



# Diabetic Prediction

**Alfrin Samraj P**

EXPOSYS DATA LAB (Intern)

St Joseph's College of Engineering, Chennai

alfrinsamrajp007@gmail.com

## Abstract

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy.

Therefore three machine learning algorithms namely Decision Tree, SVM and Logistic Regression are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Logistic Regression outperforms with the highest accuracy of 83.12% compared to other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.



## Table of Content


1. Introduction
2. Existing Method
3. Proposed method with Architecture
4. Methodology
5. Implementation
6. Conclusion

## Introduction

Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrates abnormal and raises the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, Intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar.

Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma . Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot be controlled. Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. Early identification is the only remedy to stay away from complications .

Many researchers are conducting experiments for diagnosing the diseases using various algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms work better in diagnosing different diseases. Machine learning algorithms gain their strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study .



This study focuses on pregnant women suffering from diabetes. In this work, Logistic Regression, SVM, and Decision Tree machine learning algorithms are used and evaluated on the PIDDD dataset to find the prediction of diabetes in a patient. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy .

## Existing method

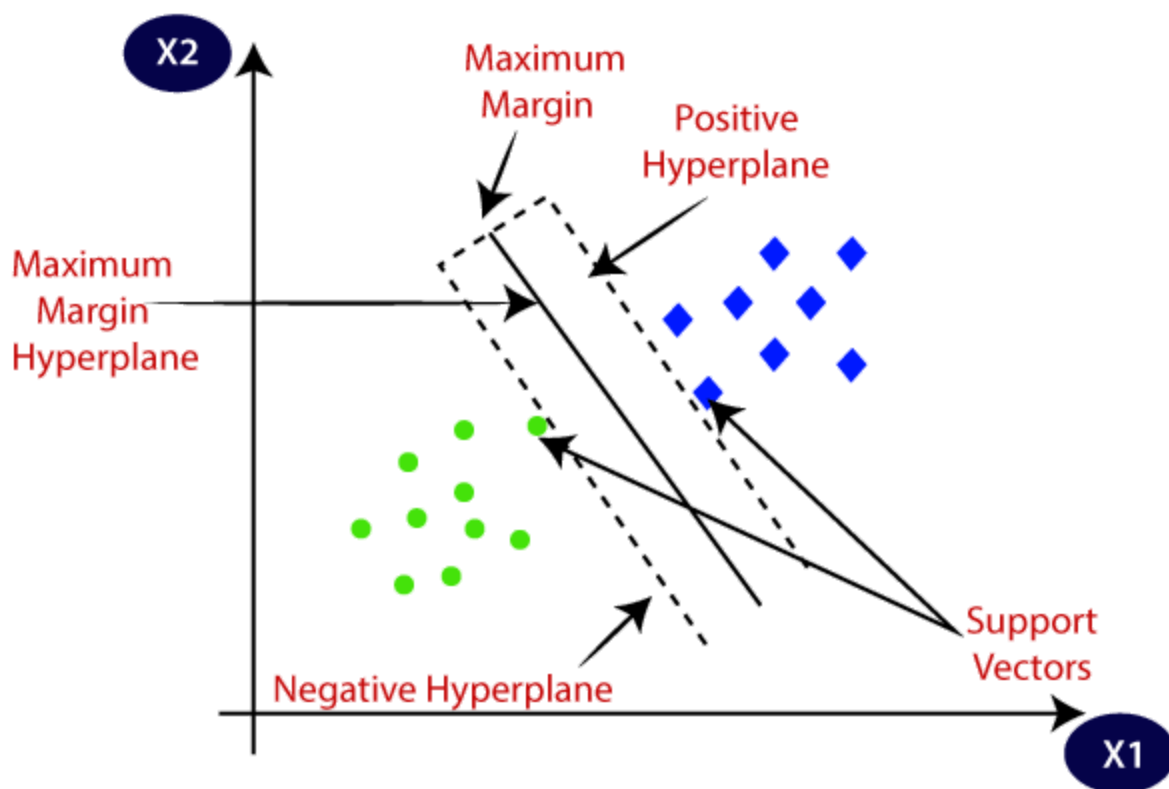
### Algorithms Used :

- Support Vector Machine (SVM)
- Logistic Regression
- Decision Tree Classifier

### Support Vector Machine (SVM)


SVM is one of the standard sets of supervised machine learning models employed in classification. Given a two-class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes. For better generalization hyperplanes should not lie closer to the data points belonging to the other class. Hyperplanes should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors .

The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by  $w^T x + b = -1$  and the hyperplane defined by  $w^T x + b = 1$ . This distance is equal to  $2/w$ . This means we want to solve  $\max 2/w$ . Equivalently we want  $\min |w|$ . The SVM should also correctly classify all  $x(i)$ , which means  $y_i (w^T x_i + b) \geq 1, \forall i \in \{1, \dots, N\}$ .



## Logistic Regression

Logistic regression models the probability of the default class (e.g. the first class).



For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height, or more formally:

$$P(\text{sex}=\text{male}|\text{height})$$

Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:

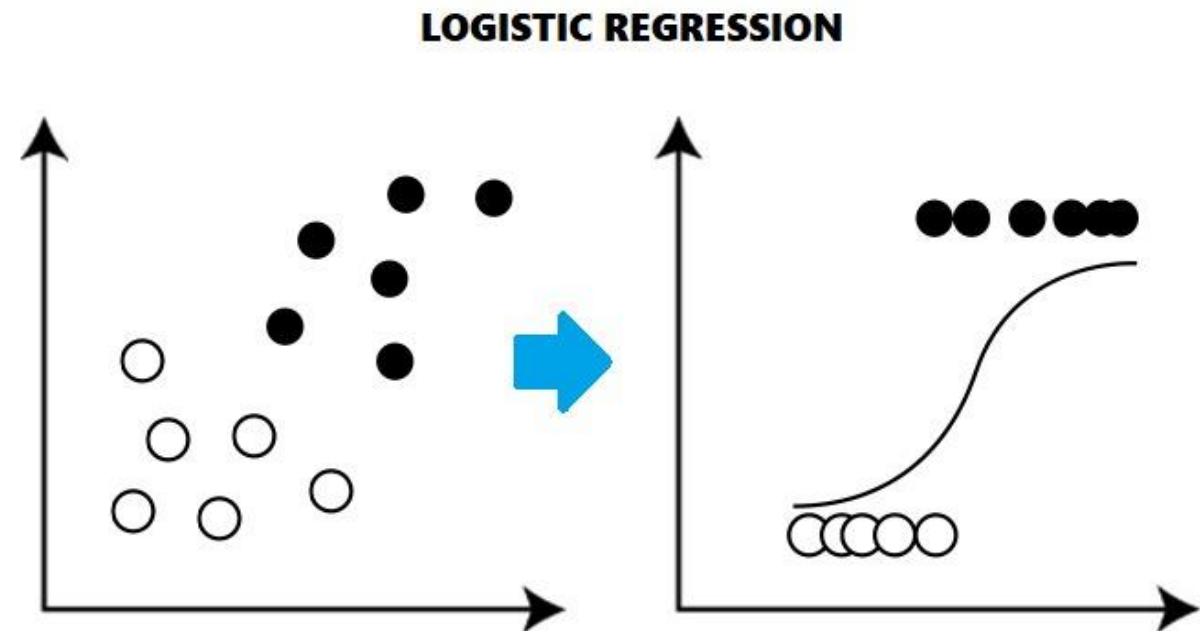
$$P(X) = P(Y=1|X)$$

We're predicting probabilities? I thought logistic regression was a classification algorithm?

Note that the probability prediction must be transformed into a binary values (0 or 1) in order to actually make a probability prediction. More on this later when we talk about making predictions.

Logistic regression is a linear method, but the predictions are transformed using the logistic function. The impact of this is that we can no longer understand the predictions as a linear combination of the inputs as we can with linear regression, for example, continuing on from above, the model can be stated as:

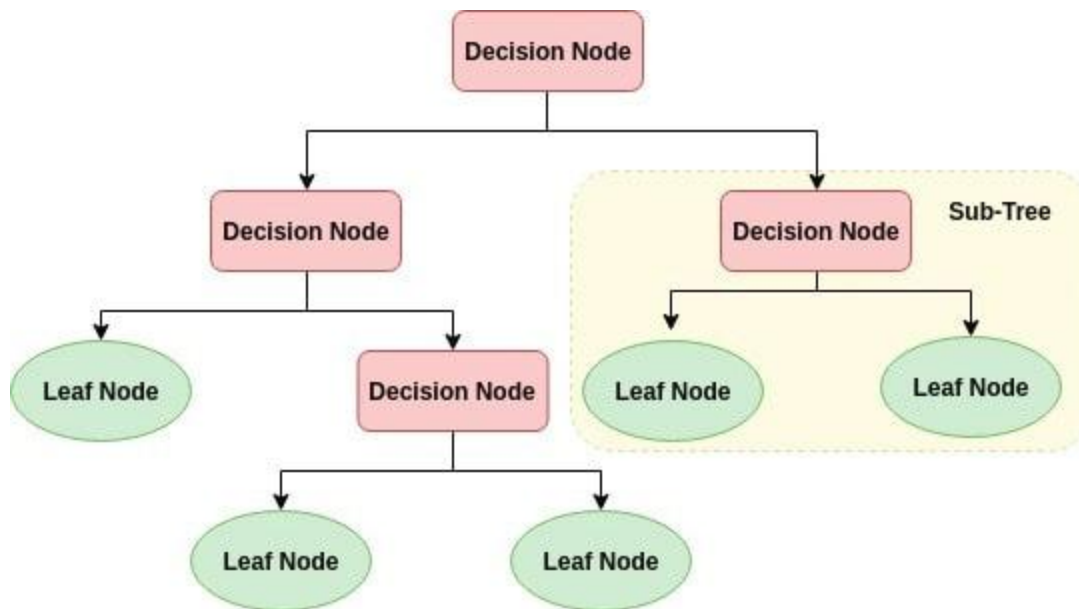
$$p(X) = e^{(b_0 + b_1 X)} / (1 + e^{(b_0 + b_1 X)})$$



## Decision Tree Classifier

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of the target class using decision rules taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, the Decision tree chooses each node by evaluating the highest information gain among all the attributes .





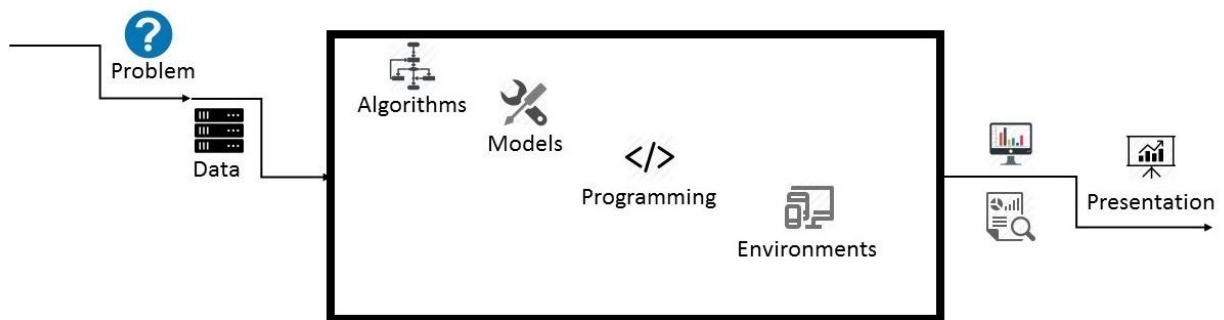
## Dataset Used

### PIDD - Pima Indians Diabetes Dataset

The proposed methodology is evaluated on the Diabetes Dataset (PIDD) , which is taken from UCI Repository. This dataset comprises medical details of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class '0' is treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes.


Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1

## Proposed method with Architecture



## Problem

Problem statement can be anything from building a prediction models, building a market segmentation, building a recommendation engine, association rule discovery for fraud detection, minimizing production costs, minimizing advertisement costs, maximizing ROI, right rewards to right users, right offers from




offline to online to right users, best deals and offers to right users, right gift to right users, or simulations to predict extreme events such as floods.

## Data

Data comes in many shapes: transactional, real-time, sensor data, unstructured data , structured data, big data, images or videos, pdfs, news, press releases, public hearings, and so on. Typically raw data needs to be identified or even built and put into databases (NoSQL or traditional), then cleaned and aggregated using EDA (exploratory data analysis). The process can include selecting and defining metrics.

## Working Model

Model means testing algorithms, selecting, fine-tuning, and combining the best algorithms using techniques such as model fitting, model blending, data reduction, feature selection, and assessing the yield of each model, over the baseline. It also includes calibrating or normalizing data, imputation techniques for missing data, outliers processing, cross-validation, overfitting avoidance, robustness testing and boosting, and maintenance. Criteria that



make a model desirable include robustness or stability, scalability, simplicity, speed, portability, adaptability (to changes in the data), and accuracy.

## Presentation


Presentation means presenting the results. Not all data science projects run continuously in the background, for instance to automatically buy stocks or predict the weather. Some are just ad-hoc analyses that need to be presented to decision makers, using Excel, Tableau and other tools. In some cases, the data scientist must work with business analysts to create dashboards, or to design alarm systems, with results from analysis e-mailed to selected people based on priority rules.

## Methodology

The process of building the machine learning algorithm model. The steps are elaborated as collecting data, importing data, data visualization, data processing, splitting dataset, model fitting, model evaluation and making prediction as follows :

### Collecting Data

Collecting data can be of Primary Source or Secondary Source. In primary sources, data is collected directly without any Third-party whereas, in Secondary sources, it takes the data



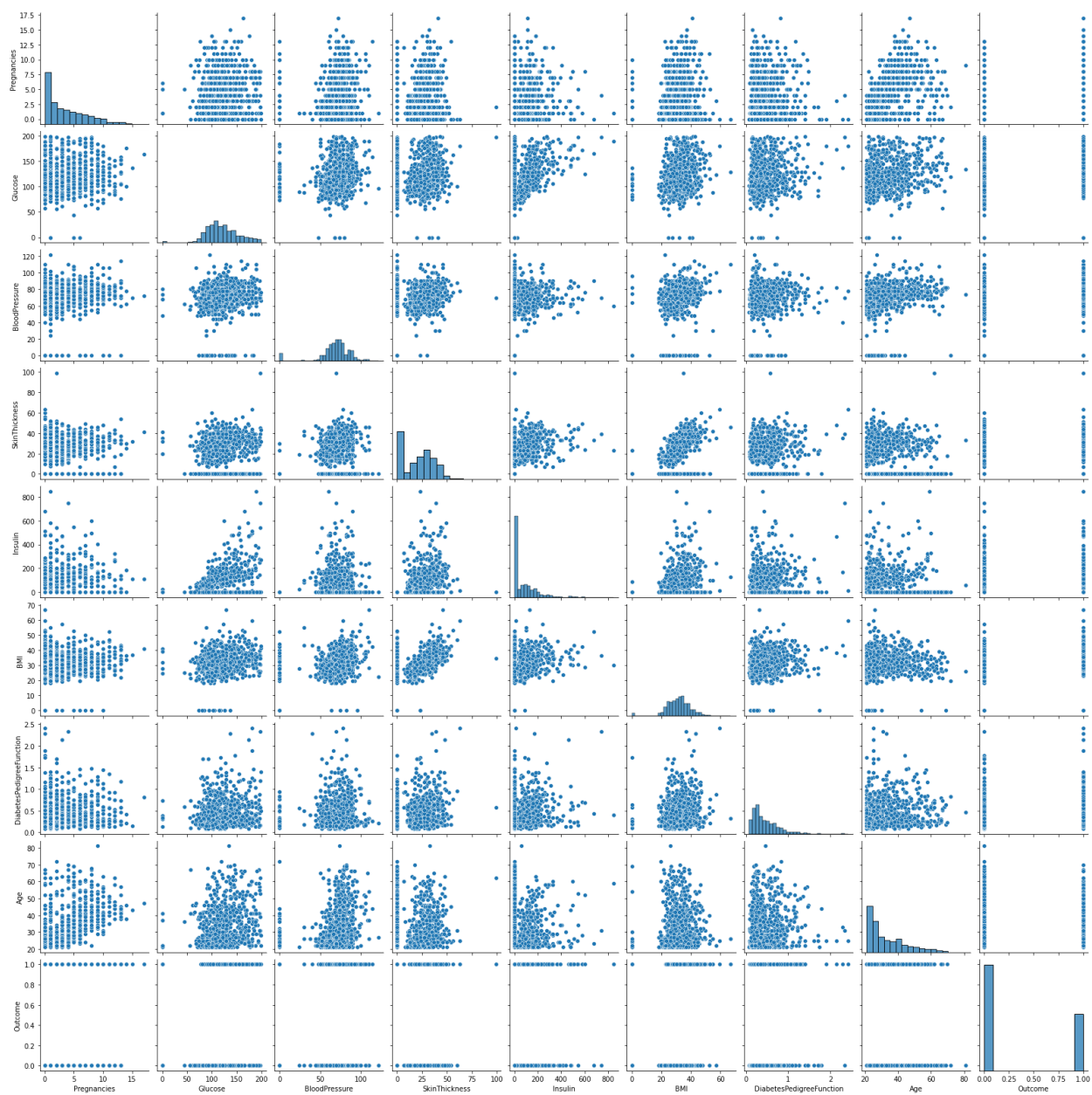
from the primary source. The dataset for this study has been downloaded from the Kaggle website and falls under secondary data. As it is a classification dataset, the dataset's features required for regression analysis have been separated from the original dataset.

## Importing Data

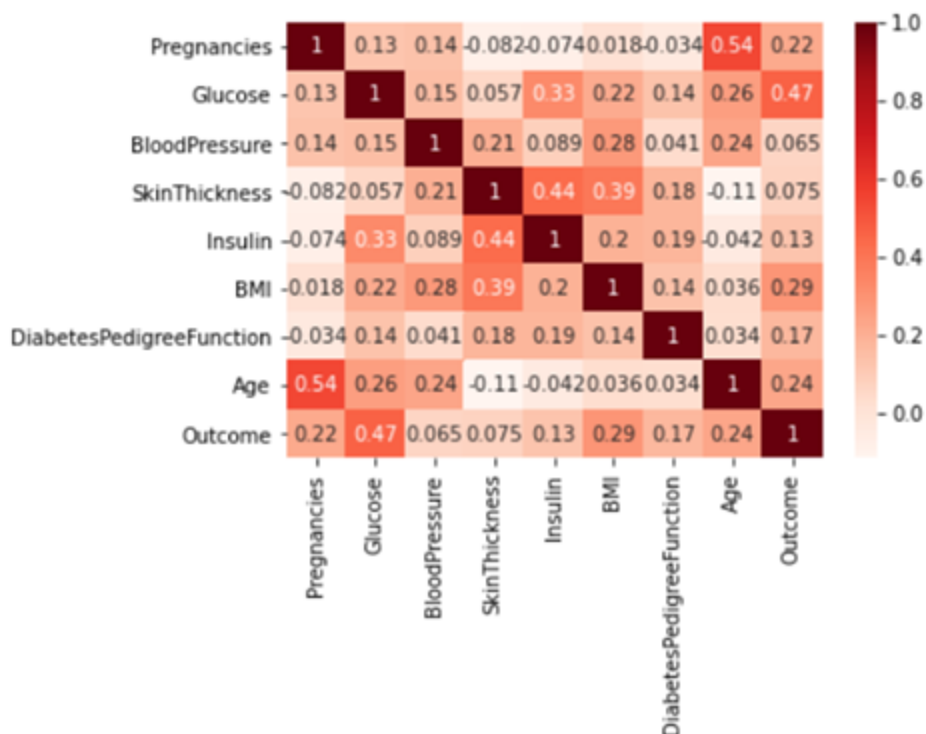
Locate the CSV file from the system and import it using the pandas library with the pre-defined function named `.read_csv("Filename.csv")`.

## Data Visualization

The main goal of data visualization is to make it easier to visualize the information and to identify patterns in the dataset. Charts and Graphs make the new data findings easier than the normal dataset.



Heat camp :




## Data Processing

Data processing refers to the technique of preparing the raw data into data that is suitable to build the model. Without this step, the model's accuracy may be reduced.

Missing values: Missing values can create problems. It is good if the missing value has been identified and replaced. These can be done with the mean/median method or by simply removing it but it may result in a dosage of data.

Feature Scaling: Feature Scaling is a technique used to normalize the range of features of the data. Feature Scaling brings all the values in the dataset to the same magnitude. In this work, Min-max normalization is one of the techniques to perform feature scaling, it rescales a feature value with a distribution value from 0 to 1.



Removing Outliers: An outlier is an observation in a dataset that is different from other observations in the feature. It may occur due to input error or data corruption.

The following features have been provided to help us predict whether a person is diabetic or not:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)<sup>2</sup>)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age:** Age (years)
- **Outcome:** Class variable (0 if non-diabetic, 1 if diabetic)



```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Pregnancies                          768 non-null    int64
1   Glucose                              768 non-null    int64
2   BloodPressure                        768 non-null    int64
3   SkinThickness                        768 non-null    int64
4   Insulin                              768 non-null    int64
5   BMI                                  768 non-null    float64
6   DiabetesPedigreeFunction             768 non-null    float64
7   Age                                  768 non-null    int64
8   Outcome                              768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB


```

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

There is no missing data. Data is cleaned before working on it. The above figures show the description and information of the dataset.

## Splitting Dataset

The dataset has been split into 2 separate sets namely the training set and test set. A training set is a subset that is used to train the model; on the other hand Test set is the one that is used to test the predictions of the model. Usually, the dataset splitting happens in the ratio of 70:30 or 80:20 which means 70% or 80% of the data is for training the model



and 30% or 20% of the data is for testing the model. The splitting can vary from dataset to dataset. In this work, the dataset has been divided into a ratio of 80% for the training dataset and 20% for the testing dataset.

## Model Fitting

The model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A well-fitted model produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is under fitted doesn't match closely enough.

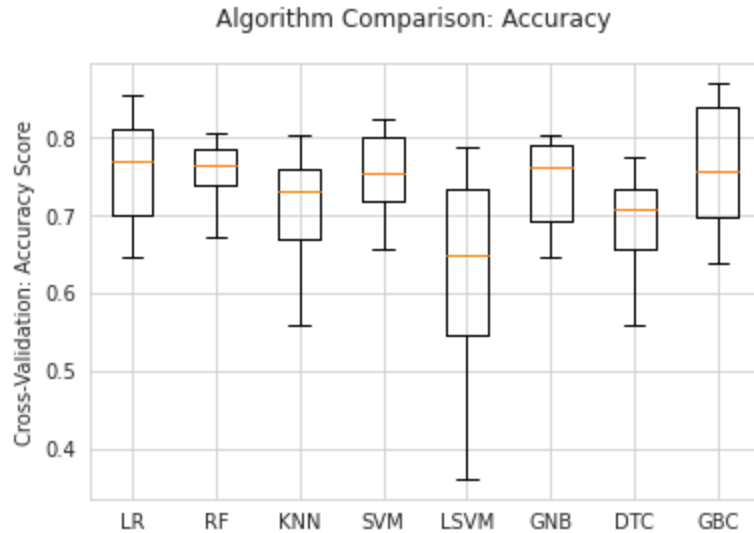
## Model Evaluation

In Model Evaluation, the model is evaluated using different performance metrics. A model with good performance metrics is considered the best deployment model.

## Prediction

The model is used for predicting the disease. Given the new input data, the model will predict the occurrence of the diabetics in a given record of the patient .

## Implementation



As per the above figure mentioned, Logistic regression (LR) outperforms in the accuracy score when compared to other algorithms like decision tree (DTC) and support vector machine (SVM) as we take it into consideration for this study. The accuracy score of the logistic regression algorithm is 83.12 %. The accuracy score of the decision tree algorithm is 75.32 %. The accuracy score of the support vector machine is 77.27 %.

	D	ND
D	100	7
N D	19	28

The above figure is the confusion matrix of the logistic regression algorithm.

The model correctly predicted those who have diabetes with a precision of 100 / 154. It also accurately predicted those who do not have diabetes with a precision of 28 / 154. The model wrongly interpreted the diabetic patients as non diabetic with a precision of 7 / 154 and also the non diabetic ones as diabetic patients with a precision of 19 / 154.

	D	ND
D	84	23
N D	15	32

The above figure is the confusion matrix of the decision tree algorithm. The model correctly predicted those who have diabetes with a precision of 84 / 154. It also accurately predicted those who do not have diabetes with a precision of 32 / 154. The model wrongly interpreted the diabetic patients as non diabetic with a precision of 23 / 154 and also the non diabetic ones as diabetic patients with a precision of 15 / 154.

	D	ND
D	91	9
N D	26	28

The above figure is the confusion matrix of the support vector machine algorithm. The model correctly predicted those who have diabetes with a precision of 91 / 154. It also accurately predicted those who do not have diabetes with a precision of 28 / 154. The model wrongly interpreted the diabetic patients as non diabetic with a precision of 9 / 154 and also the non diabetic ones as diabetic patients with a precision of 26 / 154.

Preferred algorithm : Logistic Regression > Support Vector Machine > Decision Tree

## Conclusion

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 83.12 % using the Logistic Regression algorithm. In future, the designed system with the used machine learning algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.