A PROJECT REPORT ON

Diabetic Mellitus Prediction using IBM Auto AI

Submitted By:

Prashanth Nayak

Index

1. INTRODUCTION

1.1 OVERVIEW

This is an internship offered by smartbridge on MACHINE LEARNING USING PYTHON.
We are well trained with python, data pre-processing, supervised Algorithms,unsupervised Algorithms, flask integration, accessing the IBM Cloud .The project assigned to me is an individual project of topic 'Diabetic Mellitus Prediction using IBM AutoAI'. We used the database given to us and the relevant algorithm to obtain the model using AutoAI.
In this project there are 4 tasks:

1. Data collection
    - Downloading the dataset/creating the dataset
2. IBM Cloud account
    - IBM cloud Registration
    - Login to IBM cloud
    - Create Cloud Object Storage Service
    - Create Watson Studio Platform
    - Create Machine Learning Service
3. Model building
    - Create Project In Watson Studio
    - Auto AL Experiment In Add Projects
    - Setup Your Auto Al Environment
    - Import Dataset
    - Run the Model
    - Selection of Auto Al Pipeline
    - Deploy and test the model
4. Application building
    - Create Node Red Service
    - Build UI with Nodered

The skills gained by doing this project are:
Python, Python for Data Analysis, Python for Data Visualization, Exploratory Data Analysis, IBM Cloud, IBM WATSON.

1.2 PURPOSE

The purpose of this project is to predict the Diabetic Mellitus using IBM Watson Studio. The purpose of this project is to learn about Machine Learning and the facilities available in IBM cloud and explore its wide range of services. The IBM Cloud includes Infrastructure as a service, Software as a service and Platform as a service. IBM offers tools for cloud-based collaboration, development and test, application development, analytics, business-to-business integration, and security.

The purpose of this project is to understand the insights of machine learning and Building the model for prediction of Diabetic Mellitus is useful in Medical Industry.Model identifies trends and patterns.Using the model no human Intervention is needed.Machine learning makes it easy to handle multi-dimensional and multi-variety data and the web application built can be used by everyone.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

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. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health for accurate predictions.

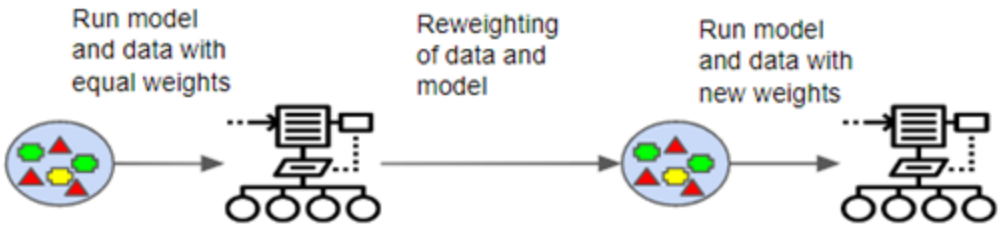
2.2 PROPOSED SOLUTION

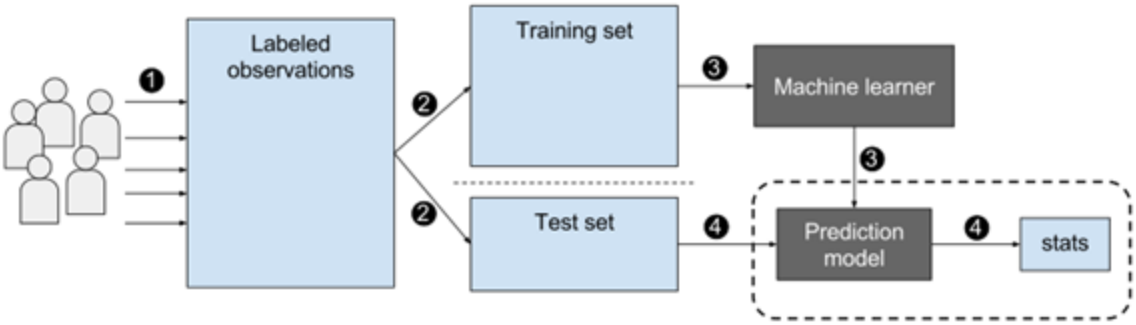
This project prevents the people from the avalanche by priory informing them there is a chance to the occurrence of avalanche or not. The model gets the data from the IOT based sensors. After that we want to process those data using a suitable algorithm, then our model display whether the avalanche occur or not and how strength it was. To analyse the data coming from different sensors we are applying various machine learning algorithms. If there is a chance of avalanche then the notification will be sent to people so that they can take decisions accordingly and the model is been built in Auto AI.

3. THEORETICAL ANALYSIS

The Algorithm Used for this Dataset is **Gradient Boosting Classifier**.

3.1 BLOCK DIAGRAM





3.2 HARDWARE/SOFTWARE DESIGNING

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
The idea of gradient boosting originated in the observation by Leo Breiman that boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed by Jerome H. Friedman, simultaneously with the more general functional gradient boosting perspective of Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean. The latter two papers introduced the view of boosting algorithms as iterative functional gradient descent algorithms. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.
GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n_classes regression trees are fit on the

negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.

GradientBoostingClassifier(*,loss='deviance',learning_rate=0.1,n_estimators=100,subsample=1.0,criterion='friedman_mse',min_samples_split=2,min_samples_leaf=1,min_weight_fraction_leaf=0.0,max_depth=3,min_impurity_decrease=0.0,min_impurity_split=None,init=None,random_state=None,max_features=None,verbose=0,max_leaf_nodes=None,warm_start=False,presort='deprecated',validation_fraction=0.1,n_iter_no_change=None,tol=0.0001, ccp_alpha=0.0)

The parameters are:
loss{'deviance', 'exponential'}, default='deviance' loss function to be optimized. 'deviance' refers to deviance (= logistic regression) for classification with probabilistic outputs. For loss 'exponential' gradient boosting recovers the AdaBoost algorithm.

learning_ratefloat, default=0.1 learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimators.

n_estimatorsint, default=100

The number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance.

subsamplefloat, default=1.0

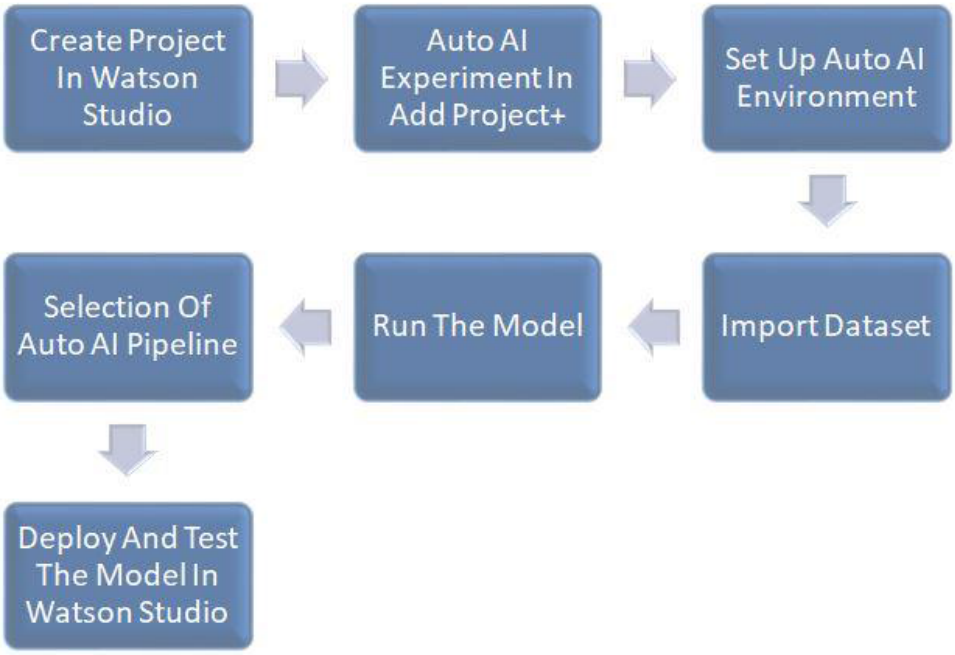
The fraction of samples to be used for fitting the individual base learners. If smaller than 1.0 this results in Stochastic Gradient Boosting. subsample interacts with the parameter n_estimators. Choosing subsample<1.0 leads to a reduction of variance and an increase in bias.

criterion{'friedman_mse', 'mse', 'mae'}, default='friedman_mse'

The function to measure the quality of a split. Supported criteria are 'friedman_mse' for the mean squared error with improvement score by Friedman, 'mse' for mean squared error, and 'mae' for the mean absolute error. The default value of 'friedman_mse' is generally the best as it can provide a better approximation in some cases.

4. EXPERIMENTAL INVESTIGATIONS

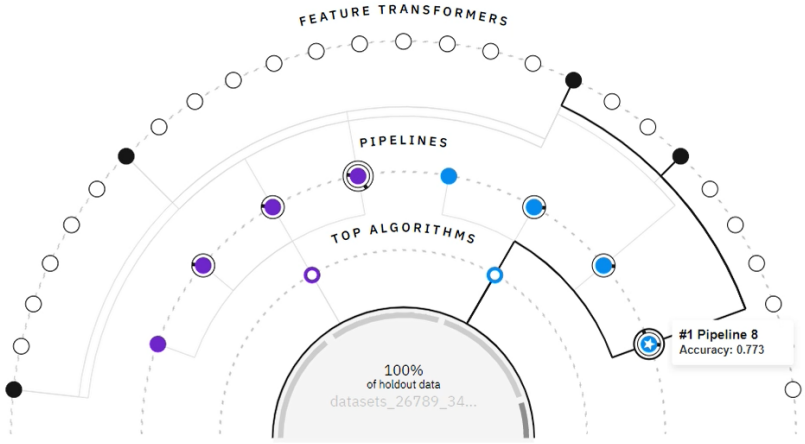
AUTO AI EXPERIMENT



- In Machine Learning, we create a project in IBM Watson Studio and we add an Auto AI Experiment to the project.
- We need to import our dataset to calculate the required output. Hence, we import/upload the 'datasets_26789_34175_pima-indians-diabetes.data' csv file and give 'Diabetic Mellitus Prediction' column of the dataset as the prediction column.
- Run the model and we get our required Auto AI Pipeline. We get the Relationship Map and Pipeline Comparison. We have to save the Auto AI pipeline as a model.
    - We should deploy the saved model.
    - Then we get to test the deployed model.

Relationship map

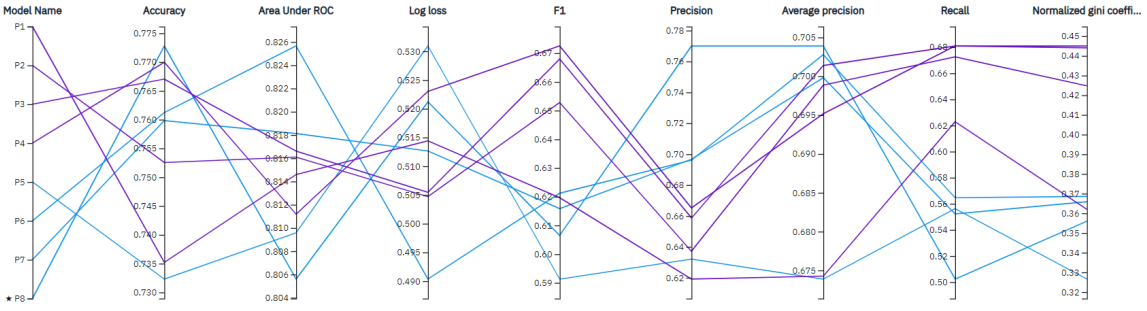
Prediction column: **class**



Rank ↑	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
★ 1	Pipeline 8	Gradient Boosting Classifier	**0.773**	HPO-1 FE HPO-2	00:00:09
2	Pipeline 4	XGB Classifier	**0.770**	HPO-1 FE HPO-2	00:01:10
3	Pipeline 3	XGB Classifier	**0.767**	HPO-1 FE	00:02:00
4	Pipeline 6	Gradient Boosting Classifier	**0.761**	HPO-1	00:00:03
5	Pipeline 7	Gradient Boosting Classifier	**0.760**	HPO-1 FE	00:00:32
6	Pipeline 2	XGB Classifier	**0.753**	HPO-1	00:00:32
7	Pipeline 1	XGB Classifier	**0.735**	*None*	00:00:01
8	Pipeline 5	Gradient Boosting Classifier	**0.732**	*None*	00:00:01

Metric chart

Prediction column: **class**



APPLICATION BUILDING(NODE-RED)

- From the catalog of IBM Cloud we should install Node-Red Application.
- In this, we have 'Nodes' to the left side of the page and we use these nodes to create a 'flow'.
- We need to install 'Dashboard Nodes' by going to the 'Manage Palette' section.

The nodes which we use in this project are:

- Form node

Adds a form to user interfaceHelps to collect multiple value from the user on submit button click as an object in msg.payload.Multiple input elements can be added using add elements button

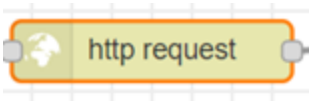
- Timestamp node

Injects a message into a flow either manually or at regular intervals. The message payload can be a variety of types, including strings, JavaScript objects or the current time.

- function node

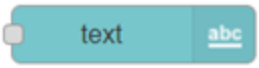
A JavaScript function block to run against the messages being received by the node. The messages are passed in as a JavaScript object called msg.By convention it will have a msg.payload property containing the body of the message. The function is expected to return a message object (or multiple message objects), but can choose to return nothing in order to halt a flow

● http request node



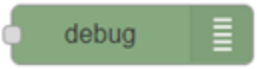
Sends HTTP requests and returns the response.

● Text node



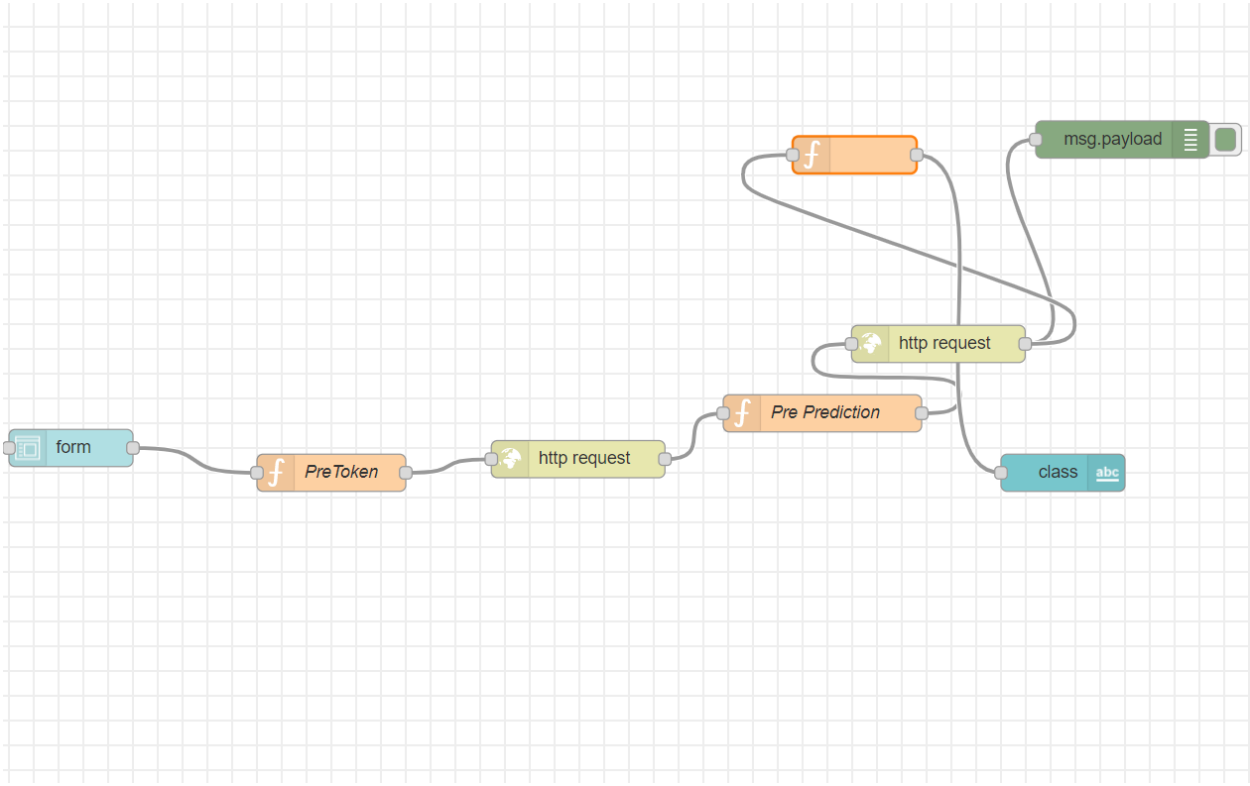
Will display a non-editable text field on the user interface.Each received msg.payload will update the text based on the provided Value Format.The Value Format field can be used to change the displayed format and can contain valid HTML and Angular filters

● Debug node



Displays selected message properties in the debug sidebar tab and optionally the runtime log. By default it displays msg.payload, but can be configured to display any property, the full message or the result of a JSONata expression.

5. FLOWCHART



- The input in the form node is passed to pre-token node.
- The pre-token node is connected to http-request node passing its input.
- The http-request node is connected to pre-prediction node.
- The pre-prediction node is connected to http-request node containing scoring-end-point url.
- The http-request node is connected to function node as well as debug node
- The function node is connected to the text node.
- FORM NODE is used to give the input columns
- TIMESTAMP NODE is not that required
- PRE-TOKEN FUNCTION NODE is used to enter the ‘API-KEY’ and set the input columns to global.
- HTTP REQUEST NODE is used to give the ‘scoring-end-Point url’.
- PRE-PREDICTION NODE is used to enter the ‘INSTANCE-ID’ and get the input columns which were set to global.
- In FUNCTION NODE we give the path of the output obtained.
- TEXT NODE is used to give a title to the output.
- DEBUG NODE is used to get the output.
- After giving all the inputs to the respective nodes, we should deploy the flow.
- Then we have to launch the dashboard where we enter random values and get the predicted output for the entered values.

6. RESULT

By predicting the flow,we obtained the required output.The predicted value obtained in IBM Watson Studio and the predicted value obtained in Node-red matches when we give the same input values in both the applications.Hence,the flow is successfully built without any errors

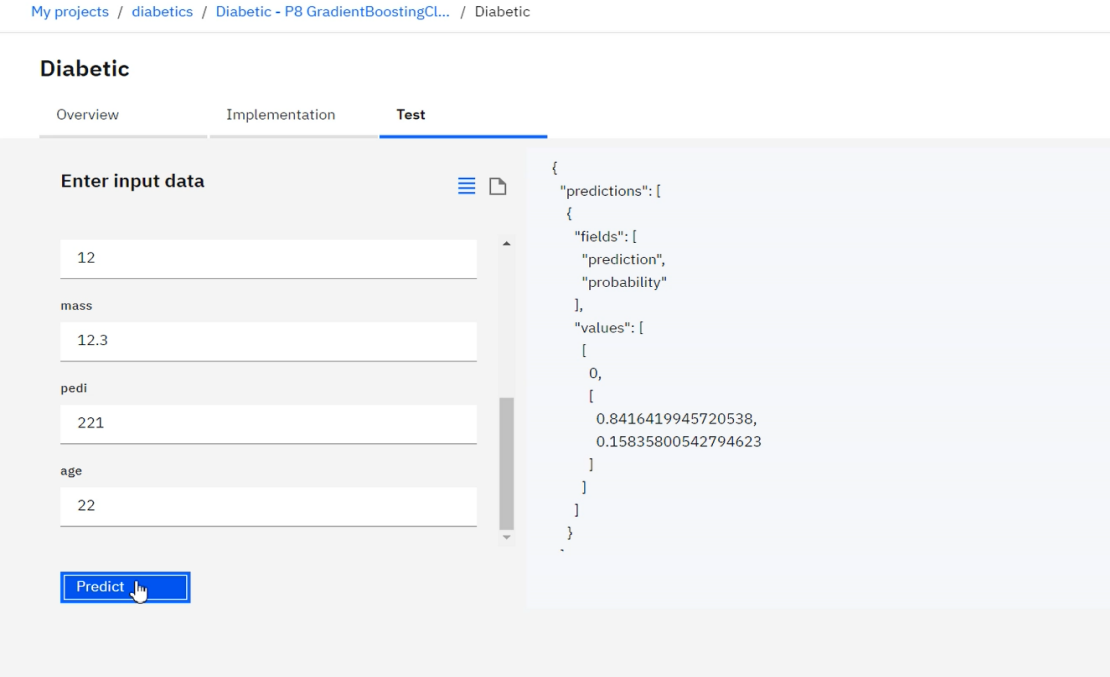


Figure: IBM WATSON Auto AI experiment

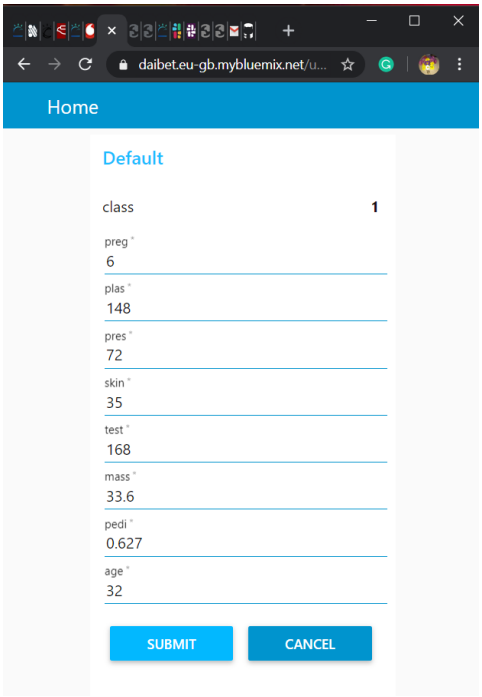


Figure: Node red Application

7. APPLICATIONS

The applications of Machine Learning are:

Virtual Personal Assistants

- Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice.

Predictions While Commuting

- We all have been using GPS navigation services. While we do that, our current locations and velocities are being saved at a central server for managing traffic. This data is then used to build a map of current traffic. While this helps in preventing the traffic and does congestion analysis, the underlying problem is that there are less number of cars that are equipped with GPS.
- When booking a cab, the app estimates the price of the ride. When sharing these services, how do they minimize the detours? The answer is machine learning.

Videos Surveillance

- The video surveillance system nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to human attendants, which can ultimately help to avoid mishaps.

Social Media Services

- From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits.

Email Spam And Malware Filtering

- There are a number of spam filtering approaches that email clients use. To ascertain that these spam filters are continuously updated, they are powered by machine learning

9. CONCLUSION

8.CONCLUSION

Using the IBM Watson Auto AI Experiment we tested the model and using node red Application in IBM cloud we created a UI where we deployed our model.The predicted value obtained in IBM Watson Studio Auto AI and the predicted value obtained in Node-red application matches each other when we give the same input values are given .Therefore model is successfully deployed can predict the Diabetic Mellitus.

9. FUTURE SCOPE

The scope of Machine Learning in India, as well as in other parts of the world, is high in comparison to other career fields when it comes to job opportunities. According to Gartner, there will be 2.3 million jobs in the field of Artificial Intelligence and Machine Learning by 2022. Also, the salary of a Machine Learning Engineer is much higher than the salaries offered to other job profiles.

10. BIBILIOGRAPHY

- https://node-red-cyder.eu-gb.mybluemix.net/ui/#!/0?socketid=wKFaCFw_-ylhSeOMAAAG
- https://www.ibm.com/in-en/cloud
- https://en.wikipedia.org/wiki/Machine_learning