Mini Project Report

A report submitted in partial fulfilment for the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

Information Technology

By

Yash Saxena (2007340130069)

Nikita Kabir (2007340130030)

Sandeep Kumar (2007340130048)

Akash Kumar (2007340130007)

To

DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, LUCKNOW

Under the Guidance of

Mr. Shantanu Shukla

Mr. Sandip Kr. Singh

Mr. Abhishek Kr. Yadav

DEPARTMENT OF INFORMATION TECHNOLOGY

RAJKIYA ENGINEERING COLLEGE

ATARRA, BANDA-210201

DEPARTMENT OF INFORMATION TECHNOLOGY

RAJKIYA ENGINEERING COLLEGE

ATARRA, BANDA-210201

This is to certify that the "Mini Project report" entitled "Diabetes Prediction System" submitted by Yash Saxena, Nikita Kabir, Sandeep Kumar, Akash Kumar and submitted during 2022 – 2023 academic year, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Information Technology in the Department of Information Technology at Rajkiya Engineering College, Banda, under Dr. A.P.J. Abdul Kalam Technical University, Lucknow.

This report is an authentic record of candidates own work carried out under our supervision. The matter embodied in this Mini Project report is original and has not been submitted for the award of any other degree.

College Mini Project Coordinator

Mr. Shantanu Shukla

Dr. Vibhash Yadav

(Associate Professor)

Head of Department

Information Technology

ACKNOWLEDGEMENT

It is indeed a great pleasure to express our sincere thanks to our august supervisor **Mr. Shantanu Shukla Sir, Mr Abhishek Kr. Yadav Sir and Mr. Sandip Kr. Singh Sir**, Department of Information Technology of Rajkiya Engineering College, Banda for their continuous support in this Mini Project. They were always there to listen and to give advice. There were always there to meet and talk about our ideas, to proofread and mark up our mini project, and to ask us good questions to help me to think through our problems. Without their encouragement and constant guidance, I could not have finished this project.

I am highly indebted to Director **Prof. Sheo Prasad Shukla Sir** for the facilities provided to accomplish this Mini Project.

I would like to thank my Head of the Department **Dr. Vibhash Yadav Sir** for his constructive criticism throughout my Mini Project.

I am thankful to my family whose unfailing love, affection, sincere prayers and best wishes had been a constant source of strength and encouragement.

I am extremely great full to my department staff members and friends who helped me in successful completion of this Mini Project.

Submitted by

Yash Saxena (2007340130069)

Nikita Kabir (2007340130030)

Sandeep Kumar (2007340130048)

Akash Kumar (2007340130007)

ABSTRACT

Diabetes Prediction: Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or imply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. 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. This project aims to predict diabetes via four different supervised machine learning methods including: SVM, Logistic regression, Random forest Classifier, Decision Tree Classifier. This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms and end to end deployment using flask.

CONTENTS

CHAPTER 1 - INTRODUCTION

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One of such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet. It is not just a malady yet, also a maker of different sorts of sicknesses like a coronary failure, visual deficiency, kidney ailments and nerve harm, and so on.

Subsequently, the identification of such chronic metabolic ailment at a beginning period could help specialists around the globe in forestalling loss of human life. Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database.

Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Types of Diabetes

Type 1

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can't use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining method is used to pre-process and select the relevant features from the healthcare data, and the machine learning method helps automate diabetes prediction. Data mining and machine learning algorithms can help identify the hidden pattern of data using the cutting-edge method; hence, a reliable accuracy decision is possible. Data Mining is a process where several techniques are involved, including machine learning, statistics, and database system to discover a pattern from the massive amount of dataset. According to Nvidia: Machine learning uses various algorithms to learn from the parsed data and make predictions.

CHAPTER 2 – RELATED WORK

Diabetes prediction is a classification technique with two mutually exclusive possible outcomes, either the person is diabetic or not diabetic. After extensive research, we came to conclusion that although numerous classification techniques can be used for the purpose of prediction, the observed accuracy varied. On careful examination of the performance of techniques used in prevalent works, logistic regression, random forest, decision tree, and support vector machine, we found the mat par when applied to our dataset. SVM technique was able to achieve 80% accuracy.

The primary factor which influenced our algorithm selection was its adaptability and compatibility with future applications.

The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit asymptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can fore see whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives

CHAPTER 3 - Data Set

The dataset collected is originally from the Pima Indians Diabetes Database is available on Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome. Independent variables include the number of pregnancies the patient has had their BMI, insulin level, age, and so on as Shown in Following Table 1:

Table 1 Dataset description

Serial no	Attribute Names	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin
6	BMI	Body mass index
7	Diabetes pedigree function	Diabetes pedigree function
8	Outcome	Class variable (0 or 1)
9	Age	Age of patient

The diabetes data set consists of 768 data points, with 9 features each.

“Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

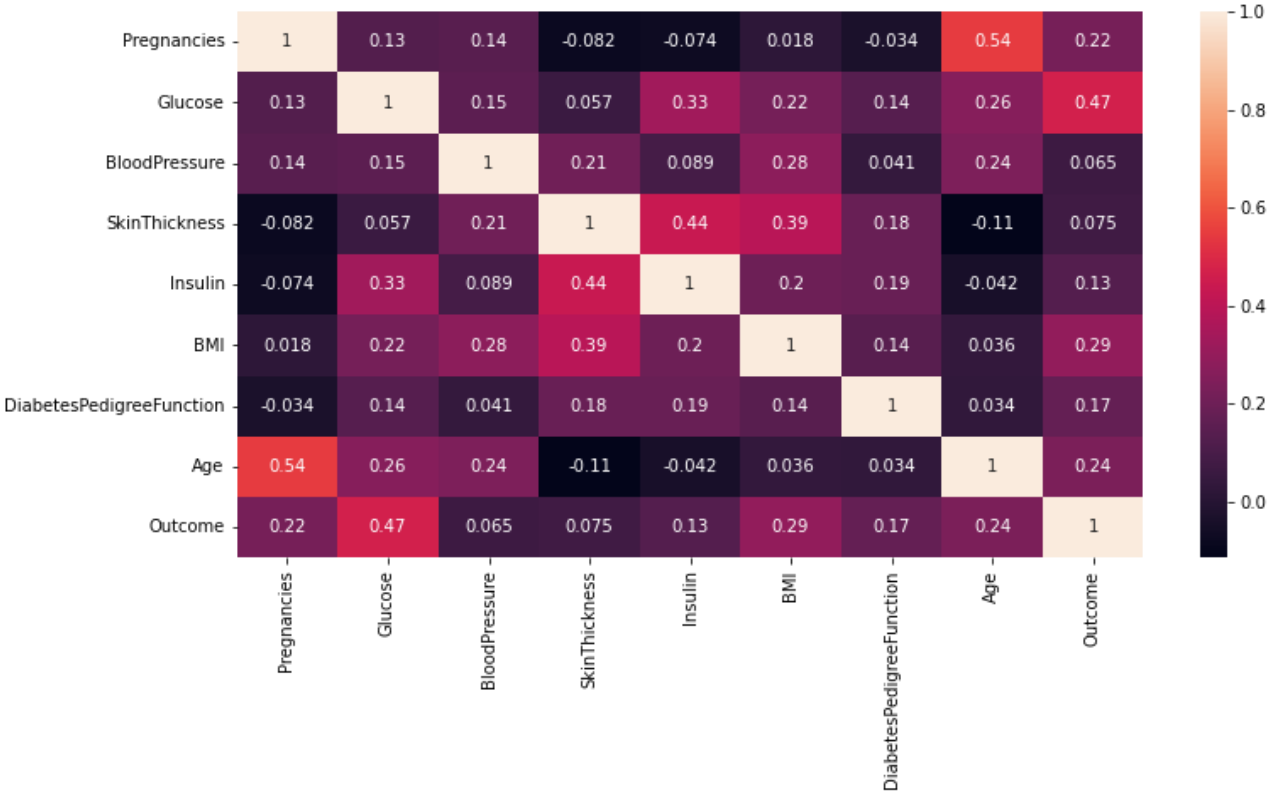
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):

#	Column	Non-Null Count	D-type
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

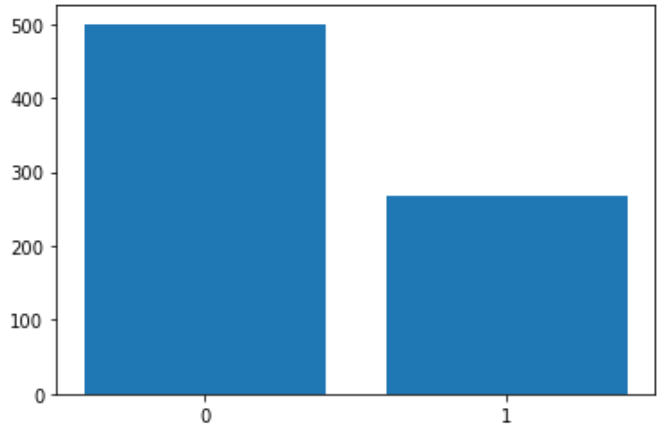
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
There is no null values in dataset.

Correlation Matrix:



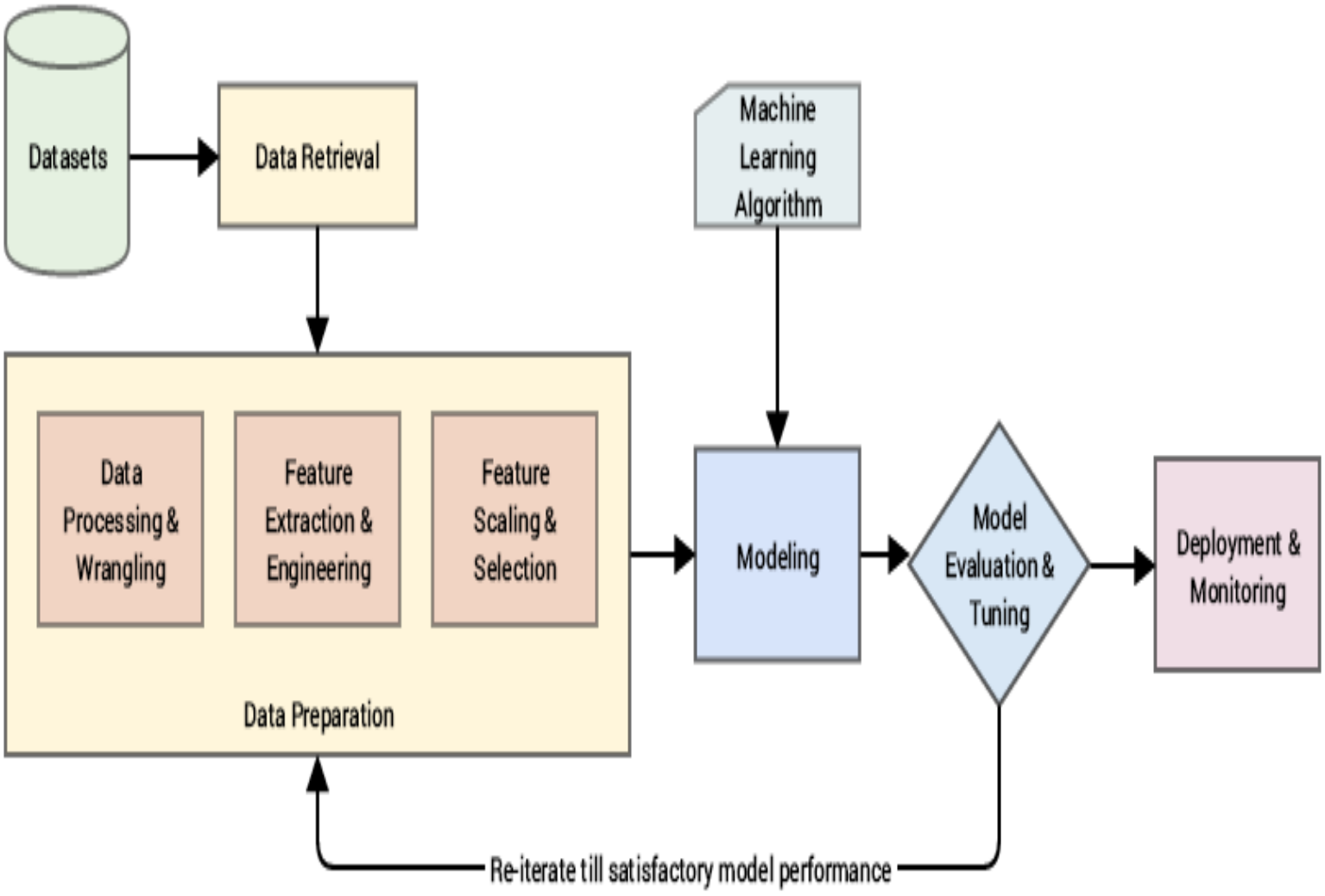
It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a vey low correlation with the outcome value and some have high correlation.

Bar Plot for Outcome Class:



The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

Flowchart of the Model



CHAPTER 4 - PROPOSED METHODS

I] Dataset collection – It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries 768 rows i.e., total number of data and 9 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome.

II] Data Pre-processing: This phase of model handles inconsistent data in order to get more accurate and precise. This dataset doesn't contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using Standard Scaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

III]Missing value identification: Using the Panda library and SK-learn, we check for the missing values in the datasets, shown below. We replaced the missing value with the corresponding mean value if we get.

Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0
dtype: int64	

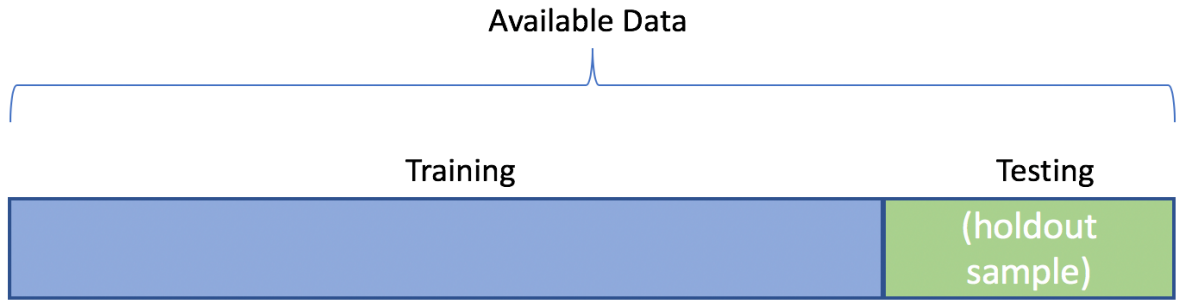
IV] Feature selection: Pearson's correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation

Correlation with Outcome

Glucose	0.466581
BMI	0.292695
Age	0.238356
Pregnancies	0.221898
DiabetesPedigreeFunction	0.173844
Insulin	0.130548
SkinThickness	0.074752
BloodPressure	0.065068

V] Scaling and Normalization: We performed feature scaling by normalizing the data f rom 0 to 1 range, which boosted the algorithm's calculation speed. Scaling means that you're transforming your data so that it fits within a specific scale, like 0- 100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM). With these algorithms, a change of "1" in any numeric feature is given the same importance.

VI] Splitting of data: After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training we took 614 sample and for testing we took 154 sample.



VII] Design and implementation of classification model: In this work, comprehensive studies are done by applying different ML classification techniques like Decision Tree, Random Forest, Logistic Regression, Support Vector Machine.

VIII] Machine learning classifier: We have developed a model using Machine learning Technique. Used different classifier and ensemble techniques to predict diabetes dataset. We have applied Support Vector Machine, Logistic Regression, Decision Tree and Random Forest. Machine learning classifier to analyse the performance by finding accuracy of each classifier All the classifiers are implemented using scikit learn libraries in python. The implemented classification algorithms are described in next section.

CHAPTER 5 - MODELING AND ANALYSIS:

5.1 - Logistic Regression: Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

5.2 - SVM: SVM is supervised learning algorithm used for classification. In SVM we have to identify the right hyper plane to classify the data correctly. In this we have to set correct parameters values. To find the right hyper plane we have to find right margin for this we have choose the gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to miss classification.

5.3 - Decision Tree: Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Information Gain for splitting the nodes.

5.4 - Random Forest: Random Forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy. Where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. Here number of trees in forest used is 100 in number and Information Gain is used for splitting the nodes.

CHAPTER 6 – MEASUREMENTS

To find the efficient classifier for diabetes prediction we have applied a performance matrices are confusion matrix and accuracy are discussed as follows:

Confusion matrix: - which provides output matrix with complete description performance of the model.

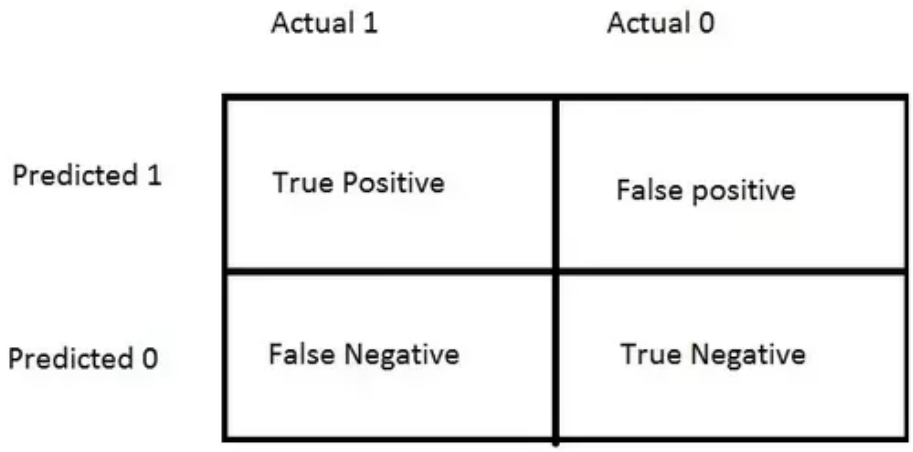
Here,

TP: True positive

FP: False positive

TN: True negative

FN: False negative



The following performance metrics are used to calculate the presentation of various algorithms.

True positive (TP) – person has disease, and the prediction also has a positive

True negative (TN) – person not having disease and the prediction also has a negative

False positive (FP) – person not having disease but the prediction has a positive

False negative (FN) – person having disease and the prediction also has a positive

TP and TN can be used to calculate accuracy rate and the error rates can be computed using FP and FN values.

True positive rate can be calculated as TP by a total number of persons have disease in reality.

False positive rate can be calculated as FP by a total number of persons do not have disease in reality.

Precision is TP/ total number of person have prediction result is yes.

Accuracy is the total number of correctly classified records.

Accuracy- We have chooses accuracy matrix to measure the performance of all the models. The ratio of number of correct predictions to the total number of predictions Made.

$$\text{Accuracy} = \frac{\text{Number of correct Prediction}}{\text{Total numbers of predictions made.}}$$

CHAPTER 7 - RESULTS AND DISCUSSION

Machine learning classification algorithms developed for prediction of diabetes in earlier stage. We used 70% of data for training and 30% of data for testing. In this ratio of data splitting Here we found that Support Vector Machine has the least root mean square error. Comparison of results of all the implemented classifiers are listed in below

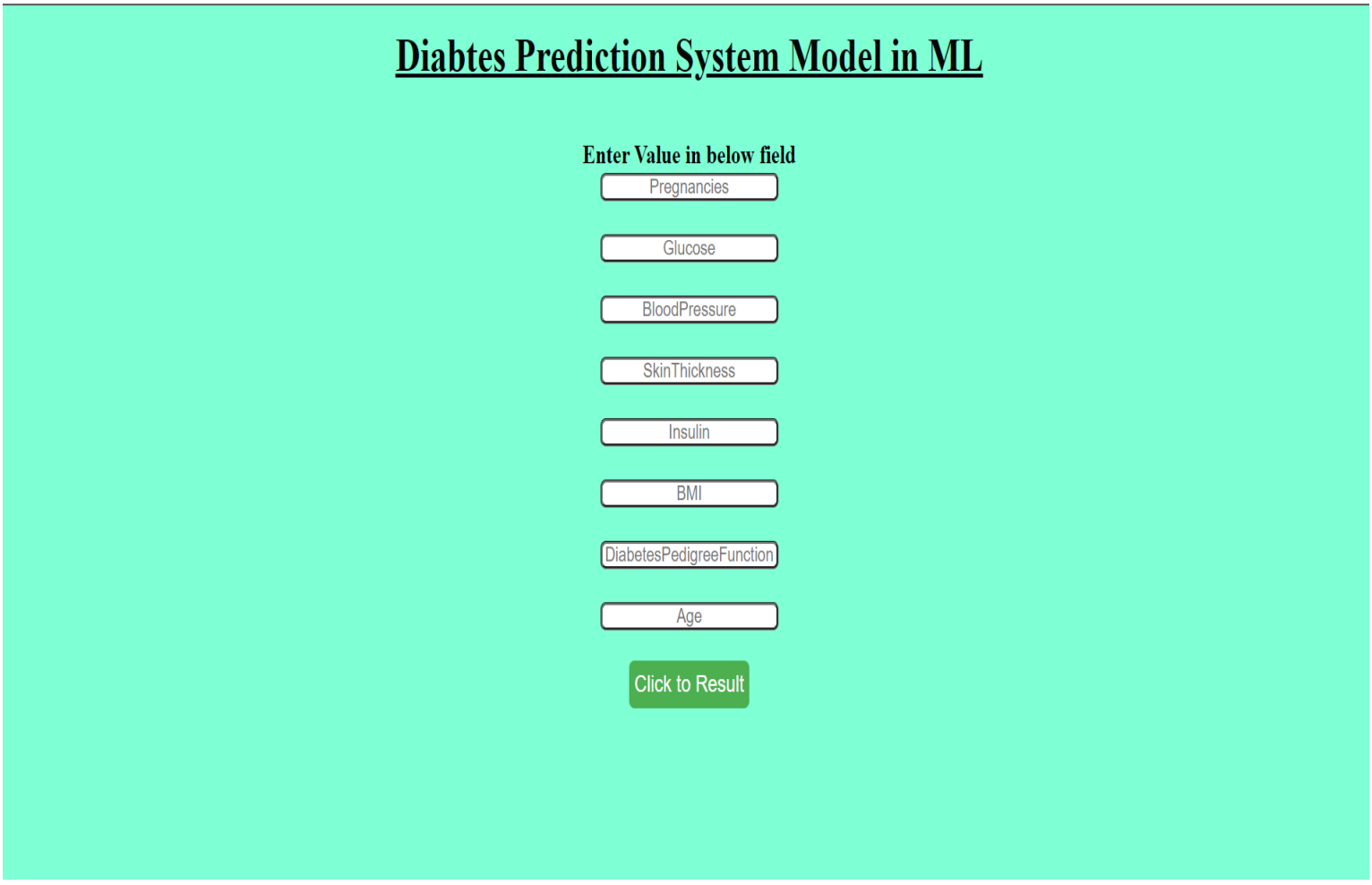
Machine Learning Algorithms	Error (Root mean square error)
Logistic Regression	0.47846366201788553
Decision Tree	0.5542214735280918
Random Forest	0.48658978444889833
Support Vector Machine	0.4715113977918053

Creating a User Interface for Accessibility:

The last part of the project is the creation of a user interface for the model. This user interface is used to enter unseen data for the model to read and then make a prediction. The user interface is created using “Flask” Web app, Hyper Text Markup Language, and Cascading Style Sheets.

CHAPTER 8 – RESULT AND ANALYSIS

The project predicts the onset of diabetes in a person based on the relevant medical details collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of 80%, which is good and reliable. Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes Our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice.



On submission of this form, data the model gives the result in the form of Table; as shown in following figures;

Patient Report

Pregnancies	3.0
Glucose	191.0
BloodPressure	68.0
SkinThickness	15.0
Insulin	130.0
BMI	30.9
DiabetesPedigreeFunction	0.299
Age	34.0
Result	**Non-Diabetic**

Patient Report

Pregnancies	6.0
Glucose	89.0
BloodPressure	66.0
SkinThickness	23.0
Insulin	94.0
BMI	28.1
DiabetesPedigreeFunction	0.962
Age	31.0
Result	**Diabetic**

CHAPTER 9 - CONCLUSION

The objective of the project was to develop a model which could identify patients with diabetes who are at high risk of hospital admission. Prediction of risk of hospital admission is a fairly complex task. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution's understanding of what is important in predicting the hospital admission risk. This project is a small contribution to the present existing methods of diabetes detection by proposing a system that can be used as an assistive tool in identifying the patients at greater risk of being diabetic. This project achieves this by analyzing many key factors like the patient's blood glucose level, body mass index, etc., using various machine learning models and through retrospective analysis of patients' medical records. The project predicts the onset of diabetes in a person based on the relevant medical details that are collected using a Web application. When the user enters all the relevant medical data required in the online Web application, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or nondiabetic. The model is developed using artificial neural network consists of total of six dense layers. Each of these layers is responsible for the efficient working of the model. The model makes the prediction with an accuracy of 80%, which is good and reliable.

CHAPTER 10 - REFRENCES

1. Sahoo, K.S., et al.: An evolutionary SVM model for DDOS attack detection in software defined networks. IEEE Access 8, 132502–132513 (2020)

2. Sahoo, K.S., et al.: A machine learning approach for predicting DDoS traffic in software defined networks. In: 2018 International Conference on Information Technology (ICIT). IEEE (2018)

3. Jakka, A., Vakula Rani, J.: Performance evaluation of machine learning models for diabetes prediction. Int. J. Innov. Technol. Explor. Eng. (IJITEE) 8(11) (2019). ISSN: 2278-3075

4. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting diabetes mellitus with machine learning techniques. Bioinform. Comput. Biol. Sect. J. Front. Genet., published: 06 2018