National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN : 2456-3307

Performance Analysis of Data Mining Classification Algorithms to Predict Diabetes

Rakesh Singh Sambyal[1], Tanzeela Javid[2], Abhinav Bansal[3]

[1]Department of Information Technology and Engineering, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir, India
[2]Department of Computer Science and Engineering, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir, India
[3]Department Of Computer Science and Engineering, PEC university Of Technology, Chandigarh, India
rssambyal@bgsbu.ac.in[1],tanzeela.javid@yahoo.com[2], abhinavbansal19961996@gmail.com[3]

ABSTRACT

Data mining refers to non-trivial extraction of valid, implicit, novel, potentially useful and ultimately understandable information patterns of data from enormous volumes of data. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. One of the most important application of data mining is in disease prediction. In this paper we present a classification model developed using cloud platform Microsoft Azure that predicts the occurrence of Diabetes in an individual on the basis of non-pathological parameters – age, gender, family history of being diabetic, smoking and drinking habits, frequency of thirst and urination, weight height and fatigue. Six different algorithms have been compared among which the model created using "Two-Class Neural Network Algorithm" has the highest accuracy of 98.3% and hence has been deployed as a web service. Finally, a GUI is been developed in python to access the web service.

Keywords: Data Mining, Diabetes, Classification and Prediction, Neural Networks, Microsoft Azure, Python diagnosis System.

I. INTRODUCTION

Diabetes is a chronic disease that contributes to a significant portion of the healthcare expenditure for a nation as individuals with diabetes need continuous medical care [4]. In order to prevent or delay the onset of diabetes, it is necessary to identify high risk populations and introduce behaviour modifications as early as possible. One of the most accurate tests of diabetes is through the analysis of fasting blood sugar, but it is invasive and costly. Furthermore, it is only useful when the individual is already displaying symptoms i.e., making a diagnosis, which is considered too late to be an effective screening mechanism. Therefore, a reliable non-invasive inexpensive test to predict high risk individuals in advance is needed. Many researchers have worked and developed systems that predict diabetes on the basis of pathological parameters that are usually obtained by conducting various clinical tests. None of the system predicts the occurrence of diabetes by using non-pathological attributes like – age, gender, family history of being diabetic, smoking, frequency of thirst and urination, weight, height and fatigue.

II. DATA MINING

Traditional data analysis techniques fail to extract information out of large data tombs. Data mining is a blend of traditional data analysis methods with sophisticated algorithms for processing huge volumes of raw data [2]. With the advent of low cost storage devices and automation of almost all fields of life, huge amounts of data is being collected with every passing day. Data mining refers to non-trivial extraction of valid, implicit, novel, potentially useful and ultimately understandable information patterns of data from enormous volumes of data [1]. This enormous data is referred as raw data .The raw material is the transactional data while data mining algorithm is the excavator that extracts potentially valuable information from it [5]. The extracted knowledge can provide insight to the data miners as well as be invaluable in tasks like decision making predictions and strategic planning [6]. Data mining can be viewed as a result of the natural evolution of information technology.

In medical field, data mining techniques can be used by the researchers for the diagnosis and prediction of various diseases [7]. Data mining techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy [3]. Data extracted is a key resource to be analyzed for knowledge extraction that enables support for cost-savings, better medical care, prediction of diseases, medical diagnosis, image analysis, drug development and decision making.

III. METHODOLOGY

A. Database Collection

In this work we use classification to mine dataset gathered from various sources including different laboratories and hospitals in Jammu and Srinagar during the year 2010 with 986 records.

These records are cleansed to remove the noise and unwanted data, thus resulting in 388 of total records. Diabetes file is an XLS file that consists of 11 attributes viz; age, gender, family history, water intake, urination, smoking, drinking, height, weight, fatigue and the 11th attributes is a class attribute. Since diabetes is a lifestyle disease thus we have selected those factors that have a huge effect on the lifestyle of an individual. These factors are verified from the endocrinologist for their huge effect on diabetes. The given dataset concerns classification into diabetic or non-diabetic person.

Data Representation

Number of records: 388.

Number of attributes: 10 and a class attribute.

Class 0: Non- diabetic.

Class 1: Diabetic.

Table I Data Representation of the Diabetes Dataset

Attribute	*Description*
1. Age	Age in years
2. Gender	Male : 1/ Female : 0
3. Family History	Yes : 1 /No : 0
4. Smoking	Yes : 1 /No : 0
5. Alcoholic	Yes : 1/ No : 0
6. Water Intake	Number of time (eg. 1,2,3...) / day
7. Urination	Number of time (eg.;1,2,3...) / day
8. Height	Height in centimeters
9. Weight	Weight in kilograms
10.Fatigue	Yes : 1 / No : 0

The classification process is carried out with soft computing technique known as Artificial Neural Networks and five other algorithms. The various factors are transformed in numeric format suitable for the experimentation. The code representing these attributes are given in the table I.

B. Tools Used

The tools that are used in this experimentation of disease prediction are as follows:

1) Microsoft Azure

Azure is a complete cloud platform that can host existing application infrastructure, provide computer-based services tailored for the application development needs, or even augment on-premises applications. Azure integrates the cloud services that need to develop, test, deploy, and manage applications —while taking advantage of the efficiencies of cloud computing.

Machine learning is a technique of data science that helps computers learn from existing data in order to forecast future behaviours, outcomes, and trends. Azure Machine Learning is a cloud predictive analytics service that makes it possible to quickly create and deploy predictive models as analytics solutions.

Predictive analytics uses algorithms that analyze historical or current data to identify patterns or trends in order to forecast future events.

2) Python 3.6.1

Python is a general-purpose interpreted, object-oriented, and high-level programming language. Guido van Rossum released it in 1991. It is an open source technology i.e. there is no cost of procurement. It can create different types of apps like Console-based, Web-based, Desktop-based etc.

C. Training Predictive Model

The complete process of designing the expert system is divided into two into two parts viz:

- Training a model
- Creating an Application

The training process is carried out using various algorithms and the model with highest accuracy is deployed as an application. Dataset is imported to the Microsoft Azure and is stored at the cloud storage. Then six experiments are carried out using six different algorithms.

The result of the best algorithm is analysed by confusion matrix and ROC curve.

- *Confusion matrix*

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning [8].

- *ROC curve*

The name ROC stands for *Receiver Operating Characteristic* [9]. An ROC curve shows the trade-off between the true positive rate or sensitivity and the false-positive rate for a given model [9]. The area under the ROC curve is a measure of the accuracy of the model.

D. Algorithms Used

Azure Machine Learning Studio provides many different state-of-the art machine learning algorithms to help build analytical models.

1) Two-Class Neural Network

One of the foremost vital models in Artificial Neural

Network is Multilayer Perceptron (MLP) [10]. A neural network is a set of interconnected layers, in which the inputs lead to outputs by a series of weighted edges and nodes. The input layer receives signals from external nodes [11]. Artificial neural networks provide a powerful tool to help doctors analyze, model, and make sense of complex clinical data across a broad range of medical applications [12-18]

To compute the output of the network for any given input, a value is calculated for each node in the hidden layers and in the output layer. For each node, the value is set applying an activation function to that weighted sum. A Neural Network (NN) consists of many Processing Elements (PEs), and weighted interconnections among the PEs. [19]

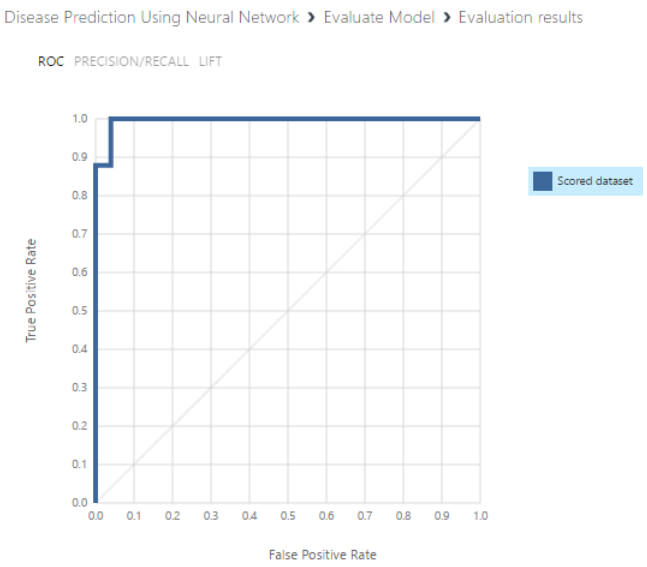


Figure 1. ROC Curve of NN Predictive Model

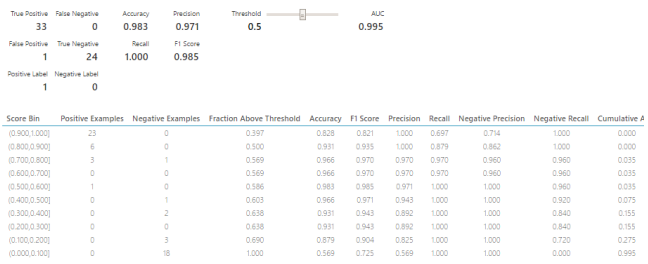


Figure 2. Various Parameters of **NN** Predictive Model

2) Class Support Vector Machine

Support vector machines (SVMs) are supervised learning models that analyze data and recognize patterns. They can be used for classification and regression tasks. Given a set of training samples labelled as belonging to one of two classes, the SVM algorithm assigns new samples into one category or the other. The samples are represented as points in space, and they are mapped so that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Support vector machines are among the earliest of machine learning algorithms. Although recent research has developed algorithms that have higher accuracy, this algorithm can work well on simple data sets when your goal is speed over accuracy.

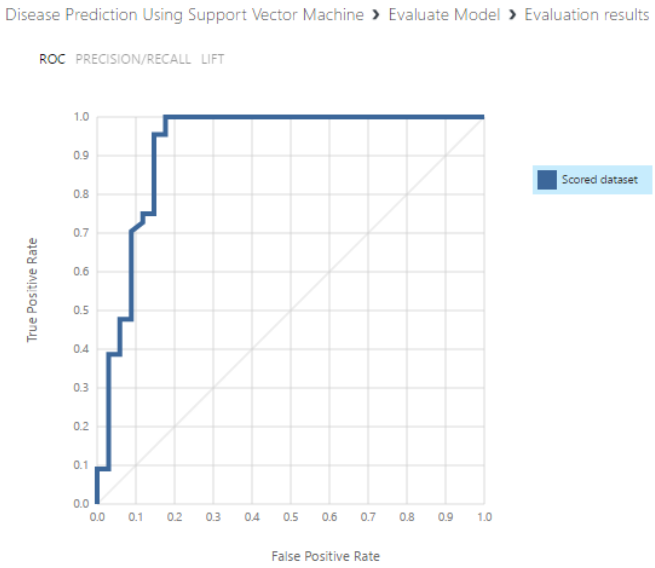


Figure 3. ROC Curve of Support Vector Machine Model

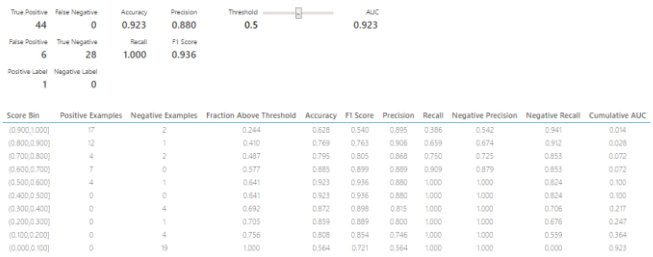
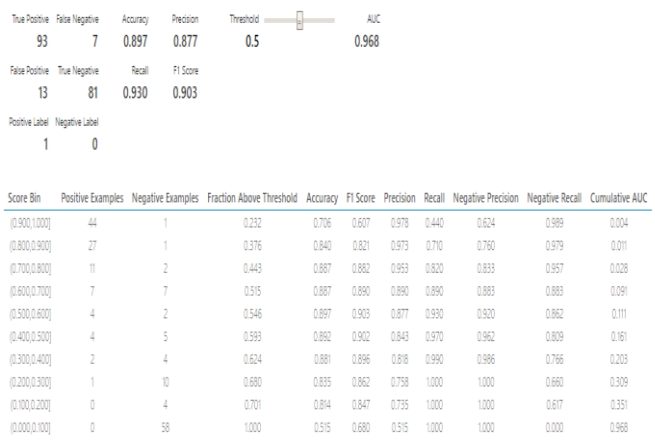


Figure 4. Various Parameters of Support Vector Machine Model

3) Two-Class Logistic Regression

Logistic regression is a well-known method in statistics that is used to predict the probability of an outcome, and is especially popular for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function. In this module, the classification algorithm is optimized for dichotomous or binary variables.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
93	7	0.897	0.877	0.5	0.968

False Positive	True Negative	Recall	F1 Score
13	81	0.930	0.903

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	44	1	0.232	0.706	0.607	0.978	0.440	0.624	0.989	0.004
(0.800,0.900]	27	1	0.376	0.840	0.821	0.973	0.710	0.760	0.979	0.011
(0.700,0.800]	11	2	0.443	0.887	0.862	0.953	0.820	0.833	0.957	0.026
(0.600,0.700]	7	7	0.515	0.887	0.890	0.890	0.890	0.883	0.883	0.091
(0.500,0.600]	4	2	0.546	0.897	0.903	0.877	0.930	0.920	0.862	0.111
(0.400,0.500]	4	5	0.593	0.892	0.902	0.843	0.970	0.962	0.809	0.161
(0.300,0.400]	2	4	0.624	0.881	0.896	0.818	0.990	0.986	0.766	0.203
(0.200,0.300]	1	10	0.680	0.835	0.862	0.758	1.000	1.000	0.660	0.309
(0.100,0.200]	0	4	0.701	0.814	0.847	0.735	1.000	1.000	0.617	0.351
(0.000,0.100]	0	58	1.000	0.515	0.680	0.515	1.000	1.000	0.000	0.968

Figure 5. Various Parameters of Logistic Regression Model

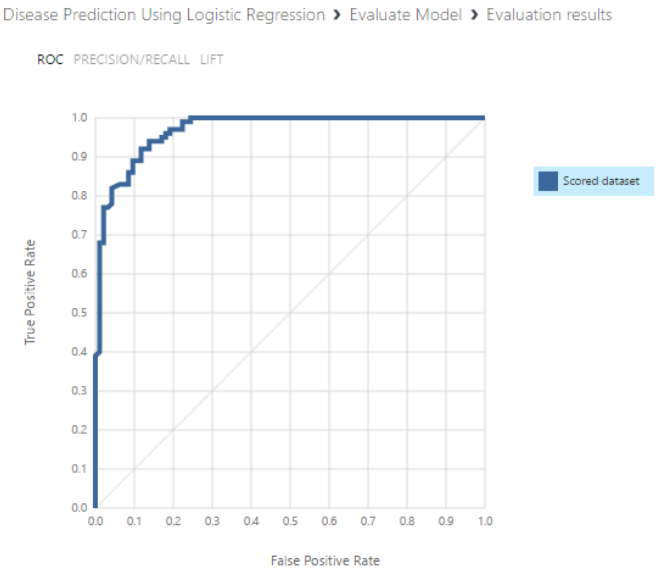


Figure 6. ROC Curve of Logistic Regression Model

4) Two-Class Averaged Perceptron

The averaged Perceptron method is an early and very simple version of a neural network. In this supervised learning method, inputs are classified into several possible outputs based on a linear function, and then combined with a set of weights that are derived from the feature vector—hence the name "Perceptron."

Perceptrons are faster, and because they process cases serially, Perceptrons can be used with continuous training.

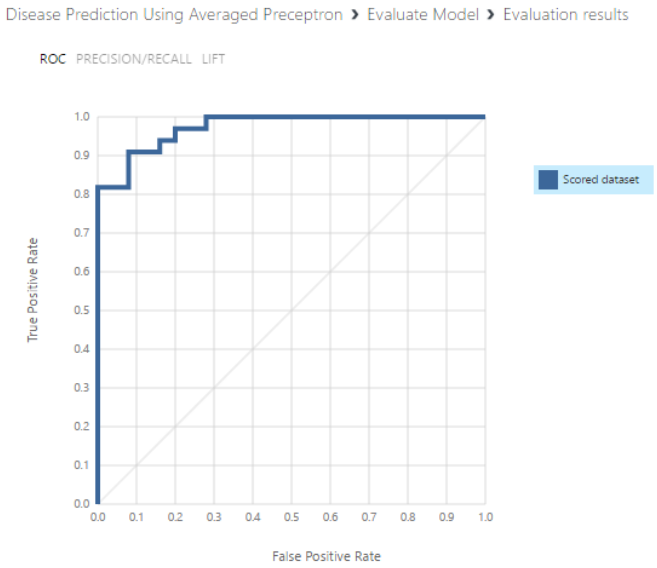
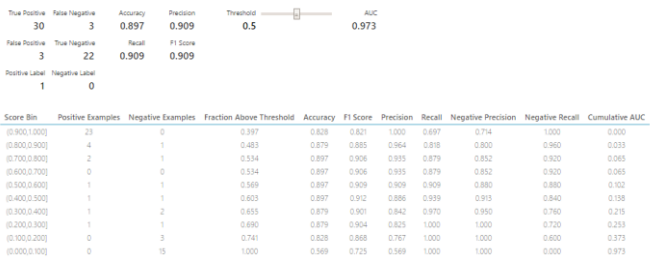


Figure 7. ROC Curve of Averaged Perceptron Model



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
30	3	0.897	0.909	0.5	0.973

False Positive	True Negative	Recall	F1 Score
3	22	0.909	0.909

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	23	0	0.397	0.828	0.821	1.000	0.697	0.714	1.000	0.000
(0.800,0.900]	4	1	0.483	0.879	0.885	0.964	0.818	0.800	0.960	0.033
(0.700,0.800]	2	1	0.534	0.897	0.906	0.935	0.879	0.852	0.920	0.065
(0.600,0.700]	0	0	0.534	0.897	0.906	0.935	0.879	0.852	0.920	0.065
(0.500,0.600]	1	1	0.569	0.897	0.909	0.909	0.909	0.880	0.880	0.102
(0.400,0.500]	1	1	0.603	0.897	0.912	0.886	0.939	0.913	0.840	0.158
(0.300,0.400]	1	2	0.655	0.879	0.901	0.842	0.970	0.950	0.760	0.215
(0.200,0.300]	1	1	0.690	0.879	0.904	0.825	1.000	1.000	0.720	0.253
(0.100,0.200]	0	3	0.741	0.828	0.868	0.767	1.000	1.000	0.600	0.373
(0.000,0.100]	0	15	1.000	0.569	0.725	0.569	1.000	1.000	0.000	0.973

Figure 8. Various Parameters of Averaged Perceptron Model

5) Two-Class Boosted Decision Tree

Two-Class Boosted Decision Tree module to create a machine learning model is based on the boosted decision trees algorithm. A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. The decision tree approach is one of the most powerful techniques in data mining [20].

Classification is a supervised learning method, and therefore requires a tagged dataset, which includes a label column.

We trained the model by providing the boosted decision tree model and the tagged dataset as an

input to Train Model. The trained model is used to predict values for the new input samples.

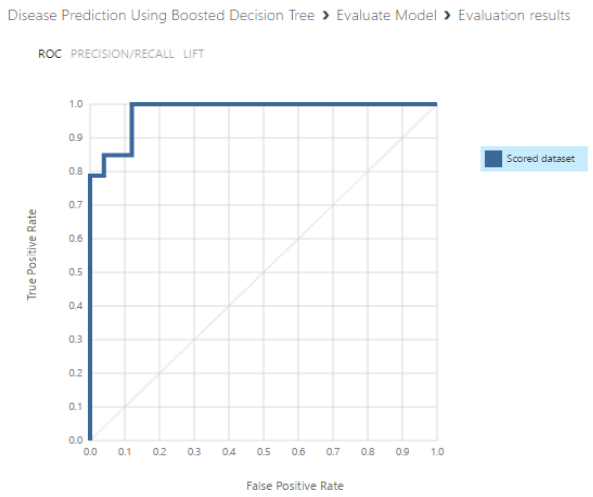


Figure 9. ROC Curve of Boosted Decision Tree Model

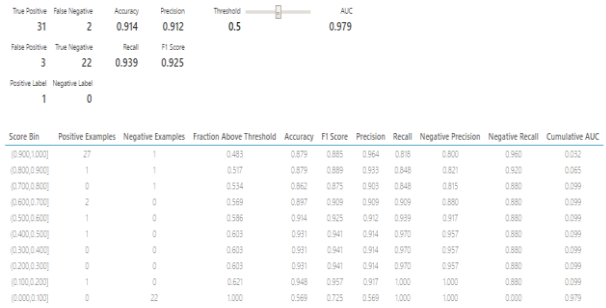


Figure 10. Various Parameters of Boosted Decision Tree Model

6) Two-Class Decision Forest

The decision forest algorithm is an ensemble learning method for classification. The algorithm works by building multiple decision trees and then voting on the most popular output class. Voting is a form of aggregation, in which each tree in a classification decision forest outputs a non-normalized frequency histogram of labels. The trees that have high prediction confidence will have a greater weight in the final decision of the ensemble.

Decision trees are non-parametric models, and they support data with varied distributions. In each tree, a sequence of simple tests is run for each class, increasing the levels of a tree structure until a leaf node (decision) is reached.

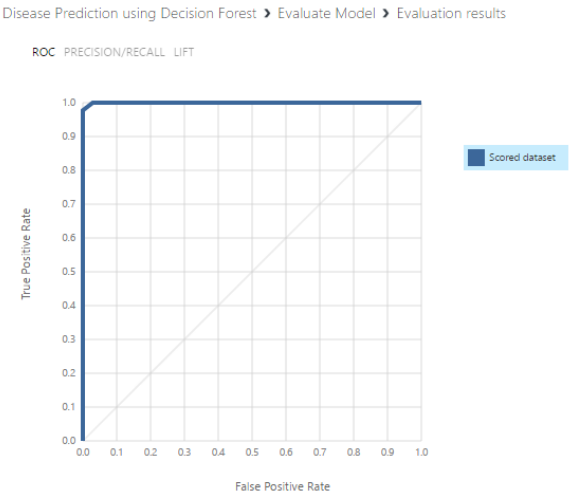


Figure 11. ROC Curve of Decision Forest Model

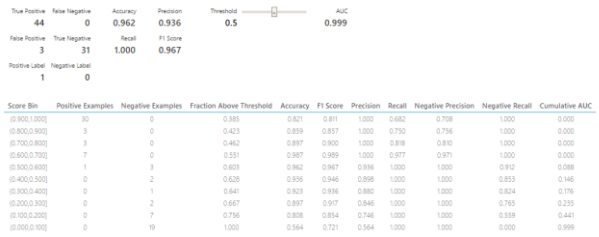


Figure 12. Various Parameters of Decision Forest Model

E. Comparison Of All Algorithms

The comparison between all the six algorithms used in the above experimentation is shown below:

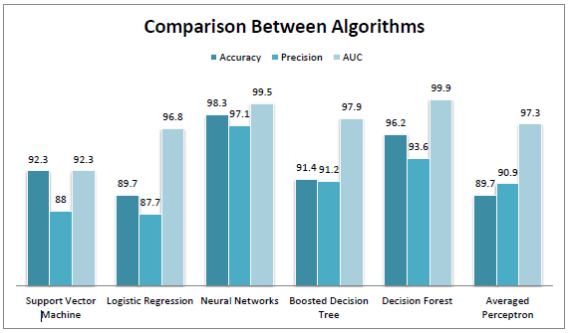


Figure 13. Comparison Between Algorithms

Based on the accuracy of the models, the model created using two -class neural network algorithm

has the highest accuracy of 98.3% and hence has been selected for deployment.

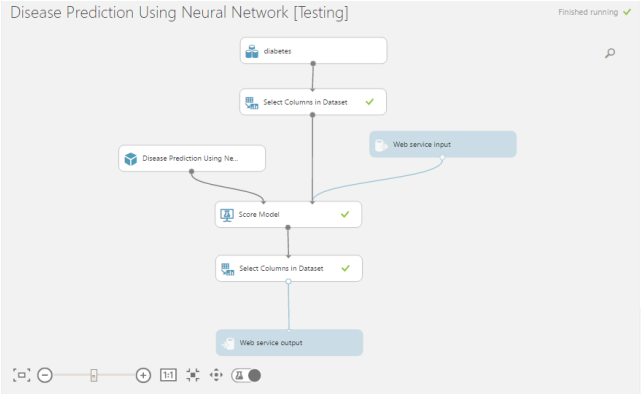


Figure 14. Neural Network Prediction model for deployment

IV. EXPERT SYSTEM DEVELOPMENT

An application has been designed for testing the real time data using Python GUI. When deployed as a Web service, Azure Machine Learning experiments provide a REST API and JSON formatted messages that can be consumed by a wide range of devices and platforms. The Web services are called by the code in Python.

Interface Module

Figure 15. Design of Interface Module

An easy way to comply with the conference paper formatting requirements is to use this document, as a template need to refer to an Internet email address or URL in your paper; you must type out the address or URL fully in Regular font.

V. CONCLUSION

Data mining techniques applied to the diabetes data set has resulted in the development of an intelligent system that can predict diabetes prior to the clinical tests that can save time and energy of people and can help them take precautionary measures if susceptible to the disease. Six different algorithms have been used to calculate the accuracy, out of which two class neural network algorithm has the highest accuracy of 98.3 % as specified in the bar chart in fig 16. The system has been trained and tested in Microsoft azure an online platform and finally the intelligent system created has been deployed as a web service using the python language. By creating this model patient can monitor his health on his own and plan preventive measures and treatment at the early stages of the diseases.

VI.REFERENCES

[1] Fayyad, U., Shapiro, G. P., Smyth, P., and Uthurusamy R., (1996d) "Advances in Knowledge Discovery and Data Mining". AAAI/MIT Press, 1996.

[2] Malik, M. B., Ghazi, M. A., Ali, R. (2012), "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on computer and Communication Technology 2012, Allahabad, India.

[3] Syed Umar Amin, Kavita Agarwal Dr. Rizwan Beg, (2013). "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors". Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)

[4] Phattharat Songthung and Kunwadee Sripanidkulchai, (2016), "Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification". 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 13-15 July 2016 Khon Kaen, Thailand.

[5] Cavoukian A., (1997), Information and Privacy Commissioner, Ontario, "Data Mining Staking a Claim on Your Privacy", www.ipc.on.ca

[6] Divanis, G. A. and Verikios, S. V. (2010), "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine 2010.

[7] Ammar Asjad Raja, Madiha Guftar, Madiha Guftar, Tamim Ahmed Khan, and Dominik Greibl, (2016). "Intelligent Syncope Disease Prediction Framework using DM-Ensemble Techniques". FTC 2016 - Future Technologies Conference 2016, San Francisco, United States.

[8] https://en.wikipedia.org/wiki/Confusion_matrix.

[9] Data Mining: Concepts Methodologies, Tools and Applications Volume 1 Edited By Management Association, Information.

[10] Girija D.K., Dr. M.S. Shashidhara, and M. Giri, (2013), "Data mining approach for prediction of fibroid Disease using Neural Networks". 2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA) 10-12 October 2013 Bangalore, India .

[11] Fundamentals of Neural Networks: Arquitectures, Algorithms, and Applications – Laurene Fausett.

[12] W. G. Baxt, (1990) "Use of an artificial neural network for data analysis in clinical decisionmaking: The diagnosis of acute coronary occlusion," Neural Comput., vol. 2, pp. 480–489..

[13] Dr. A. Kandaswamy, (1997) "Applications of Artificial Neural Networks in Bio Medical Engineering", The Institute of Electronics and Telecommunicatio Engineers, Proceedings of the Zonal Seminar on Neural Networks, Nov 20-21.

[14] Scales, R., & Embrechts, M., (2002) "Computational Intelligence Techniques for Medical Diagnostic", Proceedings of Walter Lincoln Hawkins, Graduate Research Conference.

[15] S. Moein, S. A. Monadjemi and P. Moallem, (2009) "A Novel Fuzzy-Neural Based Medical Diagnosis System", International Journal of Biological & Medical Sciences, Vol.4, No.3, pp. 146-150.

[16] D Gil, M Johnsson, JM Garcia Chamizo, (2009) , "Application of artificial neural networks in the diagnosis of urological dysfunctions", Expert Systems with Applications Volume 36, Issue 3, Part 2, Pages 5754-5760, Elsevier.

[17] Hasan Temurtas, Nejat Yumusak, Feyzullah Temurtas, (2009)" A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with Applications: An International Journal , Volume 36 Issue 4.

[18] S. M. Kamruzzaman , Md. Monirul Islam, (2006)" An Algorithm to Extract Rules from Artificial Neural Networks for Medical Diagnosis Problems", International Journal of Information Technology, Vol. 12 No. 8.

[19] Dr. K. Usha Rani (2011), "Analysis Of Heart Diseases Dataset Using Neural Network Approach". International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011.

[20] Monika Gandhi, and Dr. Shailendra Narayan Singh, (2015). "Predictions in Heart Disease Using Techniques of Data Mining". 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015), Noida, India.