PREDICTING DIABETES MELLITUS IN PREGNANCY WOMEN

ABSTRACT:

Healthcare trade contains very large and sensitive data and needs to be handled very carefully. Diabetes Mellitus is one of the growing extremely fatal diseases all over the world. Medical professionals want a reliable prediction system to diagnose Diabetes. The accessibility and availability of huge amounts of data will be able to provide us useful knowledge if certain data mining techniques are applied on it. Data mining with Classification Algorithms and ANN plays an important role in the field of medical diagnosis to diagnose the disease. There requires a model built to analyse and extract information from the data. This paper discuss about the model, which gives an accurate result to the women about chances of getting gestational diabetes according to the data given.

INTRODUCTION:

The procedure of data mining deals with recognizing valuable information from vast amount of heterogeneous data and it helps to determine the patterns and rules from the data. Now days, the health organizations producing huge amount of data in the area of cancer and other health issues so it’s very difficult to analyze the data. Medical data mining helps to extract hidden patterns, which thereby opens the door to an enormous source of analysis of medical data. Classification, Clustering, Regression etc., are some data mining techniques. Diabetes is a very severe health issue when the level of blood glucose becomes unregulated. In actual, glucose acts as fuel to our body. When the body uses glucose as a fuel, insulin is essential to get the glucose into cells. Diabetics happen because of either production of insulin is insufficient or the cells don’t

react accurately to insulin or else both. Those who are having diabetes they suffer with polyuria (increase in thirst and hunger).

Data mining is becoming popular in health field, so, many data mining tools are used for analyzing the data. For the most part information mining instruments are utilized to foresee the victories from the data collected. Developing efficient data mining tools helps to reduce the cost and time, thus increasing the efficiency.

This research is to predict the possibility of getting diabetics in pregnant ladies based on the attributes like Age, BMI, Plasma concentration and the number of times being pregnant. It provides an option for the customer to choose the classifier like Naive Bayes and SVM. When the user selects any one of the classifier it gives the output as a message whether there is a chance of getting diabetes or not for the particular user.

LITERATURE REVIEW:

Ashok Kumar et al [1] proposed Performance and Evaluation of Classification Data Mining Techniques in Diabetes. In this paper, they have used some selected classification algorithms like Support vector mechine, Regression, BayesNet, NaïveBayes and Decision Table for the classification of diabetic patient dataset. For accurate and proper results, classification techniques are widely used in the medical field. Data mining techniques are used in healthcare field for, Diagnosis and treatment. The result showed that Decision table had highest Accuracy than other Classification Algorithm.

Mukesh Kumari et al [2] proposed a prediction of diabetes using Bayesian Network. With this paper Bayesian Network classifier was utilized to gauge the people whether they are diabetic or not. The dataset has been accumulated from a medicinal center, which gathers the data of individuals with and without diabetes. They used Weka tool for the test investigation. Dataset contains every one of the subtle elements of a man like quick gtt(Glucose resistance test) value, casual Glucose tolerance test value,number of time pregnant,diastolic blood pressure (mmhg),triceps skin fold thickness(mm),serium insulin(μU/ml), body mass index (kg/m)diabetes pedigree function,age of individual.

V.Karthikeyani et al [3] has played out a near investigation of 10 Data Mining Classification Algorithm in Diabetes Disease Prediction .In this paper they have done a talk of C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA and PLSLDA. They have utilized an tool called Tanagra which is a capable device that contains grouping, supervised learning, Meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The outcomes have demonstrated The CS-CRT algorithm best among tens. Chintan shah et al [4] used Data Mining Classification Algorithms for the prediction of Breast Cancer. In this research the researchers have used three different data mining classification methods for prediction of breast cancer which are decision tree algorithm, Bayes classification algorithm, K-Nearest Neighbour classification algorithm. They focus on accuracy and lowest computing time. When compared to other algorithms Naive Bayes is more accurate (95.9943) and faster. It takes as less as 0.02 seconds which is very less when compared to other types of Data Mining algorithms.

Divya Tomar et al [5] suggested the knowledge towards prerequisites for health domain also regarding suitableness decision for available system. There is no single information mining systems which provide for steady results for different types of data in healthcare. Those execution of data mining techniques varies concerning illustration for every the dataset that we need picked for those test. Classification, clustering and mining is used by the information mining systems in healthcare to increase their ability for decision making in patient health. On the basis of the seriousness of the disease a patient can be classified into “high risk” and “low risk”. For these purposes we can use the classification methods like KNN (K-nearest neighbour), DECISION TREE (DT), and SVM (support virtual machine), NN (Neural network). The result showed that that utilizing data mining learning doctor might effectively distinguish those viable cure, patients acquire expense compelling treatments, social insurance industry manages their client What's more social insurance insurers find at whatever cases of cheating for medicinal claim.

The existing framework need diverse sorts of data mining tools would accessible in the marketplace, every with their qualities also shortcomings. As per the current day, all the tools for Data mining, where we can do pre-processing, mining, and analysis data

with a various different data sets which supports various data mining functions such as clustering, visualization, pre-processing, feature selection and regression. There are many similar Tools, which have not yet, full-fledged like WEKA, Orange etc. All the data mining tools are grouped into three categories such as dashboards, text mining tools and traditional data mining tools.

The approach used by the existing system for the detection of diabetes from the data was complex and time consuming. Those system used combination of complex algorithm to find the result. The authors categorized the drawbacks of misclassification which reduce the performance of the algorithm. The public health agencies can use these sources of error to develop techniques to refine the algorithms and improve the efficiency of the data and also reduce the errors. Additionally using these algorithms in the public health systems helps to enhance and validated the work done.

PROBLEM STATEMENT:

Data Mining is one of the most booming area of research that become increasingly popular in health organization, mostly in healthcare field may be a need for proficient explanatory procedure for identifying obscure also profitable data for health information.

Diabetes is often known as queen of all diseases as it is seen in people of all ages. Different doctors make use of different symptoms to predict the chances of a patient to be diabetic. However a standard set of parameters that directly contribute to diabetes have not yet been identified, so the proposed system focuses on identifying such a parameters set and selecting an efficient algorithm would greatly aide patients to identify their possibility of getting diabetes without the need for doctors or other expensive equipment.

DATASET DESCRIPTION:

For Each Attribute

1. Number of times pregnant: Throughout pregnancy, those placenta makes hormones that might prompt a build-up from claiming sugar in the blood. Usually, pancreas make sufficient insulin response on handle that. In not, your glucose levels will Ascent Furthermore might foundation gestational diabetes.

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test: Over fasting adults, blood plasma glucose ought further bolstering not surpass 7 mmol/l alternately 126 mg/dL. Maintained higher levels for glucose result in harm of the blood vessels what’s more of the organs they supply, prompting those difficulties for diabetes.

3. Diastolic blood pressure (mm Hg): Hosting diabetes increments your danger of Creating high blood pressure Furthermore different cardiovascular problems, in view diabetes adversely influences those arteries, predisposing them on atherosclerosis.

4. Triceps skin fold thickness (mm): Measurement of skin fold thickness is mandatory to identify diabetic patients at risk early to prevent the development of cardiovascular disease and protect them against added complications

5. 2-Hour serum insulin (mu U/ml): A lady with insulin response resistance, typical glycaemic levels might a chance to be recognized after a glucose test in view the pancreas will must discharge overabundance insulin response so as with stay with the glucose in the ordinary extend. Therefore, a single “2-hour post-glucose insulin response level” gives the idea on be a dependable pointer from claiming insulin response imperviousness to PCOS patients.

6. Body mass index (weight in kg/(height in m)^2): Overweight particularly obesity, especially In more youthful ages, significantly expands lifetime danger for diagnosed diabetes, same time their effect once diabetes risk, an aggregation expectancy, Also diabetes span diminishes with ageists.

7. Diabetes pedigree function: Gives a few information for diabetes mellitus history for relatives and the hereditary relationship for the individual’s relatives of the patient.

8. Age (years): One important measure for ensuring health in later years is controlling blood glucose levels, since high blood glucose tends to accelerate the effects of aging and increases the risk of developing diabetes complications.

9. Outcome : the probability of the test “positive” or “negative”.

METHODOLOGY:

The methodology, depicted in the following figure, has been adopted for conducting

Diabetes Mellitus in Pregnancy Women.

Exploratory Data Analysis:

Model Building

Feature Selection

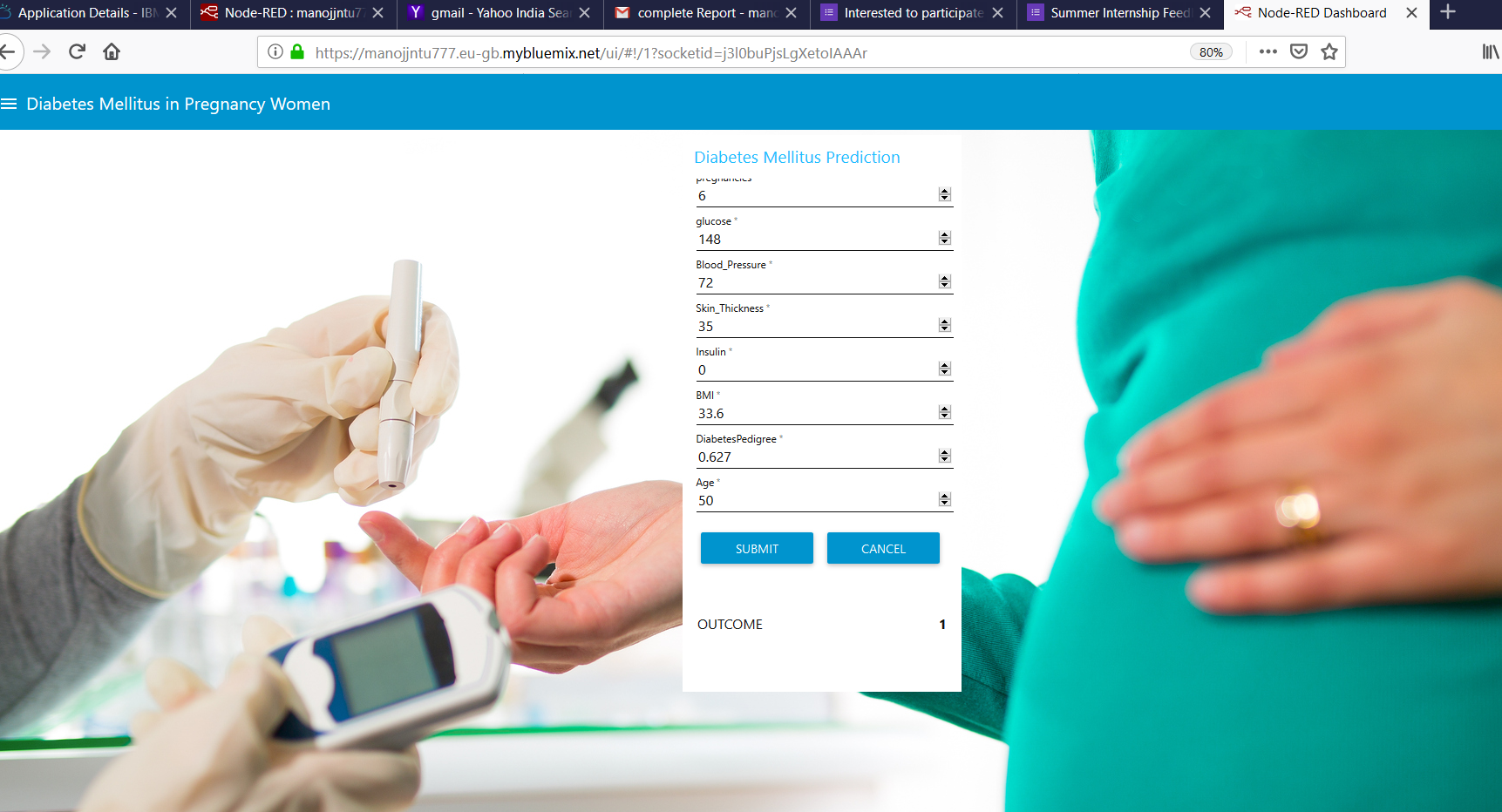
Preprocessing

Data Analysis

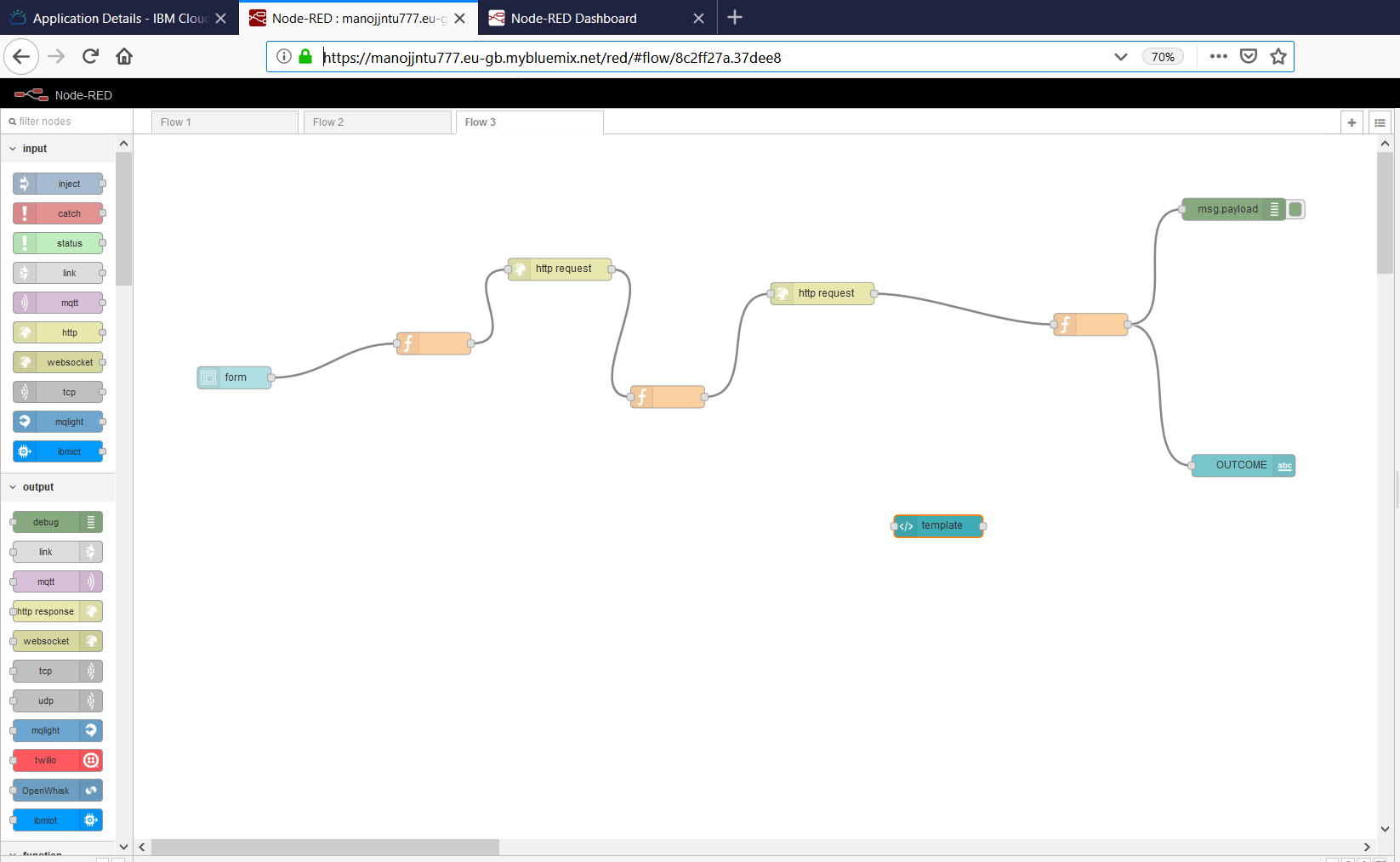
Prediction

Evaluation

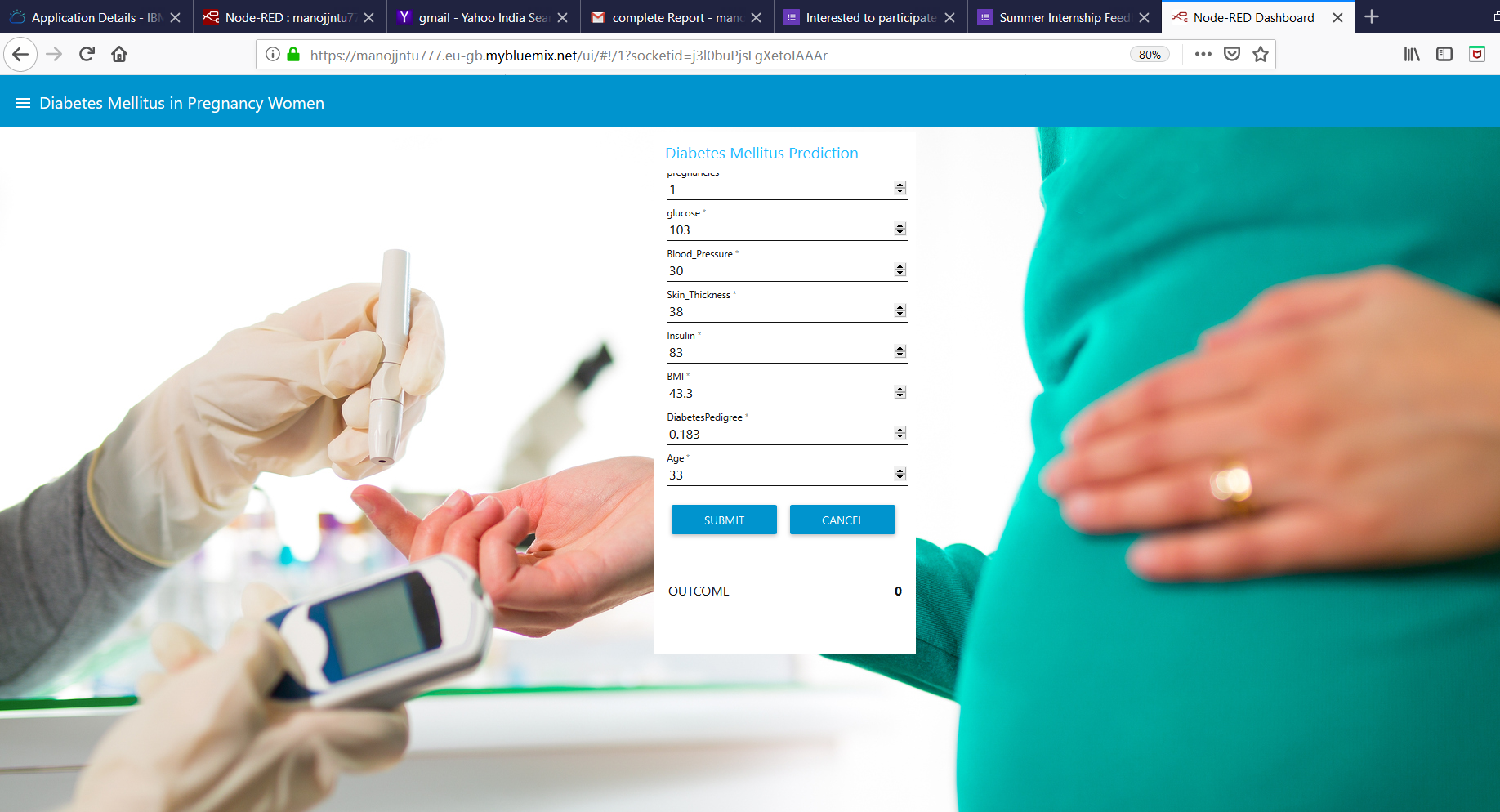
Figures:



UI Showing predicted as Diabetes Patient



Node Red Flow of model



UI Predicted not as Diabetes Patient

Conclusion:

Diabetes is one of the most common diseases in the present generation that is why this topic is chosen for this project. The project helps to analyze and predict the possibility of getting Diabetes in pregnant women based on the relevant attributes such as Age, BMI and Plasma concentration with the help of rules, which is extracted by Decision Tree and ANN algorithms.

Appendix:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import types

import pandas as pd

from botocore.client import Config

import ibm\_boto3

def \_\_iter\_\_(self): return 0

# @hidden\_cell

# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.

# You might want to remove those credentials before you share your notebook.

client\_16bcbffd4ad44876a4839ef08534f1a0 = ibm\_boto3.client(service\_name='s3',

ibm\_api\_key\_id='eZkaPQKV8gMB1QHzp9AyVSpAtWEiy--d-18k7VcaTigG',

ibm\_auth\_endpoint="https://iam.bluemix.net/oidc/token",

config=Config(signature\_version='oauth'),

endpoint\_url='https://s3.eu-geo.objectstorage.service.networklayer.com')

body = client\_16bcbffd4ad44876a4839ef08534f1a0.get\_object(Bucket='project123-donotdelete-pr-hl56fbcahtubf1',Key='diabetes.csv')['Body']

# add missing \_\_iter\_\_ method, so pandas accepts body as file-like object

if not hasattr(body, "\_\_iter\_\_"): body.\_\_iter\_\_ = types.MethodType( \_\_iter\_\_, body )

dataset = pd.read\_csv(body)

dataset.head()

x= dataset.iloc[:,0:8]

x

y= dataset.iloc[:,8].values

y

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,random\_state=0)

x\_test

import keras

from keras.models import Sequential

from keras.layers import Dense

model=Sequential()

model.add(Dense(input\_dim=8,init="random\_uniform",activation="relu",output\_dim=5))

model.add(Dense(output\_dim=5,init="random\_uniform",activation="sigmoid"))

model.add(Dense(output\_dim=1,init="random\_uniform"))

model.compile(optimizer='adam',loss='mse',metrics=['accuracy'])

model.fit(x\_train,y\_train,epochs=200,batch\_size=10)

y\_pred=model.predict(x\_test)

y\_pred1=(y\_pred>0.5)

y\_pred1

y\_test

y\_pred=model.predict(x\_test)

y\_pred

model.predict((np.array([[2,150,80,40,25,30,0.614,45]])))

x.ndim

y.ndim

from sklearn.metrics import confusion\_matrix

cm=confusion\_matrix(y\_test,y\_pred1)

from sklearn.metrics import accuracy\_score

accuracy\_score(y\_test,y\_pred1)

model.save('diabetis.h5')

get\_ipython().system(u'tar -zcvf diabetis.tgz diabetis.h5')

import watson\_machine\_learning\_client

from watson\_machine\_learning\_client import WatsonMachineLearningAPIClient

wml\_credentials={

"url": "https://eu-gb.ml.cloud.ibm.com",

"access\_key": "aVov8zRtQvhqt7XSI8kHaXasilZ9BV6qBYakIQEdsK\_J",

"username": "9cbf15c0-f92f-4b7d-80bc-495d1b3aba9d",

"password": "93a602d5-2ee2-4d0c-8f95-418df79afa11",

"instance\_id": "5fec57aa-11be-4553-a608-3a58894ebc51"

}

client = WatsonMachineLearningAPIClient( wml\_credentials )

metadata = {

client.repository.ModelMetaNames.NAME: "keras model",

client.repository.ModelMetaNames.FRAMEWORK\_NAME: "tensorflow",

client.repository.ModelMetaNames.FRAMEWORK\_LIBRARIES:[{'name':'keras','version':'2.1.3'}],

client.repository.ModelMetaNames.FRAMEWORK\_VERSION: "1.5"

}

model\_details = client.repository.store\_model( model="diabetis.tgz", meta\_props=metadata )

model\_id = model\_details["metadata"]["guid"]

model\_id

model\_deployment\_details = client.deployments.create( artifact\_uid=model\_id, name="deployment" )

scoring\_endpoint = client.deployments.get\_scoring\_url(model\_deployment\_details)

scoring\_endpoint

REFERENCES

[1]. Dr. D. Ashok Kumar and R. Govindasamy," Performance and Evaluation of Classification Data Mining Techniques in Diabetes", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6, 1312-1319, 2015.

[2]. Mukesh kumari, Dr. Rajan Vohra ,Anshul arora," Prediction of Diabetes Using Bayesian Network", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178. [3]. V.Karthikeyan, I.Parvin Begum,K.Tajudin,I.shahina Begam,"comparative of data minning classification algorithm(CDMCA)in diabetes disease prediction",International journal of computer applications(0975-8887) Volume 60No.12,December 2012.

[4]. Mr. Chintan Shah, Dr. Anjali G. Jivani," Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE – 31661

[5]. Divya Tomar and Sonali Agarwal ,"A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and BioTechnology Vol.5, No.5, pp. 241-266,2013.