AICTE Faculty Development Program

On

IBM Cloud Technologies & Hack Challenge “Guru Cool”

By

IBM in collaboration Smart Bridge

https://smartinternz.com/gurucool

Date: 28-Sep-2020 to 20-Oct-2020

Project:

Diabetes Mellitus Prediction using IBM Auto AI Service

By

Chakola Paul Johny

Email: paul.chakola@vidyaacademy.ac.in

Telephone: +91 949 675 4571

Vidya Academy of Science and Technology

AICTE Approved & Affiliated to the APJAK Technological University

Thalakkottukara PO, Thrissur - 680501, Kerala, India

Website: www.vidyaacademy.ac.in

Telephone: +91 (4885) 287751 / 287752

Contents

1. Introduction

a) Overview

Diabetes mellitus is a chronic disease characterized by hyperglycemia and can cause many complications if not detected and treated properly. Hyperglycemia refers to high levels of sugar, or glucose, in the blood. It occurs when the body does not produce or use enough insulin, which is a hormone that absorbs glucose into cells for use as energy. According to a recent study by 2040 the world's diabetic patients will reach 642 million. Thus, the need to apply technology to understand the underlying parameters leading to diabetes and using the same to predict the likelihood of its concurrence in a candidate. It has been found that many PMIA Indian women are diabetic. PIMA Indians are people who live in the Gila River area of Southern Arizona, USA. A database of their vital parameters is available which is a good source to study and develop a prediction model using machine learning techniques.

b) Purpose

PMIA Indian database is available for 768 observations of women on eight parameters viz.:

Age	: age	: Age of the candidate
Body Mass Index	: mass	: Body mass index (weight / height)
Blood Pressure	: pres	: Diastolic blood pressure
Glucose	: plas	: Plasma glucose concentration
Insulin Level	: test	: 2-hr serum insulin level
Diabetic Pedigree Function	: pedi	: Diabetic pedigree function influence
Skin Thickness	: Skin	: Triceps skin fold thickness
No. Of Pregnancies	: preg	: Number of pregnancies
Class	: class	: Diabetic (1) / non-Diabetic 0)

This data is solid base to develop a model and can be used to test likelihood of being diabetic based on these parameters which can be availed from simple blood test reports.

2. Literature Survey

a) Existing problem

Manual blood testing would prove the occurrence of being diabetic only upon crossing the threshold blood analysis. Thus, it may be too late to start any preventive actions like food intake control, exercise regimen, etc. which can assist with controlling the onset of diabetes. Thus, the existing model will tell you if you are diabetic or not diabetic. The need of the hour is to have robust model which could predict the likelihood of being diabetic in the future based on the current blood and other parameters.

b) Proposed solution

The proposed solution is to build a prediction model using the IBM Cloud suite to develop a model and deploy the same over website. The site URL can be accessed using a computer, or a mobile device to key in the basic blood test parameter results and the system would predict if the candidate is diabetic (1) or not diabetic (0).

3. Theoretical Analysis

a) Block diagram

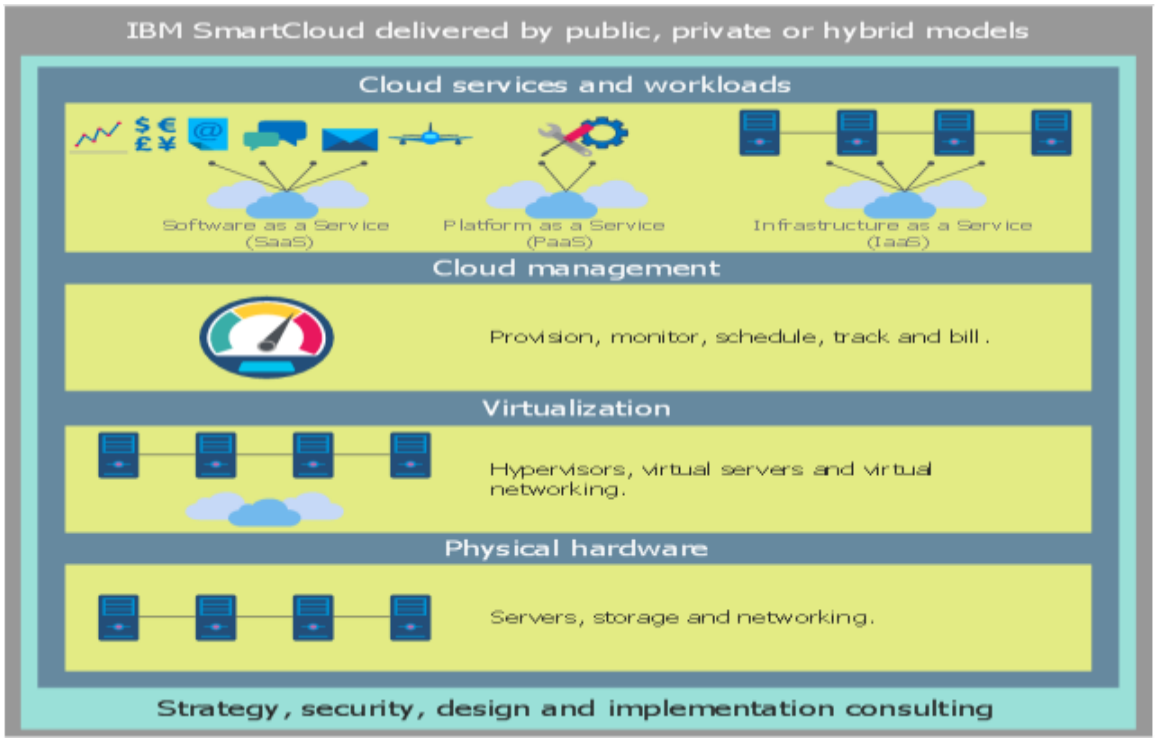

Figure: IBM Cloud high level architecture

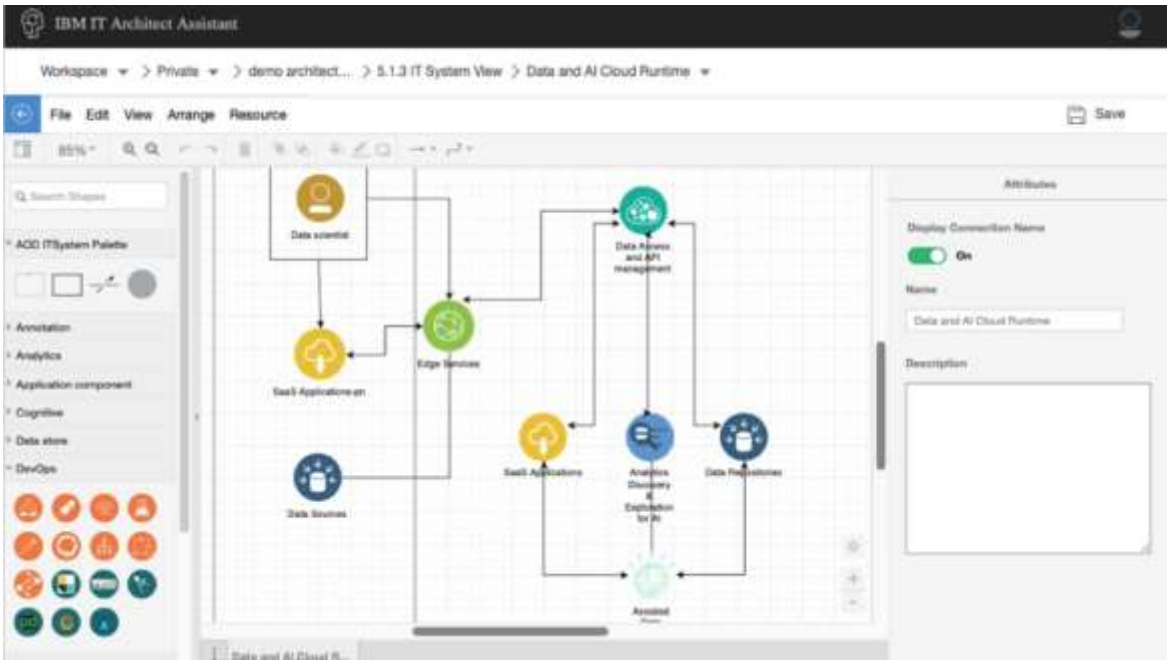

Figure: IBM Cloud service architecture

b) Hardware / software design

Hardware : IBM Cloud

- IBM Lite plan
- IBM Cloud Object Storage

Software : IBM Cloud services

- IBM Watson Studio
- IBM Watson Auto AI service
- IBM Machine Learning service
- IBM Node RED service

4. Experimental Investigation

The model was predicting the results in this case the following approach was adopted.
- Predictor variable selected was “class”
- All other variables, eight in all, and all numeric was set as the independent variables.
- Binary Classification prediction was default identified by Watson Auto AI
- The service by default executes the prediction using various algorithms and provides the details including the “accuracy” of the selected algorithm.
- The most appropriate one can be selected and the model saved.
- The deployed model also provides a test interface.
- Either by data entry or as a JSON file inputs can be given to the model.
- And the model predicts the results.
- The model result also gives the probabilities.

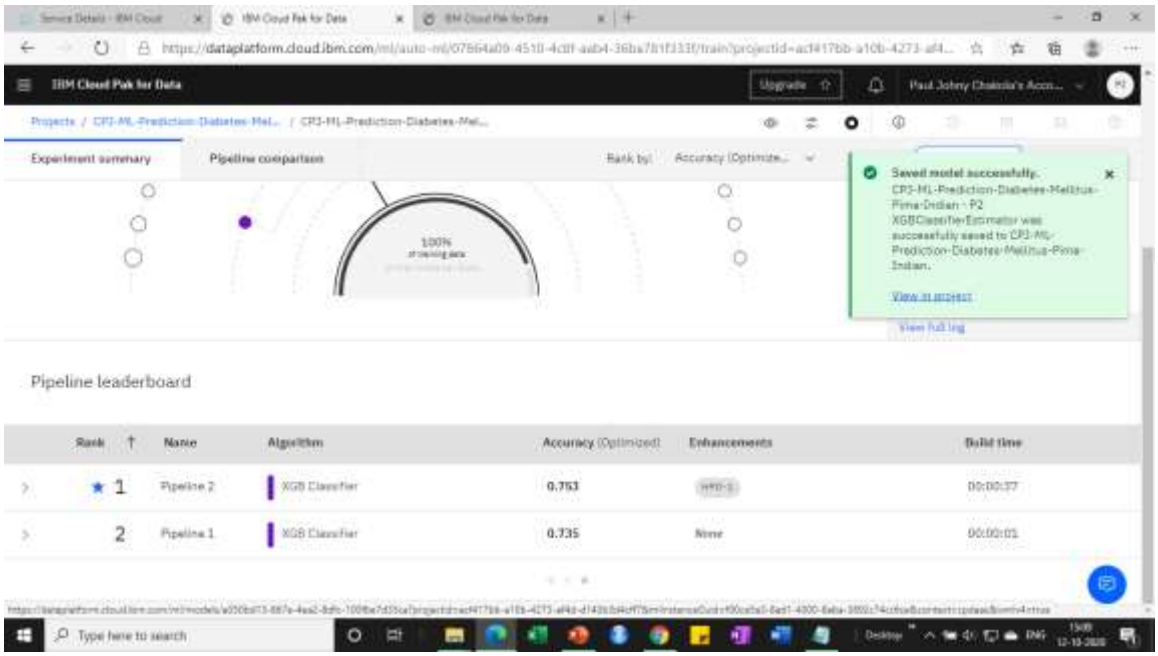

Figure: The experiment summary result to select the appropriate model.

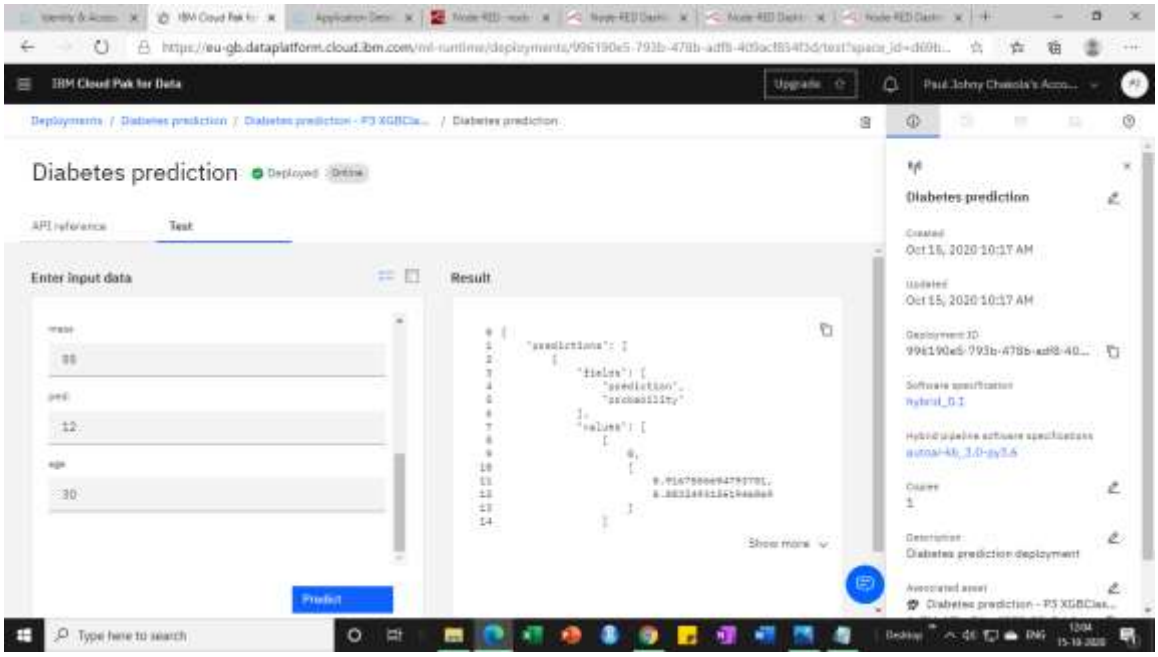

Figure: Depicting the model test interface with the prediction in the result pane.

5. Flowchart

1. Obtain IBM Cloud Account
    a) Obtain IBM ID
    b) Lite Plan (limited capacity unit hours - CUH)
2. Install the services, viz.:
    a) Configure or setup the Cloud Object Storage
    b) Watson Auto AI service
    c) Watson Studio
    d) Watson Machine Learning service
    e) Other services on need basis
3. Upload the data set
4. Train the model
5. Deploy the model
6. Test the model results
7. Create the Node RED URL
    a) Install Node RED under the Cloud Foundary service
    b) Test and verify with sample data
8. Deploy the URL

6. Result

The end users are provided a URL which is user friendly. They can provide the necessary input and upon click of submit button the application provides the predicted result.

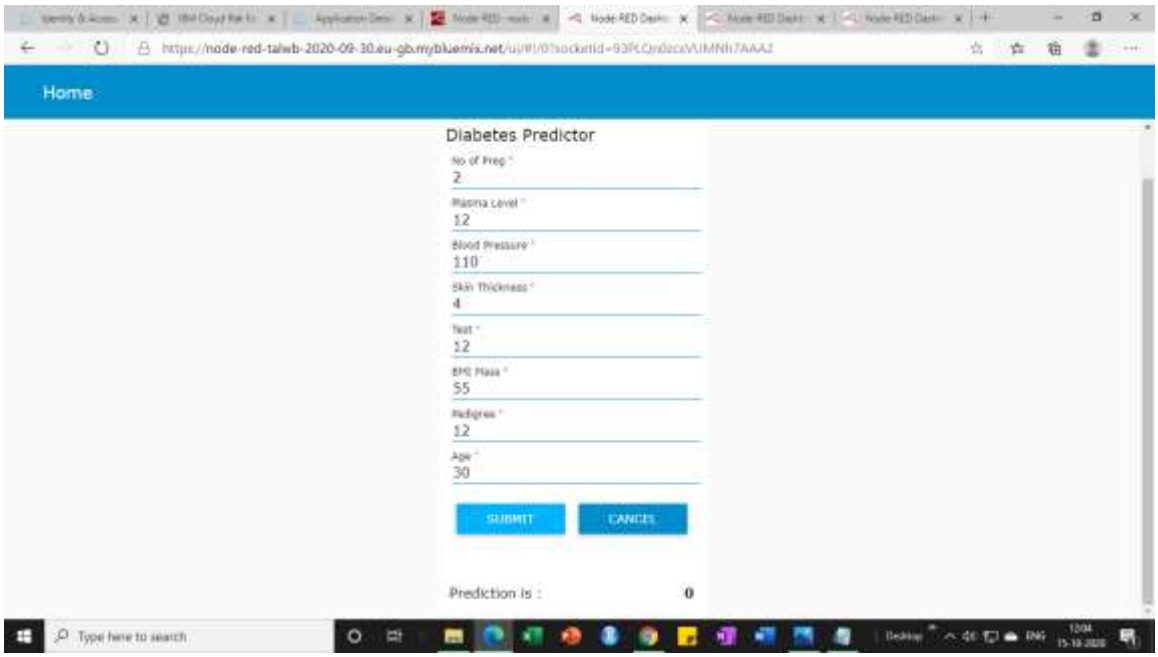

Figure: User interface showing the input values and the predicted result.

The solution URL:

https://node-red-talwb-2020-09-30.eu-gb.mybluemix.net/red/#flow/c775cd37.8678d

Note: Incidentally, from 17-Oct-20, getting the error message “This deployment cannot be processed because it exceeds the allocated capacity unit hours (CUH). Increase this compute resource for this job and try again. Thus, the URL may not work. The plan of action is to create a new account and try all over again.

7. Advantages and Disadvantages

Advantages:

- Very easy to develop a model, test it, and then deploy the same. Say in a matter of few couple of hours.
- Limited coding effort to do the machine learning job using a spectrum of algorithms.
- IBM Cloud offers a plethora of services which can be seamlessly integrated into ones project and utilized effectively.

Disadvantages:

- The model currently is based on data collected for only PIMA Indian and that too for women candidates. Can this be used to predict likelihood of diabetes in male candidates.
- The current best algorithm identified by the model is XBG Boost with an accuracy of 77%. The data set many be pruned or additional conditioning done to get the accuracy to 99% level.
- Lite plan capacity unit hours (CUH) get exhausted very quickly. This must be made available for more duration for academic users.

8. Applications

The approach can be used to develop a host application which could be regression or a classification problem. Additionally, the SPSS Model could be used to effectively analyses the problem and the solution that is being provided.

9. Conclusion

The IBM Cloud solution with all its features including cloud object storage, micro services, project space, and with the deployment feature all integrated and seamless is a very powerful concept to develop, test, and deploy scale-able application solutions in a very short time.

10. Future Scope

The model could be developed further by incorporating additional IBM Watson an associated micro service as follows:

Data wrangling : To suitable IBM Watson micro services to automatically clean and prune the train & test data set. This can be done to remove outliner values or replace them with say median or average values. Missing values could be replaced with say average value for the applicable sub-group.

Data visualization : To visually analyze the basic data to understand the parameters better. That is to verify if the data elements are normally distributed or not.

Modelling : Multiple models could be build using a different algorithm falling in high accuracy range. This could be used to validate the results across the potential spectrum of prediction algorithms.

Rationalization : The model may be rationalized to use the base PIMA data collected from women candidates to predict likelihood of diabetes in male candidates.

Re-design the test : Analyze the input parameters to rationalize the data to the most influencing parameters say BMI mass, blood pressure, glucose level and use only these parameters as inputs to predict the results.

Train the data set : Use additional data to continually train the model to predict the results better with a higher accuracy of say of 99%.

User friendly : Decode the prediction and display the results in a more user-friendly way. Rather than 0 and 1, give the predictions in a verbal mode.

11. Bibliography

https://www.medicalnewstoday.com/articles/323699#:~:text=Hyperglycemia%20refers%20to%20high%20levels,cells%20for%20use%20as%20energy.

https://www.researchgate.net/figure/IBM-cloud-computing-reference-architecture_fig1_329468734

https://www.conceptdraw.com/examples/diagrams-of-ibm-smart-cloud
https://www.ibm.com/cloud/architecture/architectures/edit

12. Appendix

a) Source Code

No explicit code was written. Cool! Please see the screen shots of the work elements.

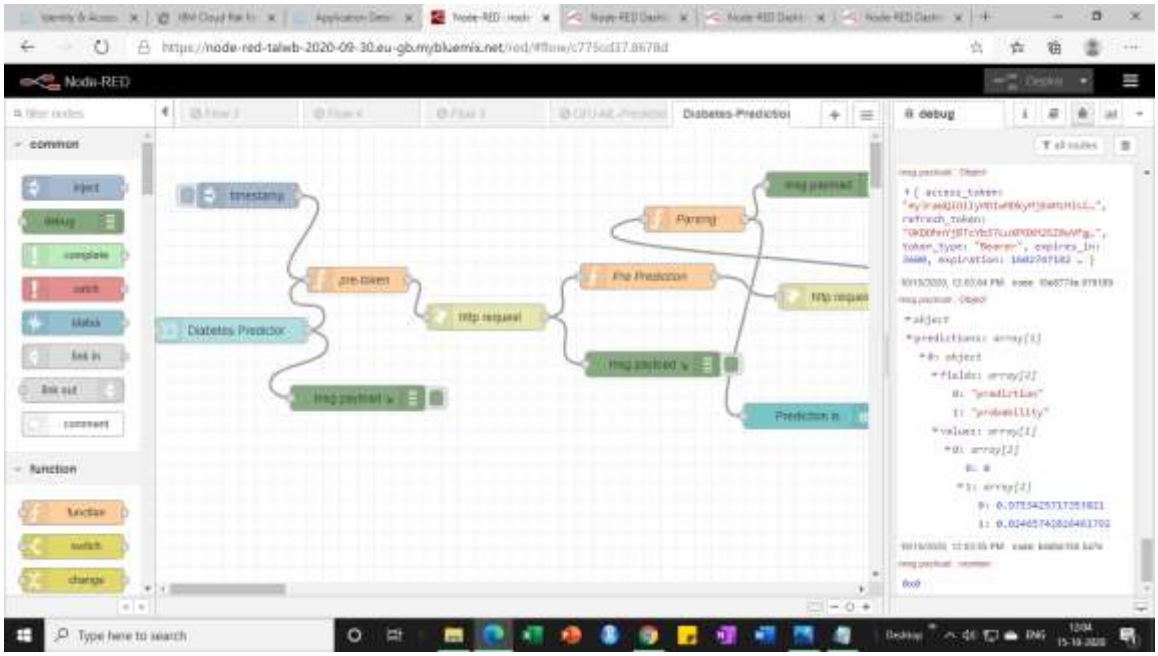

Figure: The Node RED view of the nodes and debug view of the results.

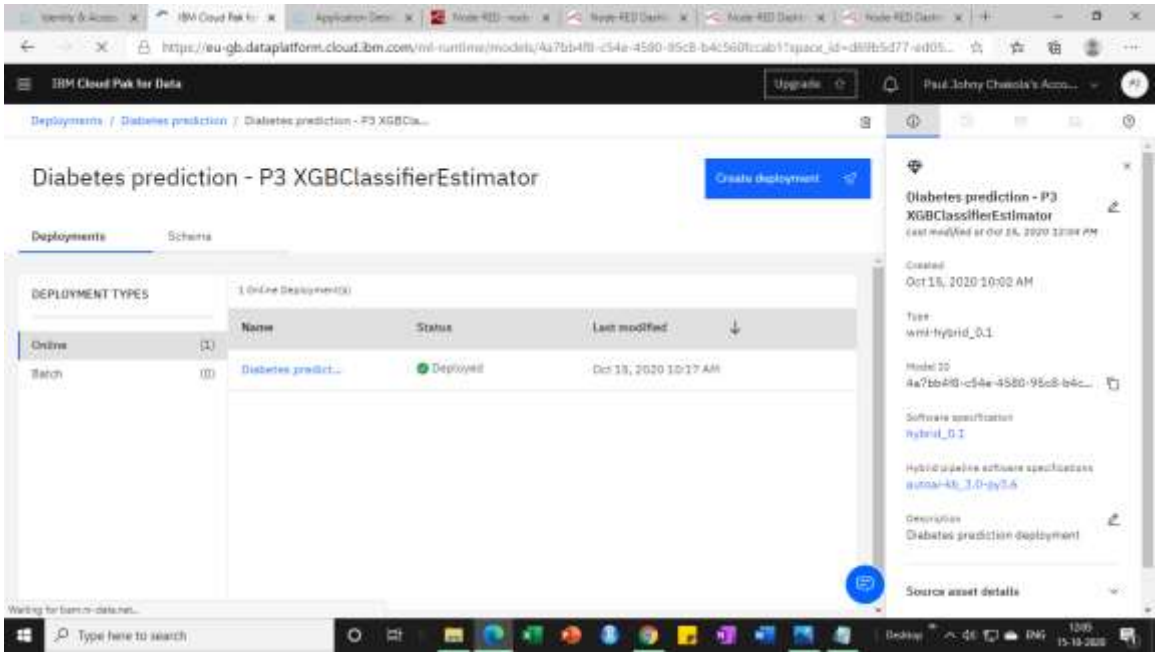

Figure: Model being taken for deployment.

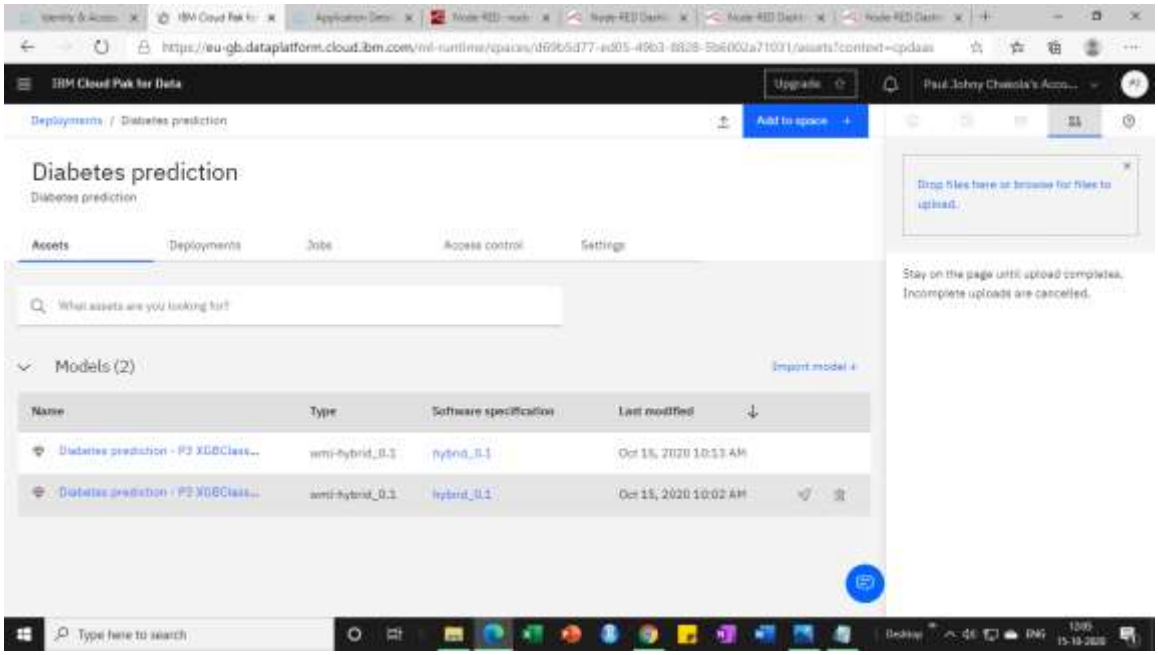

Figure: Model available the space

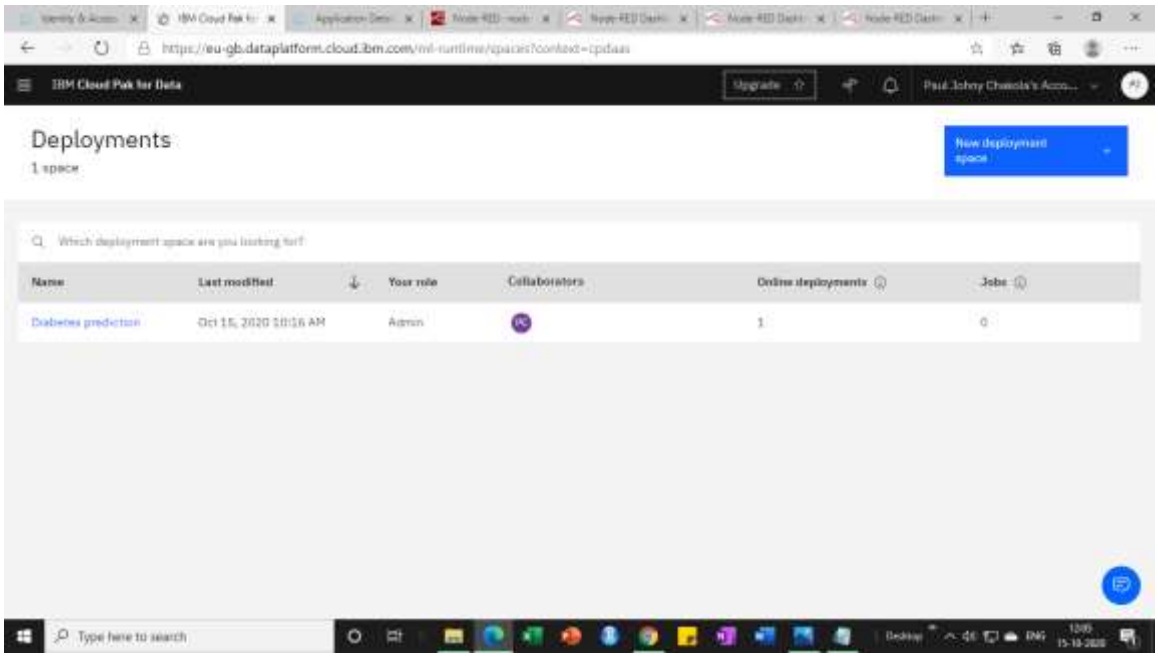

Figure: Deployment space view

b) Test Data Results:

Condition	Parameter	Validation 1	Validation 2	Validation 3
No. of Pregnancies	preg	2	2	2
Blood Plasma Level	plas	123	123	12
Blood Pressure	pres	120	220	110
Skin Thickness	Skin	4	14	4
Insulin Level	test	123	1234	12
BMI Mass	mass	80	180	55
Diabetic Pedigree Function	pedi	123	123	12
Age of the candidate	age	45	60	30
Model Prediction	class	1	1	0
Node Red URL Value Displayed	class	1	1	0