Diabetes Mellitus Prediction Using IBM AutoAI Service

Karthik B[1] , Bharathwaj Murali[2] , Kavin A K[3]
[1]karthik.b2019@vitstudent.ac.in, [2]bharathwaj.murali2019@vitstudent.ac.in, [3]kavin.ak2019@vitstudent.ac.in

Abstract--- 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.

Keywords--- Diabetes Mellitus, AutoAI, IBM Watson Studio, Node-RED, LGBM Classifier, Pipeline, HyperParameter fine-tuning, Cross validation score

1 Introduction

1.1 Overview

In this project, we will be building 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 needs to be deployed in the IBM cloud to get a scoring endpoint which can be used as API in web app building. The model prediction needs to be showcased on the User Interface.

1.2 Purpose

Diagnosis of diabetes is considered a challenging problem for quantitative research. Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension. A family history of diabetes has not been associated with BMI and insulin. However, an increased BMI is not always associated with abdominal obesity. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. Several existing techniques have not provided effective results when different parameters were used for prediction of diabetes

2 Literature Survey

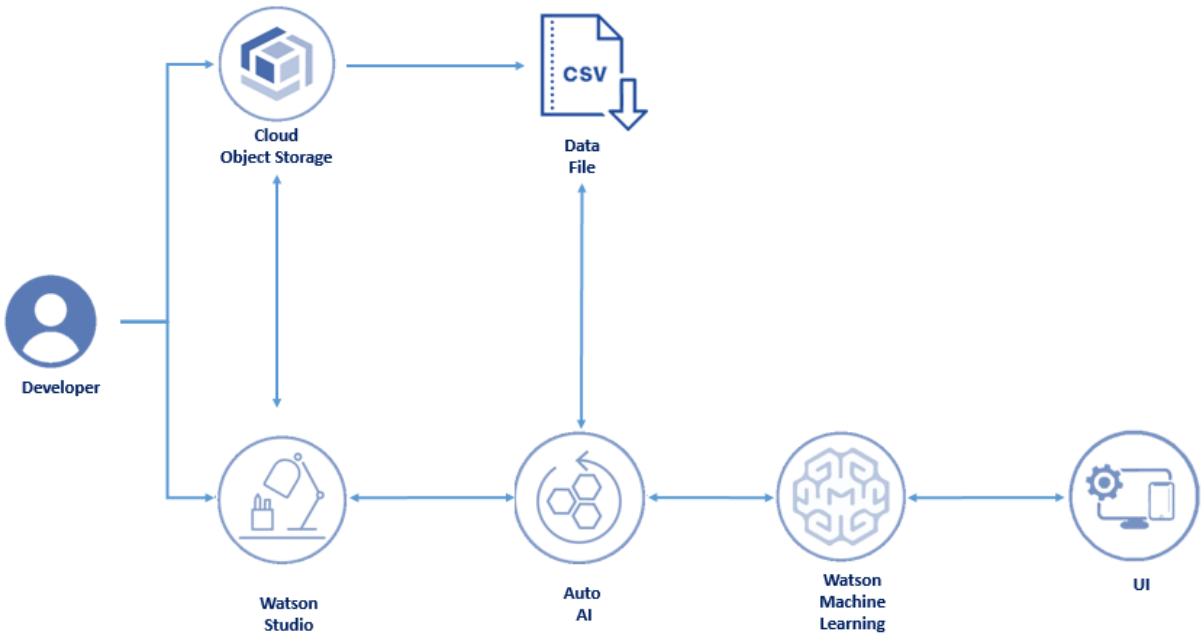
2.1 Existing problem

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.
K.VijiyaKumar et al. [1] proposed random Forest algo- rithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in ma- chine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the dia- betes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [5] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The

results are presented and compared with simi- lar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. [2] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods in- cluding: SVM, Logistic regression, ANN. This project pro- poses an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. [6] proposed diabe- tes disease prediction using data mining assemble Intelli- gent Diabetes Disease Prediction System that gives analy- sis of diabetes malady utilizing diabetes patients database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on dia- betes patients database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [7] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algo- rithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identi- fied based on previous research.

3 Theoritical Analysis

3.1 Technical Architecture



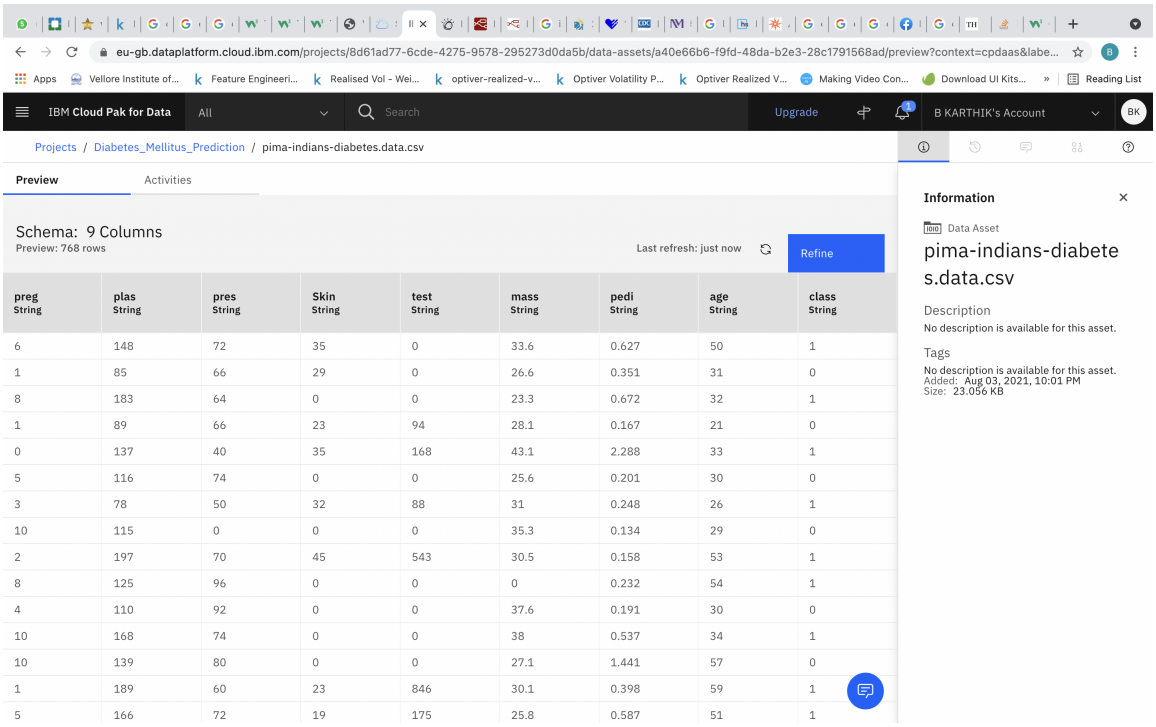
3.2 Harware/Software designing

Services Used

- IBM Watson Studio
- IBM Watson Machine Learning
- Node-RED
- IBM Cloud Object Storage
- IBM Watson Assistant

4 Proposed Method

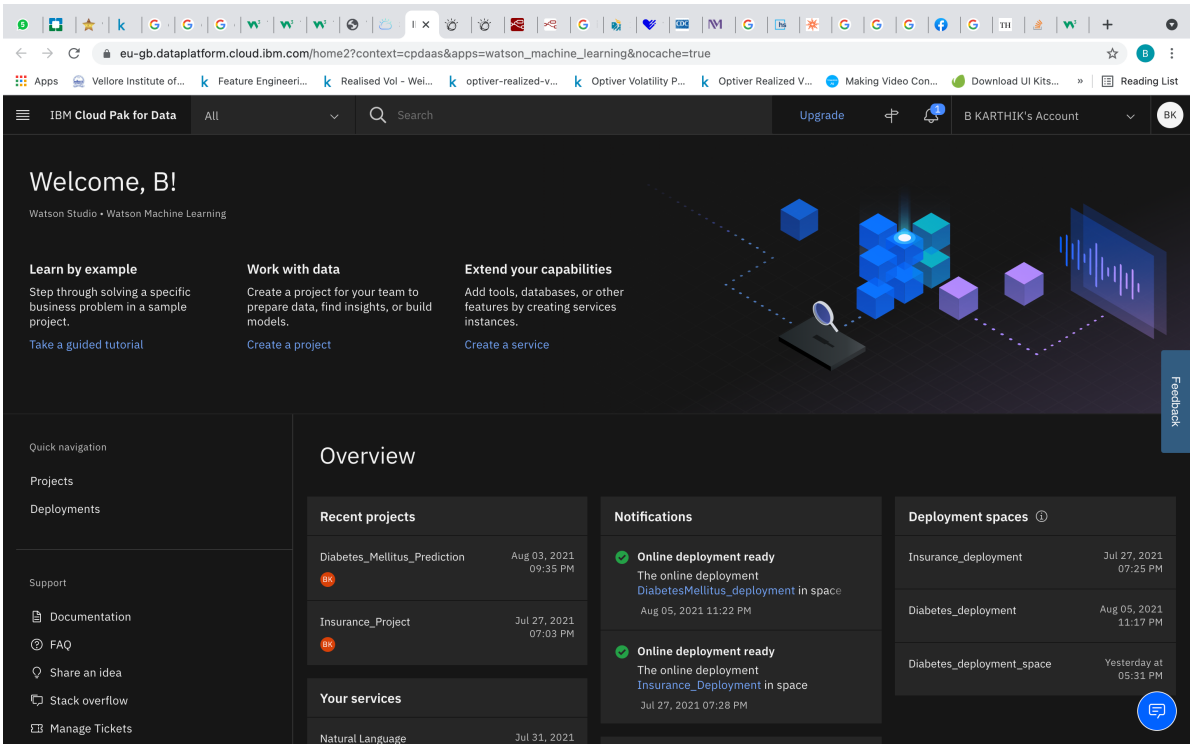
4.1 Dataset



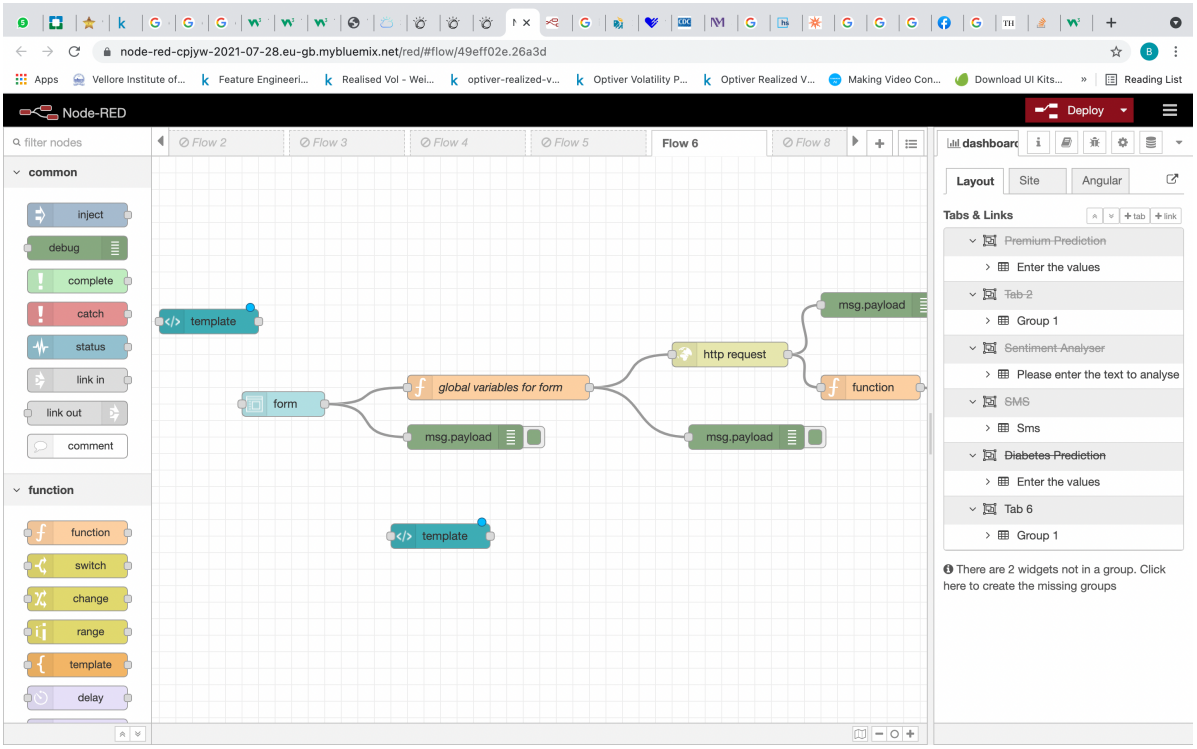
preg String	plas String	pres String	Skin String	test String	mass String	pedi String	age String	class String
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), Class variable (0 or 1)

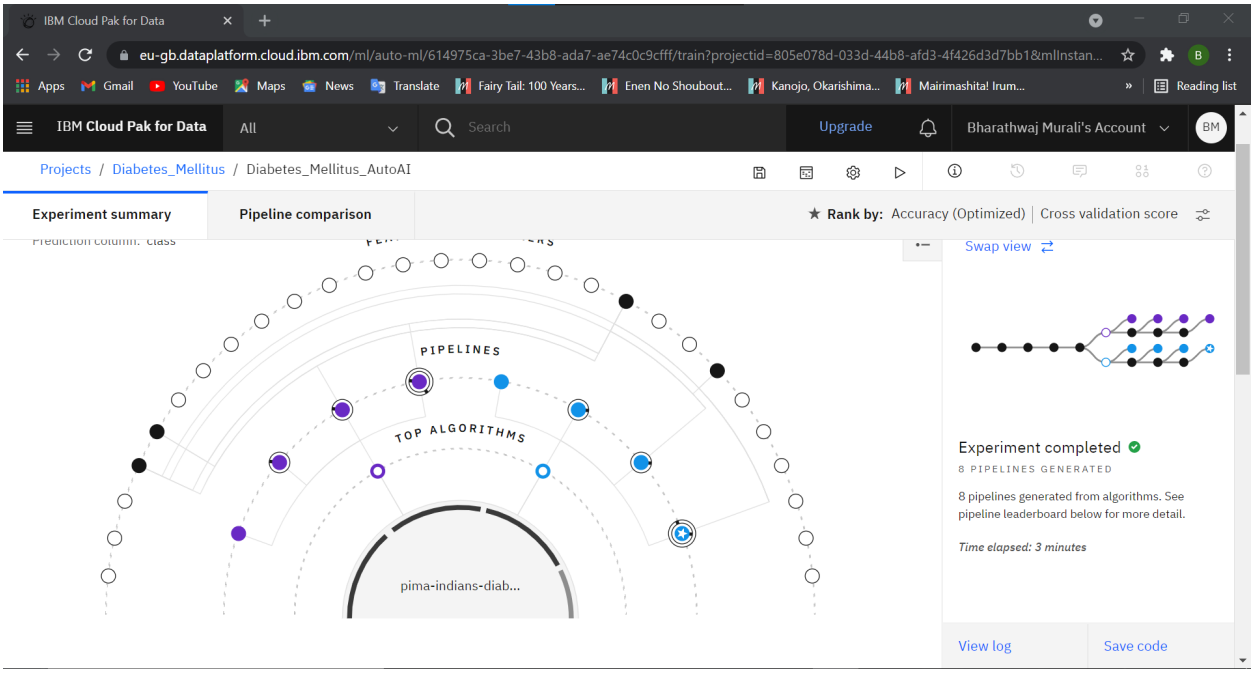
4.1 IBM Watson Studio

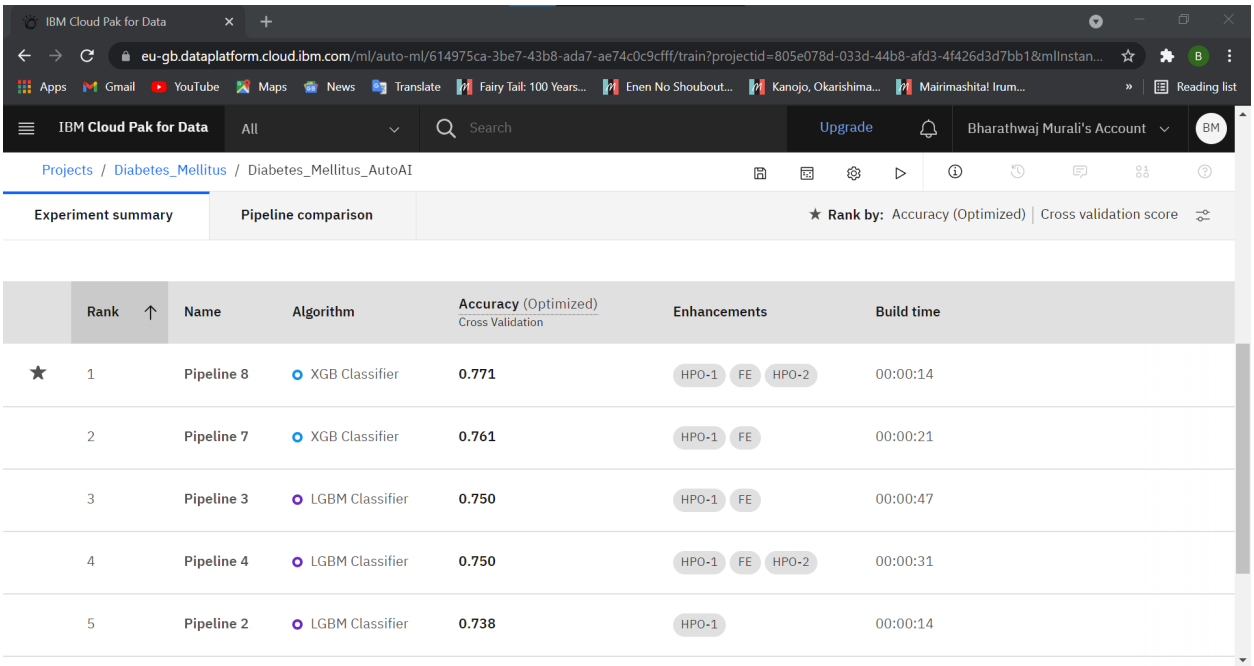


4.2 Node-RED



5 Experimental Investigations



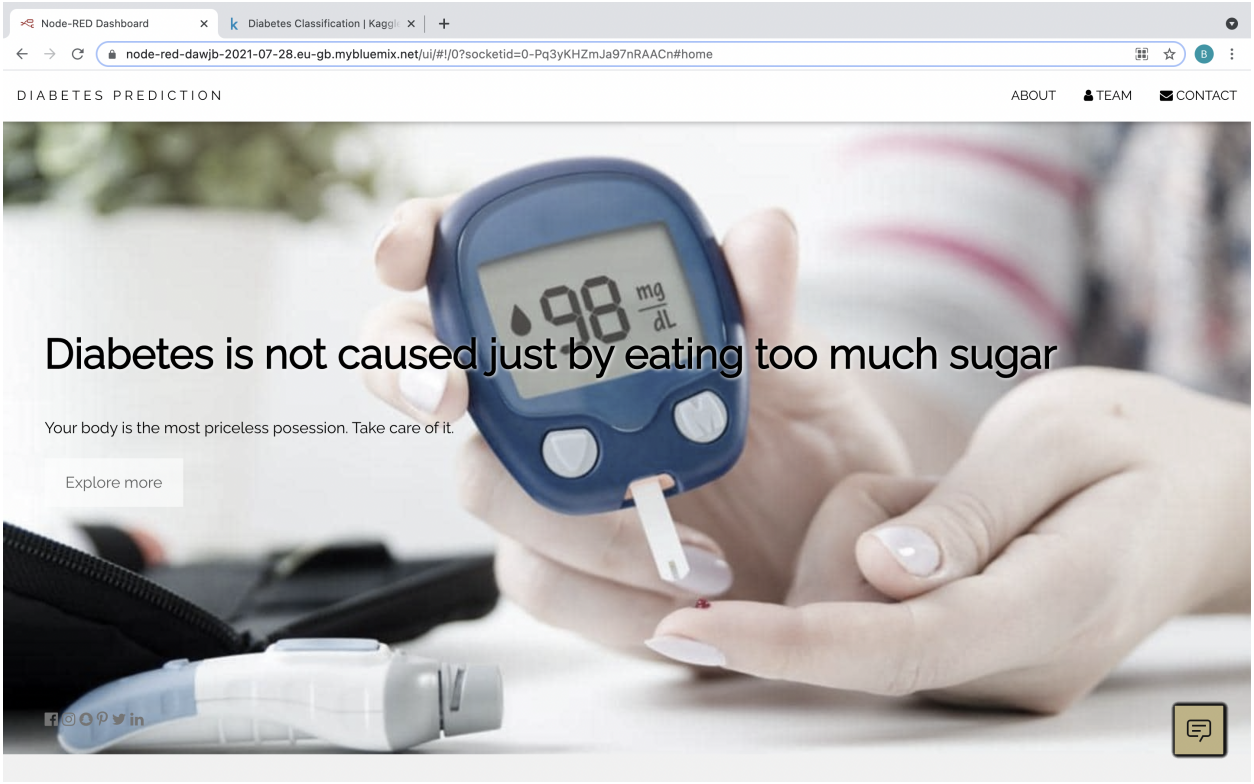


We tried different algorithms like XGBoost, LightGBM, Decisition Tree, Logistic Regression. We used 8 pipelines to compare the performance of each model. Here 85% of the data was used for training and the rest 15% for testing which yielded the best performance. Among all the models XGBoost with hyper-parameter tuning and feature engineering gave the best accuracy score of 77.1%

6 Project Flow

1. Log in to IBM account
2. Create IBM Watson Studio and Node-RED Service
3. Create a Watson studio project
4. ADD Auto AI Experiment
5. Run the Auto AI Experiment to build a Machine learning model on the desired dataset
6. Save the model
7. Deploy the model as a web server and generate scoring End Point
8. Create a WEB application Using Node-RED to take user input and showcase Prediction on U

7 Result



Node-RED Dashboard | Diabetes Classification | Kaggle

node-red-dawjb-2021-07-28.eu-gb.mybluemix.net/ui/#!/0?socketid=0-Pq3yKHZmJa97nRAACn#home

DIABETES PREDICTION

ABOUT TEAM CONTACT

One in six people with diabetes in the world is from India.

India is home to 77 million diabetics, second highest in the world

Node-RED Dashboard | Diabetes Classification | Kaggle

node-red-dawjb-2021-07-28.eu-gb.mybluemix.net/ui/#!/0?socketid=0-Pq3yKHZmJa97nRAACn#home

DIABETES PREDICTION

ABOUT TEAM CONTACT

BEST PRACTICES

How can diabetes be managed

DIET

Follow a Mediterranean diet or Dash diet. These diets are high in nutrition and fiber and low in fats and calories.

EXERCISE

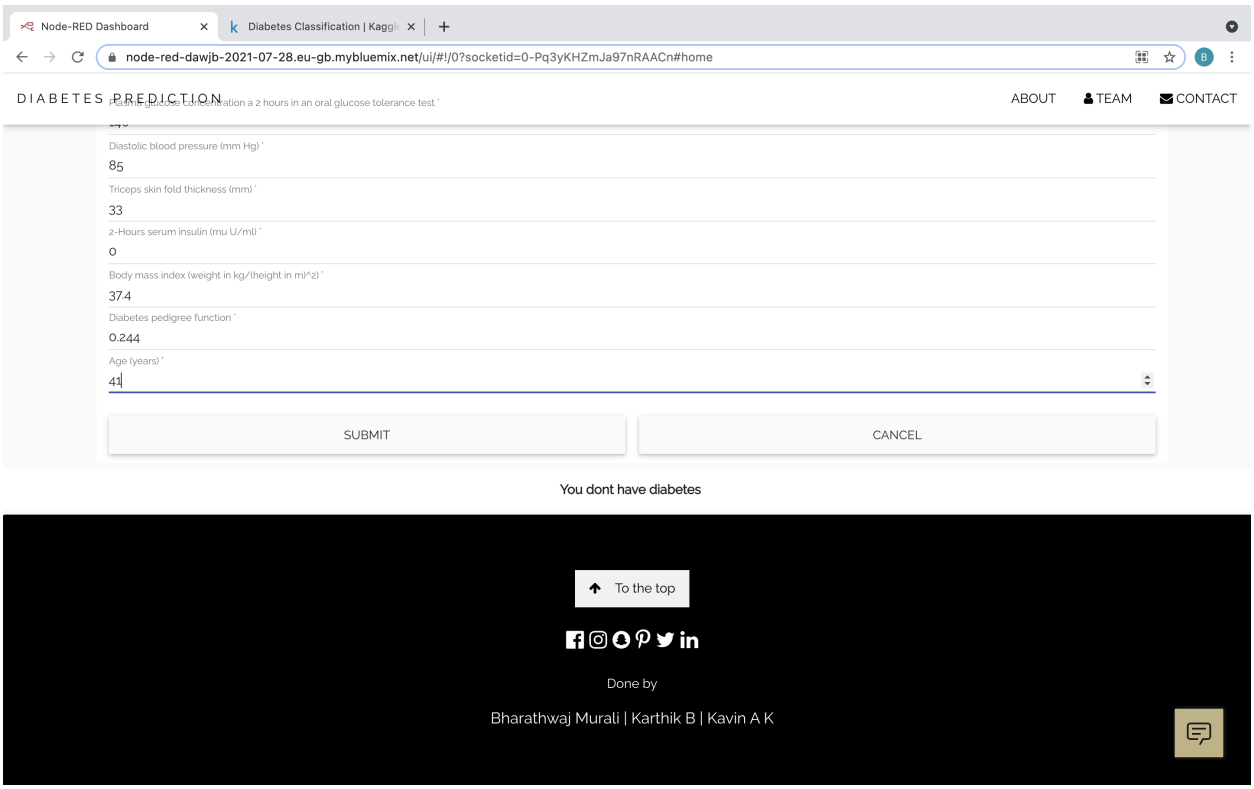
Exercising regularly. Try to exercise at least 30 minutes most days of the week. Walk, swim or find some activity you enjoy.

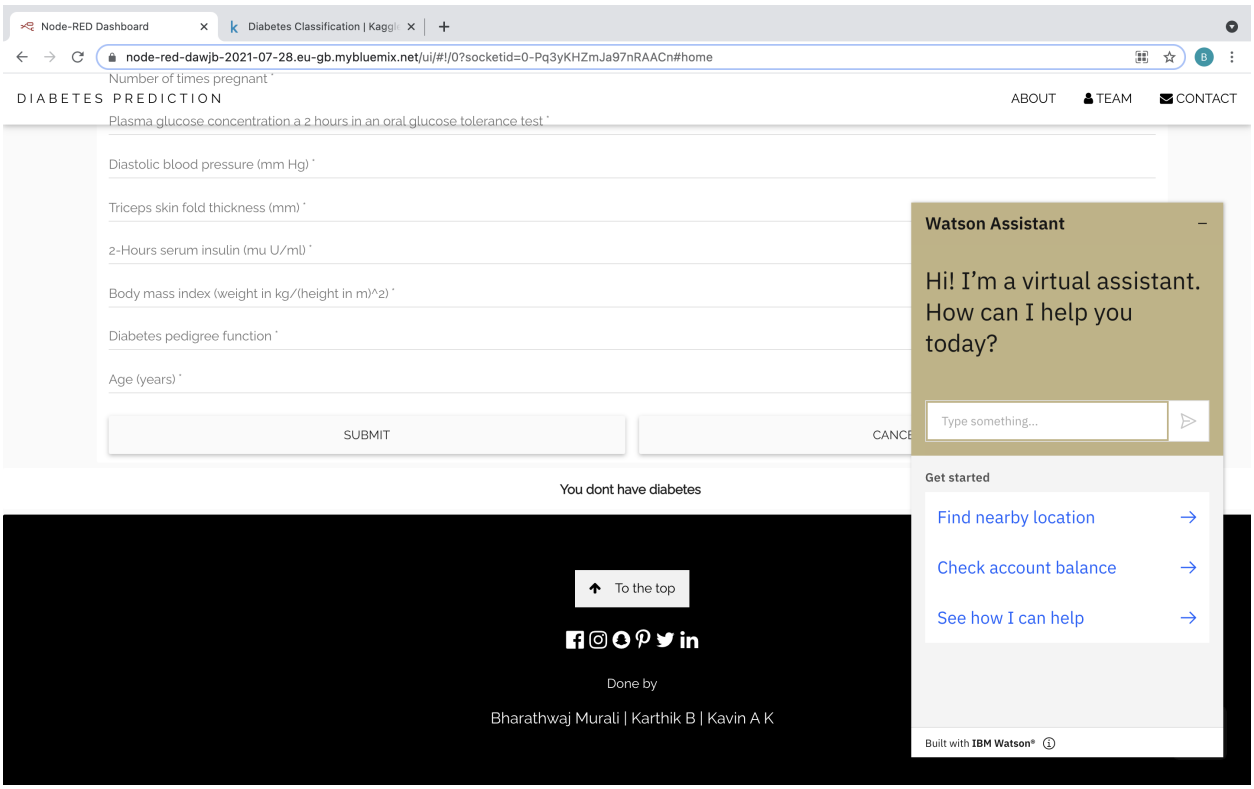
MONITOR HEALTH

Monitoring your blood glucose and blood pressure levels at home. Quitting smoking (if you smoke)

APPOINMENTS

Keeping your appointments with your healthcare providers and having laboratory tests completed as ordered by your doctor.


Node-RED Dashboard
Diabetes Classification | Kaggle
node-red-dawjb-2021-07-28.eu-gb.mybluemix.net/ui/#!/0?socketid=0-Pq3yKHZmJa97nRAACn#home
DIABETES PREDICTION
ABOUT
TEAM
CONTACT
Diastolic blood pressure (mm Hg) *
85
Triceps skin fold thickness (mm) *
33
2-Hours serum insulin (mu U/ml) *
0
Body mass index (weight in kg/(height in m)^2) *
37.4
Diabetes pedigree function *
0.244
Age (years) *
41
SUBMIT
CANCEL
You dont have diabetes
To the top
Done by
Bharathwaj Murali | Karthik B | Kavin A K


Node-RED Dashboard
Diabetes Classification | Kaggle
node-red-dawjb-2021-07-28.eu-gb.mybluemix.net/ui/#!/0?socketid=0-Pq3yKHZmJa97nRAACn#home
Number of times pregnant *
DIABETES PREDICTION
ABOUT
TEAM
CONTACT
Plasma glucose concentration a 2 hours in an oral glucose tolerance test *
Diastolic blood pressure (mm Hg) *
Triceps skin fold thickness (mm) *
2-Hours serum insulin (mu U/ml) *
Body mass index (weight in kg/(height in m)^2) *
Diabetes pedigree function *
Age (years) *
SUBMIT
You dont have diabetes
To the top
Done by
Bharathwaj Murali | Karthik B | Kavin A K
Watson Assistant
Hi! I'm a virtual assistant. How can I help you today?
Type something...
Get started
Find nearby location
Check account balance
See how I can help
Built with IBM Watson®

8 Advantages and Disadvantages

8.1 Advantages

XGBoost is an efficient and easy to use algorithm which delivers high performance and accuracy as compared to other algorithms. XGBoost is also known as regularized version of GBM. Let see some of the advantages of XGBoost algorithm:

Regularization: XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why, XGBoost is also called regularized form of GBM (Gradient Boosting Machine).

While using Scikit Learn libarary, we pass two hyper-parameters (alpha and lambda) to XGBoost related to regularization. alpha is used for L1 regularization and lambda is used for L2 regularization.

Parallel Processing: XGBoost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model.

While using Scikit Learn libarary, nthread hyper-parameter is used for parallel processing. nthread represents number of CPU cores to be used. If you want to use all the available cores, don't mention any value for nthread and the algorithm will detect automatically.

Handling Missing Values: XGBoost has an in-built capability to handle missing values. When XGBoost encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node. It then does the same when working on the testing data.

Cross Validation: XGBoost allows user to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we have to run a grid-search and only a limited values can be tested.

Effective Tree Pruning: A GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm. XGBoost on the other hand make splits upto the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.

8.2 Disadvantages

Well XGBoost (as with other boosting techniques) is more likely to overfit than bagging does (i.e. random forest) but with a robust enough dataset and conservative hyperparameters, higher accuracy is the reward. XGBoost takes quite a while to fail, that's another drawback when compared to more naive approaches. Overall though, as far as boosting goes, XGBoost is an upgrade on an idea (gradient boosting) that was itself an improvement on naive bagging techniques. Because it was created relatively recently and its design took into account the issues with existing models, it tends to outperform them based on those metrics. It's important to remember that XGBoost is essentially just regular gradient

boosting with some regularization and such (which provides some help in avoiding overfitting), so any set of circumstances which cause gradient boosting to fail could cause XGBoost to fail.

9 Applications

We can integrate this with Nodred to make it a fully working website which can be partnered with any hospital. We have also used IBM Watson Assistant to make a chatbot to interact with the users and make the feel convenient and know more about their body.

10 Conclusion

The main aim of this project was to design and impl
ement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses XGBoost classifier with hyper-parameter tuning and feature engineerng using IBM AutoAI service. And 77% classification accuracy has been achieved. The Exper- imental results can be asst health care to take early prediction and make early decision to cure diabetes and save hu- mans life.

11 Future Scope

In future, if we get a large set of diabetic dataset we can perform comparative analysis for analyzing the performance of each algorithm as well as the Hybrid algorithm so that the best one can be applied for predictive analysis. A particular method to identify diabetes is not very sophisticated way for initial diabetes detection and it is not fully accurate for predicting diseases. Thats why we need a smart hybrid predictive analytics diabetes diagnostic system that can effectively work with accuracy and efficiency. We can use data mining , neural network for exploring and utilizing to support medical decision, which improves in diagnosing the risk for pregnant diabetes. Due to the dataset we have till date are not upto the mark , we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and explore it, which may improve the accuracy of predicting diabetes. We can also study the causes of diabetes and how to avoid having diabetes.

12 Bibiliography

1. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.

2. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

3. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

4. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning

Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.

5. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

6. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.