Advances and Applications in Mathematical Sciences
Volume 20, Issue 11, September 2021, Pages 2735-2747
© 2021 Mili Publications

APPLICATION DEVELOPMENT OF DIABETES MELLITUS PREDICTION USING MACHINE LEARNING APPROACH

VIPUL MISTRY, MADHAVI DESAI and KARAN BHATT

E and C Engg. Department
SNPITRC, Umrakh, India

CSE Department
RNGPIT, Bardoli, India

Computer Engg. Department
VGEC Chandkheda, India
Gujarat Technological University
E-mail: vipul.h.mistry@snpitrc.ac.in
desaimadhavi30@gmail.com
kpbhatt@vgecg.ac.in

Abstract

World Health Organization (WHO) has mentioned that by the year 2040, there will be one person in every ten people who will be suffering from diabetes. Diabetes is responsible for many other health related complications. This has motivated us to develop an application that can predict diabetes mellitus from various health parameters of a person. This paper develops a machine learning based classification model to predict diabetes. Pima Indian Diabetes Database has been used to compare various machine learning models and best model is selected for the development of a user interface application based on Accuracy, Recall and AUC. This paper compares performance of KNN, SVM, Logistic Regression, Gradient Boost Classifier, Decision Tree, CHAID Tree and Random Forest algorithms. IBM cloud services are used to develop a Logistic Regression classification model based user interface application which can predict whether patient is having diabetes or not with more than 82% Accuracy and Recall.

I. Introduction

Machine learning in healthcare domain is thriving research field with the

2010 Mathematics Subject Classification: 60G25.

Keywords: diabetes prediction, machine learning, SVM, logistic regression, XGBoost, healthcare, classifier, dataset.

Received October 13, 2020; Accepted November 5, 2020

advancements in technology and increase in day-to-day data. There is utmost necessity of advancements in medical domain by understanding the gathered data of patients. Diabetes mellitus is a prolonged disease which occurs to persons of various ages and cut the human life at an early stage. It is a disease that results in high blood sugar due to which body cannot make enough insulin and use it effectively. There are various countries and medical organizations who are working on chronic disease control hence one can prevent early death of persons by applying prior treatments. Researchers are working with machine learning for early diagnosis of disease. There is no evidence that single machine learning method can perform best for various diseases. Performance of machine learning models will vary with various diseases. Diabetes can be divided in two major types. (i) Type 1 Diabetes is that our system is not producing insulin (ii) Type 2 diabetes system of patient doesn't respond to insulin and at later stage it may stop to produce insulin. Both type of diabetes will increase blood sugar level and dangerous disease which can shorten out human life. Both type of diabetes having some common symptoms like thrust for more frequent drinking water, frequent urination, hunger, blur vision and cuts on any organ that don't heal properly. Various parameters of human lifestyle can be affected through diabetes. Nutritional management is important part for the people who are having diabetes because it is important to keep it in stable range. It can affect various organs of body parts like kidney, eyes, heart, nerves etc.

Machine learning has applications in various fields like retail, banking, education, agriculture, health care etc. In machine learning based approach, a pre-trained model predicts the output based on features of an input instance. In health care domain with the machine learning we can do prediction of the disease at an early stage on the basis of symptoms of the patient. This paper compares different machine learning models based on their quantitative and qualitative results and recommend best model for the diabetes prediction. This best fitted model is used to develop a user interface application for the prediction of diabetes mellitus. The remaining part of the paper is organized as follows. Section II refers to the related work in this domain. Machine learning algorithms of supervised learning are explained in section III. Section IV describes the proposed methodology. Section V discusses experiment results and Section VI summarizes the paper.

II. Related Work

A good programmer can take the symptoms as an input and make a program to suggest medical treatments based on the symptoms input. However, the problem with simple programming is that with the time if symptoms get changed, there is a need to explicit programming. Compared to that, machine learning methods learn from the observations rather than explicitly programming. In prediction of diabetes, some of the input symptoms of patients like age, number of children, skin depth etc. are passed to machine learning model for training. Once the generalized model gets developed, it can classify unknown patient if patient has diabetes or not. In 2016, S. Joshi et al., [1] proposed to use back propagation neural network for detection of diabetes mellitus prediction. In 2016 P. Srikanth et al., [2] proposed critical study of classification algorithms for diabetes W. Xu et al., [3] proposed to use random forest model for risk prediction of type II diabetes in 2017. Author has proposed the model by analysing age, weight, waist, hip features of the patients. By experimental results author has proved that random forest ID3, Naïve Bayes and Adaboost algorithm are working good for the prediction of diabetes based on input health information. Role of machine learning in diabetes prediction was studied by He B. et al., [4] in 2018. Bani-Hani, D et al., [5] proposed regressive neural network on Pima Indian diabetes dataset in 2018. D. Sisodia et al., [6] performed experimental analysis of diabetes prediction with decision tree, SVM, Naïve Bayes machine learning model to detect at early stages. Author has proved that Naïve Bayes outperforms other models through Experimental results. This paper presents a comparative evaluation of various proposed machine learning algorithms on Pima Indian Diabetes Dataset [9] and finds the best model for the development of user interface application for diabetes mellitus prediction.

III. Machine Learning Models

Various machine learning models used in this paper are briefly discussed in this section.

A. Decision Tree Classifier. It is a well-developed machine learning classifier that has been used for wide range of applications. It is non-parametric machine learning model to partition datasets. This model converts large complex datasets in easy to understand graphical models. From this one can easily create set of rules for classification purpose. The main advantage of decision tree that this model has minimal requirement of data preparation and it gives robust performance even on large datasets [7]. The selection of roots and stems are depends on various attribute measures. [8]. CART, ID3 and C4.5 are three different decision tree classification algorithms with different attribute measures.

B. XGB Classifier. By using regular decision tree classification, we train a machine learning classifier on one database and afterwards trained classifier model can be used for classification. We are trying to optimize the parameters by using various attribute measures but at the end we are using single model. In contrast by using Boosting method one can get better performance because it takes an iterative approach. In this technique many models are combined together for better performance. In boosting, rather than training models individually, it trains a model in progression. Every single classifier get trained to rectify the errors which are made by an earlier classifier. It adds model in succession until no more changes are possible for the better performance. The positive impact of this gradient boosting classifier is that earlier classifier will focus on rectification of mistakes done by previous classifier.

C. KNN Classifier. K nearest neighbor classifier is a similarity or distance based classification method. This method predicts the data based on similarity with the trained data. The data with highest similarity are kept in one class and more dissimilarity will be kept in other class. Hence the key point of this classification model is distance between the data. Data instances which have smallest distance are known as neighbors. At time of testing when new data instance will arrive its distance will get measured with each instance of the trained model and placed into the class which have less distance and more similarity. For distance based classifier, the prediction for the new data is based on the mean or median of the new data values for the k nearest neighbors.

D. Linear SVM Classifier. Support Vector Machines (SVM) is a classification method that gives maximum detection accuracy. This model gives detection accuracy without over fitting the training data. It can also be used for regression. The case in which the data are not linearly not separable, SVM converts the data from lower dimension to higher dimension and hence it can be separable in higher dimensions. The principal of SVM is that it maps the data from lower dimensional data to high dimensional data hence it can categorized the data. This classifier finds the maximum marginal hyper plane (MMH) between the categories. By using this, the principal of MMH the new instance can be classified or predicted to the group into which it belongs. The Linear Support Vector Machine node uses a new algorithm with the feature space being the same as the input feature space that allows it to handle very large numbers of records and features, as well as to handle sparse data well.

E. CHAID Tree Classifier. Chi-squared Automatic Interaction Detection (CHAID) is a classification model for construction of decision trees by with chi-square data to recognize ideal splits. By internally discretizing or binning scale fields into ordinal categorical fields it can produce both classification and regression trees using input features of any measurement level.

F. Logistic Regression Classifier. To assign data instances to a discrete set of classes' logistic regression can be used. [9]. It recognizes the association between different attributes of the dataset. It returns probability value between 0 and 1 using logistic sigmoid function. If the value is less than or equal to 0.5 than it return 0 otherwise it returns 1. This classifier is better choice for binary classification. This classifier model focuses on getting best weights and regression coefficient.

G. Random Forest Classifier. It is ensemble classification model which trains numerous decision trees with bagging. It implements bootstrapping which is followed by aggregation. This method creates subsets of training datasets and trained individual decision trees on it parallel. In bootstrapping it make sure that every single decision tree classification model in the random forest is unique. By this way it reduces the overall variance of the Random Forest classifier. For the final decision this classifier aggregates the results of each individual decision tree. This model is having high generalization capability and less issues of over fitting.

IV. Methodology

The objective is to cultivate a model that predicts if a patient has diabetes or not based on the diagnostic measures of that person. This section presents the methodology used to identify the best machine learning algorithm for the prediction of diabetes. The performances of the machine learning algorithms are compared and best fitted model is used to build a user interface for the diabetes prediction system.

A. Diabetes Dataset. The dataset that has been used for the experiment is Pima Indian Diabetes Database [10]. All the patients in this dataset are at least 21 years old females who belong to Pima Indian heritage. This dataset is used as an input to evaluate the prediction model using various machine learning algorithms. The dataset contains total no of 768 patients' information. Including target class, total 9 attributes are present for each of these 768 patients. The classification is done as: diabetes positive and diabetes tested negative. There are total of 268 positive samples and 500 negative samples in the dataset. Table 1 describes the attributes listed in the dataset.

Table 1. Dataset attributes statistics and description.

Attribute	Numeric values (Min, Max)	STD	Description
Preg	0, 17	3.36	No of Pregnancies
Plas	0,199	31.97	Plasma glucose measured using oral glucose tolerance test
Pres	0,122	19.35	Blood pressure (mm Hg)
Skin	0,99	015.95	Triceps skin fold thickness (mm)
Test	0,846	115.2	Two hours serum insulin in
Mass	0.0,67.1	7.88	Body mass index
Pedi	0.078, 2.42	0.33	Probability of diabetes on the basis of family history
Age	21,81	11.76	Age of a person in years
Class	01	-	0 – Non diabetic person 1 – Diabetic person

Machine learning technique is used to build a classification model that predicts whether a person is diabetic or not based on the attributes provided as an input. Figure 1 shows the approach that is used to build the classification model for prediction of diabetes disease.

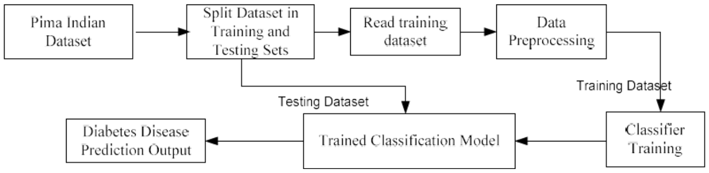


Figure 1. Diabetes prediction classification model.

B. Identification of the best Machine Learning Algorithm. The dataset used to build a model is first split in two parts: 80% is used for training and remaining 20% for testing. The training dataset instances are then given as an input to the training model. The attributes used for the model contains continuous type numeric values. The data are first given to the pre-processing stage. The pre-processed data along with the labels are used to train the model. The classifier model was built for various machine learning algorithms like decision tree, XGB classifier, gradient boosting classifier, logistic regression, random forest classifier, KNN and LSVM. Once the training is complete the testing dataset is given to the trained model for evaluation. The performance is evaluated using various metrics like confusion matrix, accuracy, precision, recall and ROC. These measures are used to select the best classification model that can predict diabetes mellitus of patients based on the given parameter about their health. In order to give the user interface to this model IBM cloud services are used. The trained model is deployed in the IBM cloud to get scoring endpoint which can be used as API in web app building. The model prediction needs to be showcased on user interface.

The diabetes prediction system uses following IBM cloud services [11] to develop user friendly diabetes prediction system.

1. IBM Watson Studio

2. IBM Watson Machine Learning

3. Node-RED

4. IBM Cloud Object Storage

The architecture of the entire system is shown in Figure 2. Node-RED is a browser based editor with a large number of nodes in the palette which can be easily connected and deployed in its run time. It is used for connecting hardware, APIs and online services. The architecture uses IBM Watson studio for the deployment of the final classification model for the user interface. The dataset is uploaded using cloud object storage. The dataset is used to train the model. Once the model is trained, the APIs are used in node-RED to make a user interface. The flow of the node-RED to interface model and user interface is shown in Figure 3.

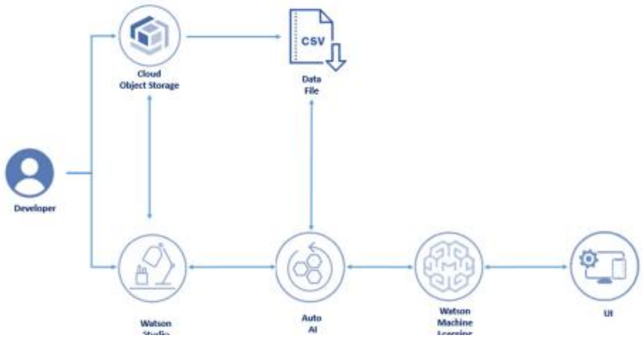


Figure 2. Diabetes Prediction system using IBM cloud services.

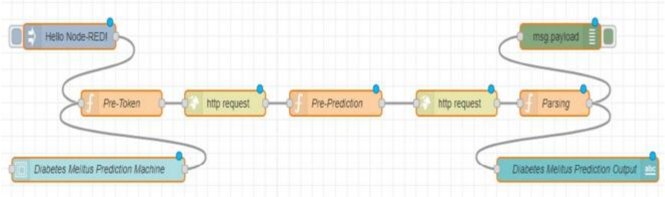


Figure 3. Node-RED flow to make user interface for diabetes mellitus prediction model.

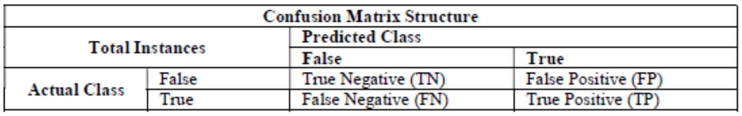
As shown in Figure 3 various nodes are used to link the trained model with the user inputs. The form node is used to accept the person's input health parameters. These inputs are linked with the model using function node and http request node. The model is accessible by API key. Function node is configured to link the trained machine learning model with the link. The outputs are passed to the link and output is displayed.

V. Experiments and Results

This section describes the experimental results which are obtained after training various machine learning algorithms. The results are compared for Linear Support Vector Machine, K-Nearest Neighbour, Decision Tree Classifier, Gradient Boosting Classifier, Random Forest, CHAID Tree and Logistic Regression. The purpose of this experiment is to identify the best machine learning model for prediction of diabetes mellitus of a person based on input health parameters.

A. Confusion Matrix. The confusion matrix or classification table contains a cross-classification of actual and predicted labels. The correct number of predictions is mentioned along the diagonal.

Table 2. Confusion matrix structure.



Confusion Matrix Structure			
Total Instances		Predicted Class	
		False	True
Actual Class	False	True Negative (TN)	False Positive (FP)
	True	False Negative (FN)	True Positive (TP)

True Positive (TP) means diabetes is correctly detected. False Positive (FP) means non-diabetes is predicted as diabetes. True Negative (TN) means non-diabetes is detected truly as non-diabetes and False Negative (FN) means diabetes is predicted as non-diabetes.

B. Accuracy. Accuracy is also used to optimize the model. Accuracy tells how correctly the trained model is predicting the diabetes mellitus based on the input attributes of the patient. It is the ratio of true detections to the total number of samples.

Accuracy = (TP+TN)/(TP+FP+TN+FN) (1)

The other evaluation measures used for the comparative analysis are as follows.

TP Rate = TP/(TP+FN) (2)

FP Rate = FP/(FP+TN) (3)

Precision = TP/(TP+FP) (4)

Recall = TP/(TP+FN) (5)

F_Measure = (2*Precision*Recall)/(Precision+Recall) (6)

Precision represents correctness over total positive detections by the model. Recall is used to measure how many out of total positive samples are detected by the model. F_measure is a weighted harmonic mean of the Precision and Recall. Many times it is used to indicate overall performance of the classification.

C. ROC Curve. The ROC (Receiver Operating Characteristic) is plotted between True Positive Rate (TPR) versus False Positive Rate (FPR). The plot is obtained by varying the threshold for positive classification over its probability range. The TPR is the total number of positives that are correctly classified which is also called sensitivity or recall. The FPR indicates total number of negative outcomes which are wrongly classified as positive. It is also known as one minus specificity (1-specificity) or False Discovery Rate (FDR). Figure 4 shows the ROC for Logistic Regression algorithm model.

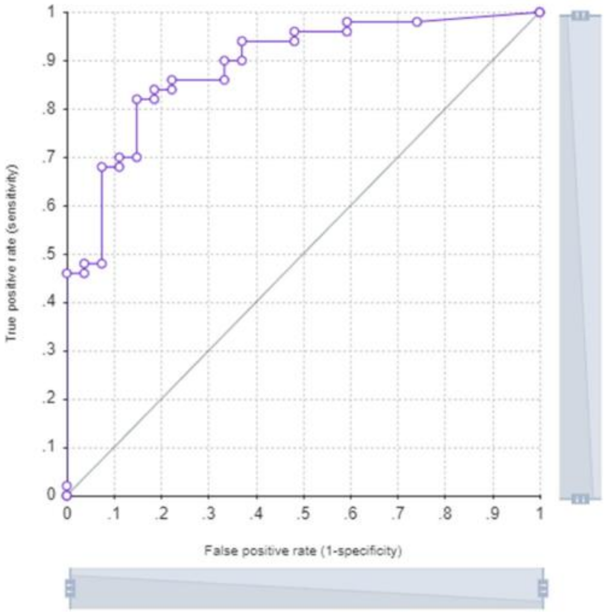


Figure 4. ROC for Logistic Regression algorithm.

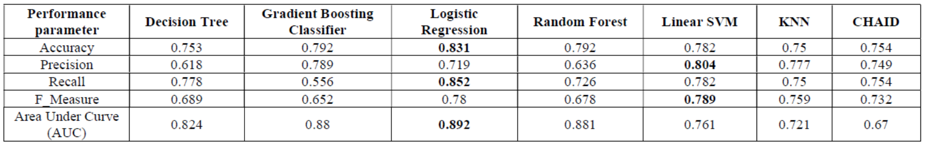
D. Precision Recall Curve

The Precision vs. Recall Curve chart plots the proportion of outcomes predicted to be positive that are positive, also known as precision, on the vertical axis, against the total number of positive outcomes which are correctly predicted, also known as recall, sensitivity or true positive rate (TPR), on the horizontal axis. The graph is plotted while changing the threshold for positive classification is varied across the predicted probability range from 1 down to 0. When the threshold is set high, few false positives will occur and precision will be high, while recall will be low. As the threshold is decreased, recall will increase and precision will generally decrease. Although there is generally a tradeoff between precision and recall, the curve may not be strictly monotonically decreasing. The area under the Prevision vs. Recall curve is generally preferred to the area under the ROC curve as an evaluation statistic for binary classification when the proportions of positive and negative observed instances are highly imbalanced.

E. Comparison of Machine Learning Methods

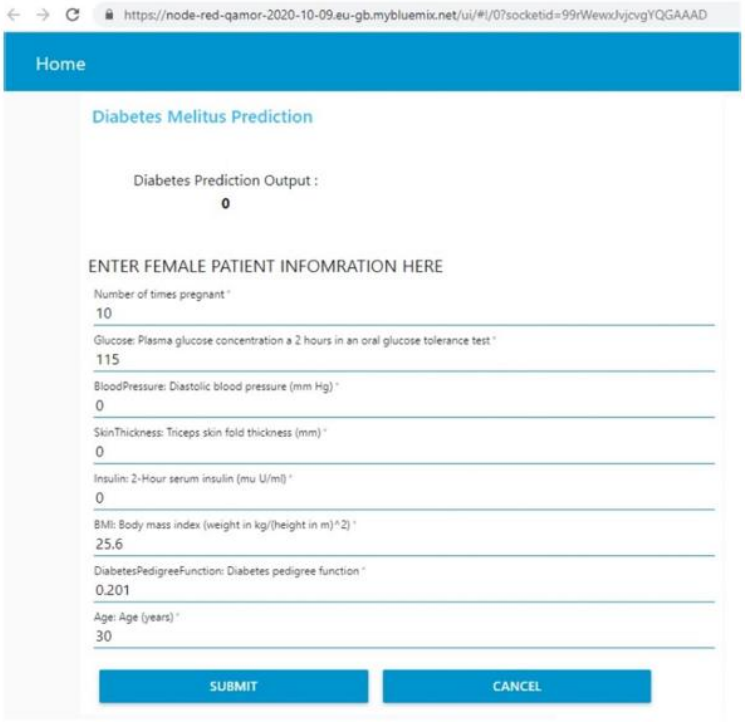
Table 3 shows the comparison between different algorithms based on different evaluation measures. The best machine learning classification model is selected based on Accuracy, Precision, Recall, F_Measure and AUC. As shown in Table 3, Logistic Regression outperforms other machine learning algorithms in terms of Accuray, Recall and AUC. Linear SVM is also providing good precision and F_Measure. However, considering Accuracy as the optimization parameter the Logistic Regression is selected as the final model for the development of user interface application for the diabetes mellitus prediction. Figure 5 shows the ROC for the Logistic Regression based classification model.

Table 3. Comparative analysis of machine learning models.



Performance parameter	Decision Tree	Gradient Boosting Classifier	Logistic Regression	Random Forest	Linear SVM	KNN	CHAID
Accuracy	0.753	0.792	**0.831**	0.792	0.782	0.75	0.754
Precision	0.618	0.789	0.719	0.636	**0.804**	0.777	0.749
Recall	0.778	0.556	**0.852**	0.726	0.782	0.75	0.754
F_Measure	0.689	0.652	0.78	0.678	**0.789**	0.759	0.732
Area Under Curve (AUC)	0.824	0.88	**0.892**	0.881	0.761	0.721	0.67

The trained logistic regression classification model for the prediction of diabetes mellitus is linked with user interface using node-RED and application page is generated. Figure 5 shows the application page for the output of the system.



https://node-red-qamor-2020-10-09.eu-gb.mybluemix.net/ui/#!/0?socketid=99rWewxJvjcvgYQGAAAD
Home
SUBMIT
CANCEL

Diabetes Melitus Prediction

Diabetes Prediction Output :
0

ENTER FEMALE PATIENT INFOMRATION HERE

Number of times pregnant *
10

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test *
115

BloodPressure: Diastolic blood pressure (mm Hg) *
0

SkinThickness: Triceps skin fold thickness (mm) *
0

Insulin: 2-Hour serum insulin (mu U/ml) *
0

BMI: Body mass index (weight in kg/(height in m)^2) *
25.6

DiabetesPedigreeFunction: Diabetes pedigree function *
0.201

Age: Age (years) *
30

Figure 5. User interface application for diabetes mellitus prediction using logistic regression based classification model.

VI. Conclusion

Diabetes is a prolonged disease which brings many other health related complications. According to medical study one person among every ten persons will be suffering from diabetes. This paper develops a user interface application for the prediction of diabetes mellitus. The machine learning model is designed using Pima Indian Diabetes Database. From the experiments it has been observed that Logistic Regression outperforms other machine learning algorithms and achieves best accuracy, Recall and AUC. Considering the severe health complications in diabetic persons it is very much significant to have high Recall rate along with good Accuracy. Logistic Regression model provides 83.1% Accuracy and 85.2% Recall when evaluated over 768 instances of Pima Indian Diabetes Database. The model is capable of an early stage prediction of diabetes mellitus which may help the patient to take precautionary steps and avoid further health complications.

References

[1] S. Joshi and M. Borse, Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network, in Proc. International Conference on Micro-Electronics and Telecommunication Engineering, Ghaziabad India, 22-23 Sept ,2016, pp. 110-113.

[2] P. Srikanth and D. Deverapalli, A Critical Study of Classification Algorithms Using Diabetes Diagnosis, in Proc. IEEE 6th International Conference on Advanced Computing , Bhimavaram India, 27-28 Feb ,2016, 245-249.

[3] W. Xu, J. Zhang, Q. Zhang, and X. Wei, Risk prediction of type II diabetes based on random forest model, in Proc International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics, Chennai India 27-28 Feb, (2017), 382-386.

[4] B. He, K. Shu and H. Zhang, Machine Learning and Data Mining in Diabetes Diagnosis and Treatment, (2018).

[5] D. Bani-Hani, P. Patel and T. Alshaikh, An Optimized Recursive General Regression Neural Network Oracle for the Prediction and Diagnosis of Diabetes, Global Journal of Computer Science and Technology (2018), 1-11.

[6] Deepti Sisodia and Dilip Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science 132 (2018), 1578-1585.

[7] Yang, Farid, Suzanne, Thornhill and Nina, Prediction of biopharmaceutical facility fit issues using decision tree analysis, Computer Aided Chemical Engineering 32 (2013), 61-66.

[8] R. Sahani, Rout C. Shatabdinalini, Badajena J. Chandrakanta, A. K. Jena and H. Das, Classification of Intrusion Detection Using Data Mining Techniques, In: P. Pattnaik, S. Rautaray, H. Das, J. Nayak (eds) Progress in Computing, Analytics and Networking, Advances in Intelligent Systems and Computing 710, Springer, Singapore, (2018).

[9] K. Shah, H. Patel, D. Sanghvi, et al., A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification, Augment Hum Res 5(12) (2020).

[10] Dataset link: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[11] IBM Cloud Services link: https:// www.cloud.ibm.com