PROJECT REPORT

ON

"Early Prediction of Diabetes Mellitus in Pregnant Women Using Machine Learning"

Submitted in the partial fulfillment of requirements

FOR

SMARTINTERNZ INTERNSHIP

Submitted by

ARPITA PM **4BD18CS011**

2021

Department of Computer Science and Engineering.

Bapuji Institute of Engineering & Technology

Davangere- 577004

1.INTRODUCTION

1.1 Overview

1.2 Purpose

2 .LITERATURE SURVEY

2.1 Existing problem

2.2 Proposed solution

3 .THEORITICAL ANALYSIS

3.1 Block diagram

3.2 Hardware / Software designing

4 .EXPERIMENTAL INVESTIGATIONS

5 .FLOWCHART

6 .RESULT

7 .ADVANTAGES & DISADVANTAGES

8 .APPLICATIONS

9 .CONCLUSION

10 .FUTURE SCOPE

11 .BIBILOGRAPHY

1.INTRODUCTION

1.1 Overview

Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves.

With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels.

1.2 Purpose

People with diabetes must take responsibility for their day-to-day care.

This includes monitoring :

a. blood glucose levels
b. dietary management
c. Blood pressure
d. insulin use via injections or pump.

2 .LITERATURE SURVEY

2.1 Existing problem

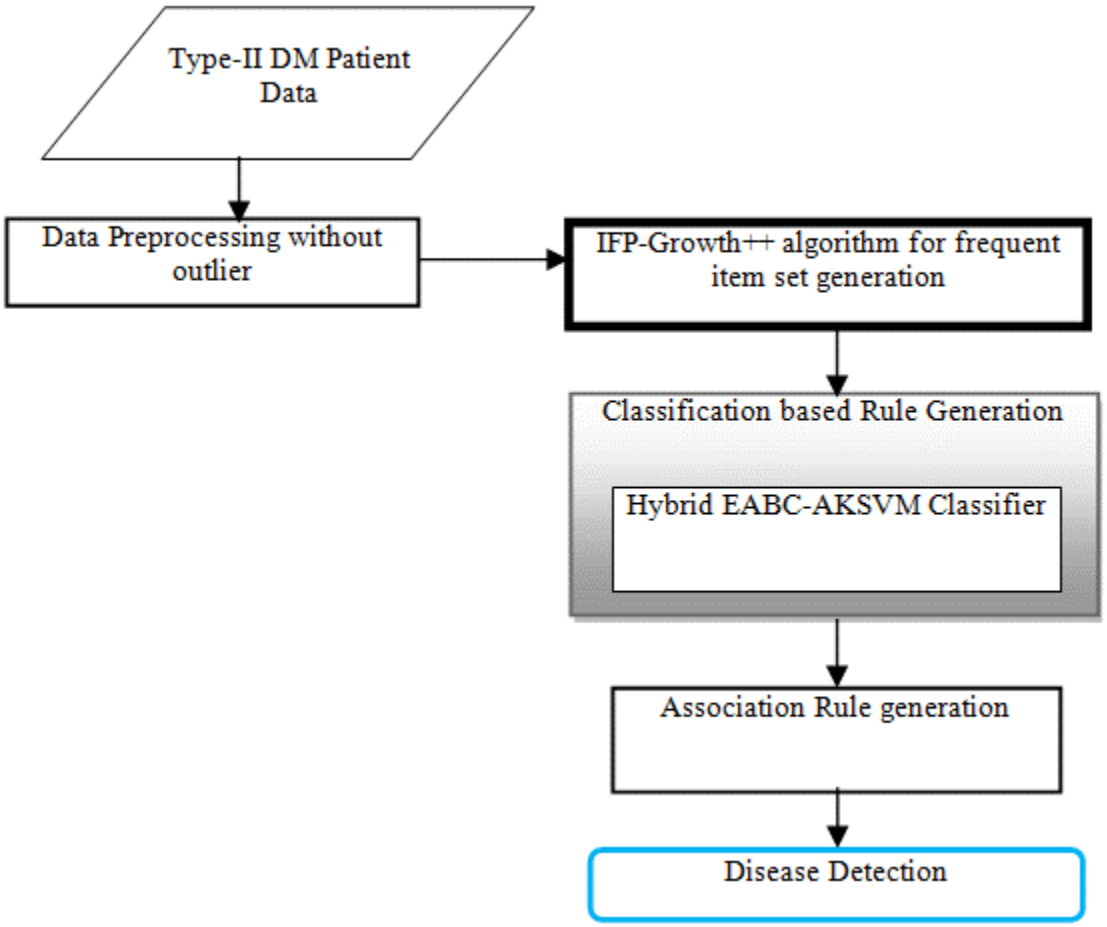
Diabetes a non-communicable disease is leading to long-term complications and serious health problems. A report from the World Health Organisationaddresses diabetes and its complications that impact on individual physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred due to the risk factors of diabetes like a cardiovascular and other diseases.

2.2 Proposed Solution

AI and ML supports various classifiers, decision support system is proposed that uses the AdaBoost algorithm with Decision Stump as a base classifier for classification. Moreover, Support Vector Machine, Naive Bayes and Decision Tree ,XGB Classifier have additionally executed as a base classifiers for AdaBoost calculation for exactness confirmation.

3 .THEORITICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

HARDWARE:

1. Processor : Intel dual core or above
2. Processor speed : 1.0 GHz or above
3. RAM : 4GB RAM or above

4. Hard Disk : 20GB hard disk or above

SOFTWARE:

1. Language : Python
2. User Interface Design: HTML
3. Web Browser : Google Chrome

4 .EXPERIMENTAL INVESTIGATIONS

Data set:

The PIMA Indians Diabetes dataset obtained from Kaggle , has been originally collected to test the ML model and predict diabetes patients.

Both datasets consist of female patients which is why the number of times of pregnancy is one of the most important features for our analysis. We have evaluated several Machine Learning classifiers on these particular datasets to predict whether a female patient is suffering from diabetes or not.

SL	Feature Name	Description	Min val	Max val	Mean
1	Number of pregnancy	Number of times pregnant	0	17	3.85
2	Glucose concentration	2-h oral glucose test (mg/dL)	0	199	120.89
3	Blood Pressure	Diastolic blood pressure (mm Hg)	0	122	69.11
4	Skin thickness	Triceps skin fold thickness (mm)	0	99	20.54
5	Serum Insulin	2-H serum insulin (mu U/mL)	0	846	79.80
6	BMI	Body mass index (kg/m^2)	0	67.10	31.99
7	Diabetes Pedigree Function	Diabetes in family history	0.08	2.42	0.47
8	Age	Age in Years	21	81	33.42

Analysis:

head() method is used to return top n (5 by default) rows of a DataFrame or series.

```
# showing the data from top 5
data.head()
```

	preg	plas	pres	Skin	test	mass	pedi	age	class
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

tail() method is used to return bottom n (5 by default) rows of a DataFrame or series.

```
# showing the data from bottom 5
data.tail()
```

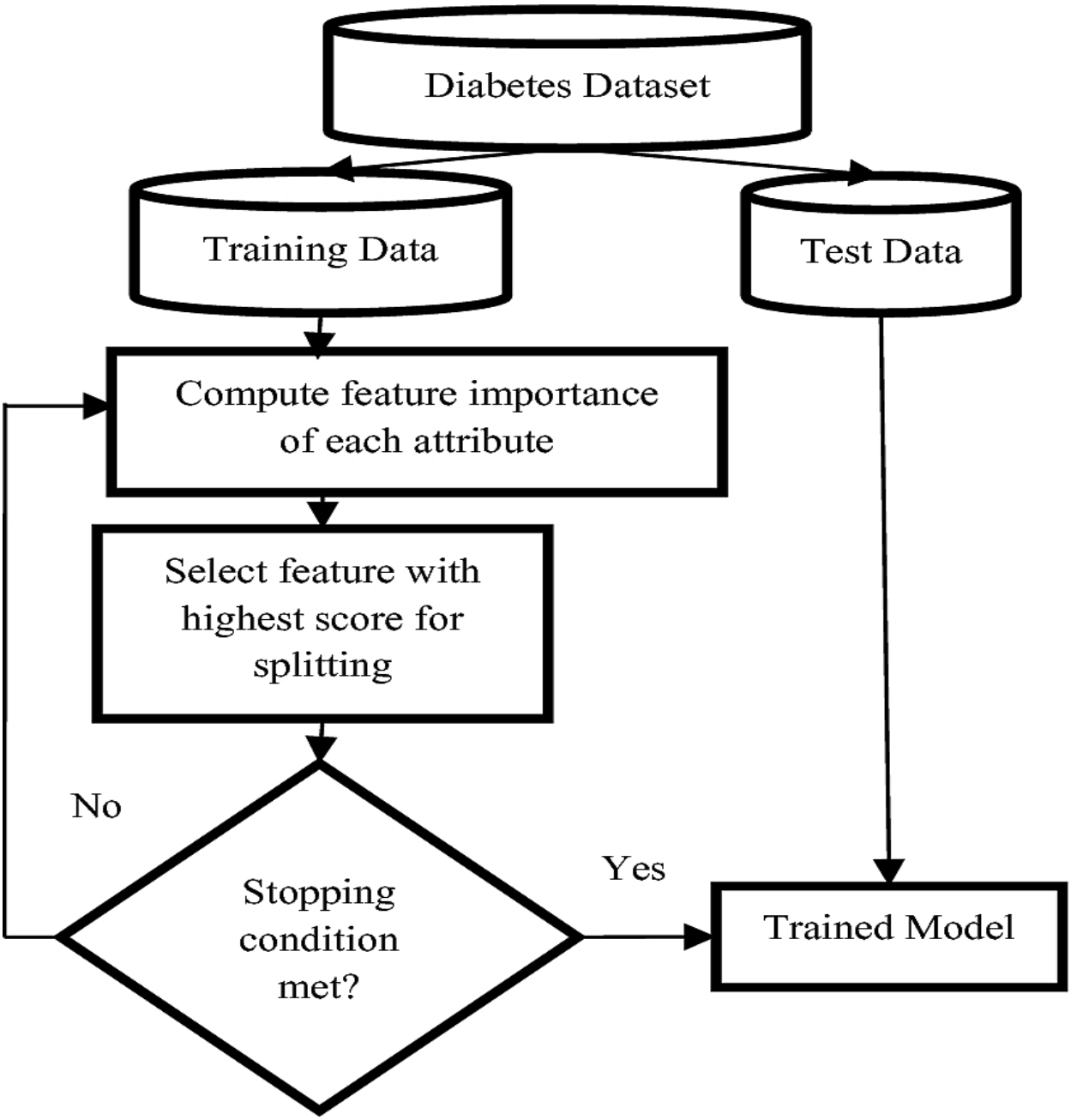
	preg	plas	pres	Skin	test	mass	pedi	age	class
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

3. describe() method computes a summary of statistics like count, mean, standard deviation, min, max, and quartile values.

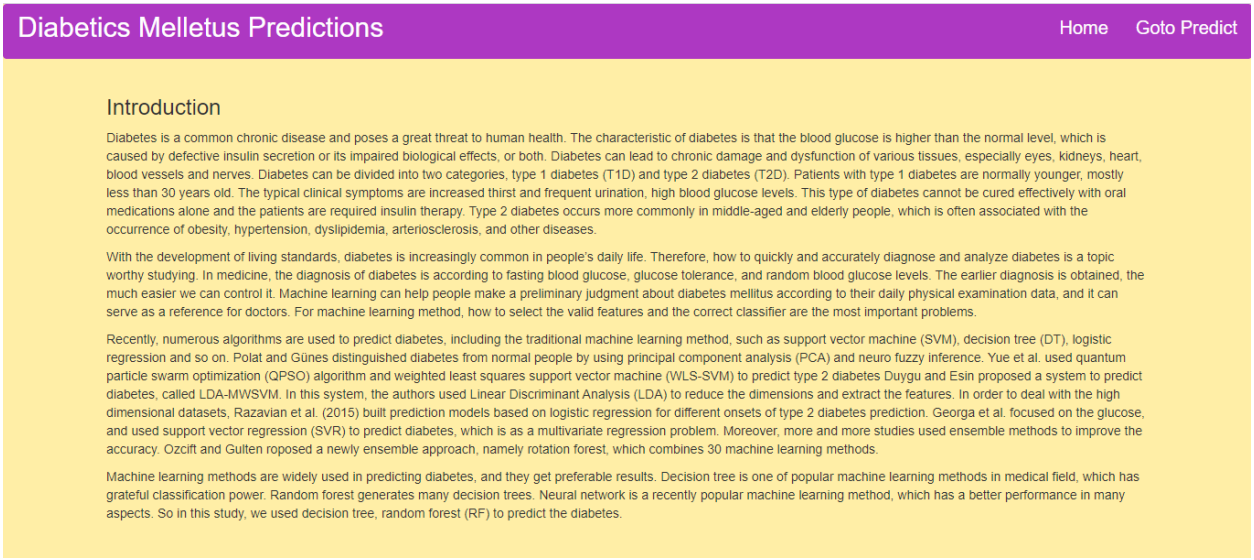
```
# Computes a summary of statistics pertaining to the DataFrame columns
data.describe()
```

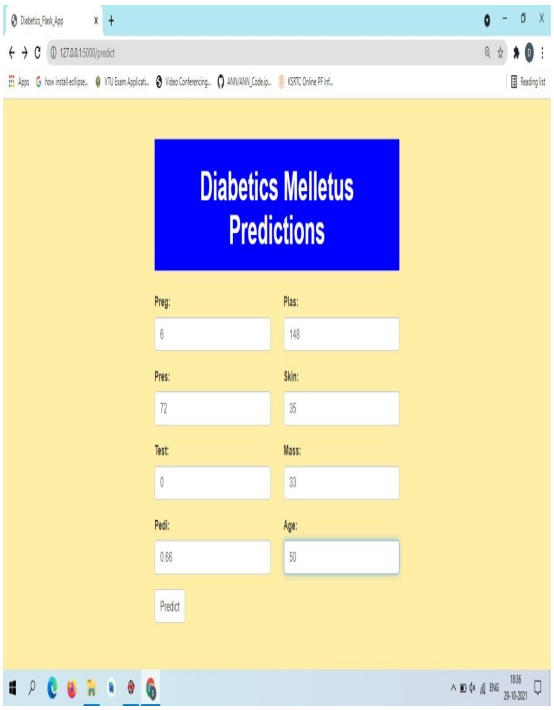
	preg	plas	pres	Skin	test	mass	pedi	age	class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

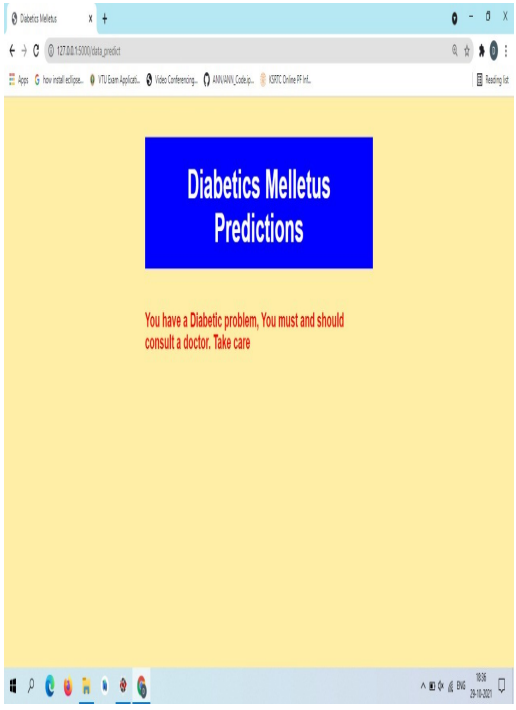
5 .FLOWCHART

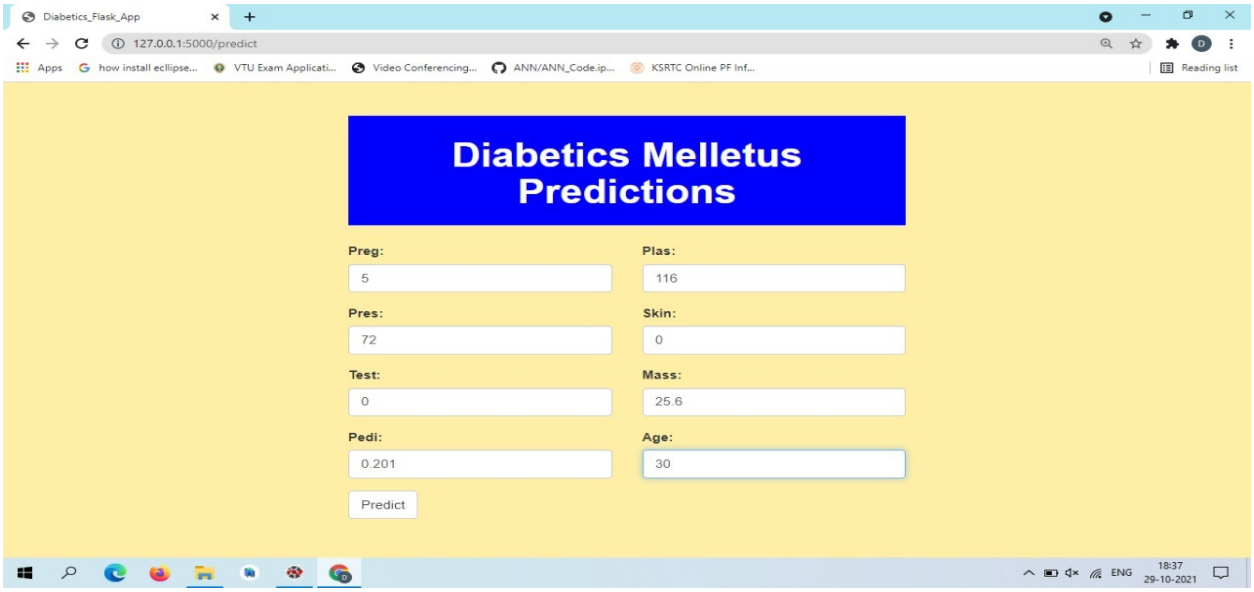

Diabetes Dataset
Training Data
Test Data
Compute feature importance of each attribute
Select feature with highest score for splitting
No
Stopping condition met?
Yes
Trained Model

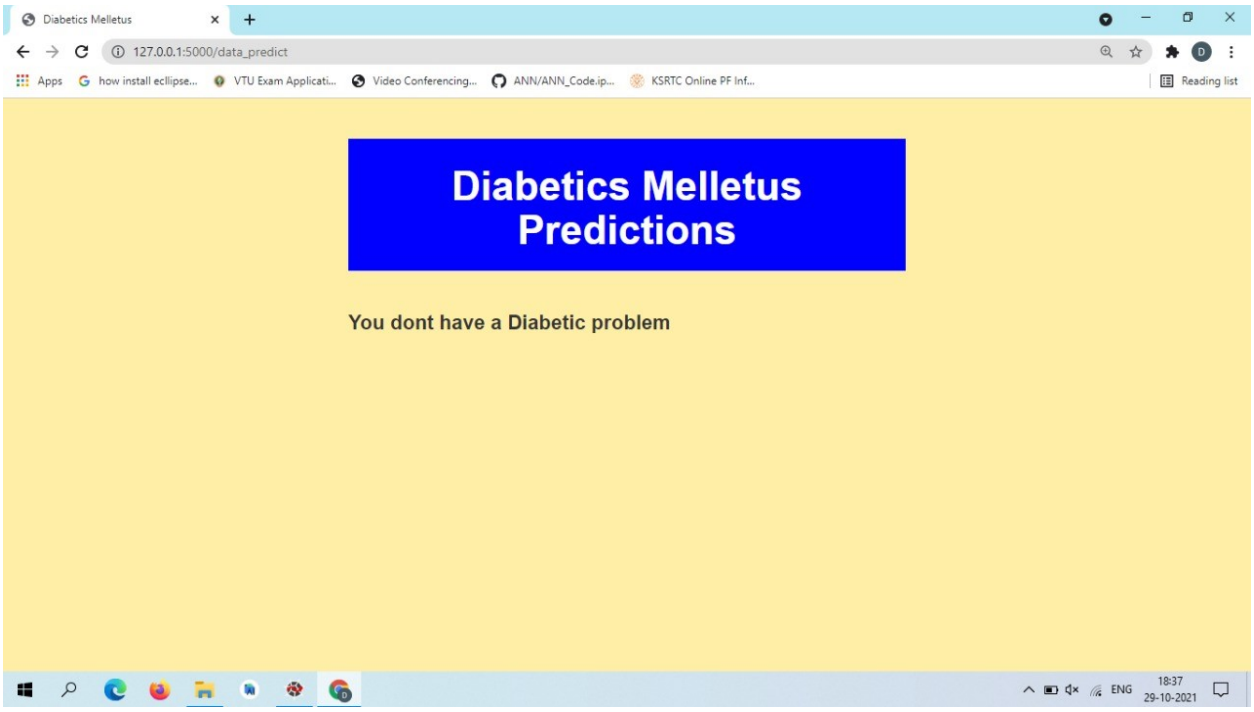
6 .RESULT

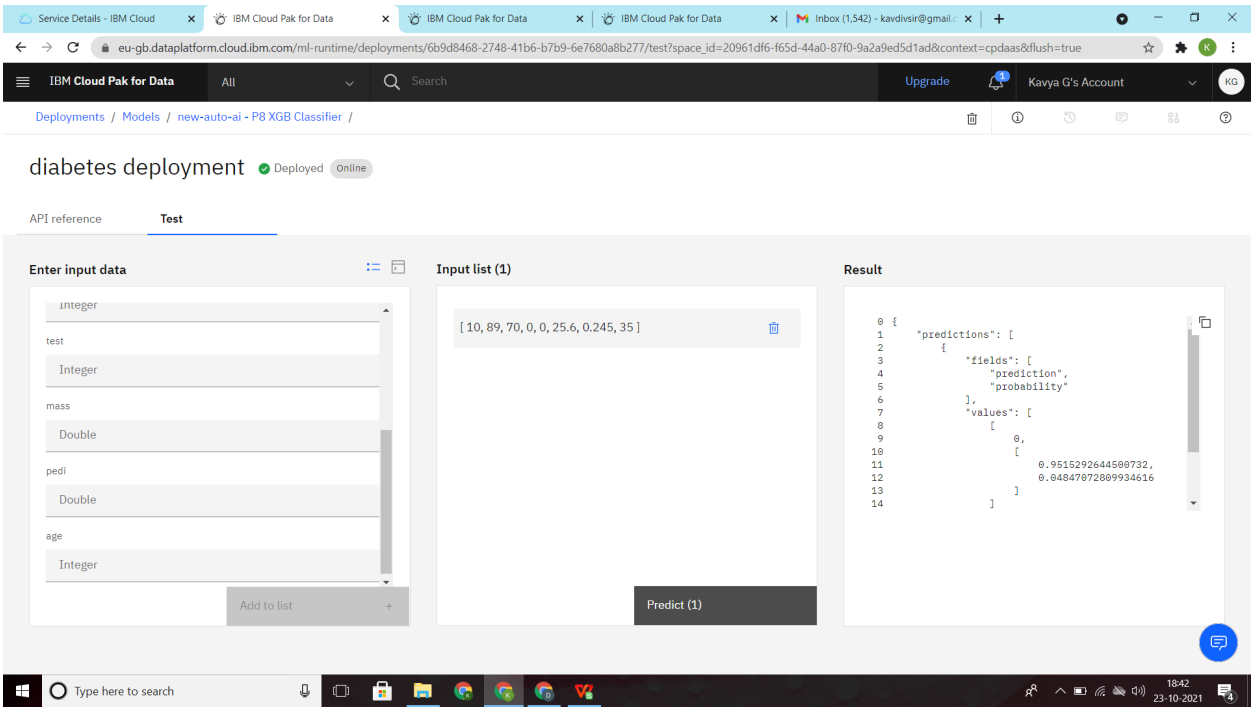

Diabetics Melletus Predictions
Home
Goto Predict
Introduction
Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both. Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels. This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases.
With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. For machine learning method, how to select the valid features and the correct classifier are the most important problems.
Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method, such as support vector machine (SVM), decision tree (DT), logistic regression and so on. Polat and Günes distinguished diabetes from normal people by using principal component analysis (PCA) and neuro fuzzy inference. Yue et al. used quantum particle swarm optimization (QPSO) algorithm and weighted least squares support vector machine (WLS-SVM) to predict type 2 diabetes Duygu and Esin proposed a system to predict diabetes, called LDA-MWSVM. In this system, the authors used Linear Discriminant Analysis (LDA) to reduce the dimensions and extract the features. In order to deal with the high dimensional datasets, Razavian et al. (2015) built prediction models based on logistic regression for different onsets of type 2 diabetes prediction. Georga et al. focused on the glucose, and used support vector regression (SVR) to predict diabetes, which is as a multivariate regression problem. Moreover, more and more studies used ensemble methods to improve the accuracy. Ozcift and Gulten roposed a newly ensemble approach, namely rotation forest, which combines 30 machine learning methods.
Machine learning methods are widely used in predicting diabetes, and they get preferable results. Decision tree is one of popular machine learning methods in medical field, which has grateful classification power. Random forest generates many decision trees. Neural network is a recently popular machine learning method, which has a better performance in many aspects. So in this study, we used decision tree, random forest (RF) to predict the diabetes.

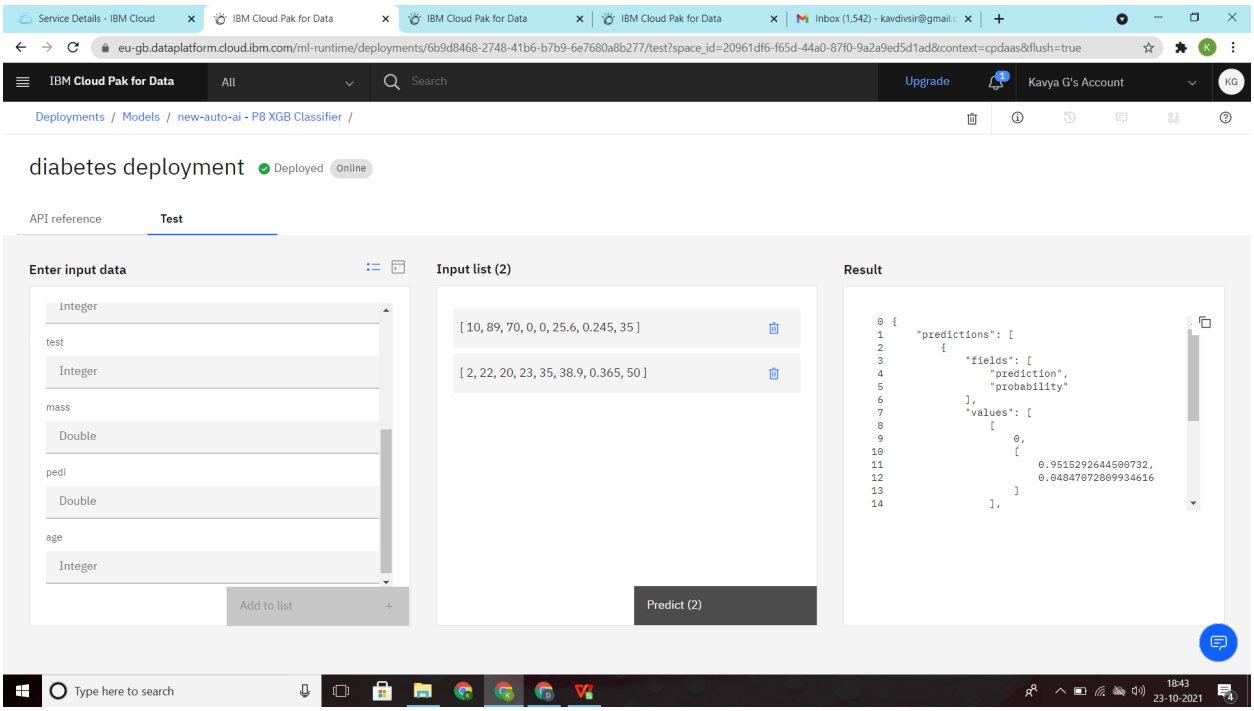

Diabetics Melletus Predictions
Preg:
6
Plas:
148
Pres:
72
Skin:
35
Test:
0
Mass:
33
Pedi:
0.66
Age:
50
Predict


Diabetics Melletus Predictions
You have a Diabetic problem, You must and should consult a doctor. Take care


Diabetics_Flask_App
127.0.0.1:5000/predict
Diabetics Melletus Predictions
Preg:
5
Plas:
116
Pres:
72
Skin:
0
Test:
0
Mass:
25.6
Pedi:
0.201
Age:
30
Predict


Diabetics Melletus
127.0.0.1:5000/data_predict
Diabetics Melletus Predictions
You dont have a Diabetic problem


Service Details - IBM Cloud
IBM Cloud Pak for Data
Inbox (1,542) - kavdivsir@gmail.c
eu-gb.dataplatform.cloud.ibm.com/ml-runtime/deployments/6b9d8468-2748-41b6-b7b9-6e7680a8b277/test?space_id=20961df6-f65d-44a0-87f0-9a2a9ed5d1ad&context=cpdaas&flush=true
IBM Cloud Pak for Data
All
Search
Upgrade
Kavya G's Account
Deployments / Models / new-auto-ai - P8 XGB Classifier /
diabetes deployment
Deployed
Online
API reference
Test
Enter input data
Integer
test
Integer
mass
Double
pedi
Double
age
Integer
Add to list
Input list (1)
[10, 89, 70, 0, 0, 25.6, 0.245, 35]
Predict (1)
Result
{
"predictions": [
{
"fields": [
"prediction",
"probability"
],
"values": [
[
0,
[
0.9515292644500732,
0.04847072809934616
]
]
Type here to search
18:42
23-10-2021


Service Details - IBM Cloud
IBM Cloud Pak for Data
Inbox (1,542) - kavdivsir@gmail.c
eu-gb.dataplatform.cloud.ibm.com/ml-runtime/deployments/6b9d8468-2748-41b6-b7b9-6e7680a8b277/test?space_id=20961df6-f65d-44a0-87f0-9a2a9ed5d1ad&context=cpdaas&flush=true
IBM Cloud Pak for Data
All
Search
Upgrade
Kavya G's Account
Deployments / Models / new-auto-ai - P8 XGB Classifier /
diabetes deployment
Deployed
Online
API reference
Test
Enter input data
Integer
test
Integer
mass
Double
pedi
Double
age
Integer
Add to list
Input list (2)
[10, 89, 70, 0, 0, 25.6, 0.245, 35]
[2, 22, 20, 23, 35, 38.9, 0.365, 50]
Predict (2)
Result
{
"predictions": [
{
"fields": [
"prediction",
"probability"
],
"values": [
[
0,
[
0.9515292644500732,
0.04847072809934616
]
],
Type here to search
18:43
23-10-2021

7 .ADVANTAGES & DISADVANTAGES

Advantages:

1. Easily identifies trends and patterns.
2. No human intervention needed
3. Continuous Improvement.
4. Handling multidimensional and multivariety data
5. Wide applications

Disadvantages:

1. High error susceptibility.
2. Data acquisition.
3. Interpretation of results.
4. Time and resources.

8 .APPLICATIONS

1. Medical Field
2. Research field

9 .CONCLUSION

Machine learning techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure. In this a few existing classification methods for medical diagnosis of diabetes patients have been discussed on the basis of accuracy. An classification problem has been detected in the expressions of accuracy. Three machine learning techniques were applied on the Pima Indians diabetes dataset, as well as trained and validated against a test dataset. The results of our model implementations have shown that Random Forest Classifier outperforms the other models. Using association rule mining, the results have shown that there is a strong association of BMI and glucose with diabetes.

10 .FUTURE SCOPE

The limitation of this study is that a structured dataset has been selected but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease. Other attributes including physical inactivity, family history of diabetes, and smoking habit, are also planned to be considered in the future for the diagnosis of diabetes.

11 .BIBILOGRAPHY

1. https://www.kaggle.com/uciml/pima-indians-diabetes-database

2. https://www.w3schools.com/html/