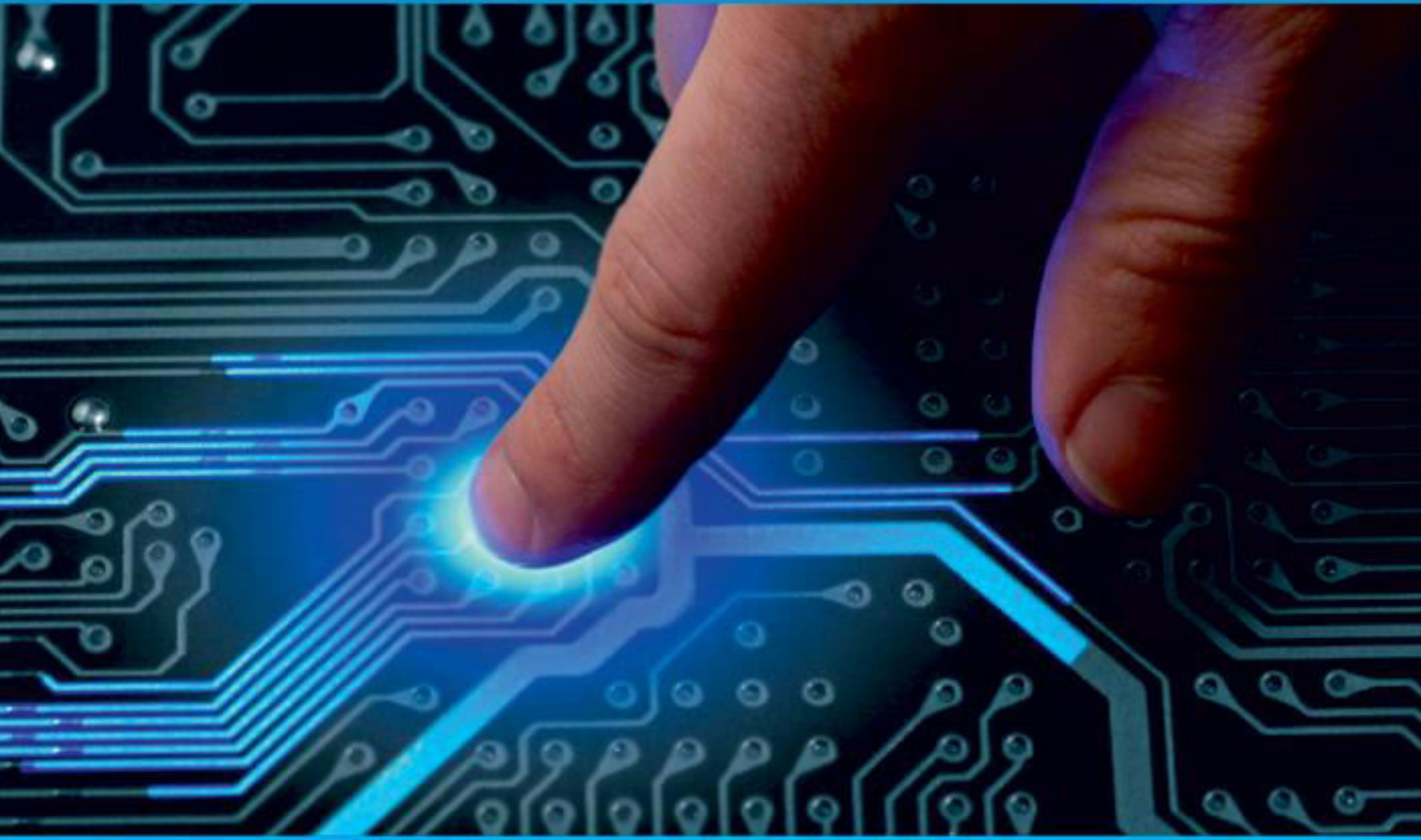
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104158 |

Diabetes Prediction Using Machine Learning Techniques

Joya Javed Shaikh , Sankita Sunil Katekar , Roshankumar Nayaku Lavate , Kedar Indrajit Sutar, Prof.Snehal Bahubali Farande

Student, Department Computer Science and Engineering, Dr. J. J. Magdum College of Engineering, Jaysingpur, India

Assistant Professor, Department Computer Science and Engineering, Dr. J. J. Magdum College of Engineering, Jaysingpur, India

ABSTRACT:
Diabetes is one of the major and deadly diseases. it is also cause of many diseases such as heart-attack, kidney diseases, blindness etc. all among the world many people are suffering from the diabetes. Diabetes can be caused by obesity, lack of exercise, bad living style, due to heredity, high blood pressure. In traditional practices in hospital required information is collected through various tests and treatment is provided on the basis of the diagnosis. Here Big data analytics helps us to find hidden patterns and information which helps us to extract knowledge from the data and predict outcomes. in this paper we have proposed a diabetes prediction model for classification of diabetes based on some regular factors such as Glucose, BMI, Insulin, Age etc. and we have tried to find maximum accuracy with the help of machine learning algorithms.

KEYWORDS: Machine learning algorithms , Dataset, Random Forest ,SVM,

I. INTRODUCTION

Diabetes is chronic disease which is caused when your blood glucose also known as blood sugar is too high. Blood glucose is the main source of energy which comes from the food you eat. A hormone called Insulin which is created by a pancreas helps the glucose from the food to get into your cells to be used for energy. Sometimes the body doesn’t make enough insulin or any Insulin then the glucose stay inside the body and causes disease such as diabetes. In the traditional process of identifying diabetes the patient has to visit the diagnostic centre again and again, go through various tests and have to wait for day or more to get their reports. It also requires lots of money for various steps. in this project we are using various machine learning techniques. Machine learning is a subset of AI which helps machines to automatically learn from previous data improve performance from past experiences and provide outcomes on the basis of learning. Machine learning contains bunch of algorithms that works on huge amount of data. This data is used to train the models and on the basis of the training the models performs specific tasks. Machine learning has various types such as;

- **SUPERVISED LEARNING:**

Supervised learning is one of the types of machine learning where labeled data is provided to the model. Labeled data contains a target variable or an output variables that answers a questions of interest. A supervised learning model is a model which learns under supervision. this supervision is provided by labeled data which contains target variables and independent variables. the model learns from the past data. The most widely used supervised machine learning algorithms are Logistic Regression, Random Forest, Gradient Boosted Trees and Support Vector Machine(SVM).

- **UNSUPERVISED LEARNING:**

Unsupervised learning is another type of machine learning**,** which acts as a complement of supervised learning. The machine learning that is deployed to find a patterns in unlabeled data is referred to as unsupervised machine learning. There is no target variables involved in unsupervised machine learning. It only works on unlabeled data. Unsupervised machine learning identifies if some pattern exists in the data. Clustering and Association Rules are two categories in unsupervised machine learning.

- **REINFORCEMENT LEARNING**:

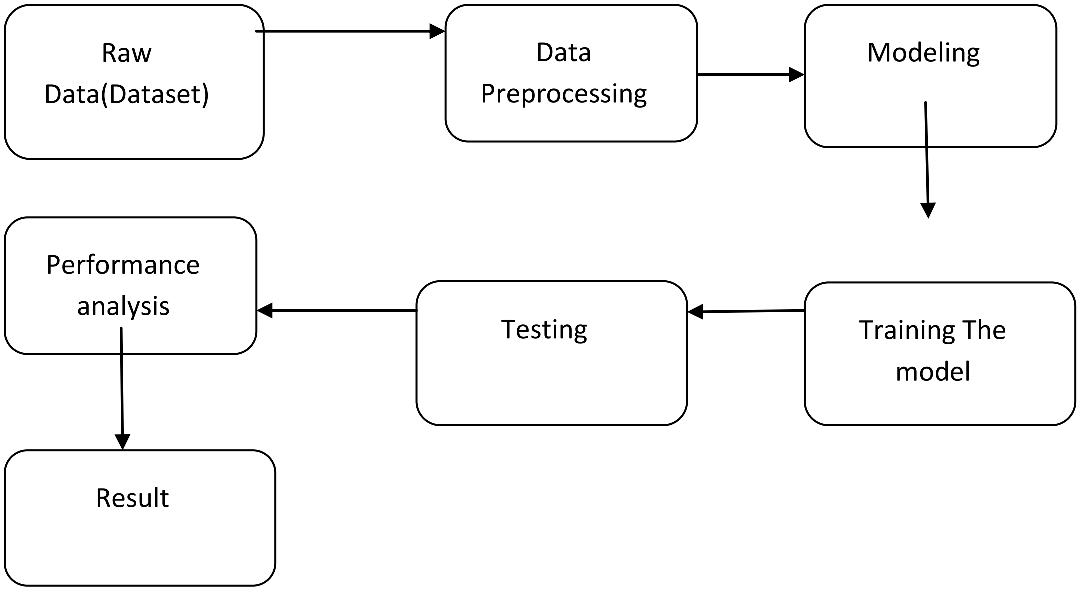
Reinforcement learning is another important area of machine learning. It is a feedback based machine learning techniques in which an agent learns to behave in an environment. By performing any particular action and seeing its results. For each good action the agent gets positive feedback and for each bad action the agent gets negative feedback. In reinforcement learning there is no labeled data like supervised machine learning, it automatically works on feedbacks.

- **DATASET:**

  For this project we have collected dataset from Pima Indians Diabetes Database which is available on Kaggle. This dataset consist of many medical analyst variables and one target variable. The objective of this dataset is to predict whether a patient has diabetes or not. This dataset has several independent variables and one dependant variable i.e. outcome. The independent variables such as pregnancies of patient, their BMI, Insulin level etc.

Sr. No.	**Attributes Names**	**Description**
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

ARCHITECTURE:

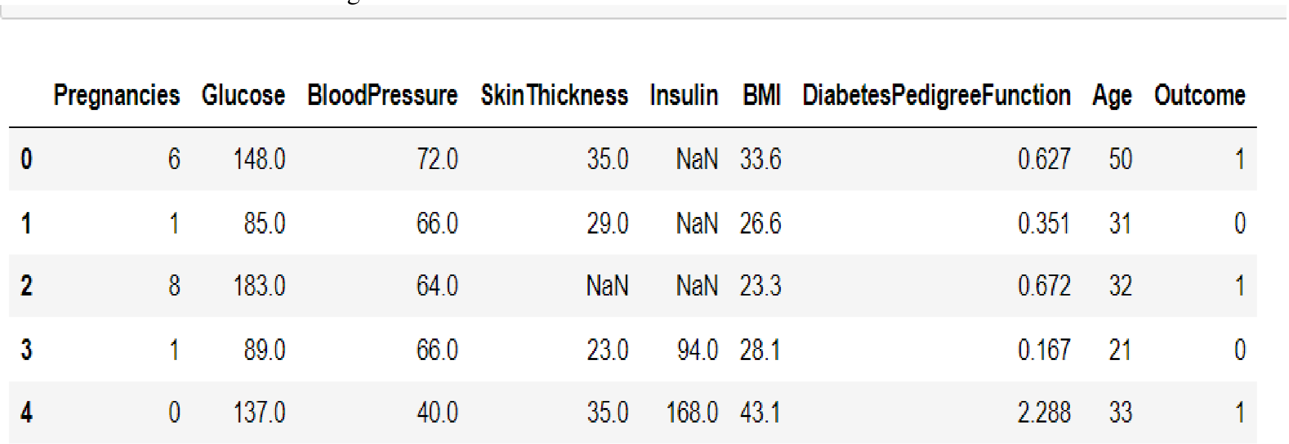


MODULES:

1. **DATASET COLLECTION:**
   Data collection is initial phase of study, where we understand the hidden patterns and trends which helps to predict outcome. This collection data has attributes such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

2. DATA PRE –PROCESSING:

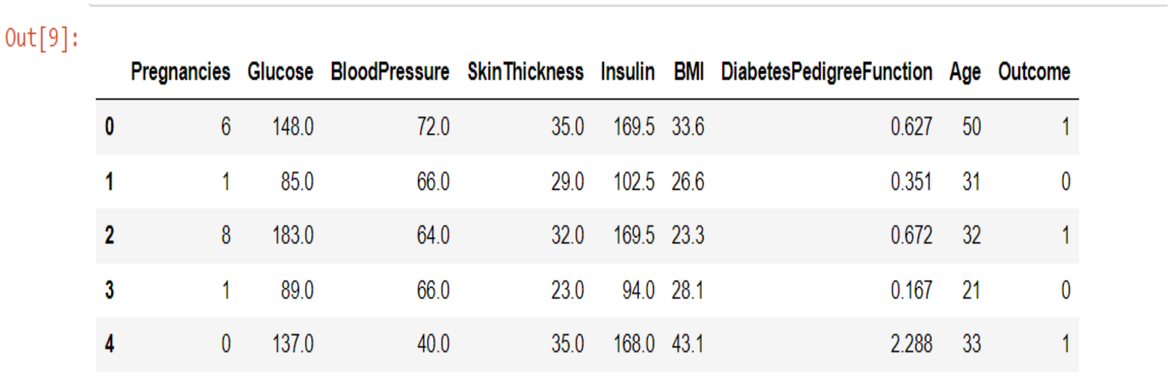
In this module we are handling inconsistent data to get more accurate results. Those figures which are not required are eliminated in this phase. This data doesn't contain missing values. The attributes cannot have zero values. Then data was scaled using StandardScalar



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

3.MISSING VALUES IDENTIFICATION:

Using the panda library and SK-Learn we got the missing values in the dataset and we replaced the missing values with the corresponding mean value.

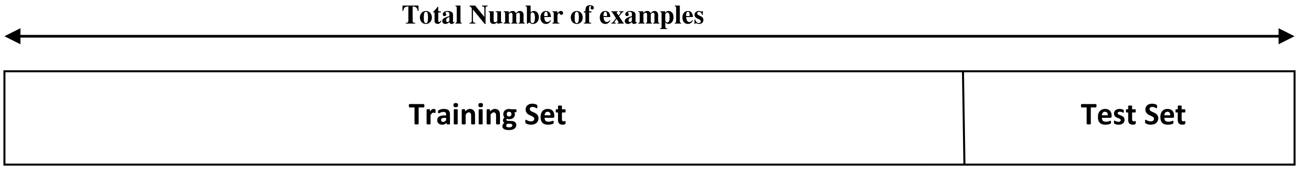


Out[9]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	169.5	33.6	0.627	50	1
1	1	85.0	66.0	29.0	102.5	26.6	0.351	31	0
2	8	183.0	64.0	32.0	169.5	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

4. SPLITTING OF DATA:

After data cleaning and pre- processing the dataset becomes ready to train and test. In splitting method we split the dataset randomly into the training and testing set.



Total Number of examples

Training Set

Test Set

5. DESIGN AND IMPLEMENTATION:

In this modules we have applied different machine learning classification techniques like Logistic Regression, Support Vector Machine (SVM),Random Forest Algorithms.

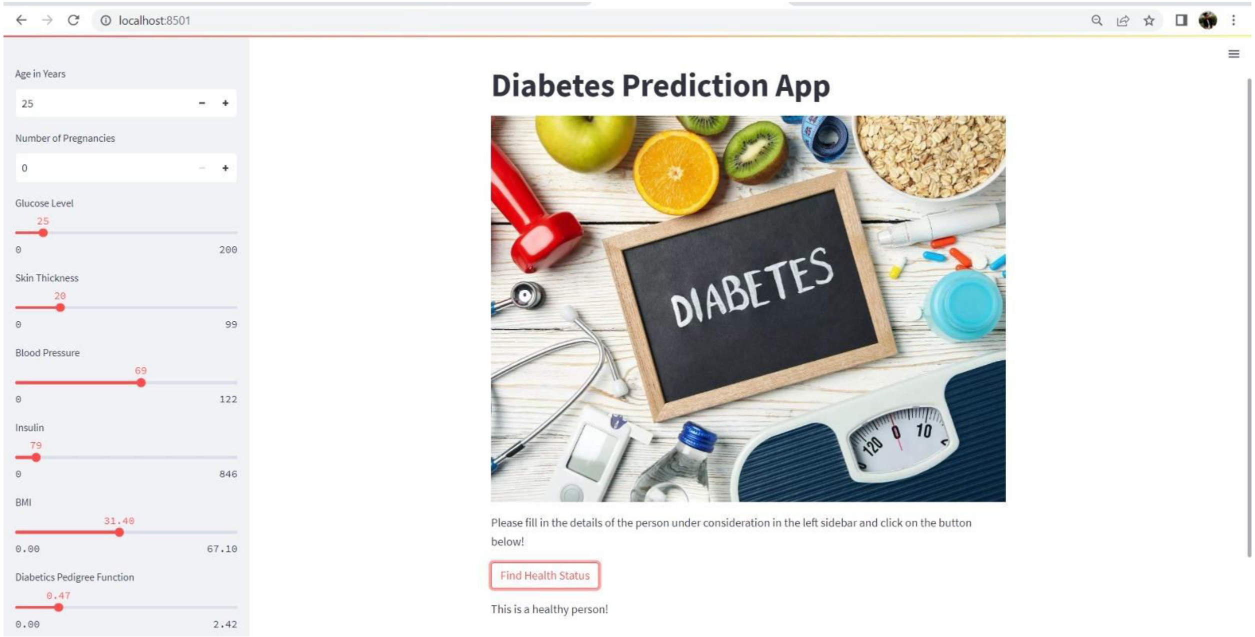
- **Logistic Regression:**
  Logistic regression is machine learning technique used when dependent variables are able to categorize. The output is obtain dependant on available feature.
- **Support Vector Machine(SVM):**
  Support Vector Machine or SVM is one of the most popular supervised machine learning techniques. SVM can be used for classification as well as regression problems.
- **Random Forest Algorithm**:
  Random Forest Algorithm is popular machine learning algorithm. It is a supervised machine learning techniques. It also can be used for Classification as well as Regression. Random Forest is made up of Decision Trees.

II. RESULT

This machine learning model is made up with the help of machine learning technique to predict diabetes in earlier stage. we used 80% of data for training and 20%of data for testing. With the help of the ratio of data here we fund that random forest classifier predicted with 90% of accuracy as highest accuracy for the dataset.

- **CREATING A USER INTERFACE FOR ACCESSIBILITY**:

The last part of this project is creating a user interface for the model. This user interface is used to enter unknown data for the model to read and make predictions. We have created this user interface with the help of streamlit by importing streamlit library in our Environment.



Age in Years: 25

Number of Pregnancies: 0

Glucose Level: 25 (0 – 200)

Skin Thickness: 20 (0 – 99)

Blood Pressure: 69 (0 – 122)

Insulin: 79 (0 – 846)

BMI: 31.40 (0.00 – 67.10)

Diabetics Pedigree Function: 0.47 (0.00 – 2.42)

Diabetes Prediction App

Please fill in the details of the person under consideration in the left sidebar and click on the button below!

This is a healthy person!

III. CONCLUSION

The objective of this project was to develop a model which could identify patient with diabetes who are at higher risk of hospitalization. There is presently a serious needs for methods that can help to detect the patients risk of hospital admission. This project will help us to find that. In this project we have used machine learning techniques to predict if the person has diabetes or not. When the person enters all the required medical data in the web application this data is passed on to the trained model for it to make predictions if the person os diabetic or non-diabetic. This model makes the prediction with an accuracy of 95% which is fairly good and reliable.

REFERENCES

1. Sahoo,K.S.a machine learning approach for predicting DDos traffic 2019n software defined network In: 2018 International Conference on Information Technology(ICIT).IEEE(2018)
2. American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064
3. WHO Global Action Plan on Physical Activity 2018-2030: More Active People for Healthier World Health Organization,Geneva,Switzerland.2019
4. M.V.D .Schaar, A.M. Alaa, A. floto et al.,” How artificial intelligence and machine learning can help healthcare system respond to COVID-19.”Machine Learning ,vol.110,no.1.pp.1-14,2021.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING