International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018) pp. 59-65

Health care System: Stream Machine Learning Classifier for features Prediction in Diabetes Therapy

D.Ramana kumar[1]
Associate professor, Department of Information Technology, Vardhaman College of Engineering, Hyderabad, India.

Orcid Id: 0000-0003-1199-0700

Dr. S.Krishnamohan Rao[2]
Professor and Principal, Gandhi Institute for Technology, Bhubaneswar, Orissa, India.

Orcid Id: 0000-0003-2878-0339

Abstract

Among the multiple diseases, the diabetes is the most deadly disease in the world because it is closely associated with other kind of diseases such as kidney diseases; blindness, and heart attacks etc., Data mining techniques are used for variety of applications like health care domain which is the most considerable factor in human life. There are many classifications of methods which analyze the medical data to find the features and take the decisions on its own based on the given data. The real time medical decision support system with stream data classifier is needed to analyze medical data streams and make real time predictions. Many algorithms have been involved that tackle the problem of classification on evolving real time data streams.

In this paper, an attempt is made to develop a novel approach to find the appropriate solution using probabilistic and Machine Learning models which have the ability to predict whether the patient is having diabetes or not. Predicting the disease in early stages leads in treating the patient before it becomes critical. The proposed model has ability to extract hidden knowledge from a huge amount of diabetes-related data - collected from Web services data repository. The evaluation experiment gives real time blood glucose level, which is, predicted on various lifetime events and it intakes insulin and measures from dynamic scenarios such as class - boosted tree algorithm as well as regularization, proportion of blood glucose levels shows diabetes positives, which are correctly predicted to 90% accurate using prediction function. Else the true negative rate, which measures the proportion of negatives that are correctly, identified percentage of NO diabetes.

Keywords: Streaming data, boosted tree, Machine learning, regularization and prediction function.

INTRODUCTION

Diabetes mellitus is a sort of metabolic diseases in which a man has high blood sugar. There are 2 general reasons behind diabetes: 1-the body does not deliver enough insulin. Just 5-10% of individuals with diabetes

have this type of the disease (type 1) [1]. With the assistance of insulin treatment and different medicines, even youthful youngsters with type1 diabetes can figure out how to deal with their condition and live sound. 2-cells don't react to the insulin that is produced (type2).

Insulin is the key hormone that manages take-up of glucose from the blood into most cells (fundamentally muscle and fat cells, yet not central nervous system cells). In this way lack of insulin or the cold- heartedness of its receptors assumes a focal part in all types of diabetes mellitus.

People are fit for processing a few sugars, specifically those most regular in nourishment; starch, and a few disaccharides, for example, sucrose, are changed over inside a couple of hours to more straight forward structures most quite the monosaccharide glucose, the vital starch vitality source utilized by the body.

On the off chance that the measure of insulin accessible is inadequate, if cells react ineffectively to the impacts of insulin (insulin insensitivity or resistance), or if the insulin itself is flawed, at that point glucose won't have its typical impact so glucose won't be ingested legitimately by those body cells that require it nor will it be put away suitably in the liver and muscles. The net impact is persistent high amounts of blood glucose, poor protein amalgamation, and other metabolic disturbances, for example, acidosis.

The dataset considered in the trial has a place with Pima Indian dataset gathered from UCI Machine Learning Repository comprising of 768 perceptions with 9 factors. In [3] Li et al., 2016 built up a customized approach that thinks about the one of the unique social, social and statistic characteristics for American Indians.

Predictivefunction $= \sum_{i=1}^{n} (\hat{y}) + \sum_{i}^{t} \Omega(f)$ ---------- 1

Where **L** is returns the Training loss function, **Ω** is the regularization, **i** is number of trees, **f** is a sample function in functional space and **t.n.** are maximum number of trees, $\hat{y}_t$ is the prediction tree from the initial trees from 1 to t values.

As it is very difficult to compute all the trees at a time to predict the diabetes from all the data attributes of the trees, we use a two class boosted tree algorithm to overcome this problem as follows.

$\hat{y}_i^{(0)} = 0$ // Initial tree

$\hat{y}_i^{(1)} = f(x_1) = \hat{y}_i^{(0)} + f(x_i)$ 1 //new tree is added to the existing tree.

$\hat{y}_i^{(2)} = f(x_1) + f(x_i) = \hat{y}_i^{(1)} + f(x_i)$ 2 //new tree is added to the existing tree.

$\hat{y}_i^{(3)} = f(x_1) + f(x_i) + f(x_i) = \hat{y}_i^{(2)} + f(x_i)$ 3 //new tree is added to the existing tree.

.....

$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x) = \hat{y}_i^{(t-1)} + f_t(x_i)$ t //new tree is added to the existing tree.

Optimized function from the equation 1 will be

Predictivefunction at stage t $= \sum_{i=1}^{n} (\hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f) + constant$

To model this machine-learning algorithm we use **Microsoft Azure machine-learning Studio**. This studio is a powerful visual drag and drop-authoring environment where no coding is necessary. It is a simple Browser-based environment and there is no need of any individual platforms and software.

LITERATURESURVEY

Proposes the Stark Assessment of Lifestyle diabetic treatment using a regression-based data mining technique on the existing datasets of NCD from UCI repository [2] which consists of only few attributes to predict the diabetes on limited data like digestive ophthalmology and density diagnosis prediction ratio is 45%, 30% 18%,10% 4%.[3] Litinskaia *et al.*, 2017 aims to improving accuracy of non-invasive blood glucose detection via optical glucometer by more than 2 times by providing a feedback between glucometer and insulin pump and developed mathematical model of blood glucose dynamics on the basis of blood glucose prediction algorithm.

Type 1 diabetes mellitus (T1DM) itself comprises of roughly around 10% of the reported cases with diabetes. T1DM usually affects children where the causes are of the disease is unknown and no known ways to prevent the T1DM. The study [4] performed and published by Jane L. Chiang et al, estimated about 80000 children showing symptoms and developing the disease each year. The motivation behind pursuing a working model in the field of T1DM is to guide the doctors and the guardians of the affected children to catch the disease in ascent stage and based on the severity of diabetes; accordingly provide treatment to the affected children.

Hyperglycemia precipitates micro and/or macro vascular complications including: retinopathy, nephropathy, neuropathy, cardiovascular diseases, peripheral vascular diseases and stroke [5]. Insulin and oral hypoglycemic agents are used in the management of type 1 and type 2 diabetes mellitus respectively [1]. This high level of blood sugar produces the symptoms of polyphagia (increased hunger), polydipsia (increased thirstiness) and polyuria (frequent urination). Type 2 diabetes mellitus (T2D) is a complex disease of major public health importance. Its incidence is rapidly increasing in the developed countries. It is estimated that by the year 2030, there will be ~366 million people affected by Type 2 diabetes (T2D) worldwide [6]. The Multi-Stream Dependency Detection (MSDD) algorithm was used on two-thirds of the dataset for training. They did not consider removing the noise data and not deleting any missing values by applying preprocess techniques. Based on the few parameters measured accuracy on the one- third for evaluation was 71.33% [8]. Michie et al. used 22 algorithms with 10 to12-fold cross validation algorithms and presented the following accuracy rates of the test case data sets: Discrim 77.5%, Quaddisc 73.8%, Logdisc 77.7%, SMART 76.8%, ALLOC80 69.9%, k-NN 67.6%, CASTLE 74.2%, CART74.5%, IndCART 72.9%, NewID 71.1%, AC2 72.4%, Baytree 72.9%, NaiveBay 73.8%, CN2 71.1%,C4.5 73%, Itrule 75.5%, Cal5 75%, Kohonen 72.7%, DIPOL92 77.6%, Backprop 75.2%, RBF 75.7%,and LVQ 72.8% [7].

METHODOLOGY

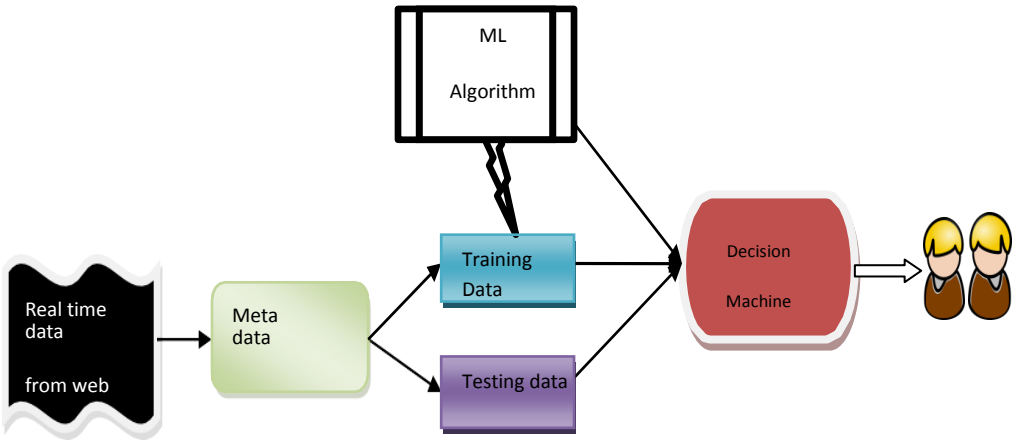

Figure 1: Block diagram of Machine learning classification on diabetes patient

By using Microsoft Azure Machine learning studio, we can depict the diabetes levels and predict the diabetes by considering various factors such as Age, Sex, Nature of work, skin fold thickness, No of times pregnant, Plasma glucose, Diastolic BP, Insulin, Body Mass Index, Diabetes Pedigree Function, Food habits, Detection.

To check diabetes every time patient has to go to the diagnostic centre, give his/ her blood and wait one day for the result, which takes a very long time and also a waste of money, as it needs to be checked regularly. As previously, there are many statistical methods, which detect diabetes, but in few cases if the patient has given the same sample in two different diagnostic centers, no patient as got the exact diabetes number, because of their approaches in detecting in them. The exact standard values based on the medical strategies shown in table- 1.

Table-1: Diabetes Dataset Range Description

Name of the Data Attribute	**Range (Min - Max)**	**Description**
No Of Times Pregnant	0 - 17	Pregnancy frequency
Serum Insulin	0 - 846	2 Hour Serum Insulin values obtained
Tri- Skin-Fold thickness	0 - 99	Thickness Fold of Triceps Skin
Diastolic BP	0 - 122	Diastolic Blood Pressure
Detection Anomaly	0 - 1	Positive and Negative
Plasma Glucose levels	0 - 199	Tolerance values of Oral Glucose Test
AGE	21 - 81	Age
Diabetes pedigree	0.078-2.42	Functions of Diabetes Pedigree
BMI	0 - 67.1	Body Mass Index

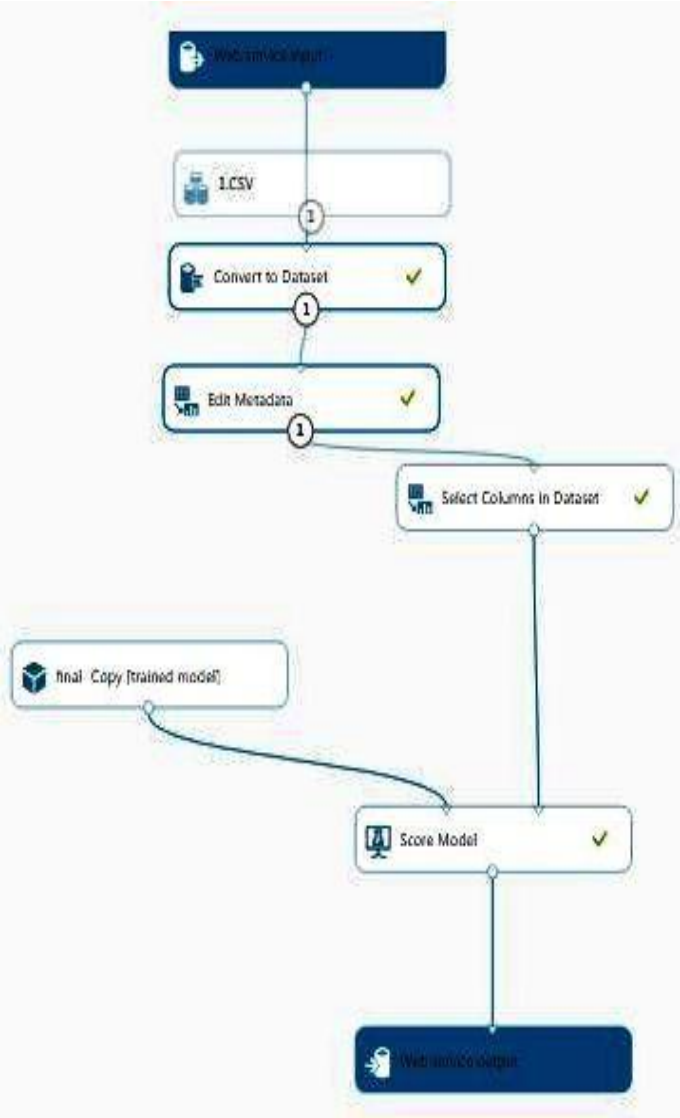

Figure 2(a): Azure MLStudio Model

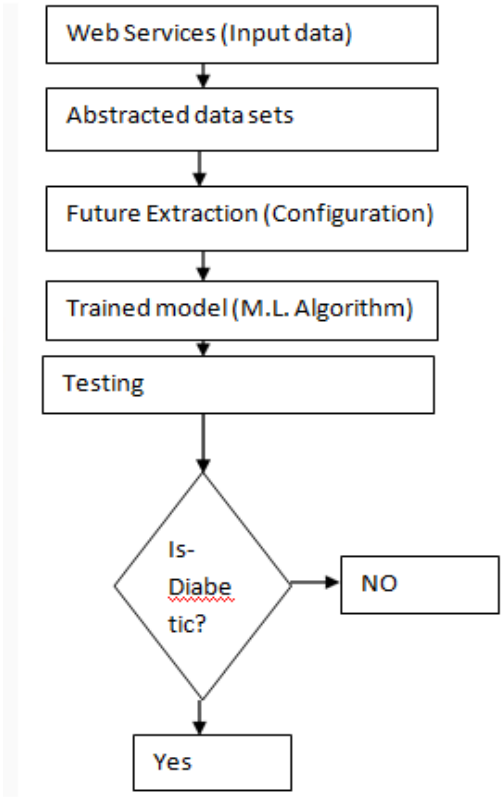

Figure 2(b): Conventional method

Figure-2. Data Flow Diagram for Machine Learning to predict the Diabetes based on all features.

By using Microsoft Azure Machine learning studio (Fig 2 (a , b)), we can depict the diabetes levels and predict the diabetes by considering various factors such as Age, Sex, Nature of work, skin fold thickness, No of times pregnant, Plasma glucose, Diastolic BP, Insulin, Body Mass Index, Diabetes Pedigree Function, Food habits, Detection.

To check diabetes every time patient has to go to the diagnostic centre, give his/ her blood and wait one day for the result, which takes a very long time and also a waste of money, as it needs to be checked regularly. As previously, there are many statistical methods, which detect diabetes, but in few cases if the patient has given the same sample in two different diagnostic centers, no patient as got the exact diabetes number, because of their approaches detecting in them.

The main aim of my paper is to develop a system, which can predict the diabetes of a patient with zero error and with higher accuracy by considering all the factors that cause diabetes.

INPUT STACK:

Web service streams input:

We follow a novel approach to predict the diabetes. We consider web services to extract the data based on which the diabetes has to be predicted, as it is a very huge diabetes database collected from the UCI data repository of the patient, which is called web services input. It just collects all the information required for the detection anomaly. By using the data mining techniques it extracts the data from the database and it will be sent for the processing of the data.

MACHINE LEARNING AND KNOWLEDGE DISCOVERY FROM DATABASE:

Machine learning is the logical field managing the courses in which machines gain as a matter of fact. For some researchers, the expression "machine learning" is indistinguishable to the expression "artificial intelligence", given that the likelihood of learning is the primary entity for a substance called intelligence in broadest way of the word. The reason for machine learning is the development of PC frameworks that can adjust and gain from their experience [9]. A more point by point and formal meaning of machine learning is given by Mitchel [10]: A computer program is said to gain as a matter of fact E concerning some class of undertakings T and performance measure P, if its performance at execution in T, as measured by P, enhances with encounter E. Knowledge discovery in databases (KDD) is a field enveloping speculations, strategies and methods, attempting to understand information and concentrate valuable information from them. It is thought to be a multistep process (choice, preprocess, change, determining, interpretation and evaluation) depicted in Fig. 1 [11]. The most vital advance in the whole KDD process is data mining, optimizing the application of machine learning algorithms in breaking down data. A total meaning of KDD is given by Fayyad et al. [11]: KDD is the nontrivial procedure distinguishing legitimate, novel, possibly helpful, and eventually reasonable examples in data.

ABSTRACTION OF DATA:

Here the data will be made separate from the collected data and made into table on all the factors that are associated for the diabetes and the statistical approach to depict the data in graphs for the comparisons. This is sent to the actual model.

ACTUAL MODEL OF PREDICTION:

Edit Metadata:

Here the actual data i.e., 1 complete data is split into 0.7 and 0.3 data sets. Then 0.3 data is considered as the trained real time data, which is kept undisturbed, as it will be the data samples for the actual model. It will be forwarded for the two input Score model as the 30% of original norms. The remaining 0.7 of the dataset is considered as the training data where it will be undergone with the two-boosted decision tree algorithm.

TESTING REAL DATA:

This data is 30% of the total data from the datasets and it will be of the original norms. This will be kept undisturbed and waits until train model is created in the Microsoft Azure

machine-learning algorithm, where it takes class boosted algorithm as the appropriate one.

TRAINING DATA:

From the complete datasets, 70% of the total data will be undergone with the two class boosted algorithm where it will check all the possibilities to get the proper prediction of the data samples from the web services input.

TWO-BOOSTED DECISION TREE ALGORITHM

To create a machine-learning model, we can also use the **Two Boosted Decision Tree** algorithm. A two class boosted decision tree algorithm is an ensemble learning method where the second tree corrects the errors of the first tree, the errors of second and first tree will be corrected by third and so forth. Predictions are made based on all ensembles of trees together, which makes the prediction.

Generally, when properly configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks. However, they are also one of the more memory-intensive learners, and the current implementation holds everything in memory; therefore, a boosted decision tree model might not be able to process the very large datasets that some linear students can handle. To train a boosted decision tree model, we must provide multiple data instances to check the accuracy.

In General, two-class boosted tree yields better outcomes when highlights are to some degree related. On the off chance that highlights have an extensive level of entropy i.e., if they are not related, they share practically no common data, and requesting them in a tree won't yield a considerable measure of predictive significance.

TRAIN MODEL:

A new model will be generated with the two class boosted machine learning algorithm, which can be formed as the train data. It will become the modified data sample, where this model undergoes with lot of variations to predict the diabetes by considering the raw data from the training data and undergone

PREDICTION

Score Model:

In this model both the samples from the train model and the testing real data (which is 30% of the original datasets) will be collected and undergone into several comparisons to predict the actual observations. But the data here will be in an inappropriate way and it will be sent to evaluate model to get the actual predictions.

EVOLUTION MODEL:

Here the data from the score model will be made to run and check the comparison with the actual model. The datasets can be compared with any combination of ensemble trees and can get the actual prediction. The exact prediction can be done to get the result. This method will get the accuracy of the data in prediction, as it undergoes with all the data samples of the patient, on which appropriate decisions can be made.

KNOWLEDGE BASE

Final Result:

Here final result can be evaluated from the evolution model by using the data mining technique and the data will be collected and stored in it as the prediction result. This will give only final result either yes or no to the patient data. It is the most appropriate data where original decision can be made to the data.

REPORTING TO THE USER:

Here without wasting the time of the patient the predicted data will be sent to the patient immediately to his mobile as a message, as we are using web services input, which is attached to the UCI data repository, all the information of the patient will be observed in it. The predicted numbers of the patient and the final result of diabetes will be sent to the patient.

RESULT AND DISCUSSION

By Microsoft Azure machine learning studio, we are going to get the final result based on the predictive analysis. After constructing the machine learning model, the patient data from the UCI database repository will be sent to the Azure studio with the help of the web services input**.** Then actual process starts. Here we have to run the model and select the appropriate datasets available with in the columns as shown in fig 3. Then it collects all the data of the each dataset of the patient. The columns contains Age, Gender, BMI, BP, weight, height, lifestyle, genetic history, frequent urination, drying of mouth, increased hunger, blurry vision, Fatigue, frequent infections, obesity and final result diabetes as shown in fig 3. of the datasets we can select any number of data sets to which the data is available to predict the diabetes.

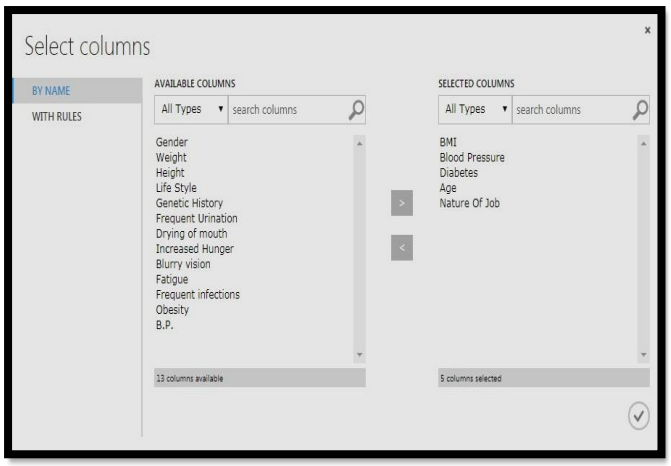

Figure 3: shows datasets selection in Azure ML Studio

GRAPHS:

After the selection of the datasets, click on the visualize on the scope of the model in the train data. Then it shows the graphs of the datasets by selecting one at a time as shown in fig 4. Then this model continuous by selecting two at time and comparisons can be made based on the simulations of the appropriate datasets. Here are some of the dataset's graphs on the patient's data based on the trained data of the diabetes with some other classifications.

Azure studio uses graphs, crosstabs, multi box plot, Histograms and scatter plot to depict the observations and comparisons. If it is single dataset, its values will be shown in Histogram. If it is a comparison the datasets will be shown in Crosstabs, multi box plot or scatter plot, where it shows based on the number of selection of the data samples. The accuracy of diabetes prediction is round to 6 decimals (0.673079 is standard) . If diabetic value is more than standard value then the patient is diabetic. The probability of accuracy in diabetes prediction is 0.9 (90%). In fig 5 it shows the graph between Age and diabetic scored probability.

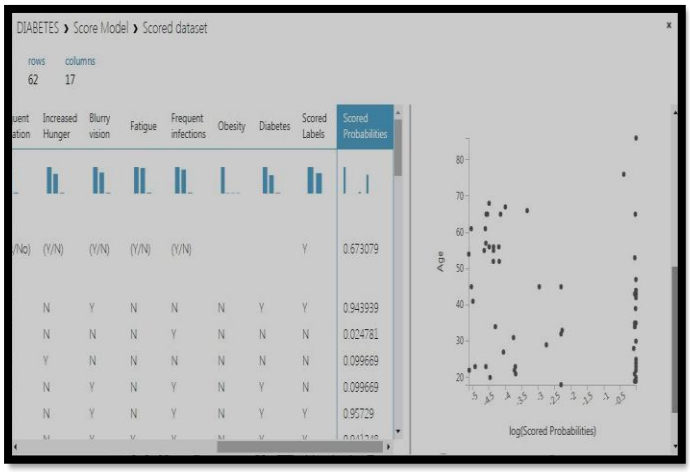

Figure 5: shows the datasets comparison with age and scored probabilities in scatter plot

CONCLUSION AND FUTURESCOPE

To get the appropriate predictions we use two class-boosted algorithms where data samples are checked continuously and rectified errors based on the final conclusions of each stage. It is developed with the concentration of systems optimization and machine's principles. The main aim of this model is to push the extreme of the computational limits of machines to provide scalability, portability and accuracy of the predicted data. This model ensures to rectify all the errors and final decisions are made accurate. In this model, by comparing all the data aspects of the decision trees, it makes the decisions and can extend optimization techniques on same studio with same data or extended data from various data streams of machine learning techniques.

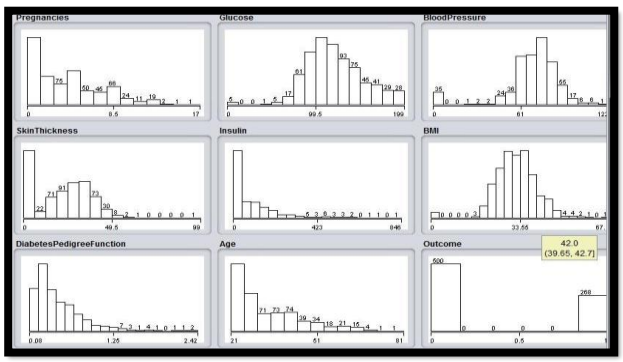

Figure 4: Graphs of considering few datasets with diabetes

REFERENCES

[1] American Diabetes Association Reports of the Experts Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes care, 2001:23:S4-19

[2] M. Sharma, G. Singh, R. Singh, Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques, In IRBM, Volume 38, Issue 6, 2017, Pages 305-324, ISSN 1959-0318.

[3] E. L. Litinskaia, N. A. Bazaev, K. V. Pozhar and V. M. Grinvald, "Methods for improving accuracy of non-invasive blood glucose detection via optical glucometer", 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, (2017), pp. 47- 49.

[4] R. Zolfaghari, "Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM", International Journal of Computational Engineering & Management, vol. 15, no. 4, (2012), pp.2230-7893

[5] Danny Meetoo, Peter McGovern, ReemaSafadi ―An epidemiological overview of diabetes across the world‖ British Journal of Nursing, 2007, pp. 1002 - 1007

[6] Derosa G ―Efficacy and tolerability of pioglitazone in patients with type 2 diabetes mellitus: comparison with other oral anti hyperglycaemic agents‖ Drugs. 2010. - pp 1945-1961.

[7] Michie, D., D. J. Spiegelhalter, et al., ―Machine learning, neural and statistical classification‖, New York, Ellis Horwood, 1994.

[8] Oates, T., ―MSDD as a Tool for Classification‖, EKSL Memorandum, Department of Computer Science, University of Massachusetts at Amherst, 1994.

[9] Wilson RA, KeilFC.The MIT encyclopaedia of the cognitive sciences.MIT Press; 1999.

[10] Mitchell T. Machine learning. McGraw Hill 0-07-042807-7; 1997 2.

[11] Fayyad U, Piatetsky-Shapiro G, Smyth P From data mining to knowledge discovery in databases. AI Mag 1996; 17:37–54.