DIABETES PRIDICTION

END TERM PROJECT

BY

NAVEEN KUSHWAHA & PRAMOD KUMAR PANDEY

Section - K20JS Section - K20JS

Roll Number – RK20JSA31 Roll Number – RK20JSA34

Department of Intelligent

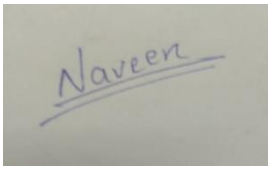
School of Computer Science Engineering

Lovely Professional University, Jalandhar

November – 2022

Student Declaration

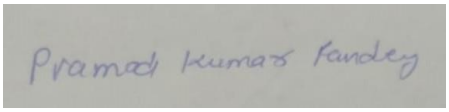
This is to declare that this report has been written by us. No part of the report is copied from other sources. All information included from other sources have been duly acknowledged. We aver that if any part of the report is found to be copied, I/we are shall take full responsibility for it.



Signature of Student:

Name of Student: Naveen Kushwaha

Roll number: RK20JSA31



Signature of Student:

Name of Student: Pramod Kumar Pandey

Roll number: RK20JSA34

Place : LPU Phagwara
Date : 08-11

TABLE OF CONTENTS

Introduction

All around there are numerous ceaseless infections that are boundless in evolved anddeveloping nations. One of such sickness is diabetes. Diabetes is a metabolic issue thatcauses blood sugar by creating a significant measure of insulin in the human body or byproducing 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 coronaryfailure, visual deficiency, kidney ailments and nerve harm, and so on.Subsequently, the identification of such chronic metabolic ailment at a beginningperiod 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 variousdomains [1, 2] we may have the option to find an answer for this issue. ML strategies andneural systems help scientists to find new realities from existing well-being-relatedinformational indexes, which may help in ailment supervision and detection. The currentwork is completed utilizing the Pima Indians Diabetes Database. The point of this frameworkis to make an ML model, which can anticipate with precision the likelihood or the odds of apatient being diabetic. The ordinary distinguishing process for the location of diabetes is thatthe patient needs to visit a symptomatic focus. One of the key issues of bio-informaticsexamination is to achieve precise outcomes from the information. Human mistakes orvarious laboratory tests can entangle the procedure of identification of the disease. Thismodel can foresee whether the patient has diabetes or not, aiding specialists to ensure thatthe patient in need of clinical consideration can get it on schedule and also help anticipate theloss of human lives.

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

Presently, with the ascent of machine learning, AI, and neural systems, and theirapplication in various domains [1, 2] 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 mutantgenes in the chromosome 6, the chromosome that affects the response of the body to variousantigens. 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 1Type 1 diabetes means that the immune system is compromised and the cells fail toproduce insulin in sufficient amounts. There are no eloquent studies that prove thecauses 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 bodycan'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 ofliving.

Data mining and machine learning have been developing, reliable, andsupporting tools in the medical domain in recent years. The data mining method is used topre-process and select the relevant features from the healthcare data, and the machinelearning method helps automate diabetes prediction . Data mining and machine learningalgorithms 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 techniquesare involved, including machine learning, statistics, and database system to discover apattern from the massive amount of dataset . According to Nvidia: Machine learninguses various algorithms to learn from the parsed data and make predictions.

Data Set

The dataset collected is originally from the Pima Indians Diabetes Database isavailable 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 datasetconsists 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:

Series no	Attribute Names	DescriptiCon
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose transformation
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 2000 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: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Fig 3.2 predictions

➔ There is no null values in dataset.

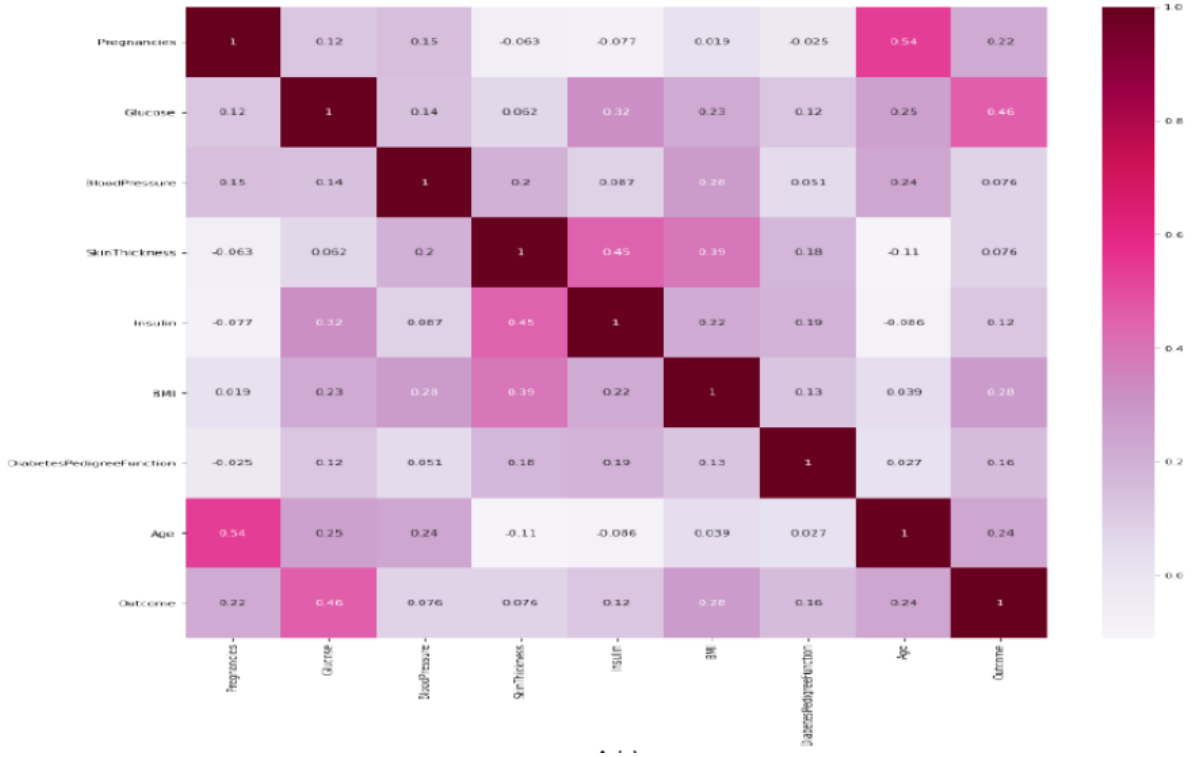


Fig: correlation matrix

It is easy to see that there is no single feature that has a very high correlation with ouroutcome value. Some of the features have a negative correlation with the outcome value andsome have positive.

Skew Of Data:

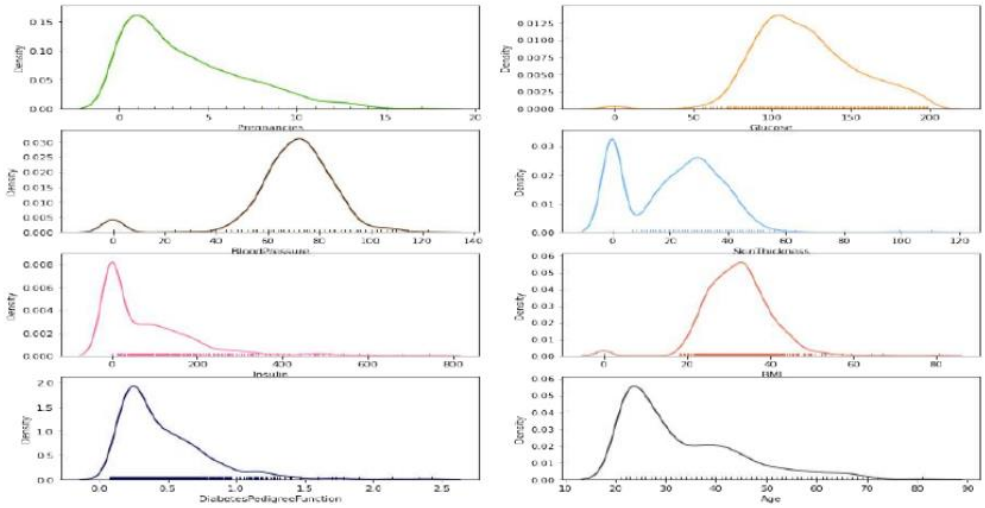


Fig3.4 skew of data

It shows how each feature and label is distributed along different ranges, which furtherconfirms the need for scaling. It basically means that each of these is actually a categoricalvariable. We will need to handle these categorical variables before applying MachineLearning. Our outcome labels have two classes, 0 for no disease and 1 for disease.

Bar Plot For Outcome class

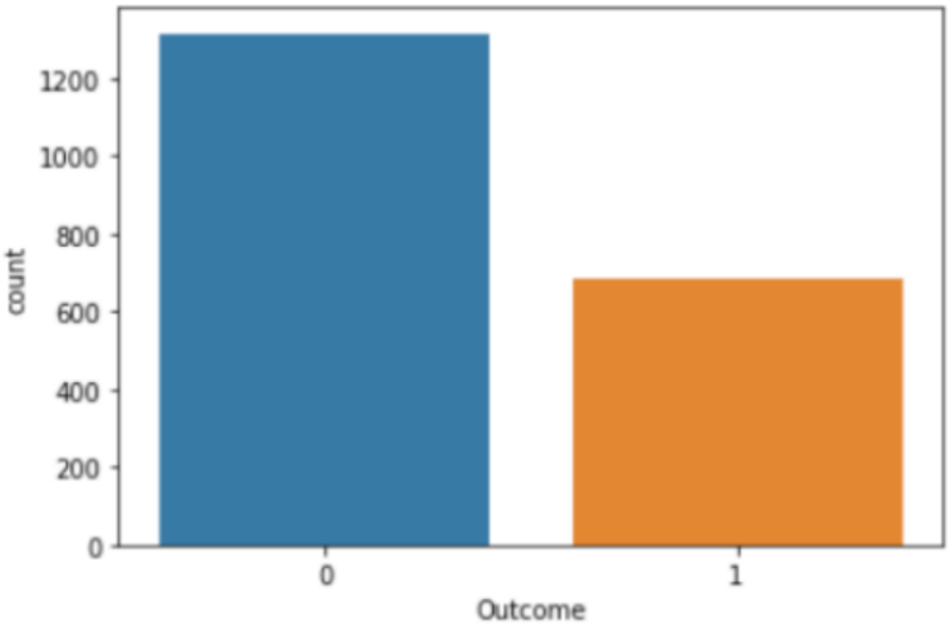
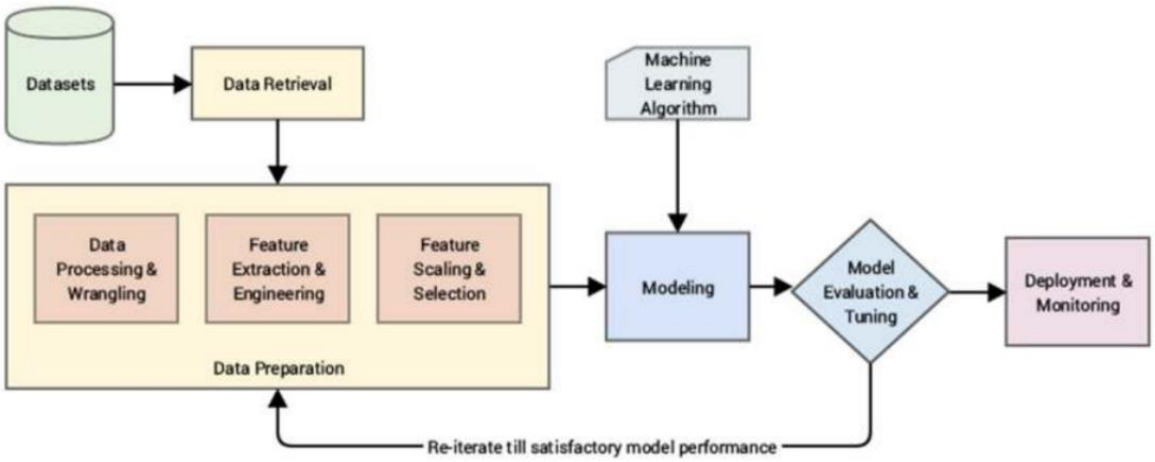


Fig:3.5 Bar plot for outcomes class

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.



PROPOSED METHODS

Dataset collection – It includes data collection and understanding the data to study thehidden patterns and trends which helps to predict and evaluating the results. Dataset carries1405 rows i.e., total number of data and 10 columns i.e., total number of features. Featuresinclude Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI,DiabetesPedigreeFunction, Age

Data Pre-processing:

This phase of model handles inconsistent data in order to get moreaccurate and precise results like in this dataset Id is inconsistent so we dropped the feature.This dataset doesn't contain missing values. So, we imputed missing values for few selectedattributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because theseattributes cannot have values zero. Then data was scaled using StandardScaler. Since therewere a smaller number of features and important for prediction so no feature selection wasdone.

Missing value identification:

Using the Panda library and SK-learn , we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.

Pregnancies	*0*
Glucose	*13*
Blood Pressure	*90*
Skin Thickness	*573*
Insulin	*956*
BMI	*28*
DPF	*0*
Age	*0*
Outcome	*0*

Feature selection:

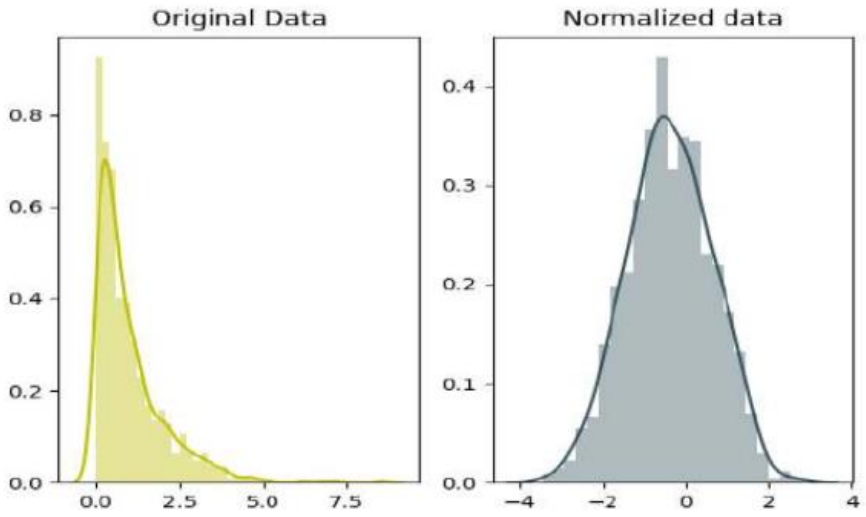
Pearson's correlation method is a popular method to find the mostvrelevant attributes/features. The correlation coefficient is calculated in this method, whichcorrelates with the output and input attributes. The coefficient value remains in the range bybetween −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, andthe zero value means no correlation

Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
Skin Thickness	0.193
BP	0.183
DPF	0.178

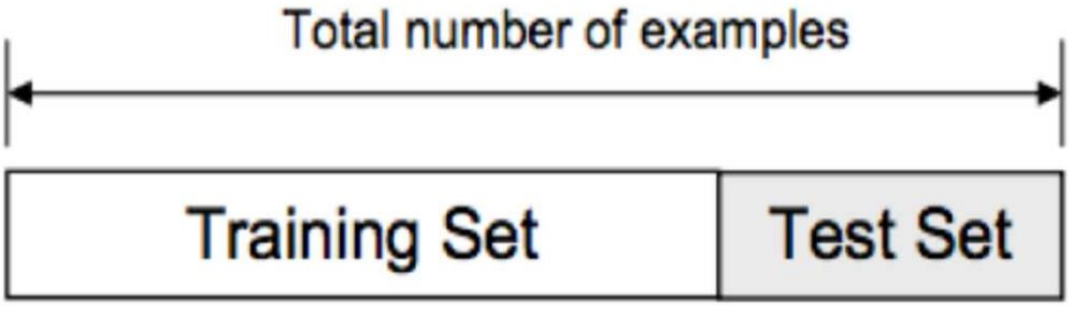
Scaling and Normalization:

We performed feature scaling by normalizing the data from 0 to 1 range, which boosted thealgorithm’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 farapart data points are, like support vector machines (SVM) or k-nearest neighbours (KNN).With these algorithms, a change of "1" in any numeric feature is given the same importance.



After data cleaning and pre-processing, the dataset becomes ready totrain and test. In the train/split method, we split the dataset randomly into the training andtesting set. For Training we took 1600 sample and for testing we took 400 sample



Design and implementation of classification model:

In this research work, comprehensive studies are done by applying different ML classification techniques like DT, KNN, RF, NB, LR, SVM.

Machine learning classifier:

We have developed a model using Machine learning Technique. Used different classifier andensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RFMachine learning classifier to analyse the performance by finding accuracy of each classifierAll the classifiers are implemented using scikit learn libraries in python. The implementedclassification algorithms are described in next section.

MODELING AND ANALYSIS:

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.

K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of newdatapoints which further means that the new data point will be assigned a value based onhow closely it matches the points in the training set. Predictions are made for a new instance (x) by searching through the entire training set forthe K most similar instances (the neighbors) and summarizing the output variable for those Kinstances.

SVM:

SVM is supervised learning algorithm used for classification. In SVM we have to identifythe right hyper plane to classify the data correctly. In this we have to set correct parametersvalues. To find the right hyper plane we have to find right margin for this we have

choose thegamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads tomiss classification.

Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes'Theorem. It is not a single algorithm but a family of algorithms where all of them share acommon principle, i.e. every pair of features being classified is independent of each other.

Decision Tree:

Decision tree is non parametric classifier in supervised learning. In this method all the detailsare represented in the form of tree, where leaves are corresponds to the class labels andattributes are corresponds to internal node of the tree. We have used Gini Index for splittingthe nodes.

Random Forest:

Random forest is an ensemble learning method for classification. This algorithm consists oftrees 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 internalnode of the tree. Here number of trees in forest used is 100 in number and Gini index is usedfor splitting the nodes

AdaBoost Classifier:

Boosting is an ensemble modeling technique that attempts to build a strong classifier fromthe number of weak classifiers. It is done by building a model by using weak models inseries. Firstly, a model is built from the training data. Then the second model is built whichtries to correct the errors present in the first model. This procedure is continued and modelsare added until either the complete training data set is predicted correctly or the maximumnumber of models are added.

AdaBoost was the first really successful boosting algorithm developed for the purpose ofbinary classification. AdaBoost is short for Adaptive Boosting and is a very popular boostingtechnique that combines multiple “weak classifiers” into a single “strong classifier”. It wasformulated by Yoav Freund and Robert Schapire. They also won the 2003 Gödel Prize fortheir work.

Measurements

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

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

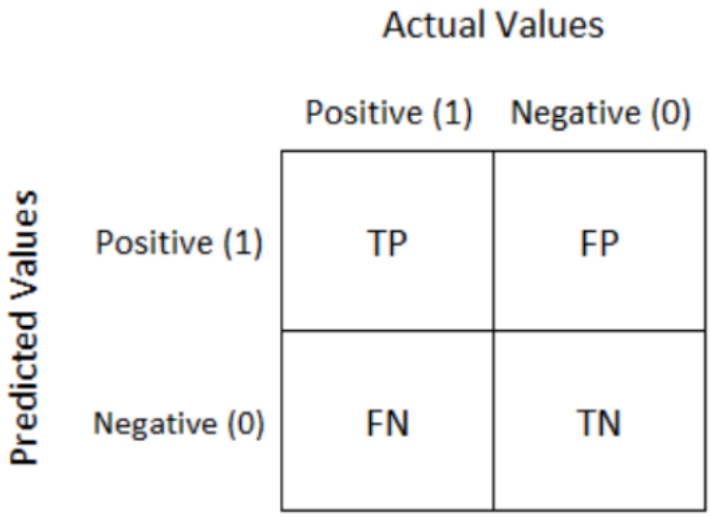
Here,

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 inreality.

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.

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

RESULTS AND DISCUSSION

Machine learning classification algorithms developed for prediction of diabetes in earlierstage. We used 70% of data for trining and 30% of data for testing. In this ratio of datasplitting Here we found that Random Forest Classifier predicted with 99% of accuracy ashighest accuracy for the dataset. Comparison of results of all the implemented classifiers arelisted in below.

Machine Learning Algorithms	Results
Logistic Regression	79.0
k-Nearest Neighbors	80.5
SVM	84.5
Naïve Bayes	76.83
Decision Tree	96.0
Random Forest	98.0
AdaBoost Classifier	81.16

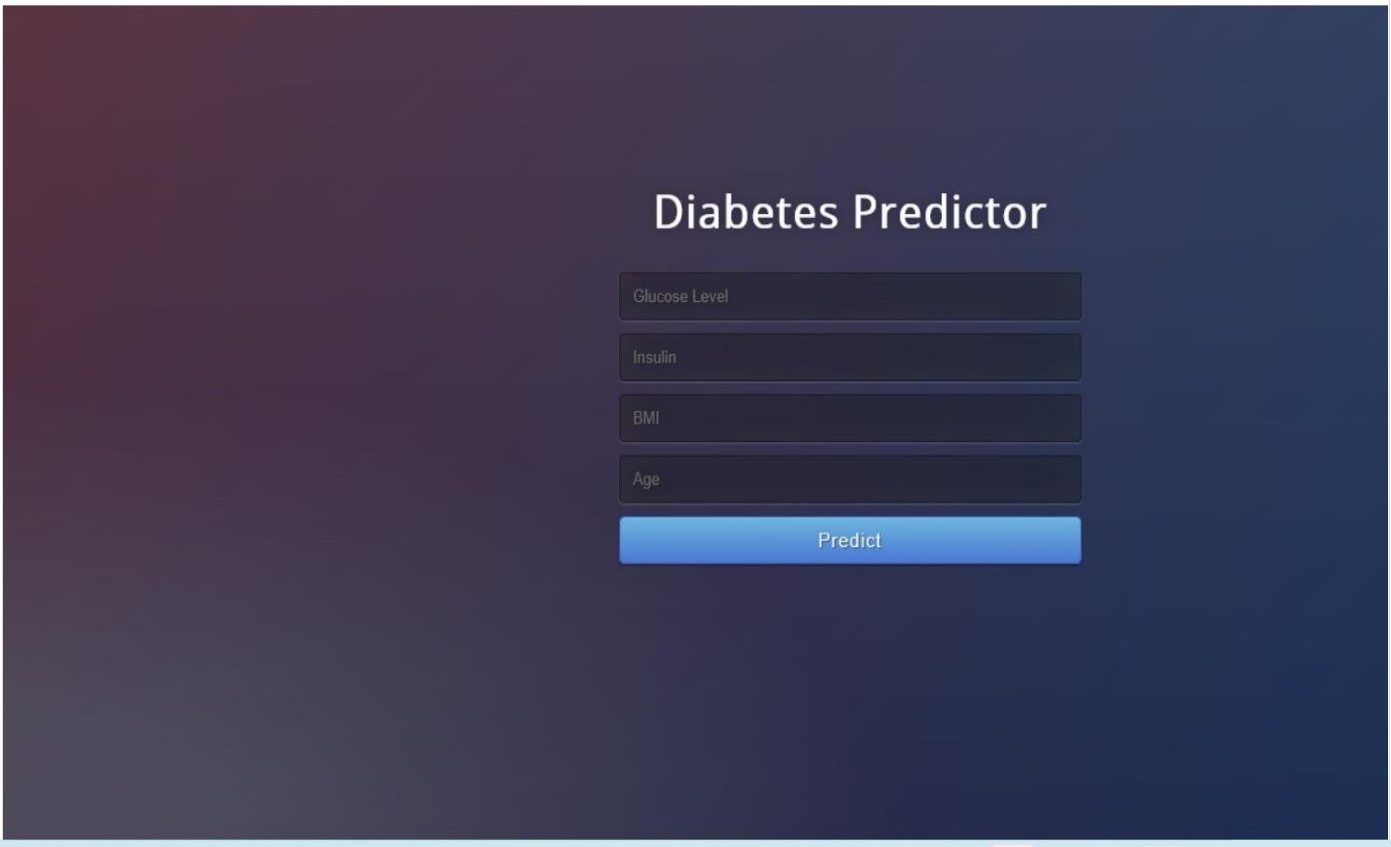
Fig fi.1 Results

Creating a User Interface for Accessibility:

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

8. Results and Analysis

The project predicts the onset of diabetes in a person based on the relevant medical detailscollected. When the person enters all the relevant medical data required in the online Webportal, this data is then passed on to the trained model for it to make predictions whether theperson is diabetic or non-diabetic the model then makes the prediction with an accuracy of98%, which is fairly good and reliable. Following figure shows the basic UI form whichrequires the user to enter the specific medical data fields. These parameters help determine ifthe person is prone to develop diabetes Our research has the added benefit of an associatedWeb app, which makes the model more user friendly and easily understandable for a novice



Conclusion

1. Diabetes is one of the ricks during Pregnancy. It has to be treat to avoid complications.
2. BMI index can help to avoid complications of diabetes a way before
3. Diabetes start showing in age of 35 – 40 and increase with person age.

THANK YOU