Project Report

on

Diabetes Prediction Using Machine Learning

Submitted to

Sant Gadge Baba Amravati University

In partial Fulfillment of the Requirement

For the Degree of

Bachelor of Engineering in

Computer Science and Engineering

Submitted by:

Prajakta Mathe (19)
Ashwini Ghate (04)
Aditi Dhote (01)
Vrushali Mange (35)
Pratiksha Patte (20)

Under the Guidance of
Prof. P. V. Deshmukh

Department of Computer Science and Engineering

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON – 444 203 (M.S.)

2022-23

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON – 444 203 (M.S.)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **Prajakta Mathe, Ashwini Ghate , Aditi Dhote, Vrushali Mange** and **Pratiksha Patte** students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute has completed the project work entitled "**Diabetes Prediction Using Machine Learning**" based on syllabus and has submitted a satisfactory account of his work in this report which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

Prof. P. V. Deshmukh
Project Guide

Dr. S. B. Patil
Head of Department

Dr. S. B. Somani
Principal

SHRI SANT GAJANAN MAHARAJ COLLEGE OF ENGINEERING,

SHEGAON – 444 203 (M.S.)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project work entitled "**Diabetes Prediction Using Machine Learning**" submitted by **Prajakta Mathe , Ashwini Ghate , Aditi Dhote, Vrushali Mange** and **Pratiksha Patte** students of final year B.E. in the year 2022-23 of Computer Science and Engineering Department of this institute, is a satisfactory account of his work based on syllabus which is recommended for the partial fulfillment of degree of Bachelor of Engineering in Computer Science and Engineering.

Internal Examiner
Date:

External Examiner
Date:

ABSTRACT

High levels of glucose in the bloodstream lead to the development of diabetes, which results in frequent urination, increased thirst, and increased hunger. It is crucial to address diabetes promptly as untreated cases may lead to severe complications in various body organs such as the heart, kidneys, blood pressure, and eyes. Predictive analytics over big data is a challenging task, particularly in healthcare. However, it can aid healthcare practitioners in making quick decisions about patients' health and treatment based on big data. The performance and accuracy of ML algorithms used in predictive Data analysis for predicting the occurrence of diabetes are compared and analyzed across various disciplines. In this study, different classification Computational methods, which may involve various algorithms, such as SVM, KNN, Logistic regression, and Random forest, were considered, and their performance metrics such as Recall, FN Measure, Precision, and Accuracy were evaluated Derived from the confusion matrix. According to the experimental results, the SVM and ontology classifiers yielded the highest accuracy for diabetes prediction

Acknowledgement

The real spirit of achieving a goal is through the way of excellence and lustrous discipline. We would have never succeeded in completing our task without the cooperation, encouragement and help provided to us by various personalities.

We would like to take this opportunity to express our heartfelt thanks to our guide ***Prof. P. V. Deshmukh****, for her esteemed guidance and encouragement, especially through difficult times. Her suggestions broaden our vision and guided us to succeed in this work. We are also very grateful for her guidance and comments while studying part of our seminar and learnt many things under her leadership.*

We extend our thanks to ***Dr. S. B. Patil*** *Head of Computer Science & Engineering Department, Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support that made us to perform consistently.*

We also extend my thanks to ***Dr. S. B. Somani,*** *Principal Shri Sant Gajanan Maharaj College of Engineering, Shegaon for their valuable support.*

Also We would like to thanks to all teaching and non-teaching staff of the department for their encouragement, cooperation and help. Our greatest thanks are to all who wished us success especially our parents, our friends whose support and care makes us stay on earth.

Ms. Prajakta Mathe
Ms. Ashwini Ghate
Ms.Aditi Dhote
Ms. Vrushali Mange
Ms.Pratiksha Patte

Final Year B. E. Sem-VIII, CSE
Session 2022-23

Contents

List of Figures

List of Tables

Abbreviations

ML	Machine Learning
DT	Decision Tree
LR	Logistic Regression
SVM	Support Vector Machine
RF	Random Forest

CHAPTER 1

INTRODUCTION

INTRODUCTION

1.1 PREFACE

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criteria to be taken into consideration for an individual to harbor this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage .

Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease.

Diabetes is a very familiar word in the present world and crucial challenges in both developed and developing countries . The insulin hormone in the body produced by the pancreas allows glucose to pass from the food into the bloodstream. The lack of that hormone due to malfunctioning of the pancreas forms diabetes which can result in coma, renal and retinal failure, pathological destruction of pancreatic beta cells, cardiovascular dysfunction, cerebral vascular dysfunction, peripheral vascular diseases, sexual dysfunction, joint failure, weight loss, ulcer, and pathogenic effects on immunity.

Diabetes is the third leading cause of death following diseases of heart and cancer. But with the rise of Machine Learning approaches, we have the ability to find a solution to this issue. The aim of Machine Learning and Data Mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. We are going to develop a Diabetes Diagnosis system using Machine Learning which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Machine Learning and Data mining has the ability to extract hidden

knowledge from a huge amount of diabetes-related data. This paper reviewed and analyzed the current studies on classification of Diabetes. Furthermore, the study has developed a classification model for diabetes using decision tree, Naïve Bayes, Support vector machine and k nearest neighbor Algorithm. The classification model is based on a dataset of 15000 cases collected from different National Institute of Diabetes and Digestive and Kidney Diseases. The results of the Naïve Bayes can be used by medical specialist to classify and diagnose diabetic patients. These results help the medical doctors in the classification process of diabetes. This study follows different machine learning algorithms to predict diabetes disease at an early stage. Such as, KNN, Naïve Bayes, Decision Tree, and Support Vector machine to predict this chronic disease at an early stage for safe human life.

1.2 BACKGROUND OF THE STUDY

Diabetes mellitus generally referred to as diabetes, is a metabolic condition causing excessive blood sugar levels (MSD Manual). The hormone insulin transfers sugar from the blood into the cells for storage or energy use. Diabetes means the body either doesn't produce enough insulin or can't use the insulin it produces effectively. Untreated high diabetes blood sugar can cause damage to your nerves, eyes, kidneys, and other organs.

Diabetes is a chronic condition that occurs either when not enough insulin is released by the pancreas, or when the body cannot use the insulin, it produces effectively. Insulin is a regulating hormone for blood sugar. Hyperglycemia, or high blood sugar, is a common effect of uncontrolled diabetes, resulting in severe damage to many of the body's systems, especially the nerves and blood vessels.

Diabetes is now a major issue in the world currently as the number of people with diabetes increases yearly and also causes the death of millions of people. It was discovered that there are majorly three types of diabetes which are type 1, type 2, and gestational diabetes.

Type 1 diabetes (previously called insulin-dependent, juvenile, or childhood-onset) is characterized by insulin production deficiency and requires daily insulin administration. The cause of type 1 diabetes is unclear, even despite current information, it is not preventable. The symptoms include excessive urinary excretion (polyuria), fatigue (polydipsia), persistent hunger, weight loss, changes in vision, and tiredness. These symptoms may occur suddenly.

Type 2 diabetes (formerly called non-insulin-dependent, or adult-onset) is the result of insufficient insulin use by the body. Type 2 diabetes is made up of most people with diabetes worldwide and is largely the result of excess body weight and physical inactivity.

Symptoms can be similar to those of type 1 diabetes, but sometimes are less pronounced. As a result, several years after the onset, once symptoms have already occurred, the disease will be diagnosed. Until recently, this form of diabetes was only seen in adults but now also occurs progressively in children.

Gestational diabetes is hyper glycemia with above-normal blood glucose levels but below the conditions of diabetes that arise during pregnancy. Women with gestational diabetes run a high risk of complications in pregnancy and childbirth. They and their children face a greater chance of type 2 diabetes in the future too. Machine learning is a data analytics tool that automates the building of analytical models (SAS.com). This is a branch of artificial intelligence focused on the premise that systems with minimal human input can learn from data, recognize trends, and make decisions.

Machine-learning algorithms use statistics in massive quantities of data to identify patterns. And data, here, involves a lot of things numbers, words, pictures, clicks, what do you have. This can be fed into a machine-learning algorithm if it can be digitally processed. An effective framework for knowledge required Machine learning has been used successfully by researchers for the prognosis and/or diagnosis of diabetes for active and accurate decision making. Therefore, this project focuses on the use of machine learning techniques on a set of data collected which is an online dataset to uncover hidden patterns and predict diabetes based on the dataset collected. Support Vector Machine and random forest are proposed for use in the prediction of diabetes in a patient to ensure that the information gotten from the system built based on these techniques are reliable.

1.3 PROBLEM STATEMENT

Diabetes is a most common disease caused by a group of metabolic disorders. It is also known as Diabetic mellitus. It affects the organs of the human body. It can be controlled by predicting this disease earlier. If diabetics patient is untreated for a long time, it may lead to increase blood sugar. Now a days, Healthcare industries generating large volume of data. Machine Learning algorithms and statistics are used

to predict the disease with the help of current and past data. Machine learning techniques helps the doctors to predict early stage for diabetics. Diabetics patient medical record and different types of algorithms are added in dataset for experimental analysis. we use logistic regression, random forest, decision tree classifier and gradient boosting to predict whether a patient has diabetes based on diagnostic measurements.

Performance and accuracy of the applied algorithms is discussed and compared.To predict the Diabetes Disease using Machine learning and study the proposed hypothesis that supervised ML algorithm can improve health care by the accurate and early detection of diseases. It affects the organs of the human body. It can be controlled by predicting this disease earlier. If diabetics patient is untreated for a long time, it may lead to increase blood sugar.

1.4 OBJECTIVES

This research work aims to analyze the Diabetes dataset, design, and implement a Diabetes prediction and recommendation system utilizing machine learning classification techniques. The specific objectives of this project work are:

i) To review existing literature along the area of diabetes diagnosis and prediction.

ii) Design and develop a model using machine learning techniques.

iii) To analyze the Diabetes dataset and use Support Vector Machine and Random forest algorithms to develop a prediction engine.

iv) To identify and discuss the benefits of the designed system along with effective applications.

1.5. SCOPE AND LIMITATIONS

1.5.1. SCOPE

1. It will Convert data into an appropriate form using various preprocessing techniques for the implementation of Machine Learning algorithms.
2. Observing and analyzing Dataset.
3. Predict and analyze prediction of Diabetes Disease.

1.5.2 . LIMITATIONS

The performance of these predictive models have shown different results

depending on the input variables, and the reproducibility of the prediction models is not guaranteed in not only established models but also other races and other populations.

1.6 ORGANIZATION OF PROJECT

Chapter 1 gives introduction to the project.

Chapter 2 provides literature survey of the project.

Chapter 3 explains materials and methods required to complete the project.

Chapter 4 Explain the working of the system .

Chapter 5 provides analysis of project.

Chapter 6 provides design phase of the project.

Chapter 7 provides how the project is implemented.

Chapter 8 provides result of the project.

Chapter 9 gives conclusion and Future work of the project .

CHAPTER 2

LITERATURE SURVEY

LITERATURE SURVEY

Arwatki Chen Lyngdoh et al. conducted research on predicting diabetes disease using 5 supervised ML Algo : KNN, Naive Bayes, Decision Tree Classifier, Random Forest, and SVM. by including current risk variables and performing cross-validation, they achieved consistent accuracy with the KNN classifier achieving a high accuracy of 76%. The main objective of the study was to identify the best outcomes for accurately predicting diabetes disease, considering accuracy and computing time.

Mitushi Soni et al. ML classification and ensemble techniques were employed to make predictions about diabetes using a dataset. They employed K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest algorithms, and found Random Forest outperformed the others in terms of accuracy.

Sivaranjani S et al. used SVM and Random Forest(RF) methods for identifying potential risks of Diabetes Related Diseases. After data preprocessing and implementing forward & backward stepwise feature selection was utilized to identify the most impactful features, they employed. Principle Component Analysis was employed to reduce dimensionality. Their study which outperformed Support Vector Machine's 81.4% accuracy.

Shejal Kale et al. applied ML Classification & Using ensemble techniques to make predictions about diabetes on a given dataset. They utilized KNN, Logistic Regression(LR), Decision Tree(DT) , SVM, Gradient Boosting(GB) , and Random Forest(RF) algorithms, and found that (RF) Random Forest had the best accuracy.

Ashwini R et al. trained ML ALGO such as KNN, Random Forest(RF), Logistic Regression(LR), and SVM using various datasets. They used preprocessing techniques to improve the accuracy of their models and prioritized risk factors by employing various feature selection approaches.

K.VijiyaKumar et al. proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly

Nonso Nnamoko et al. presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

Tejas N. Joshi et al. presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease.

Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

Muhammad Azeem Sarwar et al. proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

CHAPTER 3

METHODOLOGY

METHODOLOGY

This study aims to propose a new model for diabetic's classification. Numerous algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy.

The methods employed in this research are split by the four main phases of the research work, which are the problem formulation phase, the dataset collection phase, and the experimentation phase and the results summarizing. This research started with formulating the research problem that is reviewing of the literature and formulating of the research problem. After the research problem formulation, this research identified the scope of the research, the objectives, and limitations of the research procedure. The second phase of the study is the dataset collection. The dataset items were collected from National Institute of Diabetes and Digestive and Kidney Diseases.

The third phase of the study was the data preparation which included:

- Converting Data to Appropriate format
- Data Preprocessing
- Use Machine Learning to manipulate Data In the experimentation phase several experiments were conducted and results were collected.

3.1 DATASET

Dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. The dataset contains 768 records of female patients. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. Second is the value of Plasma glucose concentration a 2 hours in an oral glucose tolerance test and then is the Diastolic blood pressure (mm Hg), fourth in line is the Triceps skin fold thickness (mm), then is the 2-Hour serum insulin (mu U/ml), sixth is Body mass index (weight in kg/ (height in m) ^2) and then seventh is the Diabetes pedigree function and the second last value is the that of the Age (years). The ninth column is that of the Class variable (0 or 1), 0 for no diabetes

and 1 for the presence.

The Software used to visualize the entire dataset is “Jupyter Notebook” from Anaconda Navigator and the programming language used id Python 3.6

Table 3 1: Dataset Contents

Sr. Number	Attributes
I	Pregnancy Attribute
II	Glucose Attribute
III	Blood Pressure Attribute
IV	Skin thickness Attribute
V	Insulin Attribute
VI	Body Mask Index
VII	Diabetes Pedigree Function Attribute
VIII	Age Criteria

- **Pregnancies**: Those who develop gestational diabetes are at higher risk of developing type 2 diabetes later in life. The subjects with more number of pregnancies have a higher risk of developing diabetes.
- **Glucose**: The subjects were given an oral glucose test, whereby, they were administered glucose and a reading of their plasma glucose concentration was taken after 2 hours. The subjects with higher levels of glucose concentration after 2 hours have a higher risk of developing diabetes.
- **Blood pressure**: Having blood pressure over 140/90 mmHg of Mercury are linked to having increased risk of developing diabetes. Although, certain subjects having diastolic blood pressure 70 mmHg may develop diabetes.
- **Skin Thickness**: Skin thickness is primarily determined by collagen content and is increased in the case of insulin dependent diabetic patients. The subjects’ tricep skin fold were measured and results showed that having a skin thickness of 30mm of greater are at a higher risk.
- **Insulin**: Normal insulin levels after 2 hours of glucose administration is 16-166 mlU/L. Subjects having lower or higher levels than said value are at a higher risk.
- **Body Mass Index (BMI)**: Subjects having a BMI over 25 have a relatively

high risk in having diabetes.

- **Diabetes Pedigree Function**: The diabetes pedigree function provides "a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject." The higher the DPF, the more likely it is for a subject to be diabetic.
- **Age:** Diabetes is prevalent in any age group, but is commonly found in middle aged adults (45 onwards). Taking that into consideration, subjects with in the higher age group have a higher expectancy of diabetes.

1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Figure 3.1 Pima Indians Dataset

Distribution of Diabetic patient- In our attempt to develop a diabetes prediction model, we encountered a slightly imbalanced dataset. Out of the total 768 samples, around 500 were Designated as 0, denoting the nonexistence of diabetes., while 268 were designated as 1, denoting the existence of diabetes.

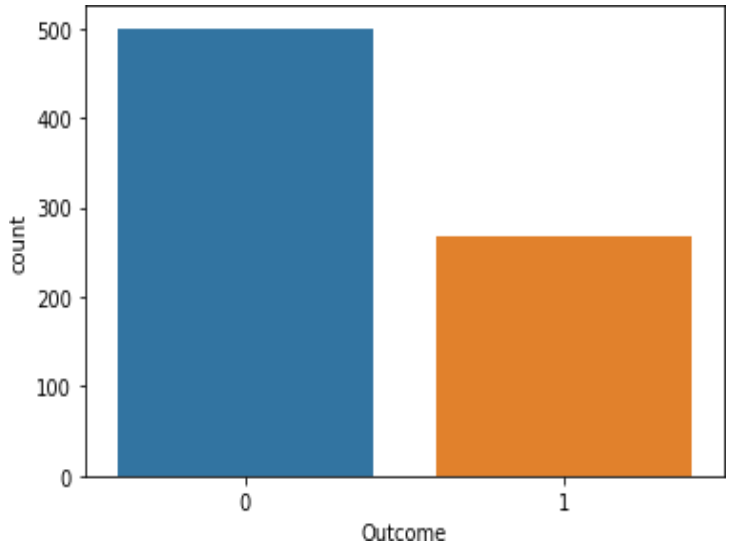

Figure 3.2 The proportion of patients with diabetes compared to those without diabetes

3.2 DATA PRE-PROCESSING

The process of preprocessing data is of utmost importance, especially for data concern with healthcare, which may contain missing values and other contaminants that may affect the effectiveness of data mining. This process is essential to achieve accurate results and successful predictions with the help of ML methodology on the CSV file. To work with the Pima Indian diabetes dataset, we require, preprocessing in couple of steps.The dataset, which is quoted above, has lapsed and have shed data. To make the dataset serviceable and obtain the knowledge from it, we have performed data preprocessing. In order to handle erroneous data, we have analyzed the dataset for the unusual entries and fixed them manually. Missing values are handled with the help of calculating the standard deviation of that particular feature and allotting it to the missing spaces. To make the dataset useful, we have used Pandas and NumPy library for handling the dataset efficiently and easy data handling throughout the experiment

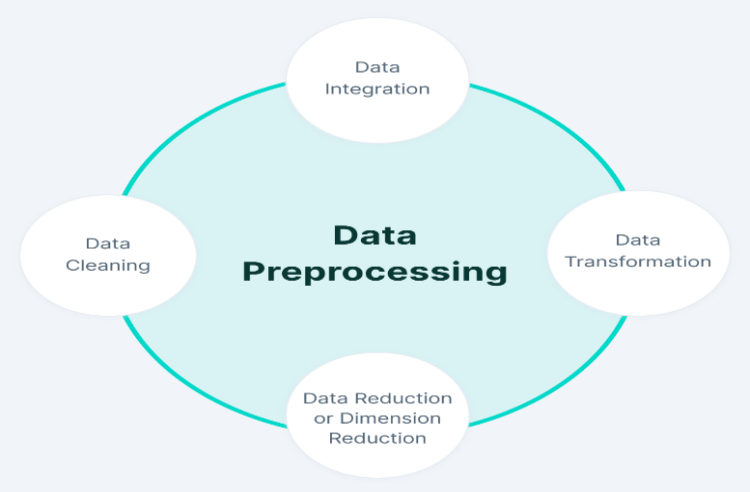

Figure3.3 Data Pre-processing

3.3 MISSING VALUE ELIMINATION

A value of zero are removed since it is not possible to have a value of zero for certain features. This process helps in feature subset selection by eliminating irrelevant features or instances, reducing the dimensionality of the data and enabling faster processing.

3.4 SPLITTING DATA

After cleaning the data, it is normalized and split into training and testing sets. The algorithm is trained on the training dataset, and the test dataset is kept aside. This training process produces a model based on logic, algorithms, and feature values in the training data. Normalization is used to bring all attributes to the same scale.

The split the modelling dataset into training and testing sets is to assign 2/3 data points to the former and the remaining one-third to the latter. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we can evaluate the performance of our model. For instance, if the training accuracy is extremely high while the testing accuracy is poor then this is a good indicator that the model is probably over fitted.

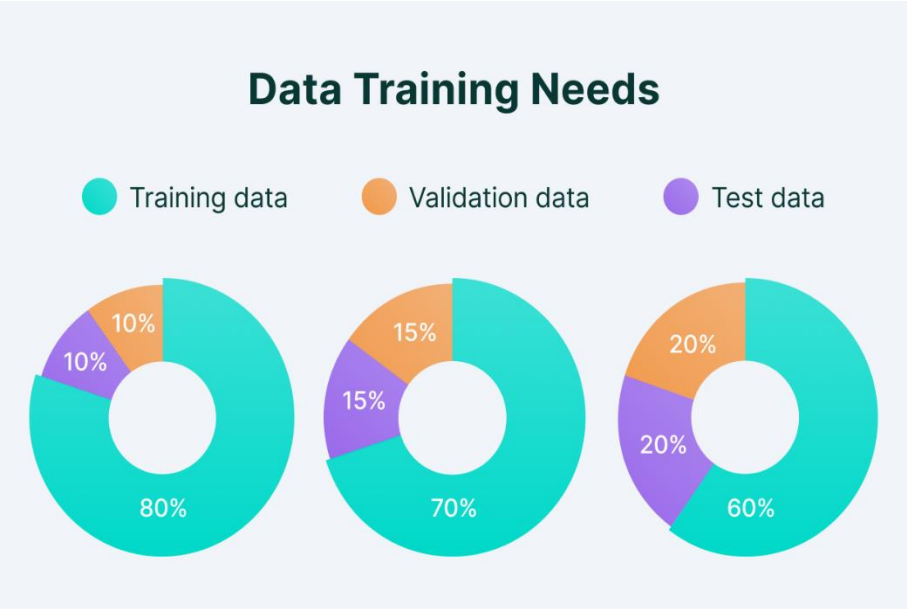

Figure 3.4 Training and Testing data

CHAPTER 4
WORKING OF SYSTEM

WORKING OF SYSTEM

4.1 SYSTEM ARCHITECTURE

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. The algorithms like K nearest neighbor, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

The presence of disease has been identified using the appearance of various symptoms. However, the methods use different features and produces varying accuracy. The result of prediction differs with the methods/measures/ features being used. Towards diabetic prediction, a Disease Influence Measure (DIM) based diabetic prediction has been presented. The method pre-processes the input data set and removes the noisy records. In the second stage, the method estimates disease influence measure (DIM) based on the features of input data point. Based on the DIM value, the method performs diabetic prediction. Different approaches of disease prediction have been considered and their performance in disease prediction has been compared. The analysis result has been presented in detail towards the development.

The first step involves importing the necessary libraries and loading the diabetes dataset. In step two, the data is pre-processed to eliminate missing values. Step three involves splitting the dataset into training and test sets using an 80-20 percentage split. Next, the machine learning algorithm we use four algorithms like Logistic regression, Support vector machine, Random forest, KNN, is selected in step four. Step five involves building the classifier model using the training set. The classifier model is then tested using the test set in step six. In step seven, a comparing and evaluating the performance results of each classifier is carried out. Finally, in step eight, after analyzing the results based on various measures, the best performing algorithm is concluded.

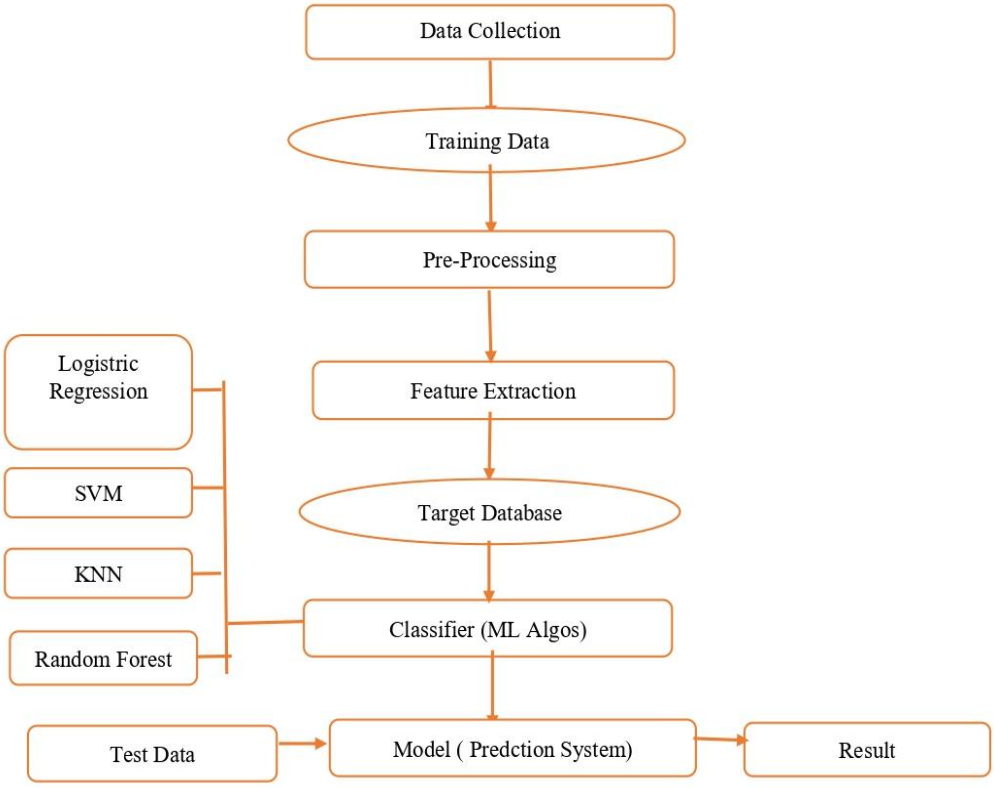

Figure 4.1 Diabetes Prediction Flow Diagram

4.2 MACHINE LEARNING

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as: Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.

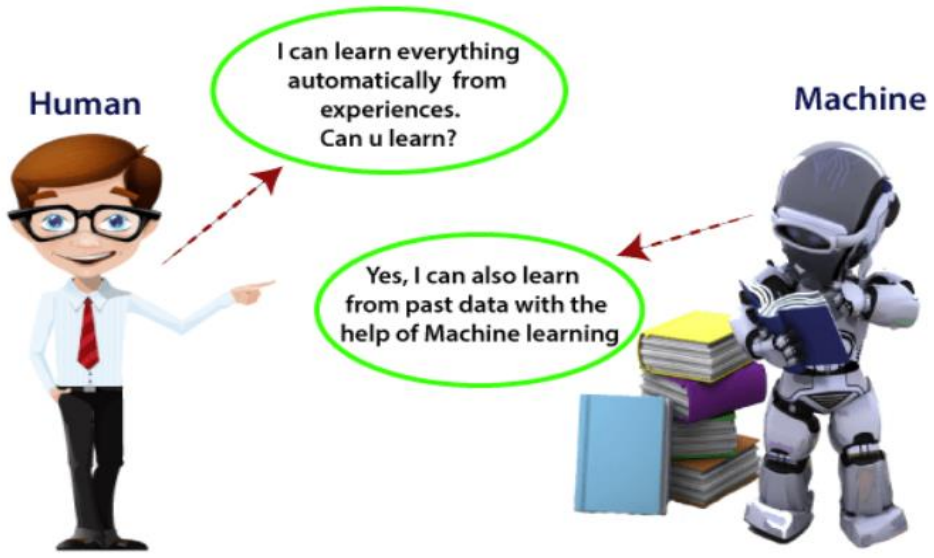

Figure 4.2 Machine Learning

4.2.1. SUPERVISED MACHINE LEARNING

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset. The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).

Categories of Supervised Machine Learning:

Supervised machine learning can be classified into two types of problems, which are given below:

a) Classification
b) Regression

a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc.

The classification algorithms predict the categories present in the dataset.

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- Simple Linear Regression Algorithm
- Multivariate Regression Algorithm
- Decision Tree Algorithm
- Lasso Regression

4.2.2. UNSUPERVISED MACHINE LEARNING

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision. The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

4.3 MACHINE LEARNING ALGORITHMS

We have also use various algorithms in our project to train the model and they are following:

1) K-Nearest Neighbors.
2) Logistic Regression.
3) Random Forest.
4) Support Vector Machine.

4.3.1 K-NEAREST NEIGHBORS:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category

that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So, for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dog's images and based on the most similar features it will put it in either cat or dog category.

KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique. KNN assumes that similar things are near to each other. Many times, data points which are similar are very near to each other. KNN helps to group new work based on similarity measure. KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure.

To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it's nearest neighbors. Here K= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e., P (p1,p2, …. pn) and Q (q1, q2,..qn) is defined by the following equation:-

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

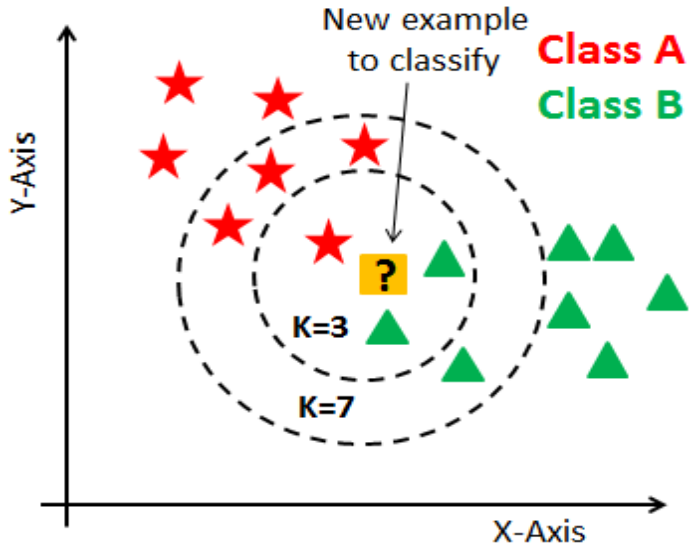

Figure 4.3 K-Nearest Neighbors

ALGORITHM:

- Take a sample dataset of columns and rows named as Pima Indian Diabetes data set.
- Take a test dataset of attributes and rows.
- Find the Euclidean distance by the help of formula :-

$$EculideanDistance = \sqrt{\sum_{i=1}^{y}\sum_{j=1}^{m}\sum_{l=1}^{n-1}(R_{(j,l)} - P_{(i,l)})^2}$$

- Then, decide a random value of K. is the no. of nearest neighbors.

- Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
- Find out the same output values.

  If the values are same, then the patient is diabetic, otherwise not.

ADVANTAGE OF KNN :

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

4.3.2 .LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

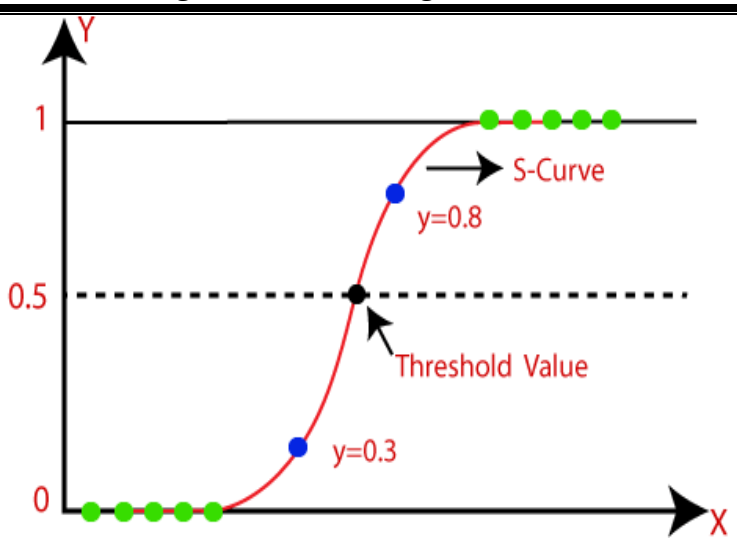

Figure 4.4 Logistic Regression

LR is a type of regression analysis. So, before we delve into logistic regression, let us first introduce the general concept of regression analysis. Regression analysis is a type of predictive modeling technique which is used to find the relationship between a dependent variable (usually known as the "Y" variable) and either one independent variable (the "X" variable) or a series of independent variables. When two or more independent variables are used to predict or explain the outcome of the dependent variable, this is known as multiple regression.

Regression analysis can be used for three things:

1. **Forecasting the effects or impact of specific changes**. For example, if a manufacturing company wants to forecast how many units of a particular product, they need to produce in order to meet the current demand.
2. **Forecasting trends and future values.** For example, how much will the stock price of Lufthansa be in 6 months from now?
3. **Determining the strength of different predictors**—or, in other words, assessing how much of an impact the independent variable(s) has on a dependent variable.

For example, if a soft drinks company is sponsoring a football match, they might want to determine if the ads being displayed during the match have accounted for any increase in sales.

Regression analysis can be broadly classified into two types: Linear regression and logistic regression.

In statistics, linear regression is usually used for predictive analysis. It essentially determines the extent to which there is a linear relationship between a dependent variable and one or more independent variables. In terms of output, linear regression will give you a trend line plotted amongst a set of data points. You might use linear regression if you wanted to predict the sales of a company based on the cost spent on online advertisements, or if you wanted to see how the change in the GDP might affect the stock price of a company.

The second type of regression analysis is logistic regression, and that's what we'll be focusing on in this post. Logistic regression is essentially used to calculate (or predict) the probability of a binary (yes/no) event occurring. We'll explain what exactly logistic regression is and how it's used in the next section.

4.3.3 RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

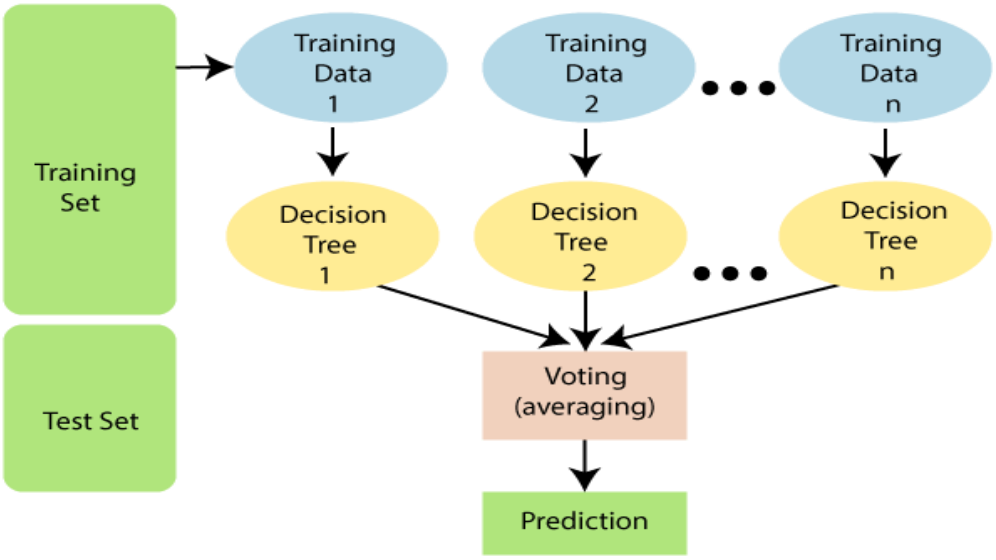

Figure 4.5 Random Forest

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

ALGORITHM

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

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.

Decision tree:

Since the random forest model is made up of multiple decision trees, it would be helpful to start by describing the decision tree algorithm briefly. Decision trees start with a basic question, such as, “Should I surf?” From there, you can ask a series of questions to determine an answer, such as, “Is it a long period swell?” or “Is the wind blowing offshore?”. These questions make up the decision nodes in the tree, acting as a means to split the data. Each question helps an individual to arrive at a final decision, which would be denoted by the leaf node. Observations that fit the criteria will follow the “Yes” branch and those that don’t will follow the alternate path. Decision trees seek to find the best split to subset the data, and they are typically trained through the Classification and Regression Tree (CART) algorithm. Metrics, such as Gini impurity, information gain, or mean square error (MSE), can be used to evaluate the quality of the split.

This decision tree is an example of a classification problem, where the class labels are "surf" and "don't surf. “While decision trees are common supervised learning algorithms, they can be prone to problems, such as bias and overfitting. However, when multiple decision trees form an ensemble in the random forest algorithm, they predict more accurate results, particularly when the individual trees are uncorrelated with each other.

Ensemble methods :

Ensemble learning methods are made up of a set of classifiers e.g. decision trees and their predictions are aggregated to identify the most popular result. The most well-known ensemble methods are bagging, also known as bootstrap aggregation, and boosting. In 1996, Leo Breiman (link resides outside ibm.com) (PDF, 810 KB) introduced the bagging method; in this method, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

After several data samples are generated, these models are then trained independently, and depending on the type of task—i.e., regression or classification—the average or majority of those predictions yield a more accurate estimate. This approach is commonly used to reduce variance within a noisy dataset.

4.3.4 SUPPORT VECTOR MACHINE

A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

An SVM algorithm should not only place objects into categories, but have the margins between them on a graph as wide as possible.

Some applications of SVM include:

- Text and hypertext classification
- Image classification
- Recognizing handwritten characters
- Biological sciences, including protein classification

It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes. Note: Don't get confused between SVM and logistic regression. Both the algorithms try to find the best hyperplane, but the main difference is logistic regression is a probabilistic approach whereas support vector machine is based on statistical approaches.

Now the question is which hyperplane does it select? There can be an infinite number of hyperplanes passing through a point and classifying the two classes perfectly. So, which one is the best?

Well, SVM does this by finding the maximum margin between the hyperplanes that means maximum distances between the two classes.

TYPES OF SUPPORT VECTOR MACHINE

Linear SVM

When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line(if 2D).

Non-Linear SVM

When the data is not linearly separable then we can use Non-Linear SVM, which means when the data points cannot be separated into 2 classes by using a

straight line (if 2D) then we use some advanced techniques like kernel tricks to classify them. In most real-world applications we do not find linearly separable datapoints hence we use kernel trick to solve them.

Now let's define two main terms which will be repeated again and again in this article:

Support Vectors: These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.

Margin: it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins hard margin and soft margin. I will talk more about these two in the later section.

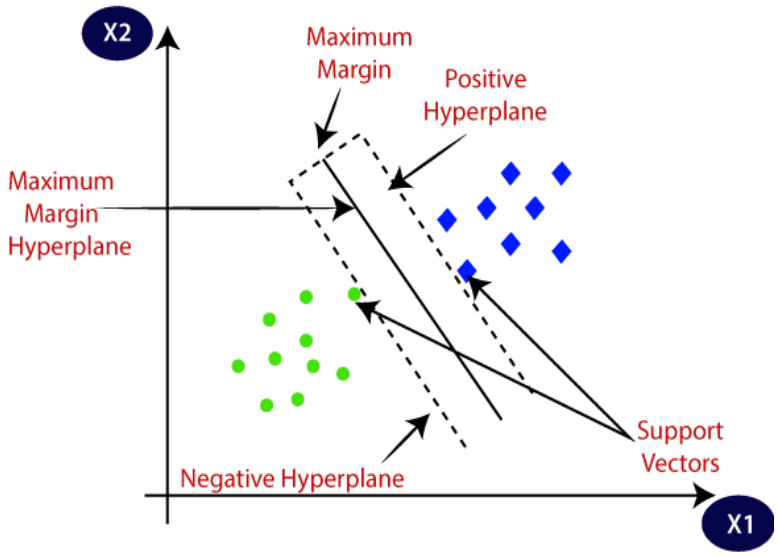

Figure 4.6 Support Vector Machine

CHAPTER 5

ANALYSIS

ANALYSIS

5.1 SYSTEM CONFIGURATION

- Hardware requirements: Processer : Any Update Processer
- Ram : Min 4GB Hard Disk : Min 100GB
- Software requirements: Operating System
- Windows family Technology : Python3.6
- IDE : Jupiter notebook

Sample Code:-

5.2 DATASET DETAILS

Dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. The dataset contains 768 records of female patients.

1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Figure 5.1 Dataset Attributes

Input dataset attributes

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- Body Mass Index(BMI)
- Diabetes Pedigree Function
- Age

5.3 PERFORMANCE ANALYSIS

In this project, various machine learning algorithms like SVM, Decision Tree, Random Forest, Logistic a used to predict diabetes. Diabetes Prediction UCI dataset, has a total of 9 attributes, out of those only 9 attributes are considered for the prediction of Diabetes Prediction. Various attributes of the patient like Pregnancies, Glucose , Blood Pressure, Skin Thickness , Insulin etc. are considered for this project. The accuracy for individual algorithms has to measure and whichever algorithm is giving the best accuracy, that is considered for the diabetes prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset. It is expressed as:

Accuracy = (TP + TN) /(TP+FP+FN+TN)

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system

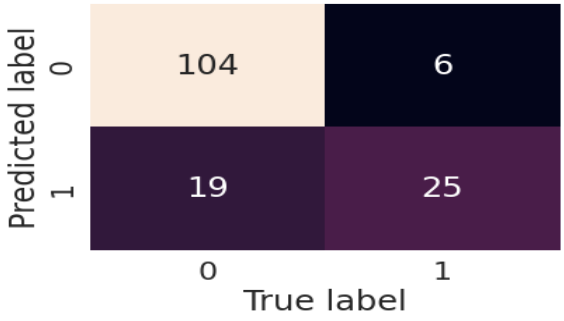

Figure 5.2 Confusion Matrix

Where TP: True positive

FP: False Positive

FN: False Negative

TN: True Negative

Correlation Matrix: The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.

Table 5.2 . Dataset

Sr.no	Attribute	Description	Type
1.	Pregnancies	Number of times pregnant	Numeric
2.	Plasma Glucose	Plasma glucose concentration of 2 hours in an oral glucose tolerance test.	Numeric
3.	Diastolic Blood Pressure	Diastolic Blood Pressure in mmHg	Numeric
4.	Triceps Thickness	Triceps Skin Fold Thickness measured in mm	Numeric
5.	SerumInsulin	2-Hour serum insulin measured in µU/ml	Numeric
6.	BMI	Body Mass Calculated using: $Weight\ in\ kg(height\ in\ meter)2$	Numeric
7.	Diabetes Pedigree	Diabetic Pedigree function – how likely the person is to have given their family history and other factors.	Numeric
8.	Age	Age of the Patient in years.	Numeric

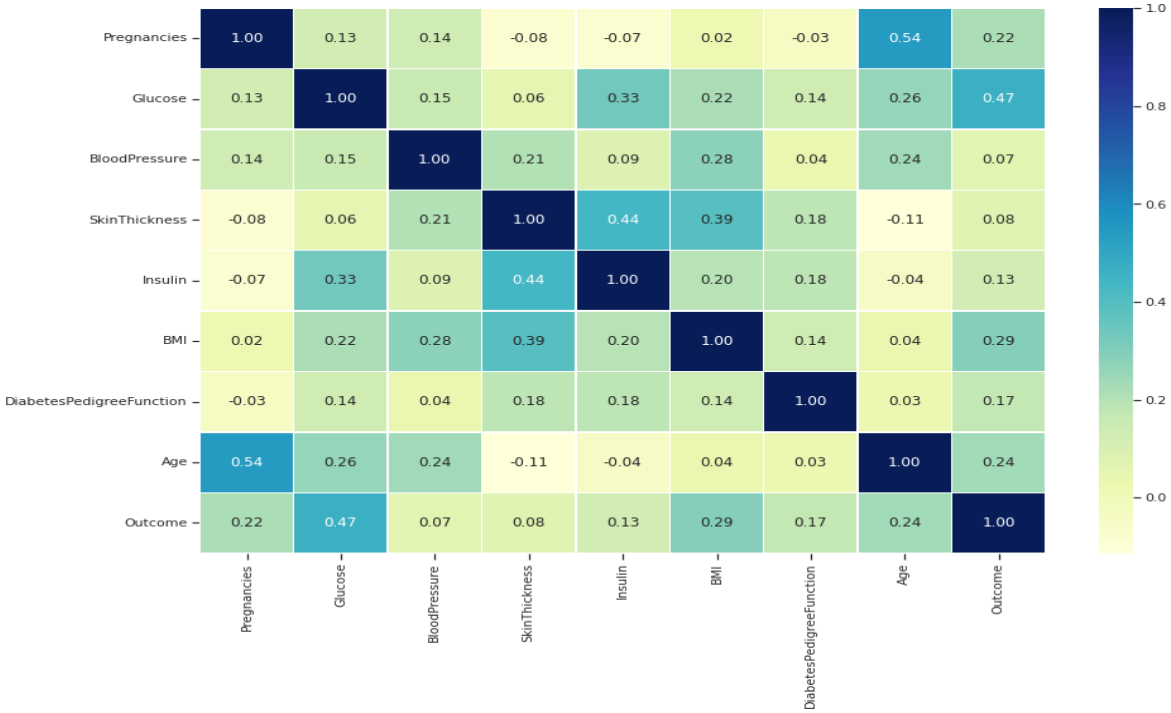

Figure 5.3 Correlation Matrix

Precision- It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: Recall-It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: F1 Score-It is the harmonic mean of Precision and Recall. It measures the test .

Chapter 6

DESIGN

DESIGN

6.1 DESIGN GOAL:

Design is a meaningful engineering representation of something that is to be built. It can be traced to a customer's requirements and at the same time assessed for quality against a set of predefined criteria for good design. In the software engineering context, design focuses on four major areas of concern: data, architecture, interfaces, and components. The design process translate requirement into representation of software that can be accessed for a quality before core generation. Design is the process through which requirement are translated to blue print for constructing into software. Initially the blueprint depicts the holistic view of software. This is the design represented at the high level of abstraction. During various stages of system development and design following goals have been setup for a complete architecture.

- Analysis
- Design
- Development
- Testing
- Implementation

6.2 DESIGN STRATEGY

System design is the process of planning a new system or to replace the existing system. Simply, system design is like the blueprint for building, it specifies all the features that are to be in the finished product. System design phase follows system analysis phase. Design is concerned with identifying functions, data streams among those functions, maintaining a record of the design decisions and providing a blueprint for the implementation phase. Design is the bridge between system analysis and system implementation.

Some of the essential fundamental concepts involved in the design of application software are:

6.2.1ABSTRACTION

Abstraction is used to construct solutions to problems without having to take account of the intricate details of the various component sub problems. Abstraction allows system designers to make stepwise refinement, which at each stage of design may hide unnecessary details associated with representation or implementation from the surrounding environment.

6.2.2 MODULARITY

Modularity is concerned with decomposing the main module into well-defined manageable units with well-defined interfaces among the units. This enhances design clarity, which in turn eases implementation, Debugging, Testing, Documenting and Maintenance of the software product. Modularity viewed in this sense is a vital tool in the construction of large software projects.

6.2.3 VERIFICATION

Verification is a fundamental concept in software design. A design is verifiable if it can be demonstrated that the design will result in implementation that satisfies the customer's requirements. Verification is of two types namely.

Verification that the software requirements analysis satisfies the customer's needs. Verification that the design satisfies the requirement analysis.

Some of the important factors of quality that are to be considered in the design of application software are:

- Reliability: The software should behave strictly according to the original specification and should function smoothly under normal conditions.
- Extensibility: The software should be capable of adapting easily to changes in the specification.
- Reusability: The software should be developed using a modular approach, which permits modules to be reused by other applications, if possible.

The System Design briefly describes the concept of system design and it contains four sections. The first section briefly describes the features that the system is going to provide to the user and the outputs that the proposed system is going to offer.

The second section namely Logical Design describes the Data Flow Diagrams, which show clearly the data movements, the processes and the data sources, and sinks, E-R diagrams which represent the overall logical design of the database, and high-level process structure of the system. The process of design involves "conceiving and planning out in the mind" and making a drawing pattern, or sketch of the system.

In software design there are two types of major activities, Conceptual Design and Detailed Design. Conceptual or logical or external design of software involves conceiving, planning out, and specifying the externally observable characteristics of a software product. These characteristics include user displays, external data sources, functional characteristics and high-level process structure for the product

Details or internal design involves conceiving, planning out, and specifying the internal structure and processing details of the software product. The goal of internal design is to specify internal structure, processing details, blueprint of implementation, testing, and maintenance activities.

One of the important fundamental concepts of software design is modularity. A modularity system consists of interfaces among the units. Modularity enhances design clarity, which in turn eases implementation, debugging, testing, documentation, and maintenance of the software product. The other fundamental concepts of software design include abstraction, structure, information hiding, concurrency and verification. The use of structuring permits decomposition of a large system into smaller, more manageable units with well-defined relationships to the other units. The system design is verifiable if it can be demonstrated that the design will result in an implementation that satisfies the customer's requirements.

Preliminary Design:

Preliminary design is basically concerned with deriving an overall picture of the system. Deriving the entire system into modules and sub-modules while keeping Cohesion and Coupling factors in mind. Tools, which assist in the preliminary design process, are Data Flow Diagrams.

Code design:

The purpose of code is to facilitate the identification and retrieval for items of information. A code is an ordered collection of symbols designed to provide unique

identification of an entity or attribute. To achieve unique identification there must be only one place where the identified entity or the attribute can be entered in the code; conversely there must be a place in the code for everything that is to be identified. This mutually exclusive feature must be built into any coding system.

The codes for this system are designed with two features in mind. Optimum human oriented use and machine efficiency. Length of the code range from length of one to length of five characteristics:

The code structure is unique; ensuring that only one value of the code with a single meaning may be correctly applied to a given entity or attribute.

- The code structure is extensible allowing for growth of its set of entities and attributes.
- The code is concise and brief for recording, communication, and transmission and storage efficiencies.
- They have a uniform size and format.
- The codes are simple so that the user can easily understand it.
- The codes are also versatile i.e., it is easy to modify to reflect necessary changes in condition, chart eristic and relationships of the encode entities.
- The codes are also easily storable for producing reports in a predetermined order of format.

The codes are also stable and do not require being frequently updated thereby promoting user efficient.

6.3 MODULE DIAGRAM

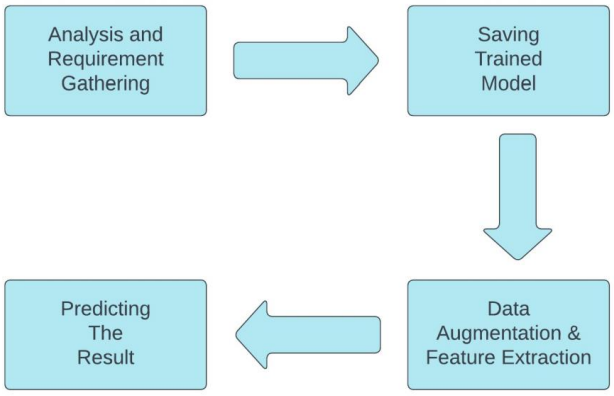

Figure 6.1 Module Diagram of System

6.4 STATE DIAGRAM

The name of the diagram itself clarifies the purpose of the diagram and other details. It describes different states of a component in a system. The states are specific to a component/object of a system. A State Chart diagram describes a state machine. State machine can be defined as a machine which defines different states of an object and these states are controlled by external or internal events. As State Chart diagram defines the states, it is used to model the lifetime of an object. State Chart diagrams are also used for forward and reverse engineering of a system. State Chart diagram is one of the five UML diagrams used to model the dynamic nature of a system. They define different states of an object during its lifetime and these states are changed by events. State Chart diagrams are useful to model the reactive systems. However, the main purpose is to model the reactive system. Following are the main purposes of using State Chart diagrams.

- To model the dynamic aspect of a system.
- To model the life time of a reactive system.
- To describe different states of an object during its life time.
- Define a state machine to model the states of an object.

State diagrams are used to give an abstract description of the behavior of a system. This behavior is analyzed and represented as a series of events that can occur in one or more possible states. State diagrams can be used to graphically represent finite state machines. Classic state diagrams require the creation of distinct nodes for every valid combination of parameters that define the state. This can lead to a very large number of nodes and transitions between nodes for all but the simplest of systems. This complexity reduces the readability of the state diagram.

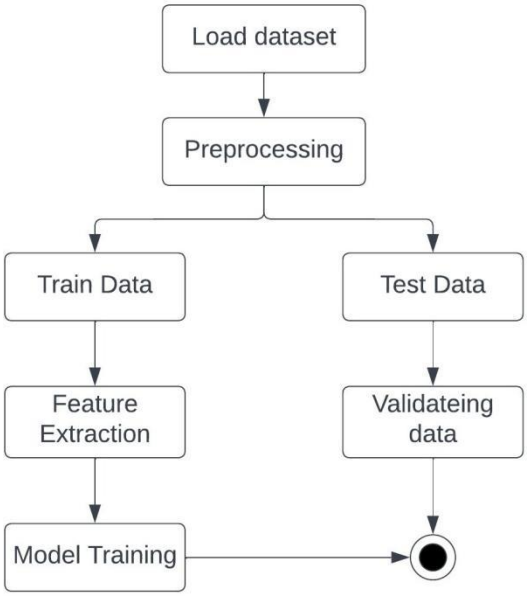

Figure 6.2 State Diagram of System

Chapter 7

IMPLEMENTATION

IMPLEMENTATION

7.1 HARDWARE PLATFORM USED:

The hardware requirement may serve as the basis for a contract for the implementation of the system and should therefore be complete and consistent in specification.

The hardware used for the system is mentioned below.

- PROCESSOR: Intel CORE i3 or above
- RAM: minimum 4.00GB
- HARD DISK: minimum 100GB

It should be noted that better the hardware facilities available, higher would-be response time of the system.

7.2 LIBRARIES AND SOFTWARE PLATFORM USED:

The software requirement document is the specification of the system. The software requirement provides a basis for creating the software requirements specification.

OPERATING SYSTEM: Windows

SYSTEM TYPE: 64-bit, intel CORE i5

SOFTWARE: Jupyter Notebook, VS Code, Anaconda

TECHNOLOGIES: Python

LIBRARIES: Flask, pandas, NumPy, pickle, sklearn, xgboost, etc

7.2.1 WINDOWS:

Windows is a proprietary operating system developed by Microsoft Corporation. It is a widely used operating system that is compatible with a wide range of software and hardware devices. Windows has a user-friendly interface and is designed to be easy to use for both personal and business use. It supports multiple languages and has built-in security features, such as Windows Firewall. Here we are using Windows OS to run the project.

7.2.2 FLASK:

Flask is a micro web framework written in Python. It is classified as a

microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

7.2.3 PYTHON:

Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant White space. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

7.2.4 SKLEARN:

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

7.2.5 TESTING:

System testing is the stage before system implementation where the system is made error free and all the needed modifications are made. The system was tested with test data and necessary corrections to the system were carried out. All the reports were checked by the user and approved. The system was very user friendly with online help to assist the user wherever necessary.

Test Plan

A test plan is a general document for the entire project, which defines the scope, approach to be taken, and schedule of testing, as well as identifying the test

item for the entire testing process, and the personnel responsible for the different activities of testing. This document describes the plan for testing, the knowledge management tool. A Test Plan is a detailed document that describes the test strategy, objectives, schedule, estimation, deliverables, and resources required to perform testing for a software product. Test Plan helps us determine the effort needed to validate the quality of the application under test. The test plan serves as a blueprint to conduct software testing activities as a defined process, which is minutely monitored and controlled by the test manager.

What is the Importance of Test Plan?

- Making Test Plan document has multiple benefits:
- Help people outside the test team such as developers, business managers, customers understand the details of testing.
- Test Plan guides our thinking. It is like a rule book, which needs to be followed.

Major testing activities are:

- Test units
- Features to be tested
- Approach for testing

Test units:

Test Case specification is a major activity in the testing process. In this project, we have performed two levels of testing.

- Unit testing
- System testing

The basic units in Unit testing are:

- Validating the user request
- Validating the input given by the user
- Exception handling

The basic units in System testing are:

- Integration of all programs is correct or not
- Checking whether the entire system after integrating is working as expected.
- The system is tested as a whole after the unit testing.

Analyze the product:

How can you test a product without any information about it? The answer is Impossible. You must learn a product thoroughly before testing it.

The product under test is educational website/system. You should research clients and the end users to know their needs and expectations from the application Who will use the website/system?

- What is it used for?
- How will it work?
- What are the software/ hardware the product uses?

CHAPTER 8
RESULT & DISCUSSION

RESULT & DISCUSSION

The table below displays the performance values of various classification algorithms, calculated using different measures. Based on the table, it is observed that Logistic regression exhibits the highest accuracy. Therefore, the Logistic regression machine learning classifier is capable of predicting the likelihood of diabetes with greater precision than other classifiers.

Table 8.3 . Accuracy Measures

Classification Algorithm	Precision
Logistic regression(LR)	0.83
Support Vector Machine(SVM)	0.76
Random Forest (RF)	0.78
KNN	0.78

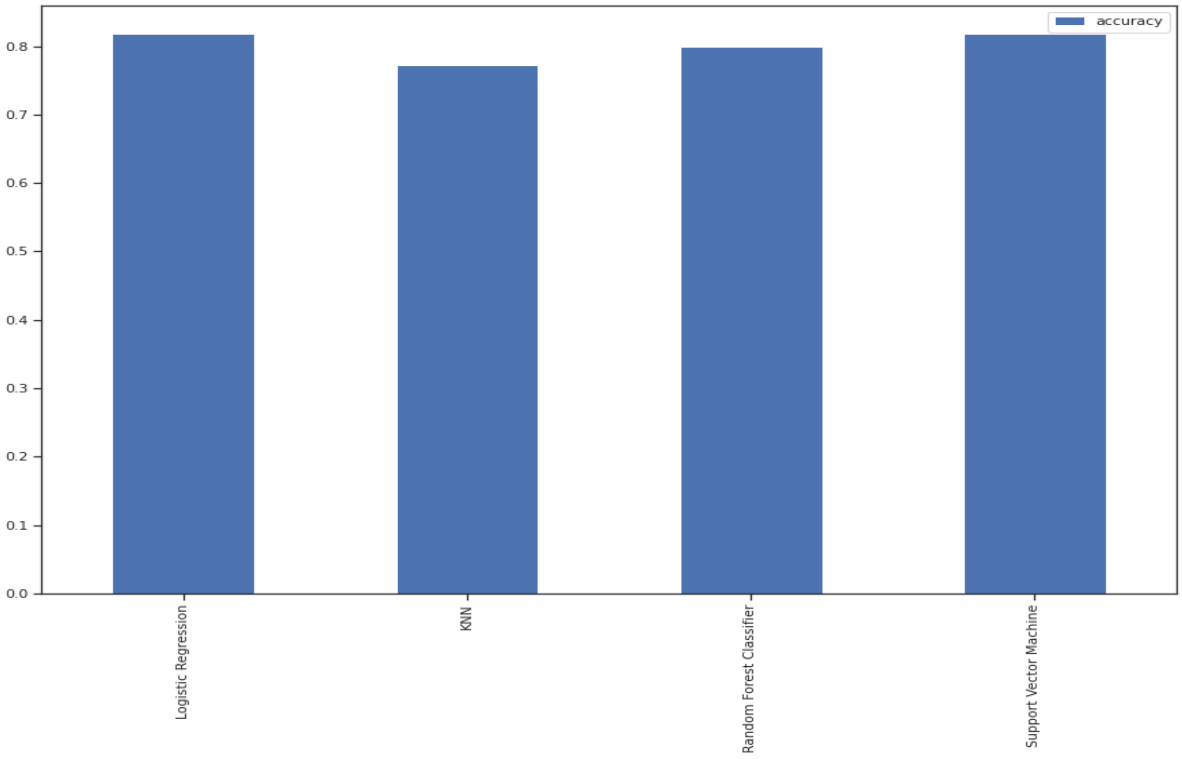

Figure 8.1 Results of the Accuracy achieved by machine learning techniques

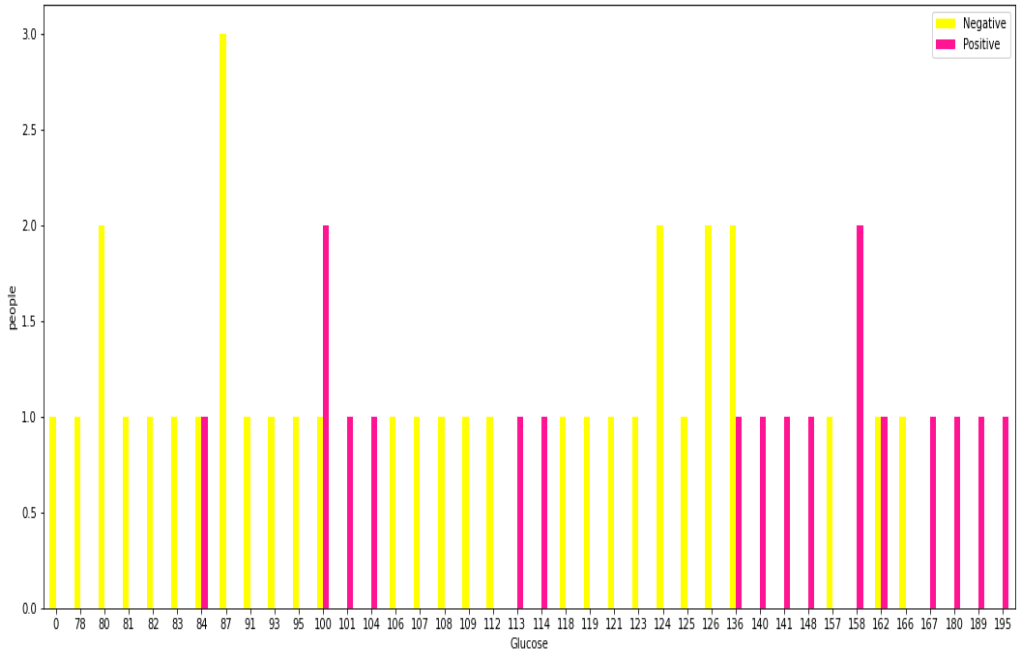

Figure 8.2 Comparing Glucose with the Outcome

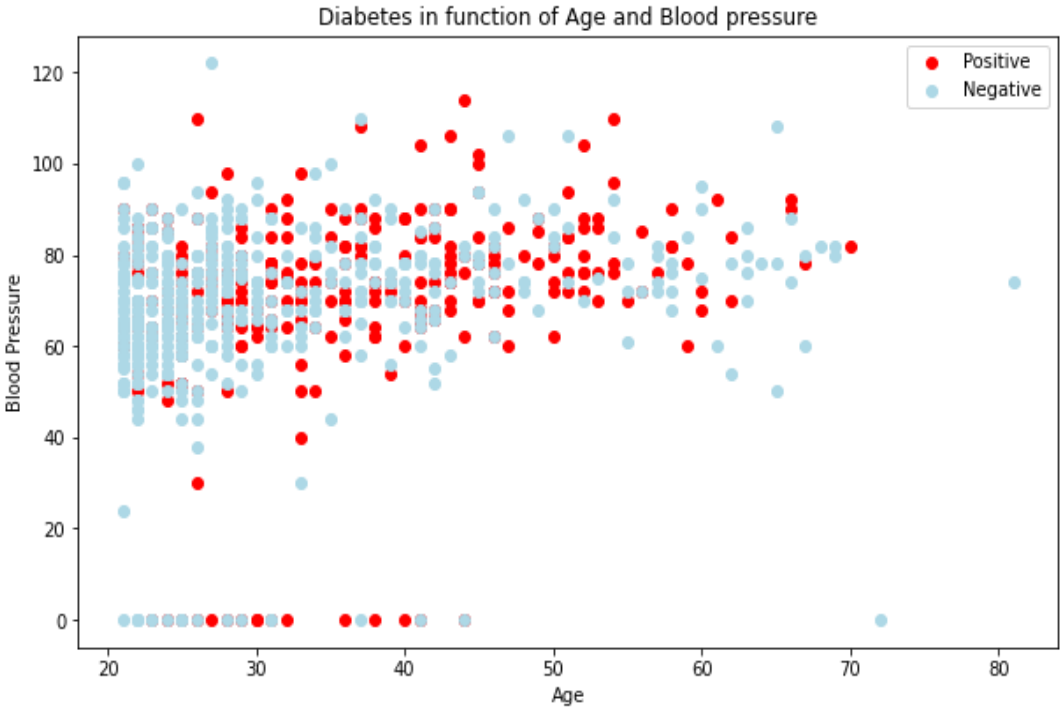

Figure 8.3 Comparing Diabetes in function of age and Blood Pressure

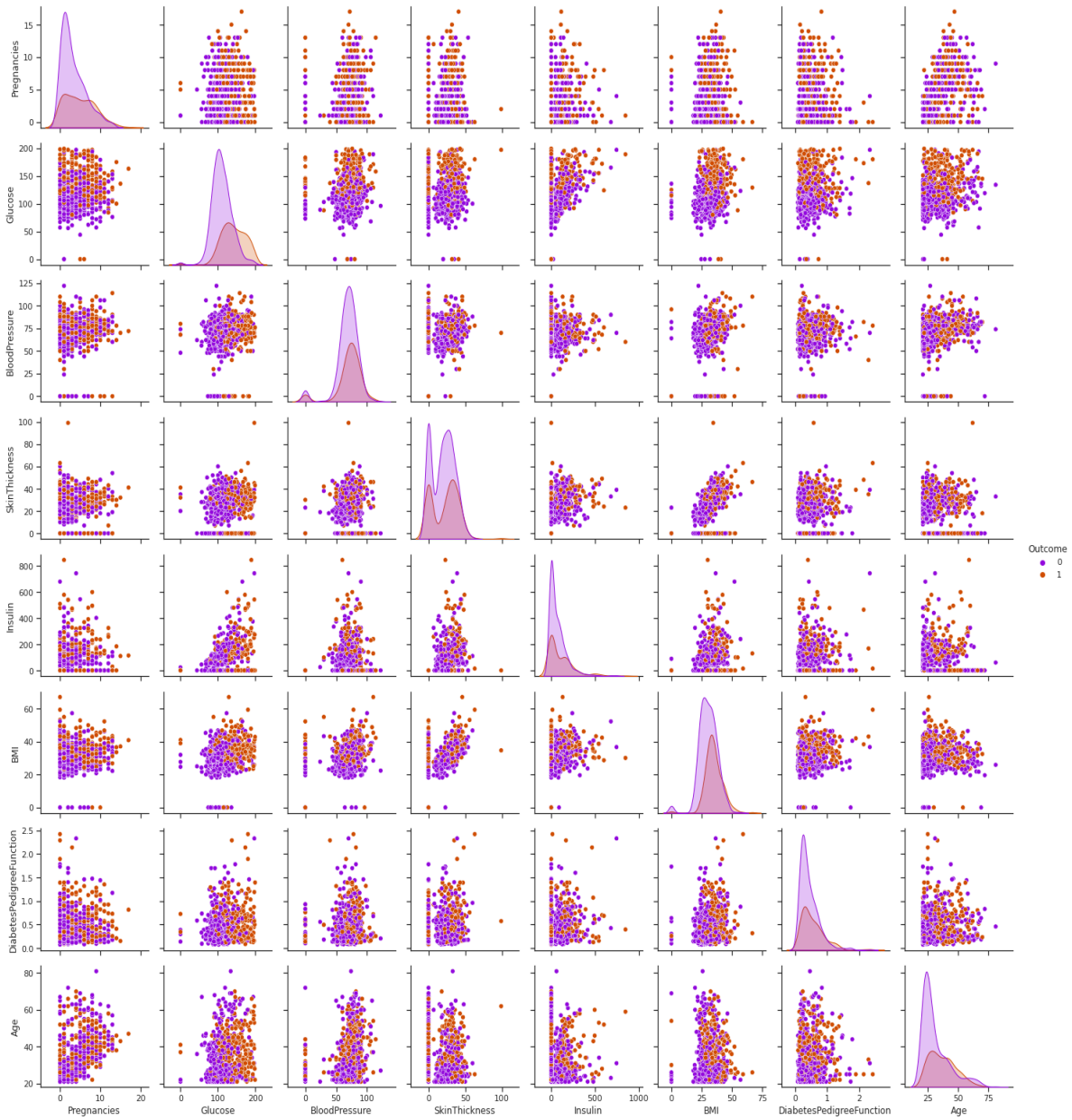

Figure 8.4 Pair plotting of data frame

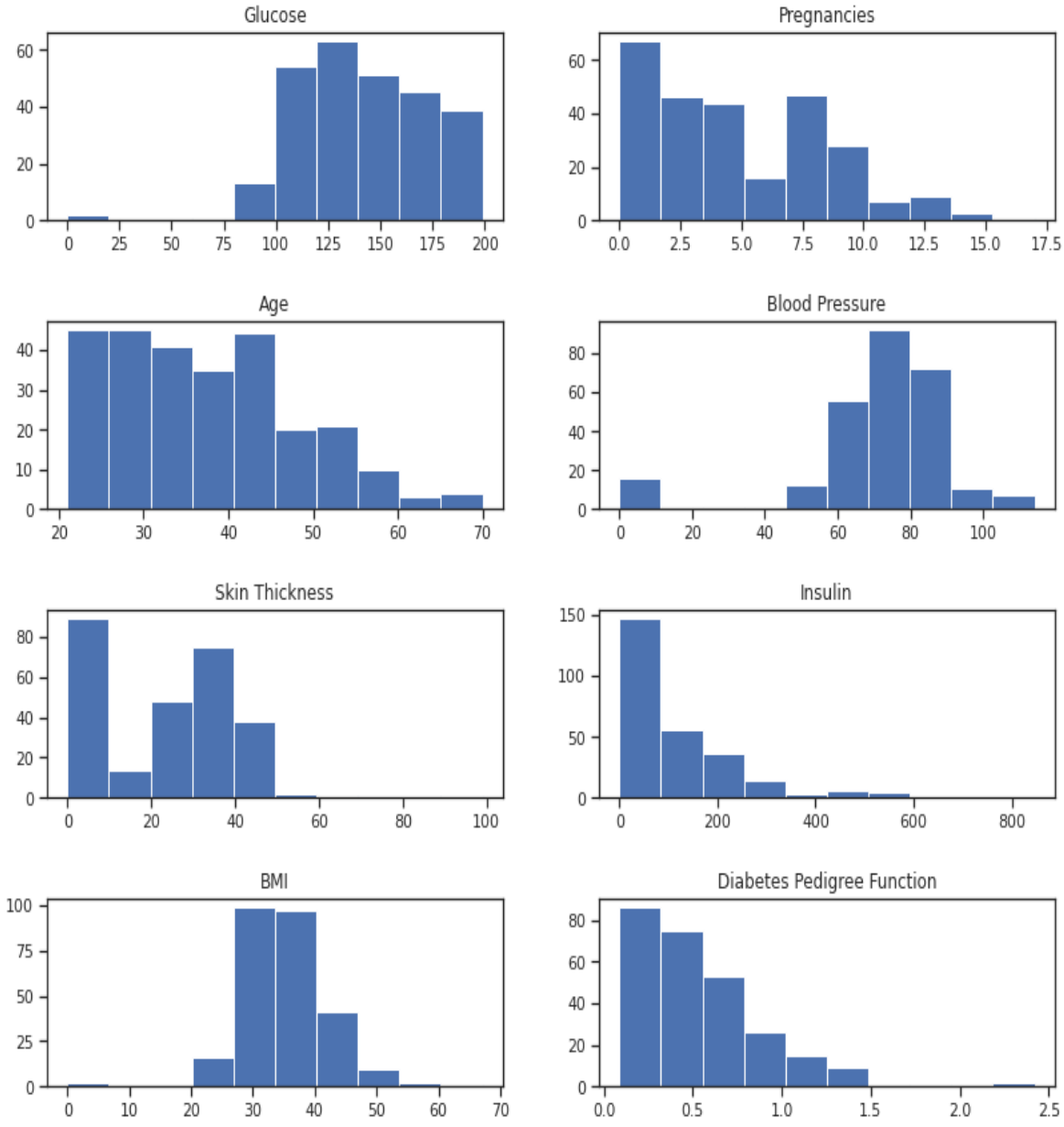

Figure 8.5 Comparing all columns when the Outcome is 1(has Diabetes)

So with the help of above mentioned procedure, we have completed our project named as Diabetes Predication System using Machine Learning to improve diabetetes detection process using machine learning. In this project, the data was formulated in different formulations and the model was trained with above 80% accuracy.

We have taken the dataset from Kaggle platform. Using Flask and Python we have created the backend and used HTML, CSS, JavaScript and Bootstrap for front-end. We used Jupyter Notebook and Visual Studio Code IDE as a free source

integrated development environment on windows for the development of our project. Visual Studio Code is high in the list for best development IDE because it is easy to use and it helps in development of various applications in no time and it works equally as good with HTML, CSS, JavaScript is being used in the development of this project The obtained results reveal that it is possible to achieve high

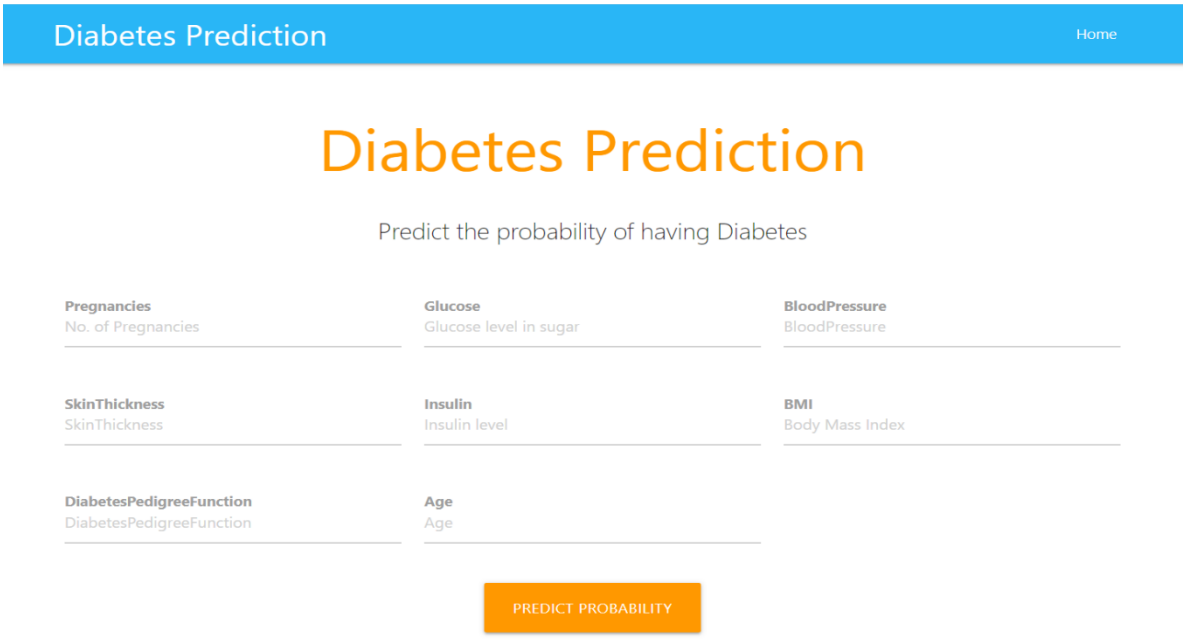

Figure 8.6 Homepage

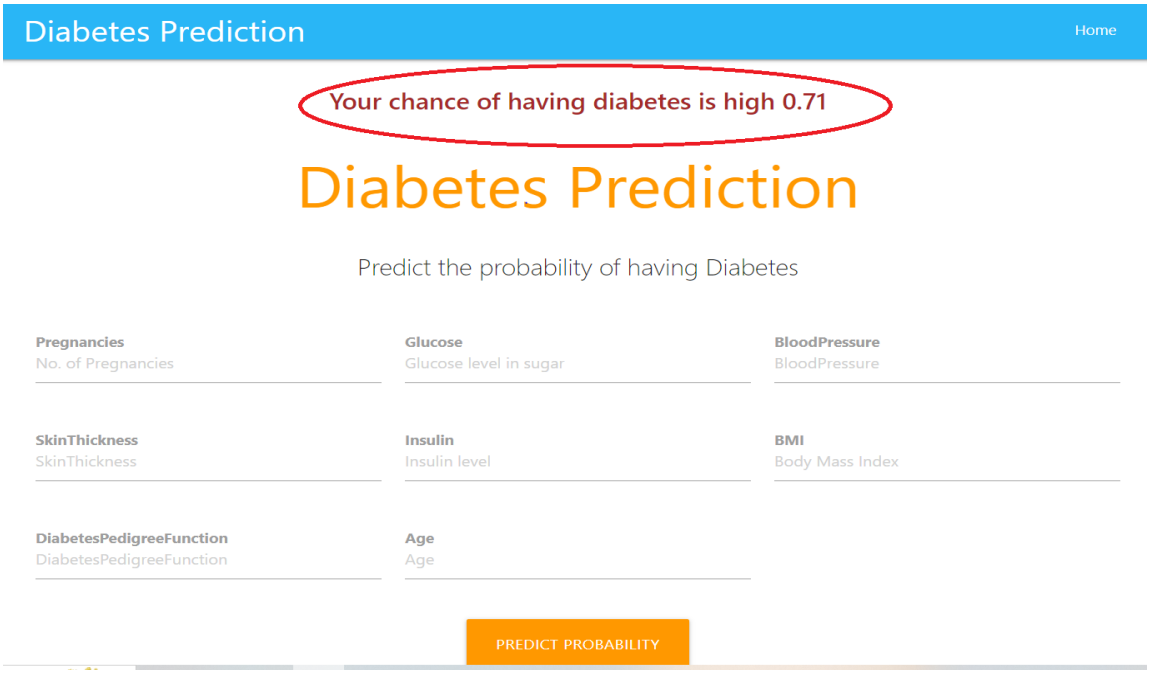

Figure 8.7 Non-Diabetes (You are Safe)

CHAPTER 9

CONCLUSION & FUTURE SCOPE

CONCLUSION

The aim of this project was to assess the effectiveness or efficiency of Logistic Regression with other linear classifiers, Examples of such classifiers include SVM , KNN, and Random Forest(RF). The results of the comparison revealed that Logistic Regression outperformed all the other classifiers. The accuracy of Logistic Regression was found to be the highest, at **0.83%.** The proposed approach utilized ensemble learning and classification methods, which resulted in high accuracy levels. These experimental results can assist healthcare professionals by enabling early predictions and informed decisions, these classifiers can aid in the treatment of diabetes and potentially save human lives.

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 83 % classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

FUTURE SCOPE

In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well .

For later work, it is far vital to conduct in hospitals actual and newest sufferers' information for regular instructing and development of our present model. The amount of the data item has massive sufficient for education and forecasting. Some further methods and portraits need to take a look of DM (data mining). To expand a chain of regulation and quality approach to stop human beings from come out of data mining. It assists to minor the increase price of blood glucose and subsequently reduce the danger of forecasting data mining.

REFERENCES

[1] Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms”, IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759 .

[2] Mitushi Soni, Dr. Sunita Varma " Diabetes Prediction using Machine Learning Techniques" , Journal of Engineering Research & Technology (IJERT) 2020,

[3] Sivaranjani S, Ananya S, Aravinth J, Karthika R, " Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”, 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021 DOI: 10.1109/ICACCS51430.2021.9441935 .

[4] Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal PatilB “Diabetes Prediction Using Different Machine Learning Approaches” IJSDR | Volume 7 Issue 5 , 2022.

[5] Ashwini r, s m aiesha afshin, kavya v, deepthi raj “diabetes prediction using machine learning” ijrti | volume 7, issue 7 | issn: 2456-3315, 2022 .

[6] G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020 .

[7] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[8] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[9] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[10] G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1009- 1014, 2020.

[11] M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, 2019.

[12] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.

[13] Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of BigData ,6(1), p.13.(2019)

[14] Sisodia,D. and Sisodia,DS,2018.Prediction of diabetes using classification algorithms. Procedia computer science,132, pp.1578-1585. (2018)

[15] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[16] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[17] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[18] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[19] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[20] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International

Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[21] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8,

[22] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[23] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[24] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[25] A.K., Dewangan, and P., Agrawal, “Classification of Diabetes Mellitus Using Machine Learning Techniques,” International Journal of Engineering and Applied Sciences, vol. 2, 2015.

[26] Azra Ramezankhani, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi and Farzad Hadaegh, ”An Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database”, Int J Endocrinol Metab, April 2015.

[27] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.

[28] A Abbasi, LM Peelen, E Corpeleijn, YT van der Schouw, RP Stolk, AM Spijkerman et al., ”Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study”, BMJ, 2012.

[29] ”Mining constrained association rules to predict heart disease”, IEEE 13th International Conference on Data Mining, pp. 433, 2010.

[30] M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, 2019.

[31] M Sabibullah, V Shanmugasundaram and Priya K Raja, ”Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends Technology in Computer Science, vol. 2, no. 6, 2013.

[32] G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1009- 1014, 2020.

[33] VZ. Tafa, N. Pervetica and B. Karahoda, ”An intelligent system for diabetes prediction”, 2015 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378-382, 2015

PUBLICATION DETAILS

PAPER TITLE	CONFERENCE NAME	CONFERENCE DURATION	ISBN NUMBER
Diabetes Prediction using Machine Learning	International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)	27 –April 2023 To 28- April 2023	2581-9429


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Ashwini Ghate
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India
HAS PUBLISHED A REVIEW PAPER ENTITLED
Diabetes Prediction using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 8, APRIL 2023
Certificate No: 042023-A2318
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9556
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Aditi Dhote
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India
HAS PUBLISHED A REVIEW PAPER ENTITLED
Diabetes Prediction using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 8, APRIL 2023
Certificate No: 042023-A2320
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9556
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Vrushali Mange
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India
HAS PUBLISHED A REVIEW PAPER ENTITLED
Diabetes Prediction using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 8, APRIL 2023
Certificate No: 042023-A2322
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9556
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Pratiksha Patte
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India
HAS PUBLISHED A REVIEW PAPER ENTITLED
Diabetes Prediction using Machine Learning
IN IJARSCT, VOLUME 3, ISSUE 8, APRIL 2023
Certificate No: 042023-A2321
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-9556
www.doi.org
www.crossref.org
7.301
www.sjifactor.com
Editor-in-Chief

PROJECT GROUP MEMBERS

Name: Prajakta Haridas Mathe
Address: Near Gajanan Ice Factory Ako,
Shegaon Tq.Shegaon Dist:Buldhana
Email id: prajaktamathe849@gmail.com
Mobile no: 8766841912

Name: Ashwini Santosh Ghate
Address: Sangawa,Tq.Shegaon Dist: Buldhana
Email id: ashwinighate28@gmal.com
Mobile no: 9511724780

Name: Aditi Vivek Dhote
Address: At. Sambhaji Nagar,Mehekar
Email id: dhoteaditi9@gmail.com
Mobile no: 9022267716

Name: Vrushali Narendra Mange
Address: At. Post Sawalapur Tq. Achalpur Dist. Amravati
Email id: vrushalimange27apr@gmail.com
Mobile no:9373485193

Name: Pratiksha Shrikant Patte
Address: At. Mothe wadgaon arni road,Yavatmal
Email id: pratikshapatte.19@gmail.com
Mobile no: 9767853482

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Diabetes Prediction using Machine Learning

Ms. P. V. Deshmukh[1], Ashwini Ghate[2], Prajakta Mathe[3], Aditi Dhote[4], Pratiksha Patte[5], Vrushali Mange[6]
Assistant Professor, Department of Computer Science and Engineering [1]
Under Graduate Students, Department of Computer Science and Engineering[2,3,4,5,6]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, India

Abstract: *High levels of glucose in the bloodstream lead to the development of diabetes , which results in frequent urination, increased thirst, and increased hunger. It is crucial to address diabetes promptly as untreated cases may lead to severe complications in various body organs such as the heart, kidneys, blood pressure, and eyes. Predictive analytics over big data is a challenging task, particularly in healthcare. However, it can aid healthcare practitioners in making quick decisions about patients' health and treatment based on big data. The performance and accuracy of ML algorithms used in predictive Data analysis for predicting the occurrence of diabetes are compared and analyzed across various disciplines. In this study, different classification Computational methods, which may involve various algorithms, such as SVM, KNN, Logistic regression, and Random forest, were considered, and their performance metrics such as Recall, F-Measure, Precision, and Accuracy were evaluated Derived from the confusion matrix. According to the experimental results, the SVM and ontology classifiers yielded the highest accuracy for diabetes prediction.*

Keywords: ML, Diabetes Prediction, SVM, KNN, Logistic Regression (LR), Random Forest.

I. INTRODUCTION

Diabetes prevalent condition in today's world and poses significant Challenges are present in both developed & developing nations. When we eat, the insulin hormone The pancreas produces a substance that permits glucose to enter the bloodstream. Pancreatic dysfunction leads to diabetes can result in several serious conditions such as coma, retinal failure , renal, destruction of pancreatic beta cells, dysfunction of the cardiovascular and cerebral vascular systems, peripheral vascular diseases, sexual and joint dysfunction, weight loss, ulcers, and negative effects on immunity. Diabetes ranks as the third leading cause of death, trailing behind heart disease and cancer. However, with the advancement of machine learning technologies, we may be able to tackle this problem. Machine learning and data mining aim to extract information from data and produce clear and understandable representations. Our goal is to utilize ML to develop the diabetes diagnosis system capable of predicting whether a Whether or not the patient has diabetes. Obesity, high blood glucose levels, and other factors can contribute to diabetes, which affects insulin and carbohydrate metabolism, leading to abnormal levels of glucose in the blood. Insufficient production of insulin by the body leads to the development of diabetes . An organization known as the World Health Organization(WHO) estimates that approximately 422 million individuals globally worldwide with the majority of individuals with diabetes reside in low or Nations with incomes considered to be in the middle range. Diabetes is categorized into type one and type two. Type one characterized by a lack of insulin production, while type two diabetes is characterized by inadequate insulin response and production.

II. LITERATURE SURVEY

Arwatki Chen Lyngdoh et al. conducted research on predicting diabetes disease using 5 supervised ML Algo: KNN, Naive Bayes, Decision Tree Classifier, Random Forest, and SVM. by including current risk variables and performing cross-validation, they achieved consistent accuracy with the KNN classifier achieving a high accuracy of 76%. The main objective of the study was to identify the best outcomes for accurately predicting diabetes disease, considering accuracy and computing time.

Mitushi Soni et al. ML classification and ensemble techniques were employed to make predictions about diabetes using a dataset. They employed K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting, and Random Forest algorithms, and found Random Forest outperformed the others in terms of accuracy.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Sivaranjani S et al. used SVM and Random Forest(RF)methods for identifying potential risks of Diabetes Related Diseases. After data preprocessing and implementing forward & backward stepwise feature selection was utilized to identify the most impactful features, they employed. Principle Component Analysis was employed to reduce dimensionality. Their study , which outperformed Support Vector Machine's 81.4% accuracy.

Shejal Kale et al. applied MLClassification & Using ensemble techniques to make predictions about diabetes on a given dataset.They utilized KNN, Logistic Regression(LR), Decision Tree(DT) , SVM , Gradient Boosting(GB) , and Random Forest(RF) algorithms, and found that (RF) Random Forest had the best accuracy.

Ashwini R et al. trained ML ALGO such as KNN, Random Forest(RF), Logistic Regression(LR), and SVM using various datasets. They used preprocessing techniques to improve the accuracy of their models and prioritized risk factors by employing various feature selection approaches.

III. METHODOLOGY

The aim of the project is to enhance the accuracy of diabetes prediction models. Our approach involved exploring various ML ALGOS like KNN , for classification and prediction. In the subsequent sections, we provide a concise overview of our methodology.

Description of the Dataset

The data utilized in this project was got From the UCI Machine Learning repo. and is known as Pima Diabetes CSV File. It comprises several features of 768 patients.

The ninth attribute in each data point represents the class variable, which indicates whether the individual is -or+for diabetes, denoted by 1 and 0, respectively.

Sr. Number	Attributes
I	Pregnancy Attribute
II	Glucose Attribute
III	Blood Pressure Attribute
IV	Skin thickness Attribute
V	Insulin Attribute
VI	Body Mask Index
VII	Diabetes Pedigree Function Attribute
VIII	Age Criteria

Table 1: Dataset Contents

1	Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Figure 1 : Pima Indians Dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Distribution of Diabetic Patient

In our attempt to develop a diabetes prediction model, we encountered a slightly imbalanced dataset. Out of the total 768 samples, around 500 were Designated as 0, denoting the nonexistence of diabetes., while 268 were designated as 1, denoting the existence of diabetes.

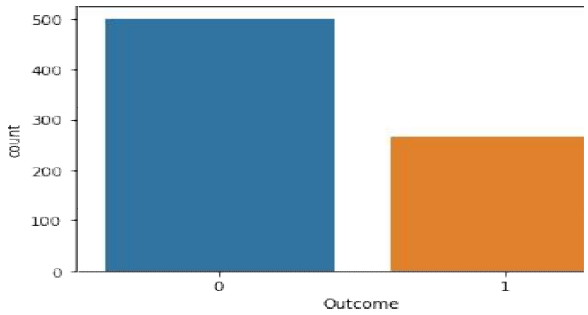

Figure 2: The proportion of patients with diabetes compared to those without diabetes.

1. **Data Pre-processing** - The process of preprocessing data is of utmost importance, especially for data concern with healthcare, which may contain missing values and Other contaminants that may affect the effectiveness of data mining. This process is essential to achieve accurate results and successful predictions with the help ofML methodology on the CSV file. To work with the Pima Indian diabetes dataset, we require, preprocessing in couple of steps.
2. **Missing Values Elimination**- A value of zero are removed since it is not possible to have a value of zero for certain features. This process helps in feature subset selection by eliminating irrelevant features or instances, reducing the dimensionality of the data and enabling faster processing.
3. **Categorization of data**- The data is in normal form and divided into training and testing sets after undergoing cleaning. The algorithm is trained on the training dataset, and the test dataset is kept aside. This training process produces a model based on logic, algorithms, and feature values in the training data. The purpose of normalization is to standardize all attributes to a consistent scale.

Apply Machine Learning

Here are the techniques to apply in Machine learning:-

- **KNN -** The algorithm learning algorithm that can tackle Tasks involving categorizing data into classes or predicting numerical values are respectively referred to as classification and regression problems. KNN adopts a lazy prediction approach and relies on the assumption that similar data points are situated near each other. By computing similarity measures, KNN groups new data and classifies them based on their similarities with existing records. The algorithm leverages a tree-like structure to measure the distance between data points. When making predictions when presented with a new data point, the algorithm identifies the K nearest neighbors in the training dataset, where K is a positive integer.
- **Random forest**, a popular machine learning algorithm coined by Leo Breiman and Adele Cutler, entails amalgamating the outcomes of numerous decision trees to generate a unified output. Given its versatility and user-friendliness, it is extensively employed to address classification and regression issues.
- **Support Vector Machine (SVM)** Refers to a learning technique that involves supervision partitions data into two distinct categories. It learns from a labeled dataset, and while it trains, it constructs the model. The aim of the SVM algorithm aims to determine the category or class that a novel data point belongs. This feature characterizes SVM as a non-binary linear classifier
- **Logistic regression** Logistic regression is a regression analysis that specializes in predicting the probability of a binary event. To understand logistic regression, it's crucial to first introduce the general concept of regression analysis. Regression analysis is a modeling technique used to establish the association between a dependent variable (usually labeled "Y") and one or more independent variables (usually labeled "X"). When multiple independent variables are utilized to predict or explain the outcome of the dependent variable, it is referred to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

as multiple regression. Regression analysis can be applied for three main objectives: projecting the impacts of specific changes, predicting future values and trends, and evaluating the efficacy of different predictors.

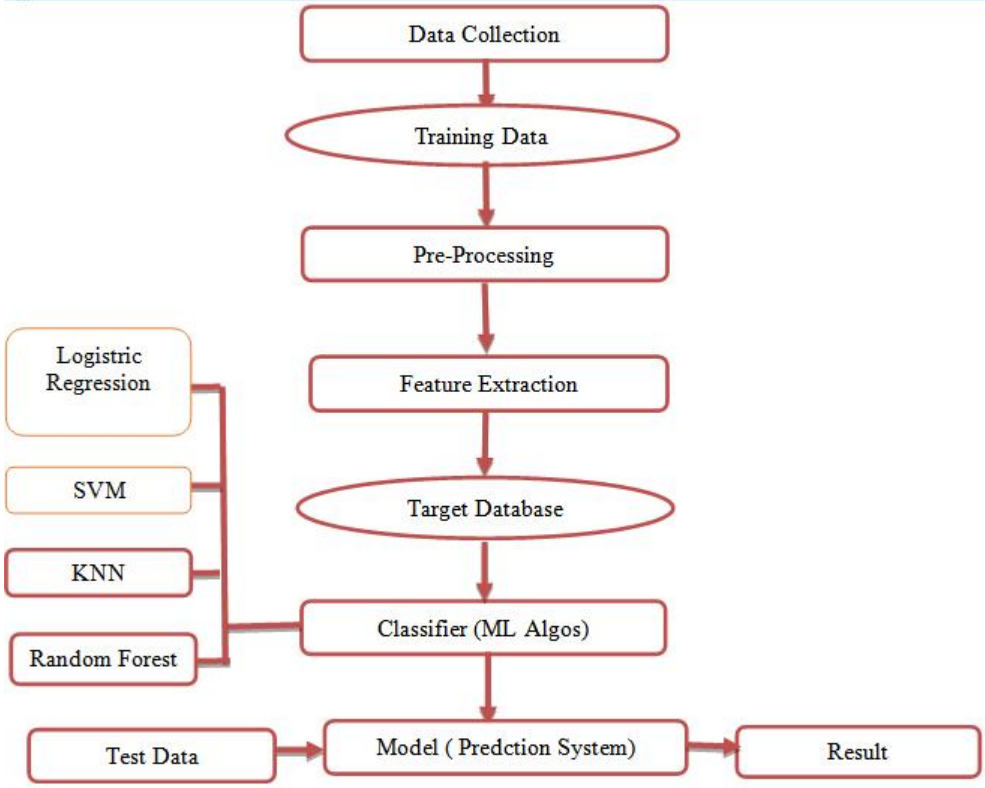

Figure 3: Diabetes prediction flow diagram

The first step involves importing the necessary libraries and loading the diabetes dataset. In step two, the data is pre-processed to eliminate missing values. Step three involves splitting the dataset into training and test sets using an 80-20 percentage split. Next, the machine learning algorithm we use four algorithm Like ML algorithms , is selected in step four. Step five involves building the classifier model using the training set. The classifier model is then tested using the test set in step six. In step seven, a comparing and evaluating the performance results of each classifier is carried out. Finally, in step eight, after analyzing the results The algorithm that performs the best is determined by evaluating various metrics..

IV. RESULT AND DISCUSSION

The table below displays the performance values of various classification algorithms, calculated using different measures. Based on the table, it is observed that Logistic regression exhibits the highest accuracy. Therefore, the Logistic regression machine learning classifier is capable of predicting the likelihood of diabetes with greater precision than other classifiers.

Classification Algorithm	Precision
Logistic regression(LR)	0.83
Support Vector Machine(SVM)	0.76
Random Forest(RF)	0.78
KNN	0.78

Table 2. Accuracy Measures

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

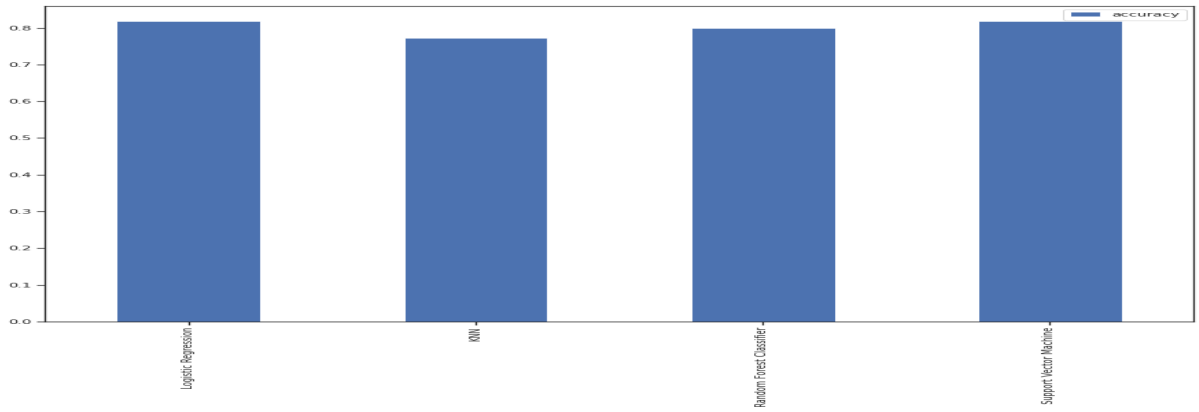

Figure 4: Results of the Accuracy achieved by machine learning techniques

V. CONCLUSION

The aim of this project was to assess the effectiveness or efficiency of Logistic Regression with other linear classifiers, Examples of such classifiers include SVM , KNN, and Random Forest(RF). The results of the comparison revealed that Logistic Regression outperformed all the other classifiers. The accuracy of Logistic Regression was found to be the highest, at **0.83%.** The proposed approach utilized ensemble learning and classification methods, which resulted in high accuracy levels. These experimental results can assist healthcare professionals by enabling early predictions and informed decisions, these classifiers can aid in the treatment of diabetes and potentially save human lives.

REFERENCES

[1]. Arwatki Chen Lyngdoh, Nurul Amin Choudhury, Soumen Moulik, “Diabetes Disease Prediction Using Machine Learning Algorithms”, IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) 2020, DOI: 10.1109/IECBES48179.2021.9398759 .

[2]. Mitushi Soni, Dr. Sunita Varma " Diabetes Prediction using Machine Learning Techniques" , Journal of Engineering Research & Technology (IJERT) 2020,

[3]. Sivaranjani S, Ananya S, Aravinth J, Karthika R, " Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction”,7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021 DOI: 10.1109/ICACCS51430.2021.9441935 .

[4]. Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal PatilB “Diabetes Prediction Using Different Machine Learning Approaches”IJSDR | Volume 7 Issue 5 , 2022.

[5]. Ashwini r, s m aiesha afshin, kavya v, deepthi raj“diabetes prediction using machine learning”ijrti | volume 7, issue 7 | issn: 2456-3315, 2022 .

[6]. G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8thInternational Conference on Reliability Infocom Technologies and Optimization (Trends and FutureDirections) (ICRITO), 2020 .

[7]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[8]. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[9]. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[10]. G. Tripathi and R. Kumar, ”Early Prediction of Diabetes Mellitus Using Machine Learning”, 2020 8th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1009- 1014, 2020.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

[11]. M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, 2019.

[12]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, ”Type 2 diabetes mellitus prediction model based on data mining”, Elsevier Informatics in Medicine, vol. 10, pp. 100-107, 2018.

[13]. Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of BigData ,6(1), p.13.(2019)

[14]. Sisodia,D. and Sisodia,DS,2018.Prediction of diabetes using classification algorithms. Procedia computer science,132, pp.1578-1585. (2018)

[15]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[16]. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[17]. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.