Project Title : Diabetes Mellitus Prediction Using IBM Auto AI Service

Project Description:

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes.

In this project, we build a machine learning model that can efficiently discover the rules to predict diabetes mellitus of patients based on the given parameter about their health. The model is deployed in the IBM cloud to get scoring endpoint which can be used as API in web app building. Finally, User Interface is created for the prediction model.

IBM Services Used:

1. IBM Watson Studio
2. IBM Watson Machine Learning
3. Node-RED
4. IBM Cloud Object Storage

Project Architecture:

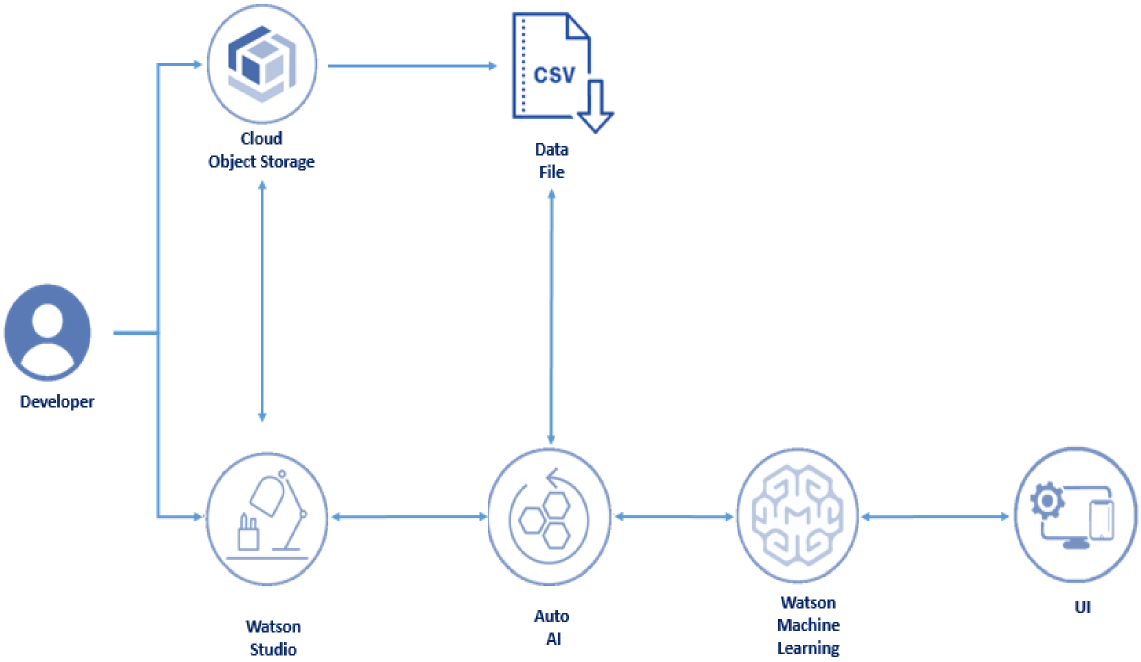


Figure 1 Project Architecture

Objectives of Project:

1. Design of Diabetes Mellitus Prediction model using IBM Auto AI Service and Pima Indian Diabetic Dataset
2. Comparative analysis of various maching learning algorithms
3. Design of User Interface

Dataset Description:

The machine learning model designed to predict the diabetes mellitus is designed using Pima Indian Diabetic patient dataset. This dataset contains 768 patients' information where 500 instances are negative and remaining 268 samples are positive. There are total 9 attributes where last attributes is class information i.e. 0- non diabetic patient and 1- diabetic patient. Table 1 shows description of various dataset attributes.

Table 1 Pima Indian Diabetic Patient Dataset

Attribute	Numeric values (Min, Max)	STD	Description
preg	0,17	3.36	No of Pregnancies
plas	0,199	31.97	Plasma glucose measured using oral glucose tolerance test
pres	0,122	19.35	Blood pressure (mm Hg)
skin	0,99	015.95	Triceps skin fold thickness (mm)
test	0,846	115.2	Two hours serum insulin in
mass	0.0,67.1	7.88	Body mass index
Pedi	0.078, 2.42	0.33	Probability of diabetes on the basis of family history
Age	21,81	11.76	Age of a person in years
Class	0,1	-	0 – Non diabetic person 1 – Diabetic person

Design of Machine Learning Model

Figure 1 describes the process used to design a machine learning model that predicts whether a patient is diabetic or not based on her health information.

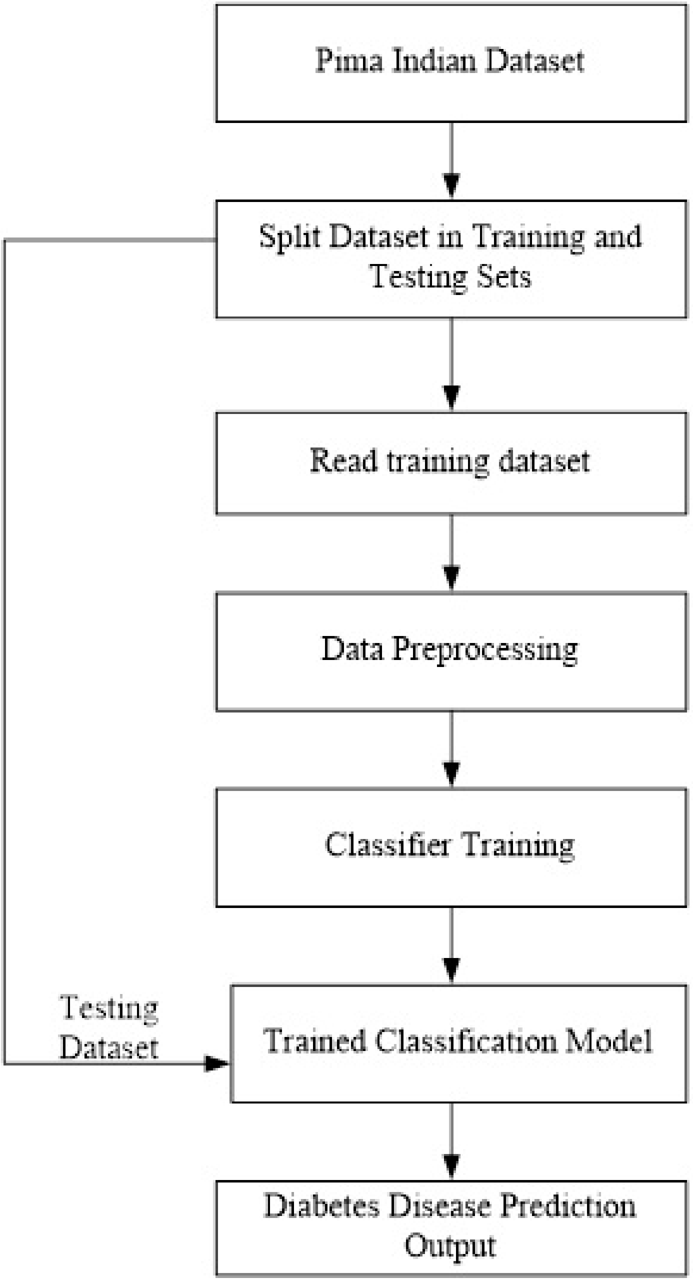

Pima Indian Dataset
Split Dataset in Training and Testing Sets
Read training dataset
Data Preprocessing
Classifier Training
Testing Dataset
Trained Classification Model
Diabetes Disease Prediction Output

Figure 1 Machine Learning Model for Diabetes Meliitus Prediction

As shown in Figure 1, the machine learning model is trained using labeled instances from dataset. The model is trained using randomly selected 80% instances and remaining 20% instances are used to evaluate the trained model. In order to find the best suitable model the quantitative comparison is done between various machine learning algorithms as mentioned below.

1. Decision Tree
2. Gradient Boosting Classifier
3. Logistic Regression
4. Random Forest
5. Linear SVM
6. KNN
7. CHAID

Comparative Analysis:

In order to analyse the performance of various machine learning models various quantitative measures are used. These measures demonstrates the effectiveness of machine learning model for the prediction of diabetes mellitus based on patient's health parameters. The quality measuers used for the comparative analysis are as follows:

1. Accuracy : Total number of correct classifications out of total test instances.

2. Precision: Correctness over the positive detections.

3. Recall: Total positive detections out of total positive instances.

4. ROC Curve: The plot between True Positive Rate (TPR) and False Positive Rate (FPR) while changing the threshold for positive classification.

5. Area Under Curve (AUC) : It is obtained from the graph ploatted between Precision and Recall.

6. F_measure: It represents the weighted harmonic mean of the Precision and Recall.

7. Confusion Matrix: Table 2 shows the confusion matrix.
   TP: True Positive i.e. Positive sample are classified as Positive
   TN: True Negative i.e. Negative samples are classified as Negative

FP: False Positive i.e. Negative samples are classified as Positive
FN: False Negative i.e. Positive samples are classified as Negative

Table 2 Confusion Matrix

Confusion Matrix Structure			
Total Instances		Predicted Class	
		False	True
Actual Class	False	True Negative (TN)	False Positive (FP)
	True	False Negative (FN)	True Positive (TP)

Table 3 shows the comparative results obtained for various machine learning models.

Performance parameter	Decision Tree	Gradient Boosting Classifier	Logistic Regression	Random Forest	Linear SVM	KNN	CHAID
Accuracy	0.753	0.792	**0.831**	0.792	0.782	0.75	0.754
Precision	0.618	0.789	0.719	0.636	**0.804**	0.777	0.749
Recall	0.778	0.556	**0.852**	0.726	0.782	0.75	0.754
F_Measure	0.689	0.652	0.78	0.678	**0.789**	0.759	0.732
Area Under Curve (AUC)	0.824	0.88	**0.892**	0.881	0.761	0.721	0.67

The best machine learning classification model is selected based on Accuracy, Precision, Recall, F_Measure and AUC.As shown in Table 3, Logistic Regression outperforms other machine learning algorithms in terms of Accuray, Recall and AUC. Linear SVM is also providing good precision and F_Measure. However, considering Accuracy as the optimization parameter the Logistic Regression is selected as the final model for the development of user interface application for the diabetes mellitus prediction. Figure 2 shows the ROC for the Logistic Regression based classification model.

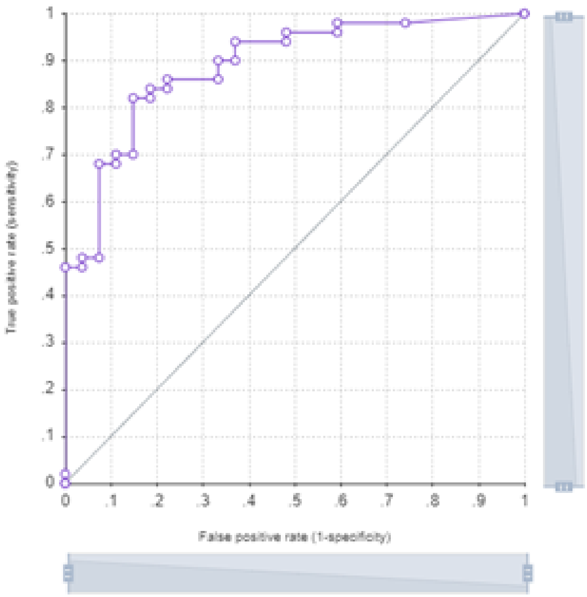


Figure 2 ROC obtained for Logistic Regression based classification model

Design of User Interface

Once the prediction model is designed using IBM Auto AI service, it is deployed using IBM Cloud storage and API is obtained. In order to make the UI another IBM service named nodeRED App is used. nodeRED contains various types of nodes which helps to design the flow and user interface. The flow of proposed model designed in nodeRED is shown in Figure 3.

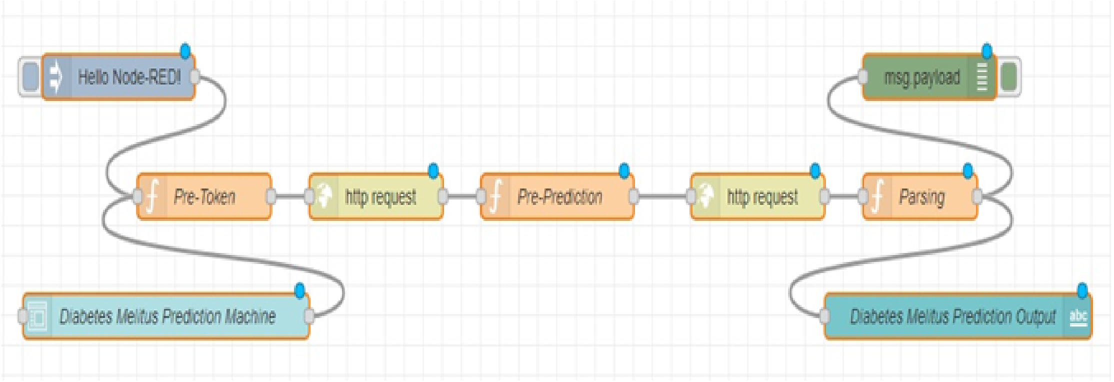
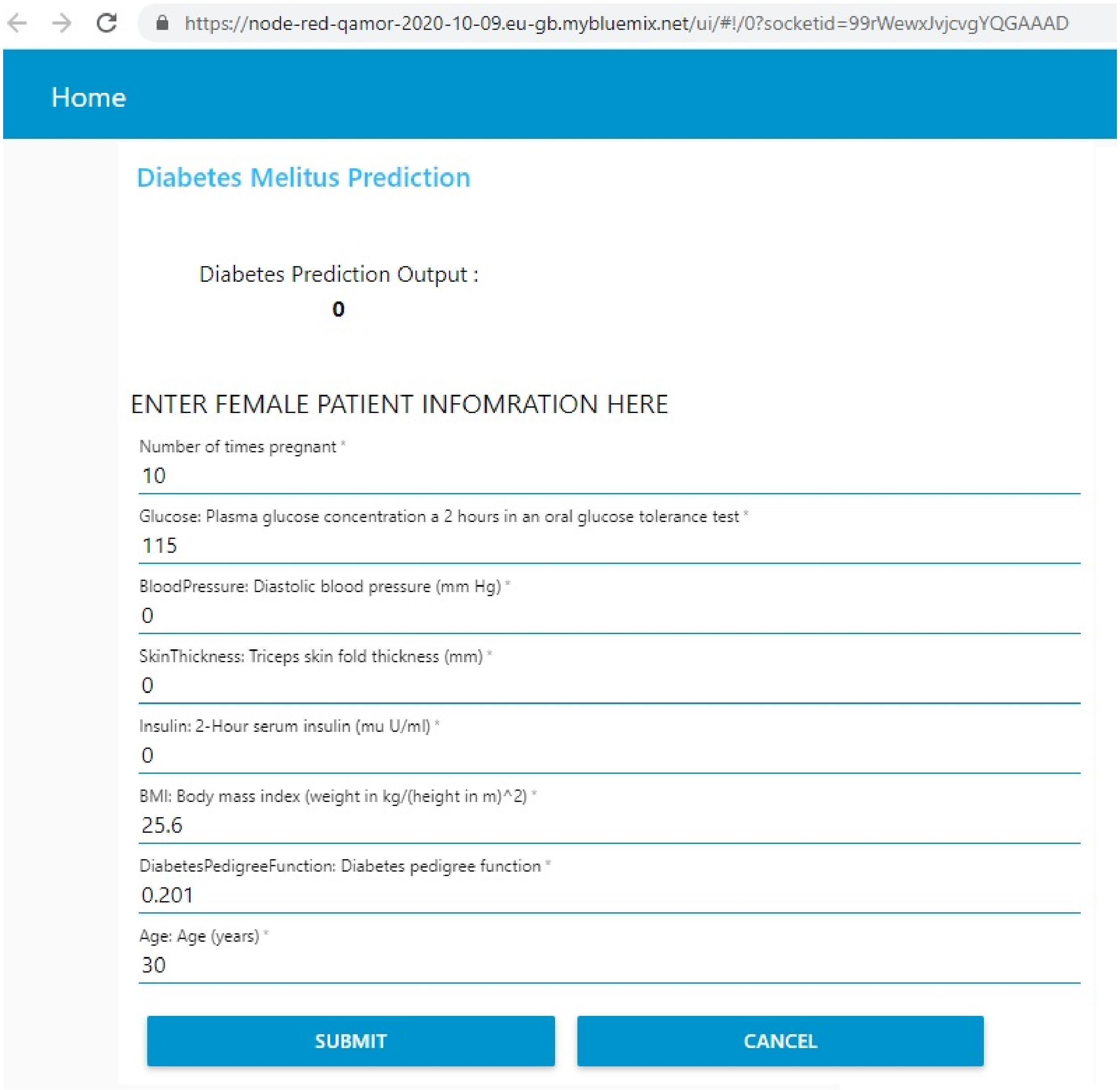


Figure 3 nodeRED flow for diabetes mellitus prediction User Interface

User Interface Output:

Figure 4 shows the user interface designed for diabetes mellitus prediction.



https://node-red-qamor-2020-10-09.eu-gb.mybluemix.net/ui/#!/0?socketid=99rWewxJvjcvgYQGAAAD

Diabetes Melitus Prediction

Diabetes Prediction Output :
0

ENTER FEMALE PATIENT INFOMRATION HERE

Number of times pregnant *
10

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test *
115

BloodPressure: Diastolic blood pressure (mm Hg) *
0

SkinThickness: Triceps skin fold thickness (mm) *
0

Insulin: 2-Hour serum insulin (mu U/ml) *
0

BMI: Body mass index (weight in kg/(height in m)^2) *
25.6

DiabetesPedigreeFunction: Diabetes pedigree function *
0.201

Age: Age (years) *
30

Figure 4 User Interface for Diabetes Mellitus Prediction

Conclusion:

Considering the severe health complications in diabetic persons it is very much significant to have high Recall rate along with good Accuracy. Logistic Regression model provides 83.1% Accuracy and 85.2% Recall when evaluated over 768 instances of Pima Indian Diabetes Database. The model is capable of an early stage prediction of diabetes mellitus which may help the patient to take precautionary steps and avoid further health complications.

Acknowledgement:

I express my since thanks to smartinternz for providing insights of using IBM cloud services during the FDP. The experts have supported immensely while designig of project on IBM Cloud platform.

Author Details:

Name: Dr. Vipul H. Mistry

Designation: Assistant Professor and Head of Electronics and Communication Engineering Department

Institute Name: S. N. Patel Institute of Technology and Research Centre, Umrakh affiliated with Gujarat Technological University, Chandkheda, Gujarat

Userid: vipul.h.mistry@snpitrc.ac.in