CS & AI Assoc. Prof. & HOD(CSE),

S R Engineering College, S R Engineering College,

Ananthasagar, Warangal. Ananthasagar, Warangal.

# ACKNOWLEDGEMENT

We express our thanks to course coordinator **Mr. Dr. E.L.N. KiranKumar, Assoc. Prof. CS and AI**, for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Sheshikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence ,Dr C.V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

# ABSTRACT

An emerging field that has received a lot of interest recently in diabetes prediction using artificial intelligence and machine learning. This research uses artificial intelligence (AI) and machine learning to predict diabetes using the input parameters of Patient number, Cholesterol, Glucose, age, HDL Chol, Chol/HDL ratio, and Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, Waist, Hip, Waist/Hip ratio The model makes use of different machine learning methods, including logistic regression, KNN Classifier, SVM, KSVM, Decision Tree, Naive Bayesian, Random Forest Classifier, Our ultimate goal is to develop a highly accurate and interpretable machine learning model that can be used to identify individuals at high risk of diabetes. The model makes precise predictions about the risk of diabetes using a variety of methods and data preprocessing. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. It could be a helpful tool for healthcare professionals to improve patient outcomes and take preventative measures.

# Table of Contents

**Chapter No. Title Page No.**

1. Introduction
   1. Overview 1
   2. Problem Statement 1-2
   3. Existing system 2
   4. Proposed system 2
   5. Objectives 2
   6. Architecture 2-3
2. Literature survey

2.1.1. Document the survey done by you 4-7

1. Data Pre-Processing
   1. Data set Description 8-11
   2. Data Cleaning 11-12

3.3Data Augmentation 12

3.4Data Visualization 12-18

1. Methodology
   1. Procedure to solve given problem 18- 29
   2. Model architecture 30
   3. Software Description 30

5.Results and Discussion 31

6.Conclusion 32

7.References 33

# LIST OF FIGURES

**Figure.No Page.No**

It is the flowchart for our project model implementation 1.6.1 3

the dataset in which it has 15 attributes and 360 rows.3.1.1 11 the heapmap correlation. 3.2.1 12 bar graph 3.4.1 13 bar graph 3.4.2 14 box plot 3.4.3 14 graphonevariablewhenanotherchanged3.4.4 15 ````````````contsstructed lineplot. result\_numeric3.4.5 16 distribution of numeric data 3.4.6 16 contsstructed lineplot. result\_numeric 3.4.7 17 result\_numeric 3.4.8 17 constructed using area plot 3.4.9 17 histogram dataset3.4.10 18 graph of logistic regression. 4.1.1 20 decision tree 4.1.2 21 k-nearestneighbourgraph 4.1.3 22 random forest classifier. 4.1.4 24 support vectore machine 4.1.5 25 naïve Bayesian 4.1.6 27 classification model accuracy. 4.1.7 28 ROC curve of all the models 4.1.8 29 barplot of acuuracy 5.1 31

# 1.INTRODUCTION

## 1.1.Overview

In both developed and underdeveloped countries, there are many recurrent infections that have no end. Diabetes is just one of these diseases. Diabetes is a metabolic disorder that raises blood sugar levels by either producing a large amount of insulin in the body or a small amount of insulin. Perhaps the deadliest disease on the planet is diabetes. It is not just a disease but also a contributor to other illnesses like heart failure, blindness, kidney problems, nerve damage, etc. Therefore , early detection of such a chronic metabolic disorder could aid medical professionals all around the world in preventing the loss of human life. Currently, with the rise of machine learning, AI, and neural systems, as well as their use in a variety of fields [1, 2], we may be able to find a solution to this problem. Scientists can discover new realities from existing health-related informational indices with the aid of machine learning (ML) algorithms and neural networks, which may aid in the monitoring and detection of illnesses.

The Pima Indians Diabetes Database is used in the current investigation. The goal of this approach is to create a machine learning model that can accurately predict the likelihood or odds of a patient having diabetes. The typical method for determining where diabetes is located is for the patient to go to a symptomatic focus. Getting precise results from the information is one of the main challenges in bioinformatics examination. Human error or a variety of laboratory tests may complicate the operation. the diagnosis of the illness. This model can predict whether a patient has diabetes or not, assisting professionals in making sure that patients who require medical attention may receive it promptly and also assisting in predicting the loss of human livers. An emerging field that has received a lot of interest recently in diabetes prediction using artificial intelligence and machine learning.

## 1.2.Problem statement

Diabetes is one of a number of metabolic abnormalities that cause this condition. Diabetic mellitus is another name for it. It has an impact on the body's organs. By foreseeing this sickness earlier, it can be managed. Long-term untreated diabetes patients run the risk of having their blood sugar levels rise. Healthcare sectors are currently producing a lot of data. With the aid of recent and historical data, statistics and machine learning algorithms are utilised to anticipate the disease. The use of machine learning techniques by doctors aids in the early diagnosis of diabetics. Based on diagnostic data, we forecast a patient's likelihood of having diabetes using logistic regression, random forest, decision tree classifiers, KNN,SVM,KSVM.

It is discussed and contrasted how well the used algorithms perform.

## 1.3.Existing system

Most existing systems predict diabetes through a blood test. Blood tests show whether your blood sugar, also known as blood sugar, is high. Although there are high-tech medical facilities for careful diagnosis and effective treatment, morality is still not well controlled. Therefore, it is very important to take early precautions in the early stages so that the symptoms and consequences can be detected early for a better diagnosis.

## 1.4.Proposed system

The proposed expert system involves developing Machine Learning techniques to detect diabetes.Not only is the patient's particular condition taken into account, but also many other parameters such as glucose level, cholesterol, HDL, BP, weight, age, hip, waist, and so on.

With proper data analysis and machine learning approaches, diabetes can be predicted early.

## 1.5.Define Objectives

The main Objective of this is to predict diabetes.This helps the people to know at which level and to take necessary treatment. Data analysis and prediction algorithms may be used to identify risk factors for diabetes and its consequences, monitor disease development, and evaluate the effectiveness of therapies, in addition to forecasting diabetes.

## 1.6.Overall architecture

Overall, the architecture for predicting diabetes using machine learning involves data collection, preprocessing, model selection, training, evaluation, deployment, and monitoring. The specific steps involved may vary depending on the data available, the machine learning algorithms used, and the intended use case. The prediction of diabetes can be accomplished using a variety of approaches, including statistical models, machine learning algorithms, and artificial intelligence techniques.

**Data collection:** Collecting data from various sources including medical records, patient history, physical examinations, and lab tests. The data collected should include features such as age, gender, family history, BMI, blood pressure, fasting glucose levels, and other relevant factors that may contribute to the development of diabetes.

**Data Preprocessing**: Once the data is collected, it needs to be preprocessed to remove any inconsistencies, missing values, or outliers. This may involve data cleaning, feature selection, feature engineering, and scaling the data.

**Model Selection:** Selecting the appropriate machine learning algorithm or model for the problem at hand. Common algorithms used for diabetes prediction include logistic regression, decision trees, random forests, support vector machines, and neural networks.

**Model Training:** Training the selected model on the preprocessed data. This involves splitting the data into training and testing sets, fitting the model to the training data, and tuning the model parameters to optimize its performance.

**Model Evaluation:** Evaluating the trained model's performance on the testing set, using metrics such as accuracy, precision, recall, and F1 score. The model's performance can also be visualized using ROC curves or confusion matrices.

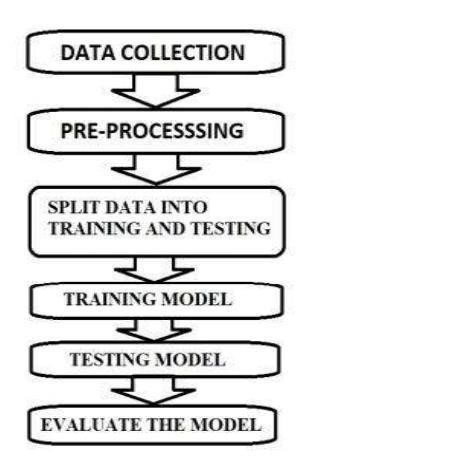


Fig:1.6.1 It is the flowchart for our project model implementation

# 2.LITERATURE SURVEY

We wanted to represent our work perfectly, that people get some benefit from our work in a proper way without any difficulty. Therefore, we decided to study this topic, and we read some researcher's work. This section deals with all of the past and present work that prediction of diabetes. We followed their work and tried to understand their demonstrated method, which was applying by them.

[1]Birjais et al experimented on PIMA Indian Diabetes (PID) data set. It has 768 instances and 8 attributes and is available in the UCI machine learning repository. They aimed to focus more on diabetes diagnosis, which, according to the World Health

Organization (WHO) in 2014, is one of the world’s fastest-growing chronic diseases. Gradient boosting, logistic regression, and naive Bayes classifiers were used to predict whether a person is diabetic or not, with gradient boosting having an accuracy of 86%, logistic regression having a 79% accuracy, and naive Bayes having a 77% accuracy.

[2]Sadhu, A. and Jadli A experimented on a diabetes data set taken from the UCI repository. There were 520 occurrences and 16 attributes in all. They attempted to concentrate their efforts on predicting diabetes at an early stage. On the validation set of the employed data set, seven classification techniques were implemented: k-NN, logistic regression, SVM, naive Bayes, decision tree, random forests, and multilayer perceptron. The random forests classifier proved to be the best model for the concerned data set, with an accuracy score of 98%, followed by logistic regression at 93%, SVM at 94%, naive Bayes at 91%, decision tree at 94%, random forests at 98%, and multilayer perceptron at 98%, according to the results of training several machine learning models.

[3]Xue et al. experimented on the diabetes data set taken from the UCI repository; there were 520 patients and 17 qualities in it. They attempted to concentrate on early detection of diabetes. They trained on the actual data of 520 diabetic patients and probable diabetic patients aged 16– 90 using supervised ML techniques such as SVM, naive Bayes classifiers, and LightGBM. The performance of the SVM is the best when comparing classification and recognition accuracy. The naive Bayes classifier is the most widely used classification algorithm, with an accuracy of 93.27%. SVM has the highest accuracy rate of 96.54%. LightGBM has an accuracy of only 88.46%. This demonstrates that SVM is the best classification algorithm for diabetes prediction. [4]Le et al .experimented on the early-stage diabetes risk prediction; the data set used in this research was taken from the UCI repository and consisted of 520 patients and 16 variables. They suggested a ML approach for predicting diabetes patients’ early onset. It was a new wrapper-based feature selection method that employed grey wolf optimizer (GWO) and adaptive particle swarm optimization (APSO) to optimize the multilayer perceptron (MLP) and reduce the number of needed input attributes. They also compared the results obtained with this method to those obtained via a variety of traditional machine learning algorithms, including SVM, DT, k-NN, naive Bayes classifier (NBC), random forest classifier (RFC), and logistic regression (LR). LR achieved a 95% accuracy rate. k-NN had a 96% accuracy rate, SVM a 95% accuracy rate, NBC a 93% accuracy rate, DT a 95% accuracy rate, and RFC had a 96% accuracy rate. The suggested methods’ computational findings show that not only are fewer features required but also that higher prediction accuracy may be attained (96% for

GWO–

MLP and 97% for APSO–MLP). This research has the potential to be applied in clinical practice and used as a tool to assist doctors and physicians.

[5]Julius et al. used the Waikato Environment for Knowledge Analysis (Weka) application platform to test a data set collected from the UCI repository. There were 520 samples in the data set, each with a collection of 17 attributes. The goal of this study was to use machine learning classification approaches based on observable sample attributes to predict diabetes at an early stage. The k-NN, SVM, functional tree (FT), and RFCs were employed as classifiers. k-NN had the highest accuracy of 98%, followed by SVM at 94%, FT at 93%, and RF at 97%.

[6]Shafi et al. reported that because diabetes is a serious illness, early detection is always a struggle. This study used machine learning classification methods to develop a model that could solve any problem and that could be used to identify diabetes development early on. The authors of this research made concerted efforts to develop a framework that could accurately predict the likelihood of diabetes in patients. As part of this study, the three ML approach classification algorithms—DT, SVM, and NBC— were studied and assessed on various measures. In the study, the PID data set acquired from the UCI repository was used to save time and produce precise findings. The experimental results suggested that the NBC approach was adequate, with a 74% accuracy, followed by SVM with a 63% accuracy and the DT with a 72% accuracy. In the future, the built framework, as well as the ML classifiers used, could be used to identify or diagnose other diseases. The study, as well as several other ML methodologies, could be extended and improved for diabetes research, and the scientists intended to classify other algorithms with missing data.

[7]Khanam et al.experimented with diabetes illness prediction. Diabetes is a condition with no known cure; therefore early detection is essential. In this study, data mining, ML techniques, and neural network (NN) methodologies were utilized to predict diabetes. They developed a technique that could accurately predict diabetes. They used data from the UCI repository’s PID data set. The data set included information on 768 patients and their 9 attributes. On the data set, they utilized seven ML methods to predict diabetes: DT, k-NN, RFC, NBC, AB, LR, and SVM. They used the Weka tool to preprocess the data. They discovered that a model combining LR and SVM is effective at predicting diabetes. They created a NN model with two hidden layers and varied epochs and found that the NN with two hidden layers gave 88.6% accuracy. ANN scored 88.57%, LR scored 78.85%, NBC scored 78.28%, and RFC scored 77.34%.

[8]Sisodia et al.used the PID data set available on the UCI repository. This data set contained 768 patients and 8 attributes. They employed three ML classifications to identify diabetic patients: DT, SVM, and NBC. NBC had the highest accuracy (76.30%) when compared to the other models.

[9]Agarwal et al. used the PID data set of 738 patients as well in their study. To analyze the effectiveness of this data set for identifying diabetic patients, the authors applied models such as SVM, k-NN, NBC, ID3, C4.5, and CART. The SVM and LDA algorithms were the most accurate, with an accuracy of 88%.

[10]Rathore et al. Employed classification techniques like SVM and DTs to predict diabetes mellitus. The PID data set provided the data for this investigation. PIMA India prioritizes women’s health. The SVM has an accuracy of 82%.To predict diabetes mellitus, Hassan et al. Employed classification approaches such as the DT, k-NN, and SVM. The SVM outperformed the DT and KNN methods with a maximum accuracy of 90.23%.

[11]Kandhasamy and Balamurali investigated the prediction accuracy of J48, k-NN,

RFC, and SVM on the diabetes data set. Before preprocessing the data, the author discovered that the J48 method had a higher accuracy than others, at 73.82%. After preprocessing, k-NN and RFC demonstrated improved accuracy.

[12]Kumari and Chitra used SVM, RFC, DT, MLP, and LR, as well as four k-fold crossvalidations (k = 2,4,5,10) in their research. According to the researchers, MLP with four-fold cross-validation achieves the best accuracy, at 78.7%. They discovered that MLP outscored all other algorithms.To predict diabetes, Kavakiotis et al.employed NBC, RFC, k-NN, SVM, DT, and LR methods. The algorithms were applied using a ten-fold cross-validation technique.

SVM had the best accuracy of all the approaches, measuring 84%, according to the study.

[13]Perveen et al. used a data set from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database to do their research. The study employed the AdaBoost and bagging ensemble techniques using the J48 (C4.5) DT as a base learner and standalone data mining methodology J48 to categorize patients with diabetes mellitus based on diabetes risk indicators. This categorization was done across three separate ordinal adult groups in the CPCSSN. In terms of overall performance, the AdaBoost ensemble method surpassed both bagging and a single J48 DT, according to the findings.

[14]Mujumdar and Vaidehi presented a diabetes prediction model for better diabetes classification that included a few extrinsic factors that caused diabetes, as well as regular components such as glucose, BMI, age, insulin, and so on. The new data set enhanced classification accuracy when compared to the old data set. Multiple ML approaches were used on the data set, and classification was done with a variety of algorithms, with LR yielding the highest accuracy at 96%. The AdaBoost classifier was found to be the most accurate, with a 98.8% accuracy rate. They used two separate data sets to compare the accuracy of ML techniques. When compared to the existing data set, it was clear that the model improved diabetes prediction accuracy and precision. the PIMA Indian community in Phoenix, Arizona, to evaluate the method.

# 3.DATA PREPROCESSING

## 3.1 Dataset Discription

In this section our main concern to collect the accurate data. We have collected our necessary data from various sources. The Diabetes.csv file contains the data for this project. The number of instances (rows) in the data set is 390, and the number of variables (columns) is 16. The number of input variables, or attributes for each sample, is 16. The input variables are numericvalued, binary, and categorical. The number of target variable is 1 or 0,1 represents diabetes and 0 represent no diabetes.

Attributes or Parameters are:

1. Patient number
2. Cholesterol
3. Glucose
4. HDL
5. Chol/HDL
6. Age
7. Gender
8. Height
9. Weight
10. BMI
11. Waist
12. Hip
13. Waist/Hip
14. Diabetes

**Patient number:**

Patient number used to identifies patients by number.

**Cholesterol:**

Cholesterol is a type of fat that is produced by the liver and found in certain foods. It is essential for the normal functioning of the body, but too much cholesterol can increase the risk of heart disease. There are several factors that can affect cholesterol levels, including genetics, diet, exercise, and lifestyle choices. A healthy diet, regular exercise, and not smoking can help to maintain healthy cholesterol levels. In some cases, medication may be prescribed to help lower cholesterol levels, particularly in individuals who are at a high risk of heart disease. The normal level of cholesterol in the body can vary depending on the individual's age, sex, and other health factors. In general, total cholesterol levels below 200 milligrams per deciliter (mg/dL) are considered desirable.

**Glucose:**

Glucose is a type of sugar that is the primary source of energy for the body's cells. It is a simple sugar that is produced when the body breaks down carbohydrates, which are found in foods such as bread, pasta, rice, fruits, and vegetables. The normal level of glucose in the body varies depending on the time of day and whether or not a person has eaten recently. In general, the normal range for glucose in the bloodstream is between 70 and 99 milligrams per deciliter (mg/dL).

**HDL:**

There are two types of cholesterol: high-density lipoprotein (HDL) cholesterol, also known as "good" cholesterol, and low-density lipoprotein (LDL) cholesterol, also known as "bad" cholesterol. HDL cholesterol helps to remove LDL cholesterol from the bloodstream and carry it back to the liver, where it can be processed and removed from the body. LDL cholesterol, on the other hand, can build up in the walls of arteries and form plaques, which can lead to heart disease.Optimal levels of HDL,for men less than 40 mg/dL is at risk and 60 mg/dL or above is Desirable. Coming to the Women less than 50 is at risk and 60 mg/dL or above is Desirable. **Chol/HDL**  Ratio of total cholesterol to good cholesterol. Desirable result is<5. **Age**

Which tells about age of a people.

**Gender**

Dataset Contains 162 males,228 females.

**Height&weight**

Height measured in inches and weight in kgs.

**BMI(Body Mass Index)**

BMI! stands for Body Mass Index, and it's a measure of body fat based on your weight and height.

To calculate your BMI, you can use the following formula:

BMI = weight (kg) / height (m)² where weight is your weight in kilograms, and height is your height in meters. A BMI between 18.5 and 24.9 is considered healthy, while a BMI between 25 and 29.9 is considered overweight, and a BMI of 30 or above is considered obese. However, it's important to note that BMI is not always an accurate indicator of health.

**Waist,Hip and waist/hip**

Waist and hip measurements are often used as parameters for predicting the risk of developing type 2 diabetes. This is because the distribution of body fat, particularly in the abdomen and hip area, can be an important indicator of insulin resistance and metabolic dysfunction, which are key factors in the development of type 2 diabetes.Hip circumference is also used as a parameter because it can reflect the amount of subcutaneous fat in the lower body, which is generally considered less harmful than visceral fat. Studies have shown that a higher waist-tohip ratio (WHR), which is calculated by dividing the waist measurement by the hip measurement, is associated with an increased risk of developing type 2 diabetes.



Fig.3.1.1 The above figure contains the dataset in which it has 15 attributes and 360 rows.

## 3.2 Data cleaning

Data cleaning is a critical step in preparing data for predictive modeling, including for the prediction of diabetes. Here are some key steps in data cleaning for diabetes prediction.

1. Identify and remove duplicate records: Duplicate records can introduce bias and skew the results of predictive models. It is important to identify and remove them from the dataset.
2. Handle missing data: Missing data can occur for a variety of reasons, and it is important to handle it appropriately to avoid bias. One approach is to impute missing data using statistical techniques, such as mean or median imputation, or using machine learning algorithms.
3. Identify and remove outliers: Outliers are data points that are significantly different from the rest of the data and can have a significant impact on the results of predictive models. It is important to identify and remove them from the dataset, or handle them using techniques such as winsorization or transformation.
4. Normalize or standardize variables: Normalizing or standardizing variables can help to ensure that variables with different scales and units are treated equally in the predictive model.
5. Check for data quality issues: Check for data quality issues, such as data entry errors, incorrect data types, or inconsistent data, and correct them as needed.
6. Ensure data consistency: Ensure that the data is consistent across all variables and that there are no discrepancies or contradictions in the data.

**Removing Null values:**

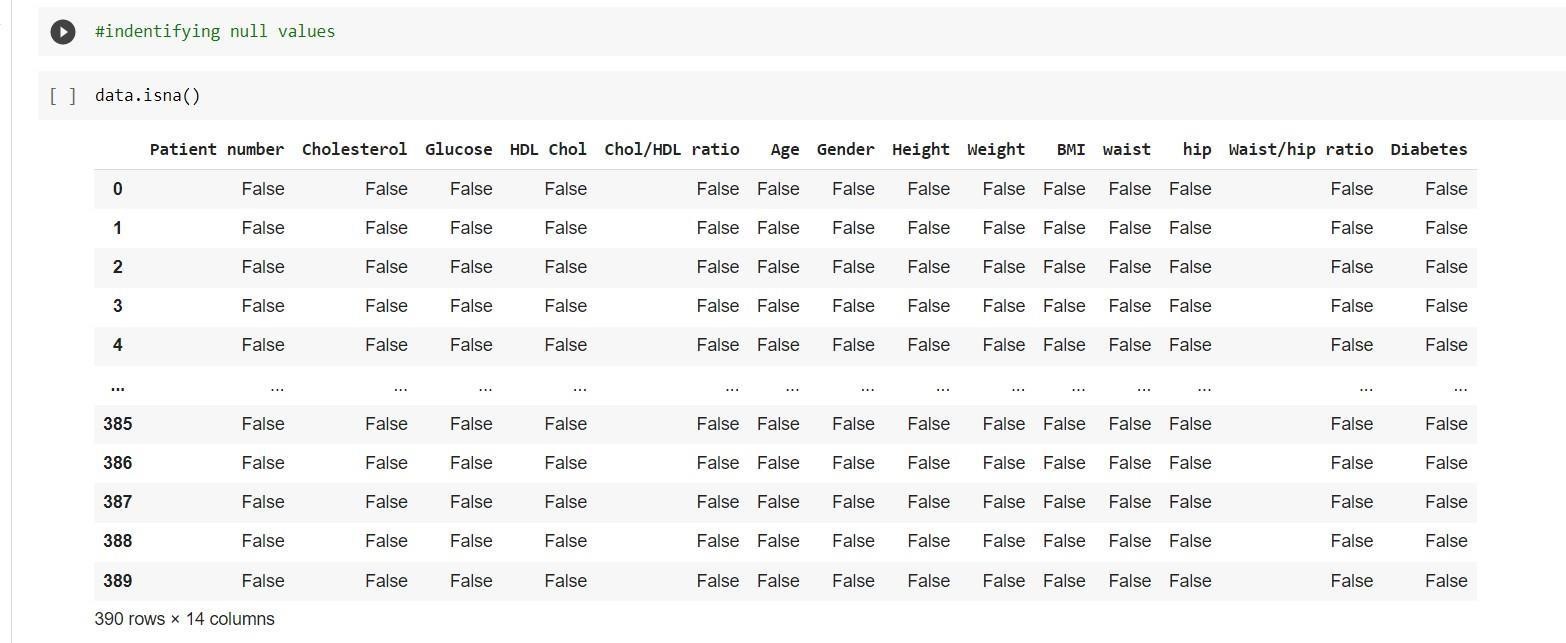


Fig .3.2.1 figure shows no null values

* 1. **Data Augmentation**

Data augmentation is a set of techniques to artificially increase the amount of data by generating new data points from existing data. This includes making small changes to data or using deep learning models to generate new data points. In this dataset we add new every month, its like updating of dataset. We haven’t perform any data augmentation techniques.

* 1. **Data Visualization**

Data visualization is the graphical representation of information and data in a pictorial or graphical format (Example: charts, graphs, and maps). Data visualization tools provide an accessible way to see and understand trends, patterns in data, and outliers.The concept of using pictures is to understand data that has been used for centuries. General types of data visualization are Charts, Tables, Graphs, Maps, Dashboards.Data Visualization Methods are Bar chart, Pie chart, chart, Scatter plot, line chart,Histogram, heatmap, Box plot, treemapping, Table, infographic, Area chart, Line graph, Violin plot.

**Heatmap**: A heatmap (aka heat map) depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges like a bar chart or histogram, and each cell’s color indicates the value of the main variable in the corresponding cell range.sns.heatmap(data.corr())

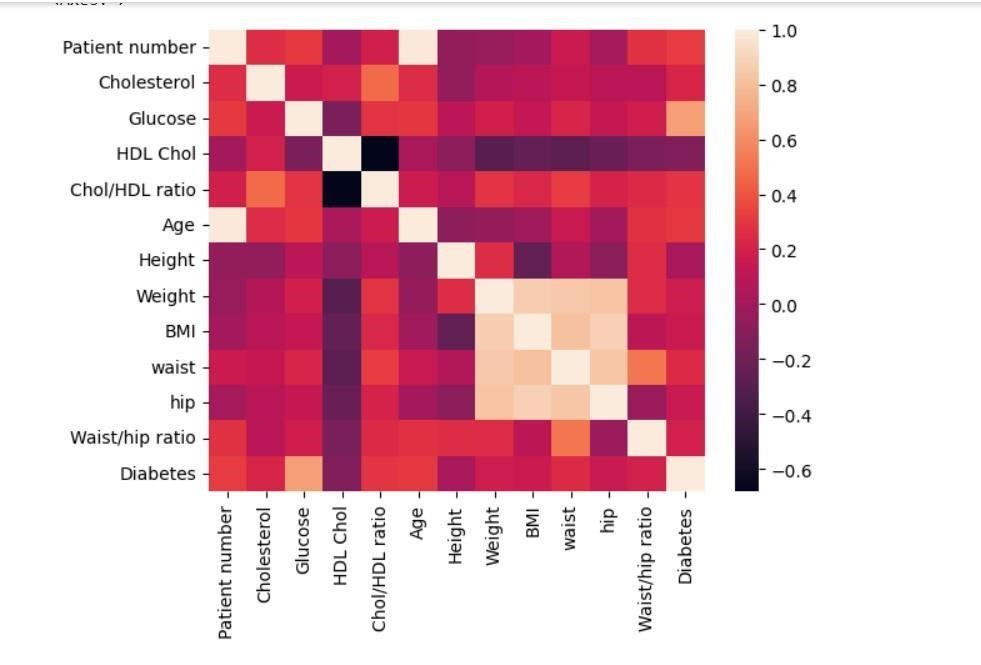


Fig.3.4.1The above figure is constructed using the heapmap,which shows the correlation.

**Barplot:**

Bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

data.count().plot.bar()

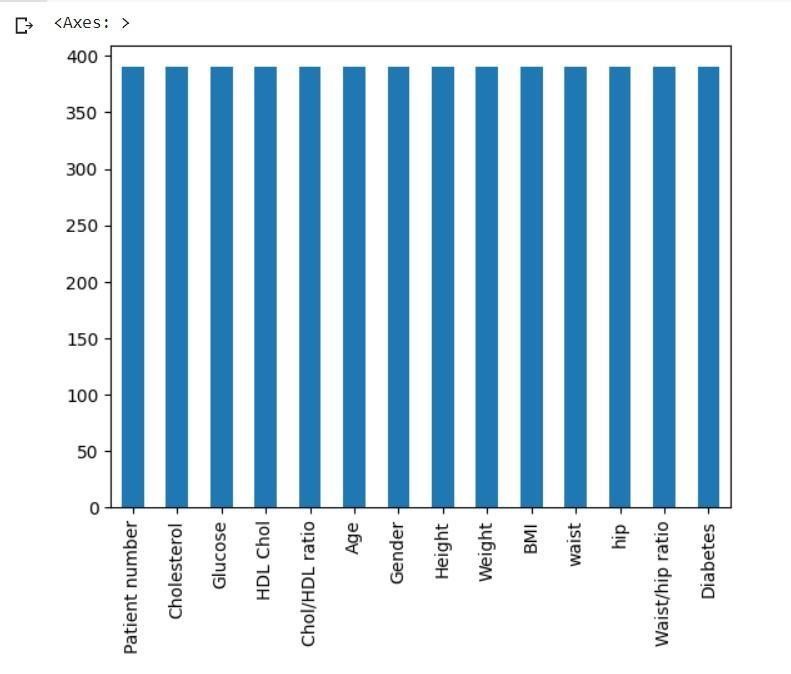


Fig.3.4.2 The above fig constructed using bar graph which tells about variables in dataset.

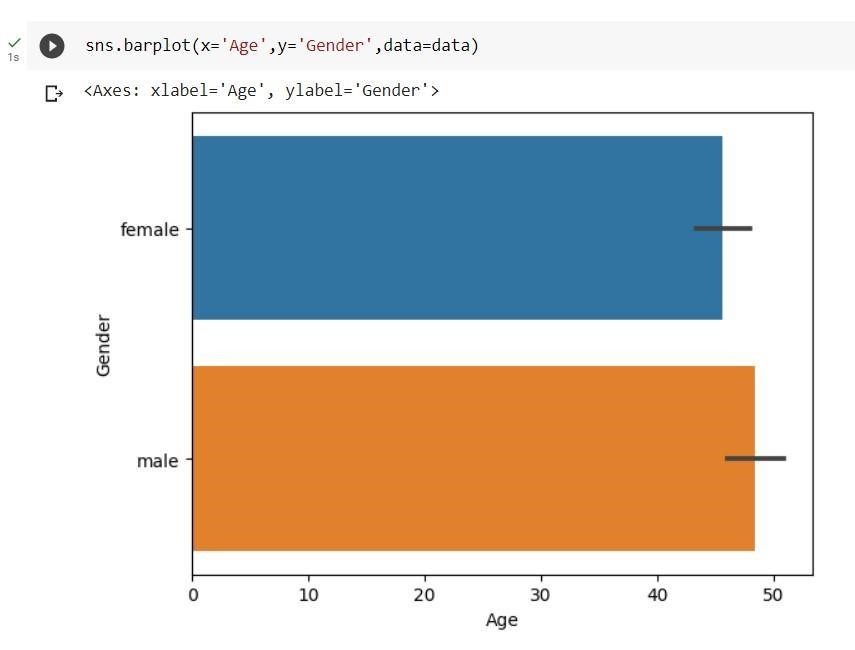


Fig.3.4.3.above fig constructed using bargraph between age on x-axis and gender on yaxis.

**Boxplot:**

Box plots give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value.

We use these values to compare how close other data values are to them.

sns.boxenplot(data=data.Age)

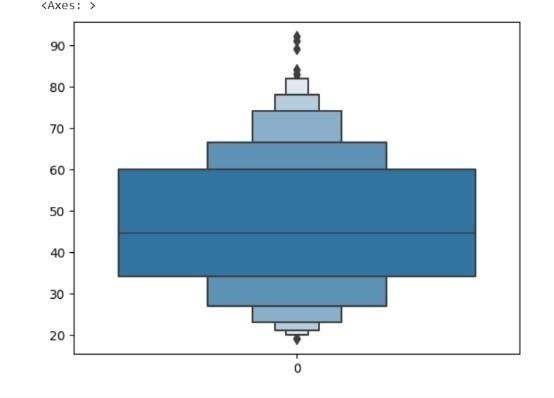


Fig.3.4.4 the above graph constructed using box plot.

**ScatterPlot:**

A scatter plot is composed of a horizontal axis containing the measured values of one variable (independent variable) and a vertical axis representing the measurements of the other variable (dependent variable). The purpose of the scatter plot is to display what happens to one variable when another variable is changed.

sns.scatterplot(data=data,s=160)

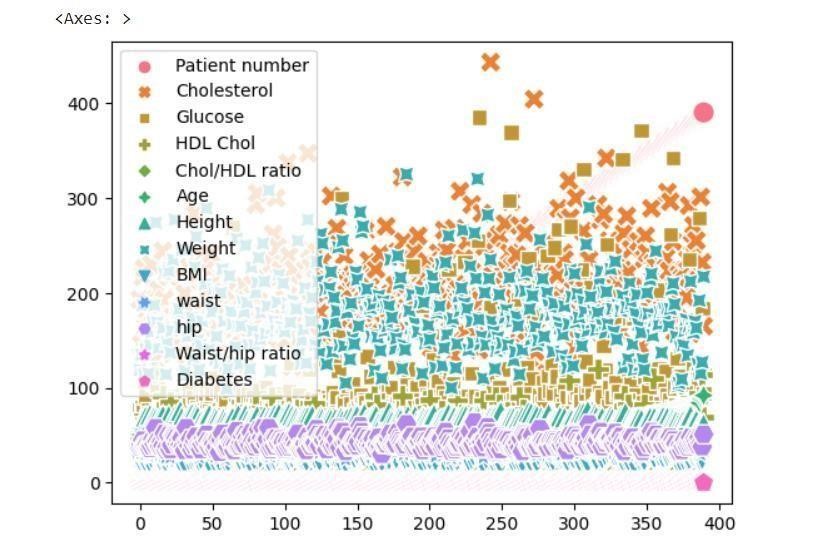


Fig.3.4.5 The above graph tell us about what happens to one variable when another changed.

**Violin Plot:**

violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data. It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable. ax=sns.violinplot(data.Glucose)

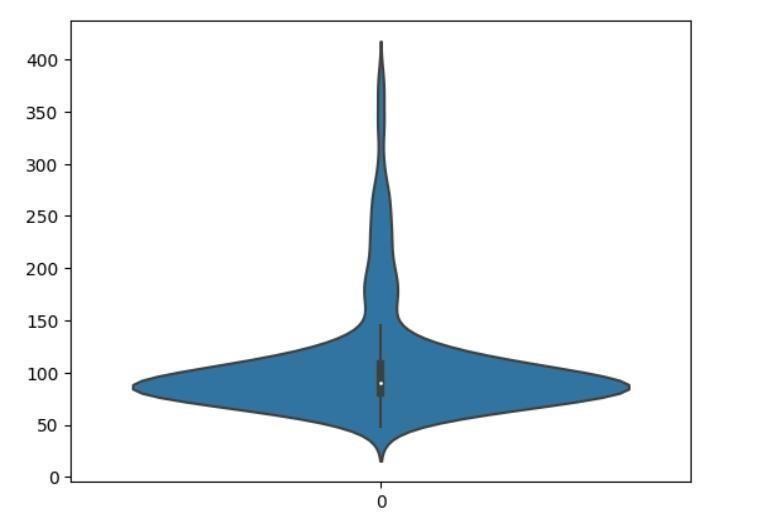


Fig.3.4.6 The above graph tell us about only distribution of numeric data

**Lineplot:** line chart or line graph, also known as curve chart, is a type of chart which displays information as a series of data points called 'markers' connected by straight line

segments. It is a basic type of chart common in many fields

sns.lineplot(x='Cholesterol',y='Diabetes',data=data)

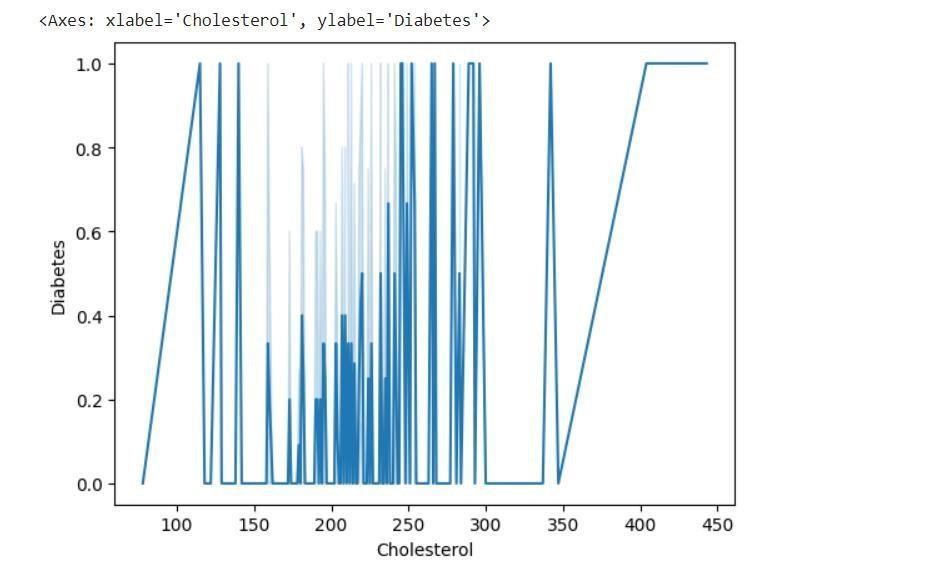


Fig.3.4.7 The above fig contsstructed using lineplot.It shows series of data points of result\_numeric connected by line.

**Density Plot:**

Density plots are used to observe the distribution of a variable in a dataset. It plots the graph on a continuous interval or time-period. This is also known as Kernel density plot.

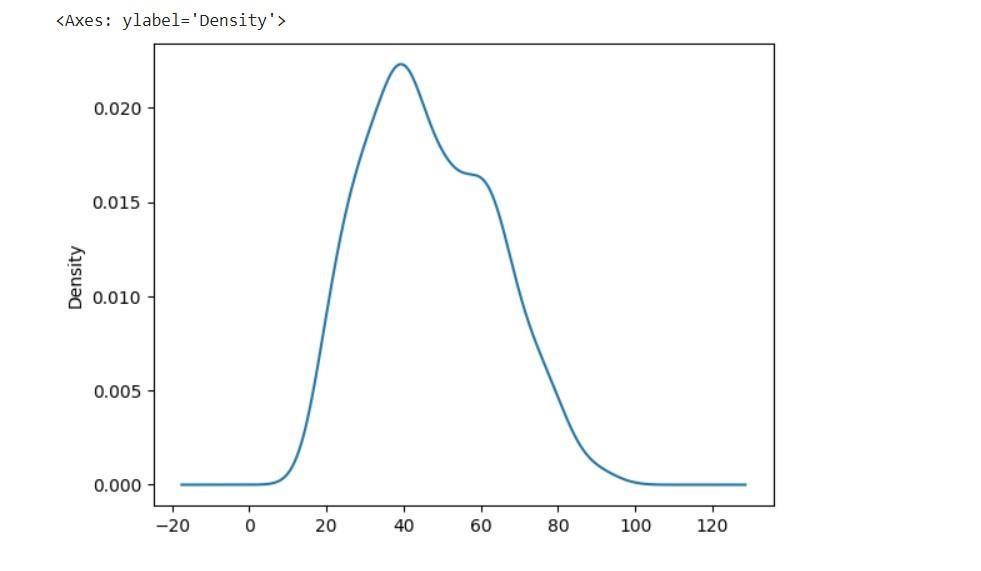
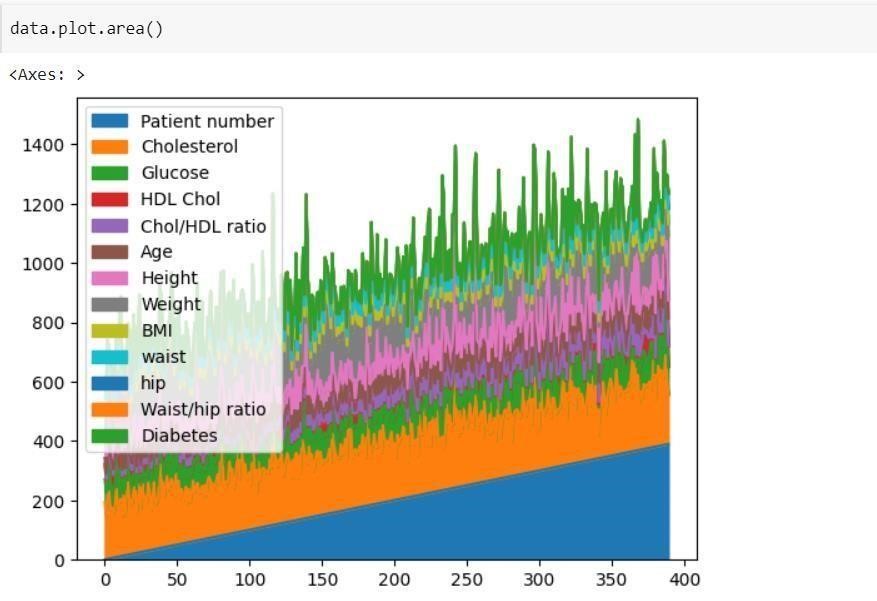


Fig.3.4.8 The above fig tell us result\_numeric are more in 20-60 **Area plot:**

Fig.3.4.9

The above fig is constructed using area plot.



An area graph is a specialized form of the line graph, where instead of simply connecting our data points with a continuous line, we also fill in the region below that line with a solid color.

**Histogram:**

A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars. Each bar typically covers a range of numeric values called a bin or class; a bar's height indicates the frequency of data points with a value within the corresponding bin.

data.hist() array([[<Axes: title={'center': 'Patient number'}>,

<Axes: title={'center': 'Cholesterol'}>,

<Axes: title={'center': 'Glucose'}>,

<Axes: title={'center': 'HDL Chol'}>],

[<Axes: title={'center': 'Chol/HDL ratio'}>,

<Axes: title={'center': 'Age'}>,

<Axes: title={'center': 'Height'}>,

<Axes: title={'center': 'Weight'}>],

[<Axes: title={'center': 'BMI'}>,

<Axes: title={'center': 'waist'}>,

<Axes: title={'center': 'hip'}>,

<Axes: title={'center': 'Waist/hip ratio'}>],

[<Axes: title={'center': 'Diabetes'}>, <Axes: >, <Axes: >,

<Axes: >]], dtype=object)

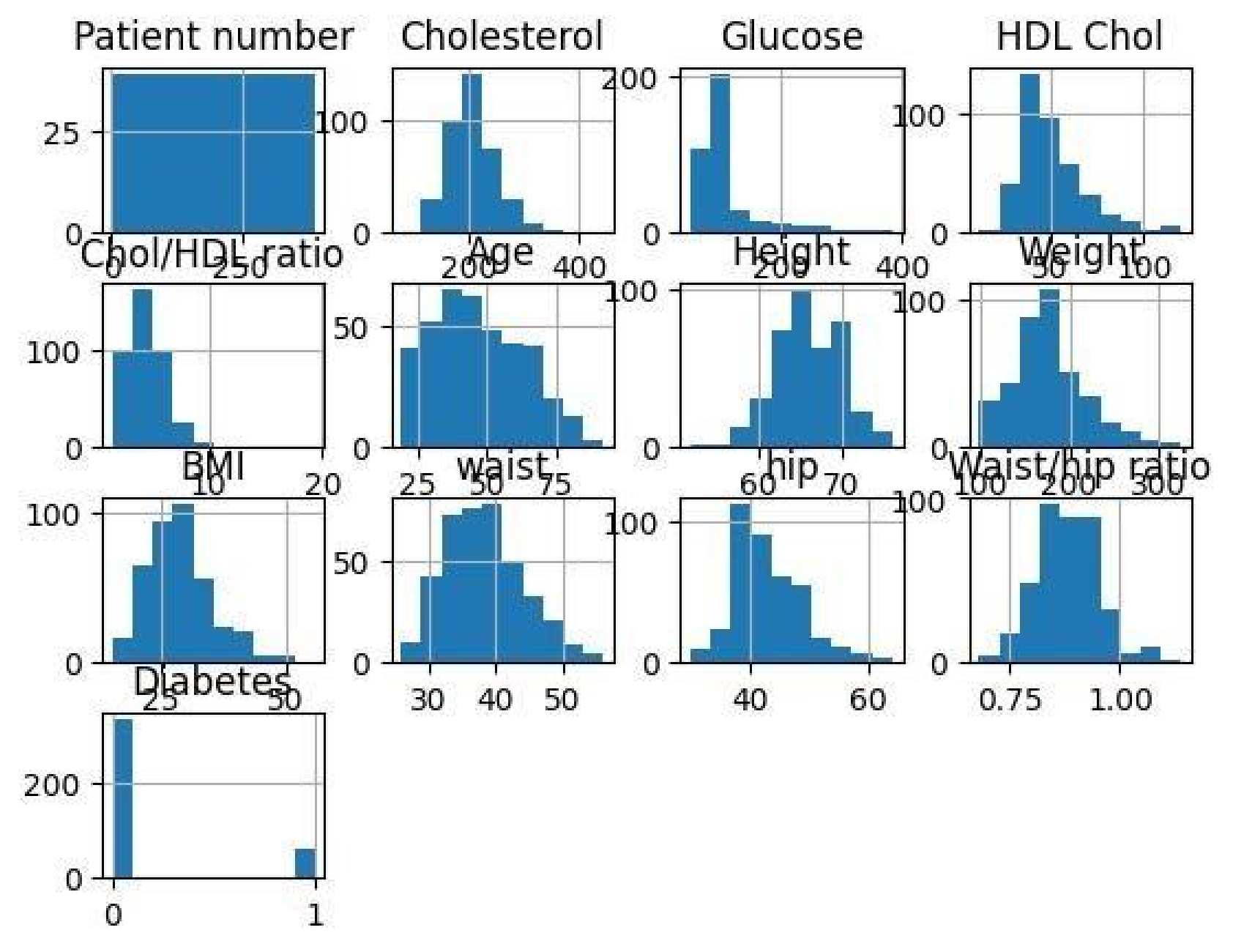


Fig.3.4.11.The above constructed using histogram for every numeric variable in our dataset.

# 4.METHODOLOGY

## 4.1 Procedure to solve your given problem

We followed following steps in our project to get our ultimate goal of predicting diabetes:

1. **Importing Necessary Libraries**

Import Python libraries such as pandas,matplotlib,seaborn,Numpy for reading and

Visualizing of dataset. import pandas as pd import numpy as np import matplotlib.pyplot as plt import seaborn as sns

1. **Loading Dataset**

First and foremost convert the dataset into csv file. And read the data data=pd.read\_csv("/content/Diabetes\_Classification.csv")

1. **Pre-Processing**

This Phase of Model handles inconsistent data in order to get more accurate and precise results.We have checked for null values and we removed those columns.And also we removed two irrelevant columns like systolic BP and diastolic BP,We applied encoding

techniques to change dependent values into ‘0’ an ‘1’.

1. **Exploratory Data Analysis**

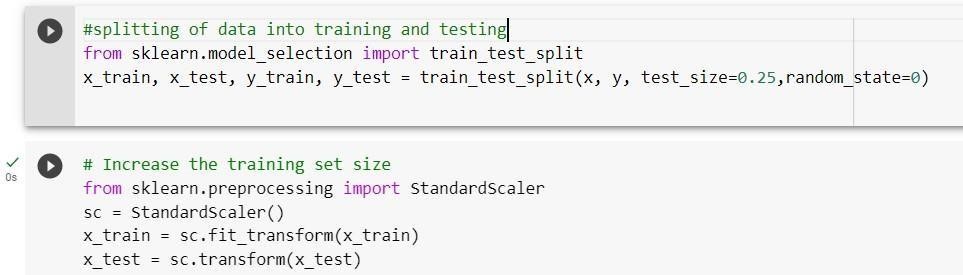
EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects.

Steps in EDA:

* Observe your dataset.
* Find any missing values.
* Categorize your values.
* Find the shape of your dataset.
* Identify relationships in your dataset.
* Locate any outliers in your dataset.
* Organizing a dataset.
* Understanding variables.

1. **Splitting of Data into training set and testingset**

In a basic two-part data split, the training data set is used to train and develop models. Training sets are commonly used to estimate different parameters or to compare different model performance. The testing data set is used after the training is done. We had divided data into training and testing dataset .75% of data for training and 25% for testing.



1. **Model Building**

We used logistic regression, KNN classifier, Decision Tree, Random forest Classifier, Naïve Bayesian, Svm,Kernel Linear. We came into conclusion by comparative study.

**Logistic Regression:**

Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm. Logistic regression uses Maximum Likelihood Estimation (MLE) approach i.e., it determines the parameters (mean and variance) that are maximizing the likelihood to produce the desired output.

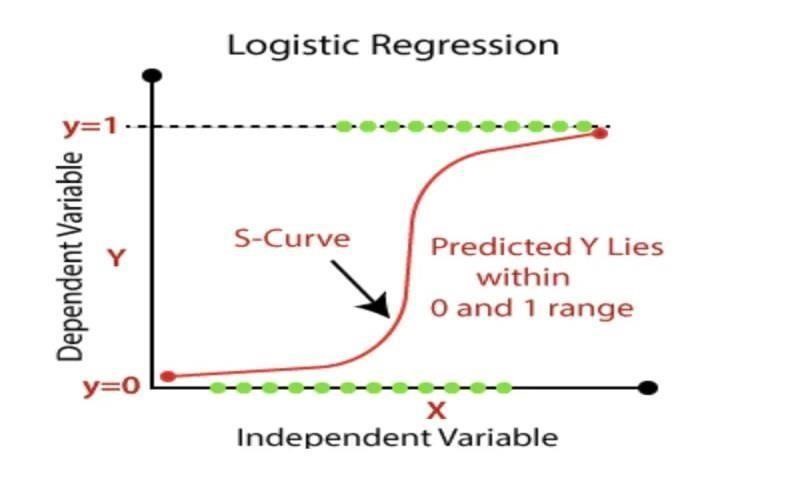


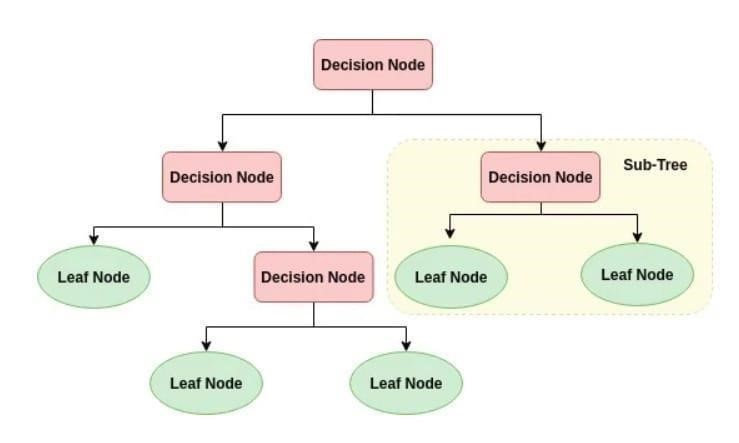
Fig.4.1.1.it is a graph of logistic regression.



**Decision Tree:**

It is a non parametric method used for supervised learning method used for both classification and regression. It uses tree representation to solve the problem. As deeper the tree goes, the more complex the decision rules and the fitter the model. Entropy and Information Gini are used to calculate root node among all the nodes. Hence, an optimal tree can be formed. From given data a tree can be formed and using entropy & information gini we can calculateaccuracy.Tree is a hierarchical representation .ENTROPY: Entropy is the measure of uncertainity of a random variable. The higher entropy results in more information. Entropy=summation(-p\*log(p ))

Information Gini=E(T)-E(T,X)



### Fig.4.1.2 it is a decision tree



**K-Nearest Neighbour:**

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

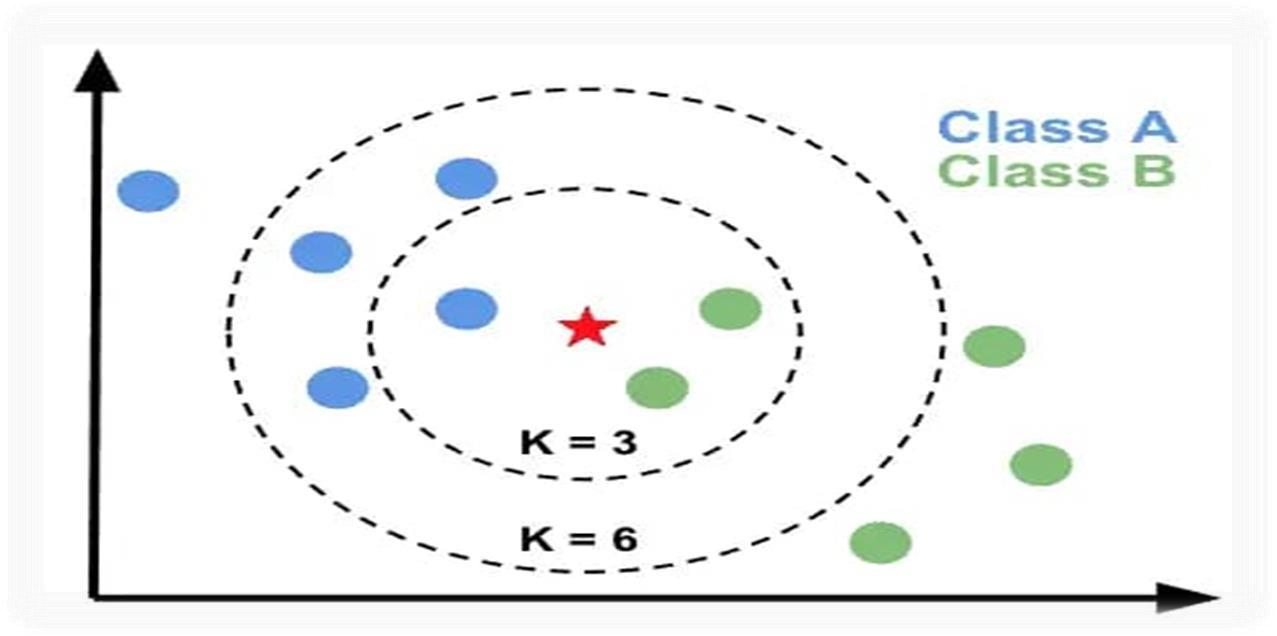


Fig.4.1.3 It is a k-nearest neighbour graph.



**Random Forest:**

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we’ll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged.

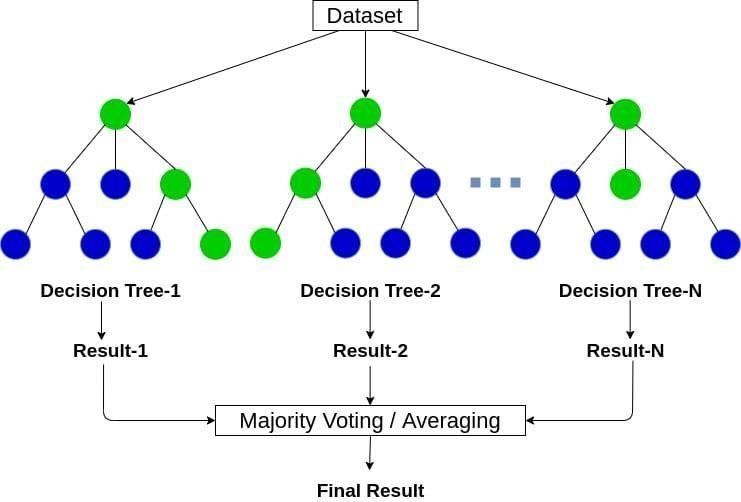


Fig.4.1.4 the above fig is a graph of random forest classifier.



**Support Vector Machine:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. In the SVM method, we plot each data item as a point in n-dimensional space(where n is no. of features you have). We perform classification by finding the hyperplane that differentiates the two classes verywell. We have three hyper planes(A,B and C). Now, identify the right hyper plane to classify star and circle.We want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.The distance between the hyperplane and the nearest data point from either set is known as margin. The goal is to choose a hyperplane with the greatest possible margin.There will never be any data point inside the margin.

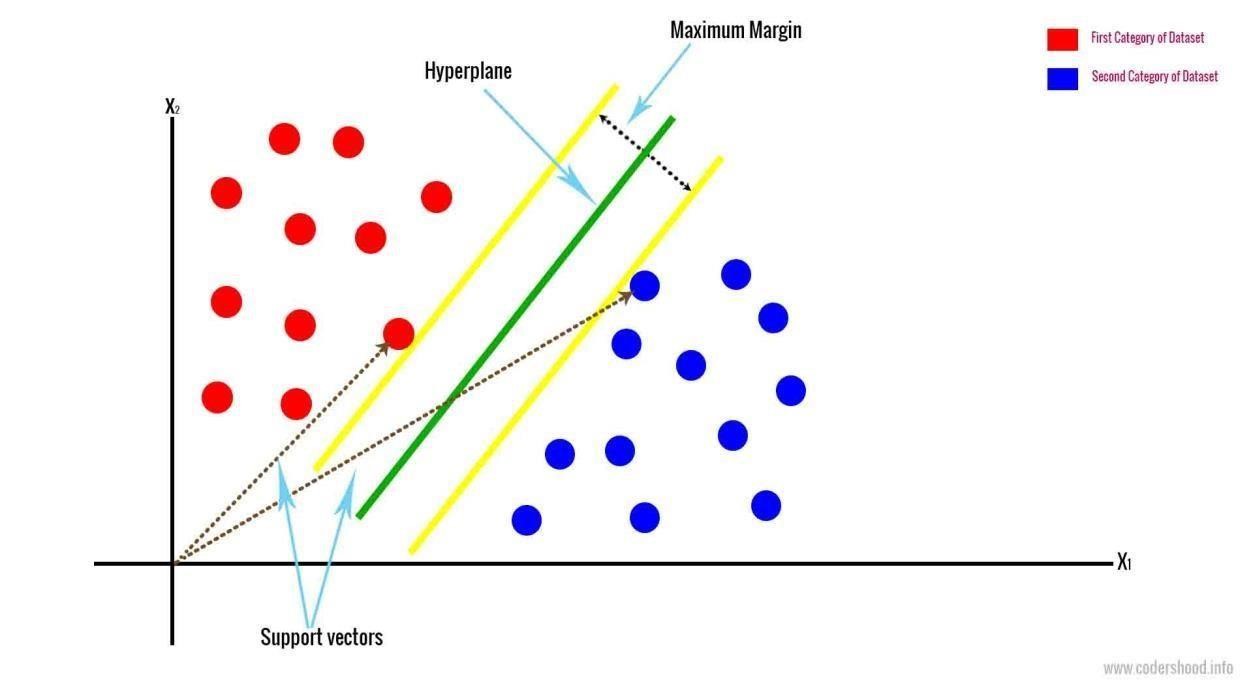


Fig.4.1.5 The above fig is a support vectore machine.



**Naive Bayes:**

The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

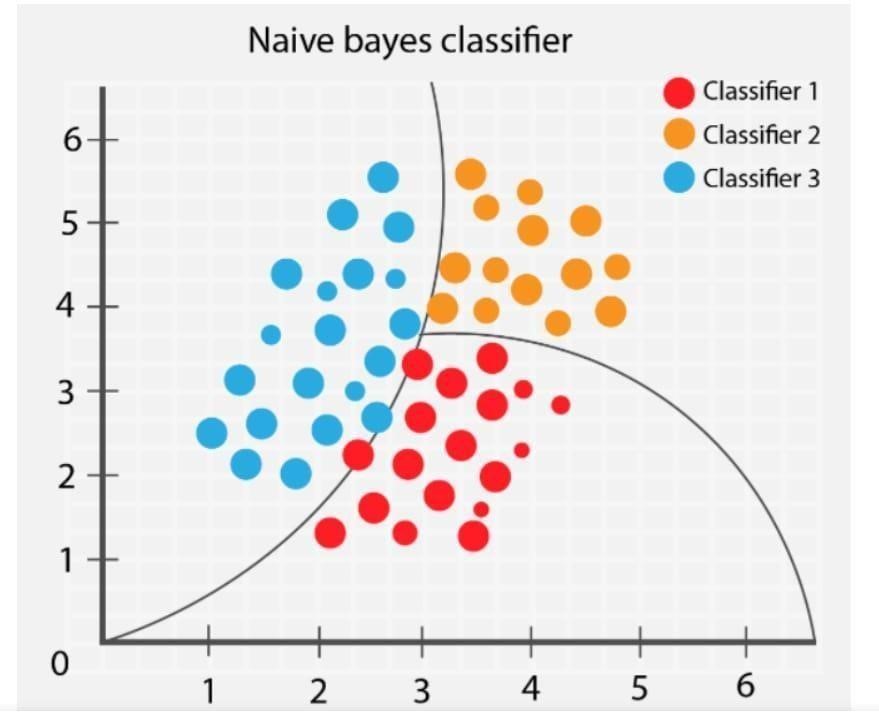


Fig.4.1.6 it a graph of naïve Bayesian



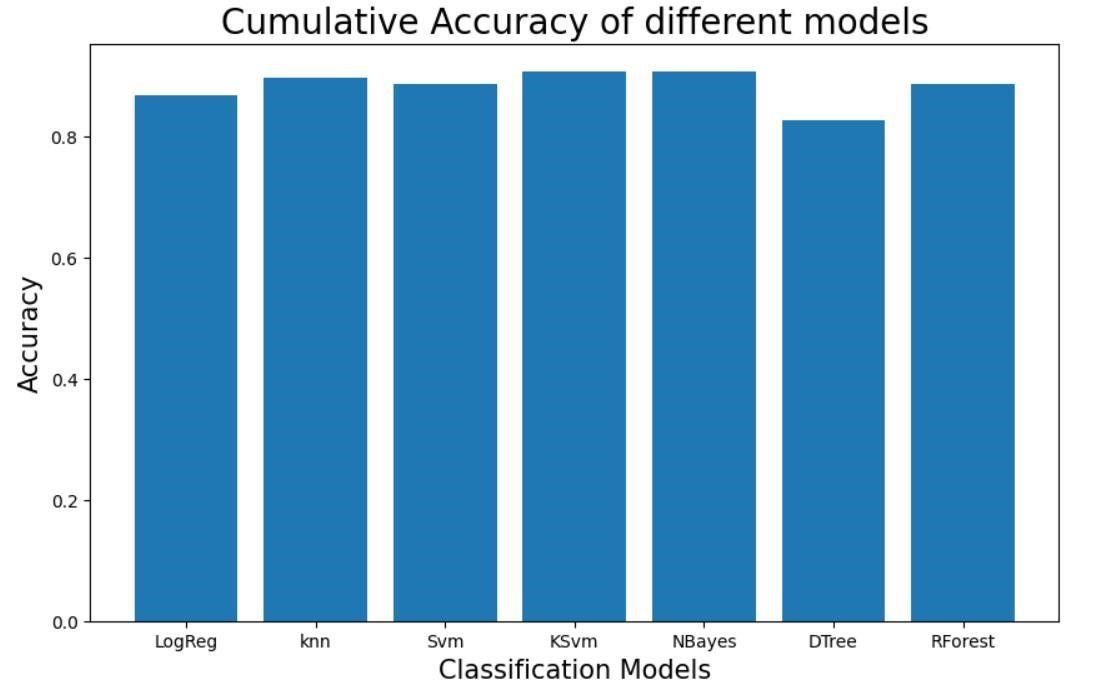


Fig.4.1.6 figure shows the plot of all the classification model accuracy.

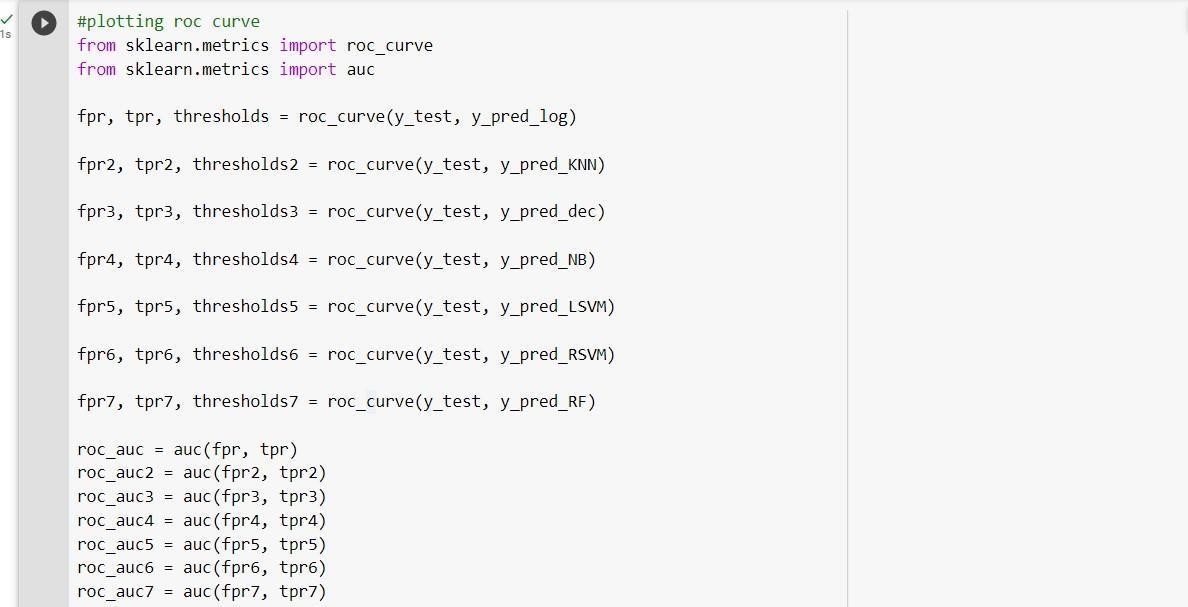
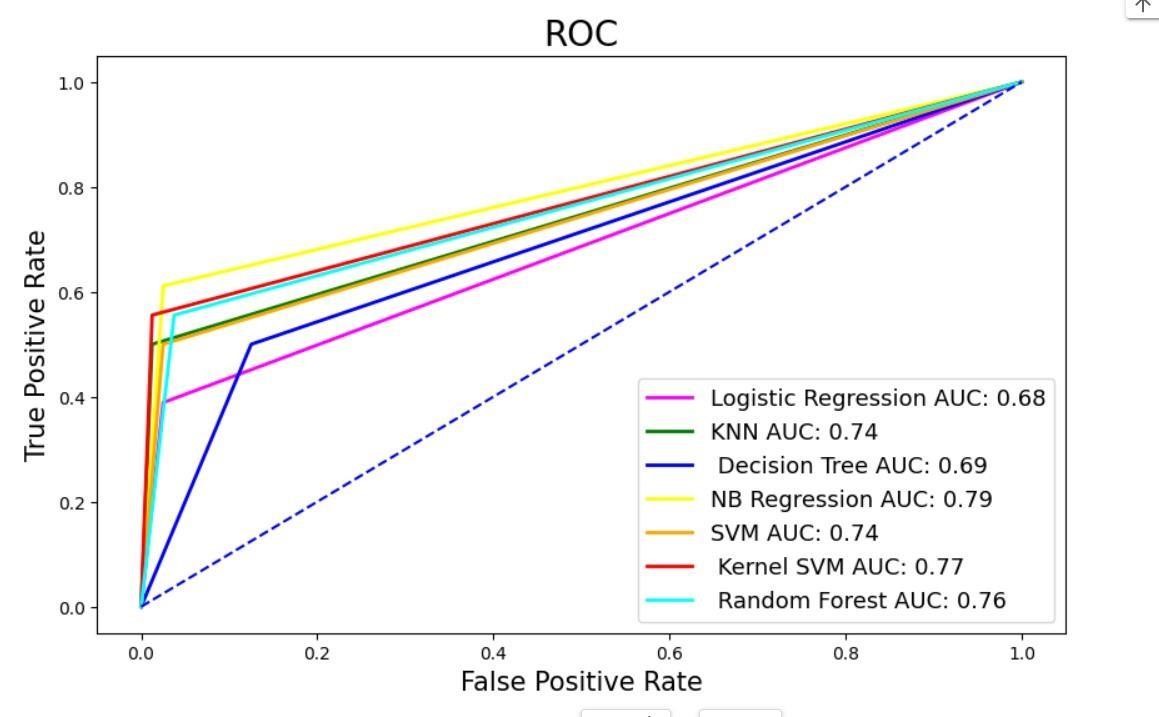
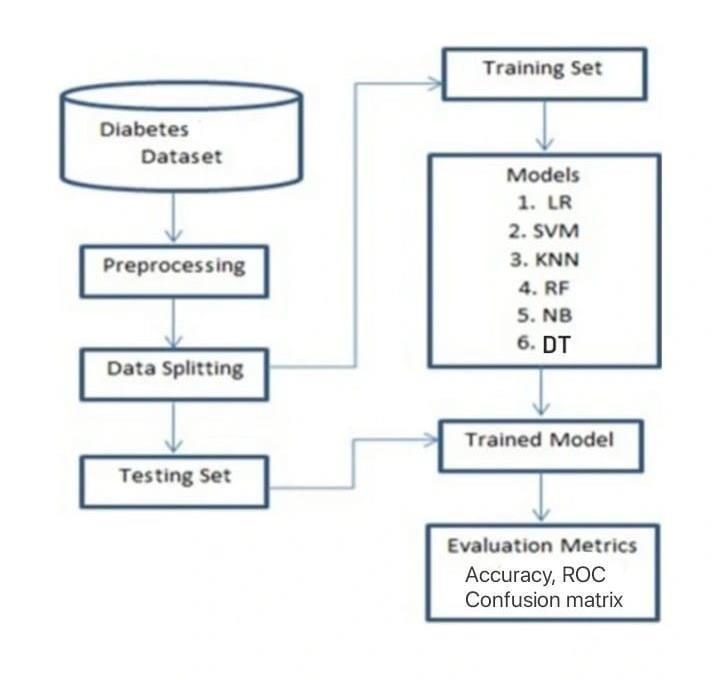


Fig.4.1.8 above figure shows the ROC curve of all the models.



## 4.2. Model architecture



## 4.3. Software description

Python: Python is an interpreted, object-oriented, high-level programming language with dynamic semantics developed by Guido van Rossum. It was originally released in 1991.Python is used for server-side web development, software development, mathematics, and system scripting, and is popular for Rapid Application Professionally, Python is great for backend web development, data analysis, artificial intelligence, and scientific computing. Developers also use Python to build productivity tools, games, and desktop apps.

Google colab: Google Colaboratory, popularly known as Colab, is a web IDE for python that was released by Google in 2017. Colab is an excellent tool for data scientists to execute Machine Learning and Deep Learning projects with cloud storage capabilities. Google Colab is a free Jupyter notebook that allows to run Python in the browser without the need for complex configuration. It comes with Python installed and has all the main Python libraries installed. It also comes integrated with free GPUs

# 5.RESULTS AND DISCUSSIONS

|  |  |
| --- | --- |
| **Machine Learning**  **Algorithms** | Accuracy |
| Logistic Regression | 0.8673469387755102 |
| K-Nearest Neighbour | 0.8979591836734694 |
| Decision Tree | 0.826530612244898 |
| Naïve Bayes | 0.9081632653061225 |
| Random Forest  Classifier | 0.8877551020408163 |
| Support Vector Machine-  Linear | 0.8877551020408163 |
| Support Vector Machine-rbf | 0.9081632653061225 |

By Using Machine Learning Algorithms we Classified Whether the Person has Diabetes or Not.

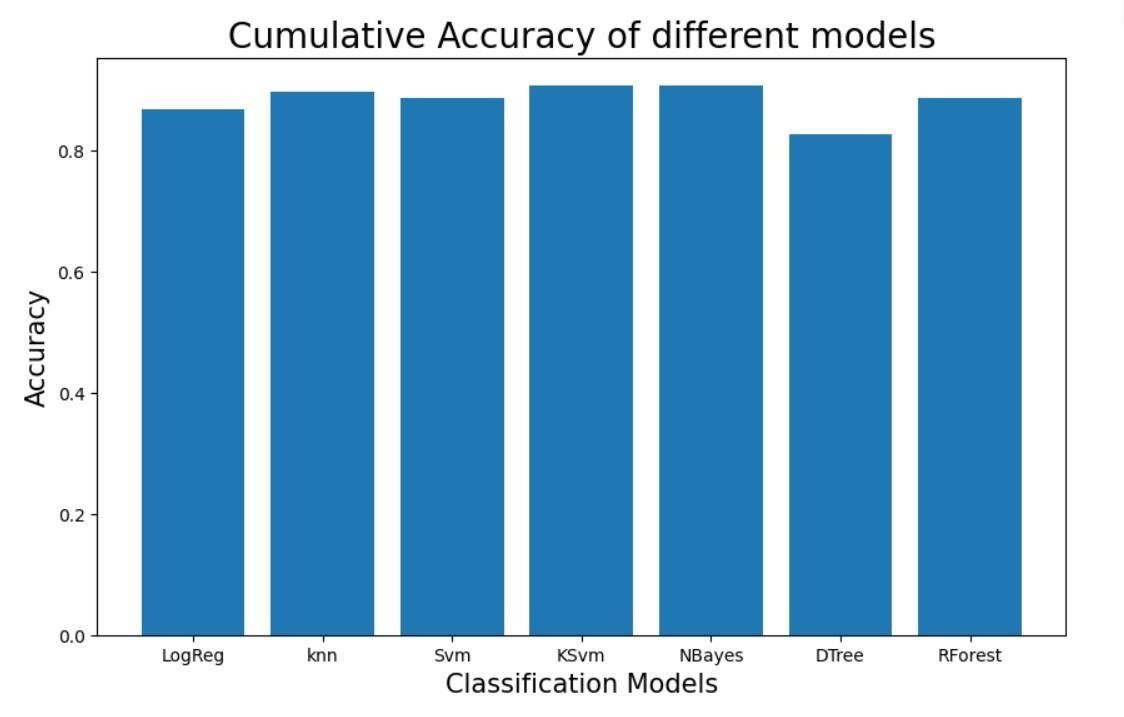


Fig.5.1 Above figure shows that barplot of acuuracy of all models.

# 6.CONCLUSION AND FUTURE SCOPE

**CONCLUSION:**

In Prediction of diabetes we are Classifying Whether a person has diabetes or not based on glucose levels ,cholesterol levels ,Age ,Weight,BMI ,hip ,waist .The person with 200(mg/dl) cholesterol level ,70-99(mg/dl) glucose level ,and BMI between 18.5 and 24.9 considered as Healthy. After using all these patient records, we are able to build a machine learning model (support vector machine and naïve bayes – best one) to accurately predict whether or not the patients in the dataset have diabetes or not In addition, through data analysis and visualisation, we were able to draw a few conclusions from the data. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning technique.

**FUTURE SCOPE:**

As millions of people worldwide are affected by diabetes, an important global health difficulties. Healthcare providers can identify people who are at a high risk of getting diabetes by using prediction models that are accurate and dependable thanks to AIML approaches .Better. As there is more data available, machine learning algorithms can be trained on larger datasets, which produces more accurate prediction models. To create more sophisticated and precise models, emerging AI approaches like deep learning could be used.Predictions that are specific to an individual may be made with the aid of personalised health data, including genetic data and lifestyle characteristics.

These models would take into account each person’s particular risk factors. Early detection: AIML algorithms may be used to create models that may spot diabetic symptoms before they become more serious, allowing for immediate treatment and intervention.

.

# 6.REFERENCES

[https://data.world/akshayginodia/diabetes-datase](https://data.world/akshayginodia/diabetes-dataset)[t https://data.world/informaticsedu/diabetesprediction](https://data.world/informatics-edu/diabetes-prediction)

* B. S. Kim, M. Ahn, W.-W. Cho, G Gao, J. Jang, and D.-W. Cho, “Engineering of diseased human skin equivalent using 3D cell printing for representing pathophysiological hallmarks of type 2 diabetes in vitro,” Biomaterials, vol. 272, Article ID 120776, 2021.

View at: Publisher Site | Google Scholar

* K. E. Minges, P. Zimmet, D. J. Magliano, D. W. Dunstan, A. Brown, and J. E. Shaw, “Diabetes prevalence and determinants in Indigenous Australian populations: a systematic review,” Diabetes Research and Clinical Practice, vol. 93, no. 2, pp. 139– 149, 2011.

View at: Publisher Site | Google Scholar

* Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus with machine learning techniques,” Frontiers in Genetics, vol. 9, p. 515, 2018. View at: Publisher Site | Google Scholar
* S. Gautam, N. Shrestha, S. Mahato, T. P. A. Nguyen, S. R. Mishra, and G. Berg-

Beckhoff, “Diabetes among tuberculosis patients and its impact on tuberculosis treatment in South Asia: a systematic review and meta-analysis,” Scientific Reports, vol. 11, no. 1, pp. 2113–2212, 2021. View at: Publisher Site | Google Scholar

* J. C. Won, J. H. Lee, J. H. Kim et al., “Diabetes fact sheet in Korea, 2016: an appraisal of current status,” Diabetes & metabolism journal, vol. 42, no. 5, pp. 415–424, 2018. View at: Publisher Site | Google Scholar
* S. Wei, X. Zhao, and C. Miao, “A comprehensive exploration to the machine learning techniques for diabetes identification,” in Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), pp. 291–295, Singapore, May 2018. View at: Publisher Site | Google Scholar

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |