SPS-6531-Diabetics-Prediction-System-based-on-Life-Style

1. 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.

Artificial Intelligence is a wide domain which contains wide spectrum of methods for modeling variety of solutions to problems. It's applications in healthcare domain, very effectively used in prediction of diseases. Various life style and other parameters have an impact progression of this disease. These parameters and the outcome for number of patients already available in the domain can be used to m from and predict with good amount of accuracy when new/unseen data points are posed to it. This would open new avenues for match of healthcare that would assist the healthcare stakeholders to make informed decisions by studying large number of patient data. This strategizes customized treatment methods to better meet the varying disease symptoms in case of chronic diseases like diabetes.

2. OVERVIEW

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.

3. PURPOSE

To review the long-term effects of the diabetic pregnancy on the offspring among the Pima Indians of Arizona.

4. HARDWARE/SOFTWARE

4.1 PROJECT REQUIREMENTS

1. A Classification algorithm with maximum accuracy to be trained and tested on the dataset.
2. The Dataset consists of 8 columns excluding the predicting column i.e. Class

4.2 SOFTWARE REQUIREMENTS

3. IBM Cloud
4. IBM Watson Studio
5. Node-red App

5. EXPERIMENTAL INVESTIGATION

5.1 Data Understanding: Data Set Properties: 9 attributes presenting 769 Pima Female Indians.

5.2 Data Set Info: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

5.3 Dataset Attributes:

1 Pregnancies: Number of times pregnant

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

3 Blood Pressure: Diastolic blood pressure (mm Hg)

4 Skin Thickness: Triceps skin fold thickness (mm)

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

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

7 Pedigree Function: Diabetes pedigree function
8 Age: Age (years)
9 Outcome: Class variable (0 or 1)"

5.4 Data Preparation:

Analyzed 0 value records across all attributes. There were several attributes where a 0 value did not make sense and I think were used for missing data. (There were 0's for Blood Pressure, Skin Thickness, Insulin, BMI, Glucose). I used mean substitution for these cases. 0 Value Analysis in Dataset.

Analyzing the 0 values in the dataset					
Attributes	Records with 0 Value	Total Records	% of Total Records	Does a 0 value make sense?	Mean substitution for 0 Values that don't make sense
Pregnancies	111	768	14%	Yes, there is no requirement for females to have kids.	0
Glucose	5	768	1%	No	122
BloodPressure	35	768	5%	No	72
SkinThickness	227	768	30%	No	29
Insulin	374	768	49%	No	156
BMI	11	768	1%	No	32
DiabetesPedigreeFunction	0	768	0%	No	0
Age	0	768	0%	No	0
Outcome	500	768	65%	Yes, 0 value represents patients not diagnosted with diabetes.	0

There were not too many outliers in the data. None of the data was removed.

6. MODELING AND EVALUATION

Here, Gradient Boosting Classifier is used with 7 features for best RMSE scores.

Model Information ⓘ

TARGET : OUTCOME

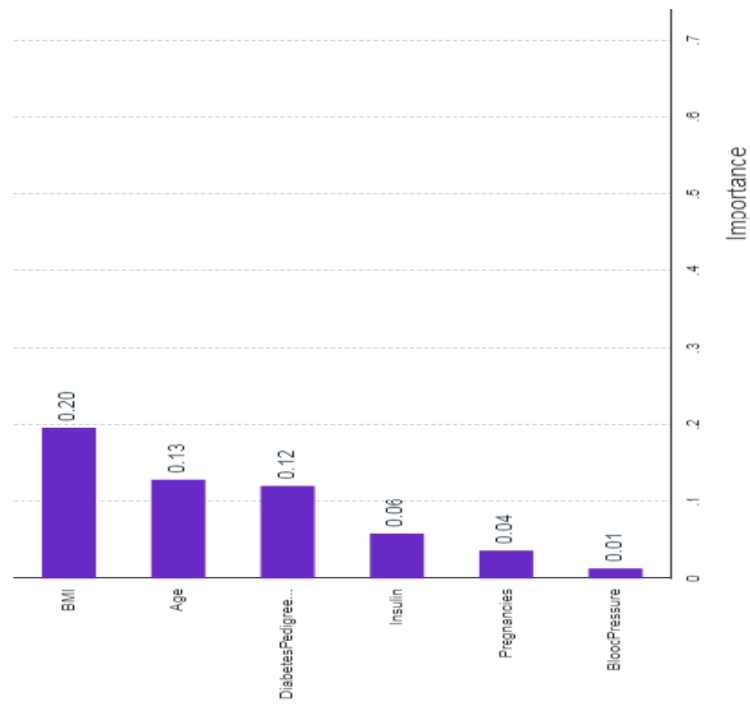| Label (Target)	Outcome
Model Type	Gradient Boosting Classifier
Number of Features	8
Number of Evaluation Instances	77
Created At	10/28/2020, 5:02:02 PM

Fig 2: Model information

Fig 3: Feature Importance

Progress map ⓘ

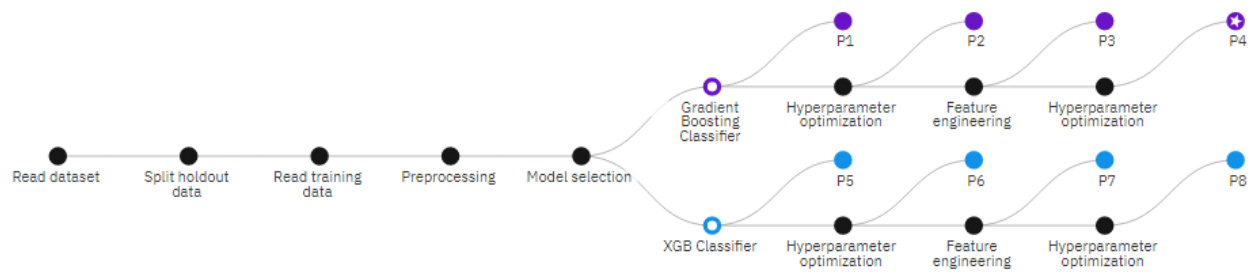Prediction column: **Outcome**

Fig 4 Progress Map Chart

Relationship map ⓘ

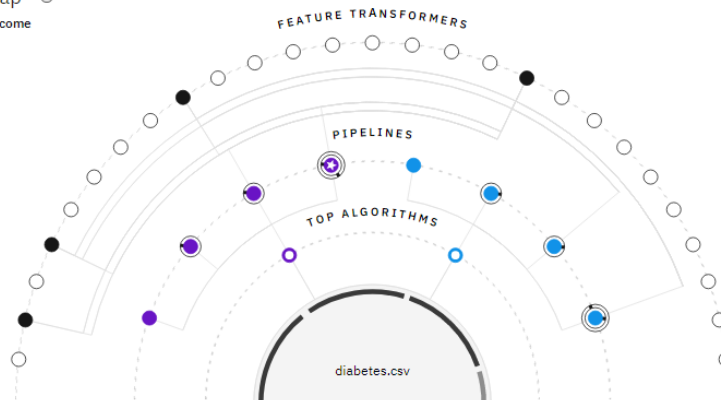Prediction column: **Outcome**
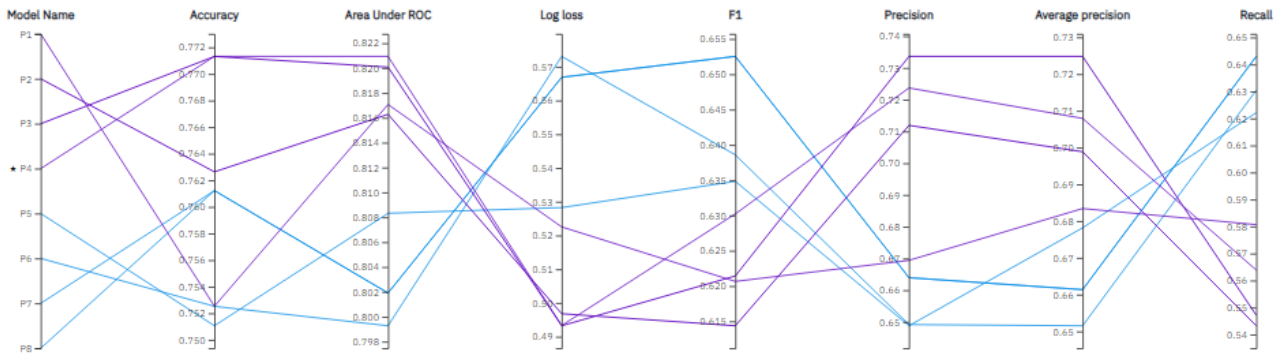

Fig. 5 Relation Ship Map

Metric chart ⓘ
Prediction column: **Outcome**

Fig. 6 Metric Chart

Rank ↑	Name	Algorithm	Accuracy (Optimi...	Enhancements	Build time
★ 1	Pipeline 1	Gradient Boosting Classifier	0.844	None	00:00:01
2	Pipeline 4	Gradient Boosting Classifier	0.844	HPO-1 FE HPO-2	00:00:21
3	Pipeline 2	Gradient Boosting Classifier	0.818	HPO-1	00:00:09
4	Pipeline 3	Gradient Boosting Classifier	0.818	HPO-1 FE	00:00:44
5	Pipeline 5	XGB Classifier	0.805	None	00:00:01
6	Pipeline 7	XGB Classifier	0.805	HPO-1 FE	00:02:32
7	Pipeline 8	XGB Classifier	0.805	HPO-1 FE HPO-2	00:01:03
8	Pipeline 6	XGB Classifier	0.779	HPO-1	00:00:30

Fig. 7 Pipeline Leaderboard

## 7. Auto AI Experiment Results

Grading Boosting Classifier is selected by the Auto AI experiment as the best performing model after fine tuning all the hyper-parameters. It is found to give about 77.1% accuracy with 90% training set size and 10% test set size. The Area Under the Curve (AUC) is also satisfactory which depicts the TPR (sensitivity) and FPR (specificity). The models having higher AUC are said to

perform better.
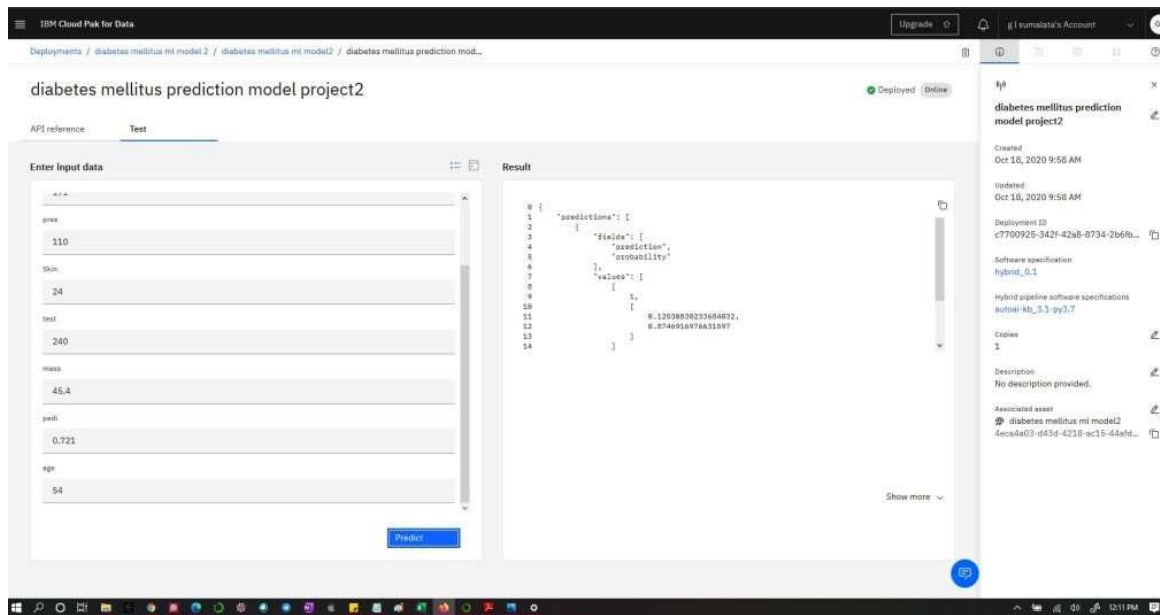
7.1 Node Red Flow

Node-RED is a flow-based development tool for visual programming developed originally by IBM for wiring together hardware devices, APIs and online services. The flows created in Node-RED are stored using JSON.
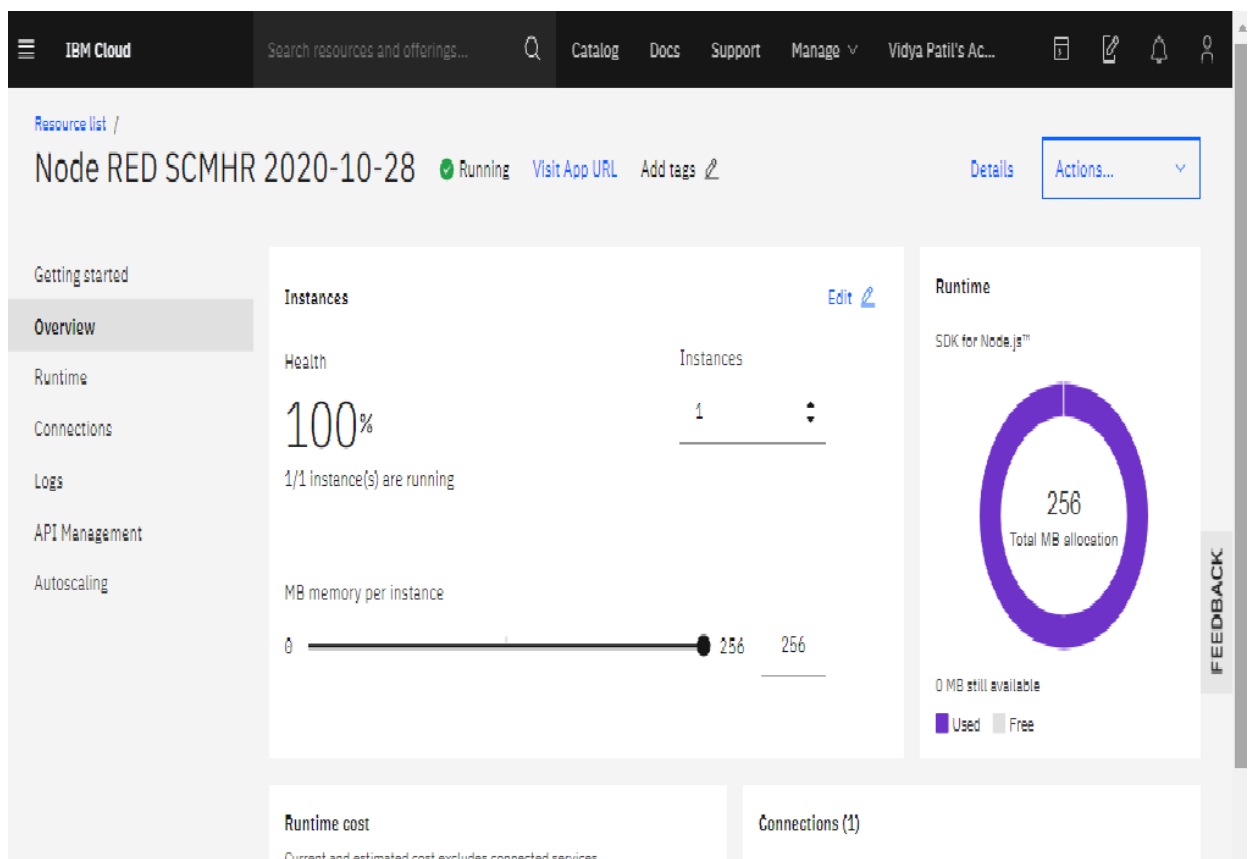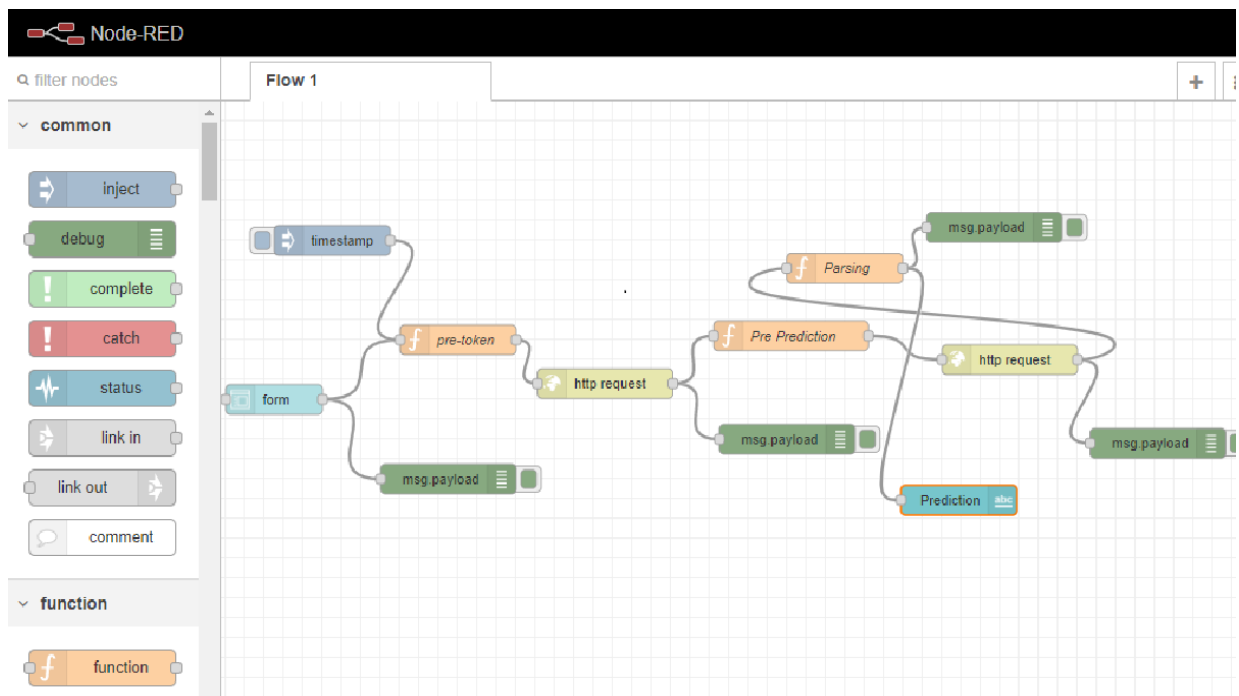
8 Results:
Cross Validation:

Experiment summary	Pipeline comparison		Accuracy (Optimize... ∨)	Cross validation	Holdout

Rank ↑	Name	Algorithm	Accuracy (Optimized)	Average precis...	F₁	Log loss	Precisio
★ 1	Pipeline 4	Gradient Boosting ...	0.771	0.708	0.630	0.493	0.724
2	Pipeline 3	Gradient Boosting ...	0.771	0.725	0.621	0.493	0.734
3	Pipeline 2	Gradient Boosting ...	0.763	0.699	0.614	0.497	0.712
4	Pipeline 7	XGB Classifier	0.761	0.661	0.653	0.567	0.664
5	Pipeline 8	XGB Classifier	0.761	0.661	0.653	0.567	0.664
6	Pipeline 1	Gradient Boosting ...	0.753	0.683	0.621	0.523	0.669
7	Pipeline 6	XGB Classifier	0.753	0.652	0.639	0.573	0.649
8	Pipeline 5	XGB Classifier	0.751	0.678	0.635	0.528	0.649

Fig 8. Cross Validation Value for Gradient Boosting Classifier

The offspring of women who had diabetes during pregnancy, on average, were more obese and had higher glucose concentrations and more diabetes than the offspring of women who developed diabetes after pregnancy or who remained non diabetic. Although no new analyses were attempted, several of the older publications were updated by repeating the analyses on later, expanded data sets.

New

Enter No of Pregnancies *
6

Enter Glucose *
50

Enter Blood Pressure *
72

Enter SkinThickness *
35

Enter Insulin *
0

Enter your BMI *
33.6

Diabetes Pedigree Function *
0.627

Age *
50

[SUBMIT]   [CANCEL]

Prediction **Higher Chances of Dibetic**

8.2 Application Result: Node red Application Ui:

https://node-red-scmhr-2020-10-28.eu-gb.mybluemix.net/ui/#!/0?socketid=lxi1GiJHTZbLx-YfAAAh

8.3 YouTube Link:

https://youtu.be/nubZ3f-KOP4

9. Conclusions:

The diabetic pregnancy, in addition to its effects on the newborn, has effects on the subsequent growth and glucose metabolism of the offspring. These effects are in addition to genetically determined traits.