ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 7, Issue 10, October 2020

Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach

[1] Prateek Mishra, [2] Dr.Anurag Sharma, [3]Dr.Abhishek Badholi
[1][2][3] Computer Science and Engineering, MATS University, Raipur,India

Abstract: Diabetes may be a major disorder which may affect entire body system adversely. Undiagnosed diabetes can increase the danger of cardiac stroke, diabetic nephropathy and other disorders. everywhere the planet many people are suffering from this disease. Early detection of diabetes is extremely important to take care of a healthy life. This disease may be a reason of worldwide concern because the cases of diabetes are rising rapidly. Machine learning (ML) may be a computational method for automatic learning from experience and improves the performance to form more accurate predictions. within the current research we've utilized machine learning technique in Pima Indian diabetes dataset to develop trends and detect patterns with risk factors using R data manipulation tool. To classify the patients into diabetic and non-diabetic we've developed and analyzed five different predictive models using R data manipulation tool. For this purpose, we used supervised machine learning algorithms namely linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, k-nearest neighbor (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR).

Keywords— Machine learning, Multifactor dimensionality reduction (MDR) Support vector machine (SVM), k-nearest neighbor (kNN), (ANN), Artificial neural network

I. INTRODUCTION

Diabetes may be a quite common metabolic disease. Usually onset of diabetes happens in time of life and sometimes in adulthood. But nowadays incidences of this disease are reported in children also. There are several factors for developing diabetes like genetic susceptibility, weight, food habit and sedentary lifestyle. Undiagnosed diabetes may end in very high blood glucose level referred as hyperglycemia which may cause complication like diabetic retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer. So, early detection of diabetes is extremely important to enhance quality of lifetime of patients and enhancement of their anticipation [1].Machine Learning cares with the event of algorithms and techniques that permits the computers to find out and gain intelligence supported the past experience. it's a branch of AI (AI) and is closely associated with statistics. By learning it means the system is in a position to spot and understand the input file, in order that it can make decisions and predictions supported it [2]. The learning process starts with the gathering of knowledge by different means, from various resources. Then subsequent step is to organize the info, that's pre-process it so as to repair the info related issues and to scale back the dimensionality of the space by removing the irrelevant data (or selecting the info of interest) [3]. Since the quantity of knowledge that's getting used for learning is large, it's difficult for the system to form decisions, so algorithms are designed using some logic, probability, statistics, control theory etc. to research the info and retrieve the knowledge from the past experiences [4]. Next step is testing the model to calculate the accuracy and performance of the system. and eventually, optimization of the system, i.e. improvising the model by using new rules or data set [5]. The techniques of machine learning are used for classification, prediction and pattern recognition. Machine learning are often applied in various areas like: program, website ranking, email filtering, face tagging and recognizing, related advertisements, character recognition, gaming, robotics, disease prediction and traffic management [6]. The essential learning process to develop a predictive model.

Now days, machine learning algorithms are used for automatic analysis of high dimensional biomedical data [7].Diagnosis of disease, skin lesions, cancer classification, risk assessment for disorder and analysis of genetic and genomic data are a number of the samples of biomedical application of ML [8,9]. For disease diagnosis. has successfully implemented SVM algorithm [10]. so as to diagnose major clinical depression (MDD) supported EEG dataset have used classification models like support vector machine (SVM), logistic regression (LR) and Naïve Bayesian (NB) [11]. Our novel model is implemented using supervised machine learning techniques in R for Pima Indian diabetes dataset to know patterns for knowledge discovery process in diabetes. This dataset discusses the Pima Indian population’s medical history regarding the onset of diabetes. It includes several independent variables and one variable class value of diabetes in terms of 0 and 1. during this work, we've studied performance of 5 different models based upon linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine (SVM-RBF), k-nearest neighbor (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR) algorithms to detect diabetes in female patients[12].

II. Related Material and Method

Dataset of female patients with minimum twenty-one-year age of Pima Indian population has been taken from UCI machine learning repository. This dataset is originally owned by the National institute of diabetes and digestive and kidney diseases. during this dataset there are total 768 instances classified into two classes: diabetic and non-diabetic with eight different risk factors: number of times pregnant, plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic vital sign , triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age[13]. We have investigated this diabetes dataset using powerful R data manipulation tool Feature engineering is a crucial step in applications of machine learning process. Modern data sets are described with many attributes for practical machine learning model building. Usually most of the attributes are irrelevant to the supervised machine learning classification. Preprocessing phase of the data involved feature selection, removal of outliers and k-NN imputation to predict the missing values [14]. There are various methods for handling the irrelevant and inconsistent data. during this work, we've selected the attributes containing the highly correlated data. This step is implemented by feature selection method which may be done by either 'manual method' or Boruta wrapper algorithm. Boruta package provides stable and unbiased selection of important features from an data system whereas manual method is error prone. So, feature selection has been through with the assistance of R package Boruta. the tactic is out there as an R package [15]. This package provides a convenient interface for machine learning algorithms. Boruta package is meant as a wrapper built around random forest classification algorithm implemented within the R. Boruta wrapper is run on the Pima Indian dataset with all the attributes and it yielded four attributes as important. With these attributes, the accuracy, precision and recall and other parameters are calculated [16].

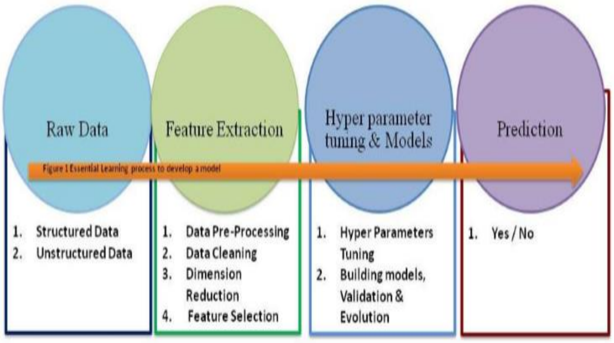


Figure1: Essential Learning process to develop a predictive model.

There are a couple of machine learning techniques which will be wont to implement the machine learning process. Learning techniques like supervised and unsupervised learning are most generally used. Supervised learning technique is employed when the historical data is out there for a particular problem. The system is trained with the inputs and respective responses then used for the prediction of the response of latest data [17]. Common supervised approaches include artificial neural network, back propagation, decision tree, support vector machines and Naïve Bayes classifier. Unsupervised learning technique is employed when the available training data is unlabeled. The system isn't given any prior information or training [18]. The algorithm has got to explore and identify the patterns from the available data so as to form decisions or predictions. Common unsupervised approaches include k-means clustering, hierarchical clustering, and principle component analysis and hidden-Markov model [19]. Supervised machine learning algorithms are selected to perform binary classification of diabetes dataset of Pima Indians. For predicting whether a patient is diabetic or not, we've used five different algorithms: linear kernel and radial basis function (RBF) kernel support vector machine (SVM), k-nearest neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR) in our machine learning predictive models which details are given below:

A. Support Vector Machine

Support vector machine (SVM) is employed in both classification and regression. In SVM model, the info points are represented on the space and are categorized into groups and therefore the points with similar properties falls in same group.

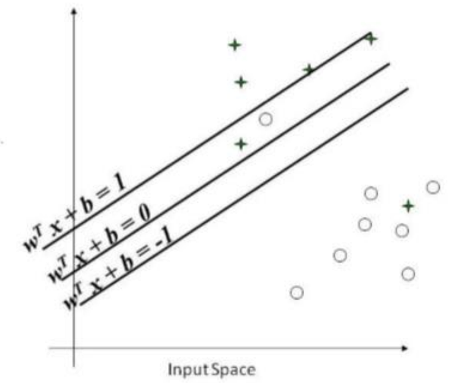


Figure 2: Representation of Support Vector Machine

In linear SVM the given data set is taken into account as p-dimensional vector which will be separated by maximum of p-1 planes called hyper-planes [20]. These planes separate the info space or set the boundaries among the info groups for classification or regression problems as in Figure 2. the simplest hyper-plane are often selected among the amount of hyper-planes on the idea of distance between the 2 classes it separates. The plane that has the utmost margin between the 2 classes is named the maximum-margin hyper-plane [21].

For n data points is defined as:

(X_1,Y_1)…….,(X_n,Y_n)…………………………….1

Where X_1 is real vector and Y_1 can be 1 or -1, representing the class to which X_1 belongs.

A hyper-plane can be constructed so as to maximize the distance between the two classes y=1 and y=-1, is defined as:

W. X- b = 0 …………………………………2

Where W is normal vector and b is offset of hyper-plane along

.

B. Radial Basis Function (RBF) Kernel Support Vector Machine

Support vector machine has proven its efficiency on linear data and nonlinear data. Radial base function has been implemented with this algorithm to classify nonlinear data [21].

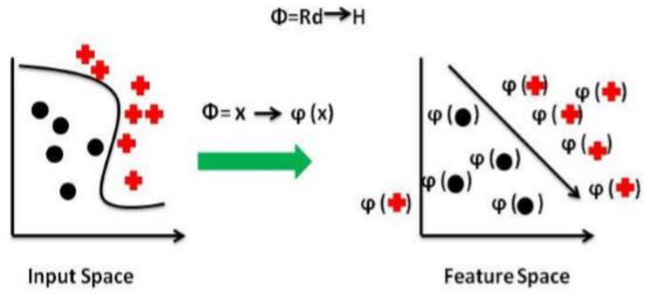


Figure 3: Representation of Radical Basis Function (RBF) Kernel support vector machine.

Kernel function plays very important role to put data into feature space. Mathematically, kernel trick (K) is defined as:

$$\mathrm{K}(x_1, x_2) = exp\left(-\frac{|x_1-x_2|^2}{2\sigma^2}\right)$$ ……….3

A Gaussian function is also known as Radial basis function (RBF) kernel. In Figure 3, the input space separated by feature map (Φ). By applying equation 1& 2 we get:

$$f(X) = \sum_i^N \alpha_i y_i k(X_i, X) + b$$ ………….4

By applying equation 3 in 4 we get new function, where N represents the trained data.

$$f(X) = \sum_i^N \alpha_i y_i exp\left(-\frac{|x_1-x_2|^2}{2\sigma^2}\right) + b$$ ………5

C. k-Nearest Neigh bour (k-NN)

k- Nearest neighbour may be a simple algorithm but yields excellent results. it's a lazy, nonparametric and instance-based learning algorithm. This algorithm are often utilized in both classification and regression problems. In classification, k-NN is applied to seek out out the category, to which new unlabeled object belongs. For this, a ‘k’ is set (where k is number of neighs bours to be considered) which is usually odd and therefore the distance between the info points that are nearest to the objects is calculated by the ways like Euclidean’s distance, Hamming distance, Manhattan distance or Minkowski distance. After calculating the space, ‘k’ nearest neighbours are selected the resultant class of the new object is calculated on the idea of the votes of the neighbours. The k-NN predicts the result with high accuracy [22].

D. Artificial neural network (ANN)

Artificial neural network mimics the functionality of human brain. It is often seen as a set of nodes called artificial neurons. All of those nodes can transmit information to at least one another. The neurons are often represented by some state (0 or 1) and every node can also have some weight assigned to them that defines its strength or importance within the system. The structure of ANN is split into layers of multiple nodes; the info travels from first layer (input layer) and after passing through middle layers (hidden layers) it reaches the output layer, every layer transforms the info into some relevant information and eventually gives the specified output [23].Transfer and activation functions play important role in functioning of neurons. The transfer function sums up all the weighted inputs as:

$$z = \sum_{x=1}^{n} w_i x_i \; + w_b b \qquad \ldots\ldots\ldots\ldots\ldots 6$$

Where b is bias value, which is usually 1.

The activation function basically flattens the output of the transfer function to a selected range. It might be either linear or nonlinear. the straightforward activation function is:

$$f(z) = z \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 7$$

Since this function does not provide any limits to the data, sigmoid function is used which can be expressed as:

$$a = \sigma(z) = \frac{1}{1+e^{-z}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

D. Multifactor Dimensionality Reduction (MDR)

Multifactor dimensionality reduction is an approach for locating and representing the consolidation of independent variables which can somehow influence the dependent variables. it's basically designed to hunt out the interactions between the variables which can affect the output of the system. It doesn't depend on parameters or the type of model getting used, which makes it better than the other traditional systems. It takes two or more attributes and converts it into one. This conversion changes the space representation of data. This results in improvement of the performance of system in predicting the category variable. Several extensions of MDR are utilized in machine learning. variety of them are fuzzy methods, odds ratio, risk scores, covariates and much more [24].

III. Predictive Model

In our proposed predictive model (Figure 4), we've done pre- processing of data and different feature engineering techniques to urge better results. Pre-processing involved removal of outliers and k-NN imputation to predict the missing values. Boruta wrapper algorithm is employed for feature selection because it provides unbiased selection of important features and unimportant features from a data system. Training of data after feature engineering features a significant role in supervised learning. we've used highly correlated variables for better outcomes [25]. input file, here indicates to check data used for predict and confusion matrix.

Early diagnosis of diabetes are often helpful to enhance the standard of lifetime of patients and enhancement of their anticipation. Supervised algorithms are wont to develop different models for diabetes detection. gives a view of the various machine learning models trained on Pima Indian diabetes dataset with optimized tuning parameters. All techniques of classification were experimented in “R” programming studio. the info set are partitioned into two parts (training and testing). We trained our model with 70% training data and tested with 30% remaining data. Five different models are developed using supervised learning to detect whether the patient is diabetic or nondiabetic. For this purpose, linear kernel support vector machine (SVM-linear), radial basis

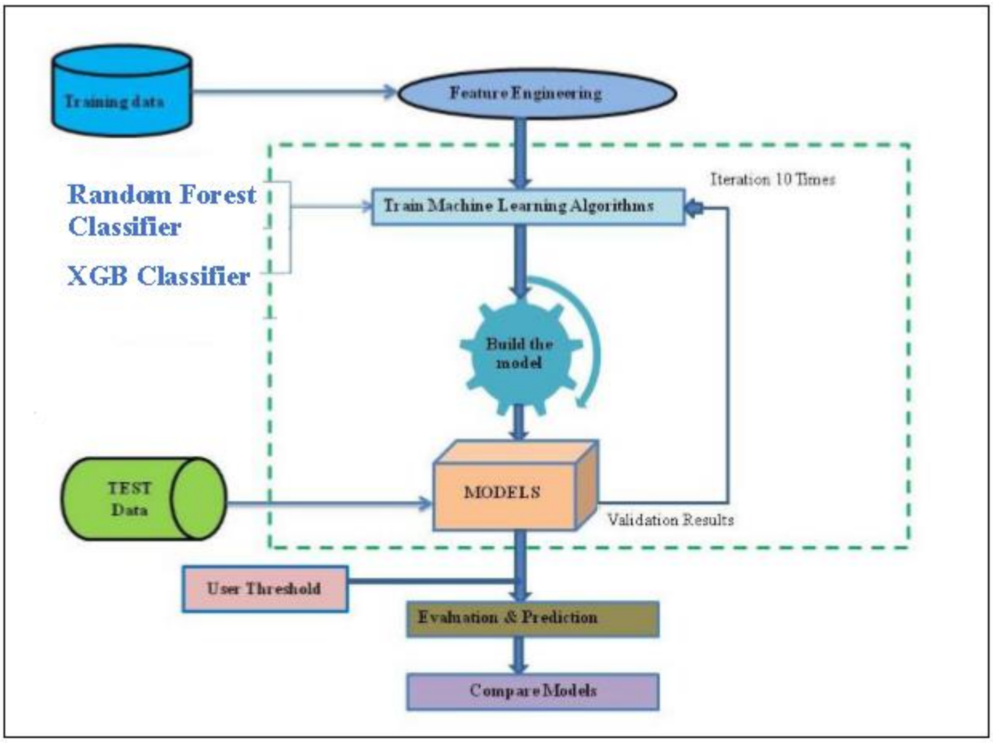


Figure 4: Framework for evaluating Predictive Model.

function (RBF) kernel support vector machine, k-NN, ANN and MDR algorithm are used. To diagnose diabetes for Pima Indian population, performance of all the five different models are evaluated upon parameters like precision, recall, area under curve (AUC) and F1 score. so on avoid problem of over fitting and under fitting, tenfold cross validation is completed. Accuracy indicates our classifier is how often correct in diagnosis of whether patient is diabetic or not. Precision has been used to determine classifier's ability provides correct positive predictions of diabetes. Recall or sensitivity is used in our work to hunt out the proportion of actual positive cases of diabetes correctly identified by the classifier used. Specificity is getting want to compute classifier's capability of determining negative cases of diabetes. because the weighted average of precision and recall provides F1 score so this score takes into account of both. The classifiers of F1 score near 1 are termed as best one [18]. Receiver operating characteristic (ROC) curve could also be a documented tool to ascertain performance of a binary classifier algorithm [19]. it's plot of true positive rate against false positive rate because the edge for assigning observations are varied to a selected class. Area under curve (AUC) value of a classifier may lie between 0.5 to1. Values below 0.50 indicated for a gaggle of random data which couldn't distinguish between true and false. An optimal classifier has value of area under the curve (AUC) near 1.0. If it's near 0.5 then this value is like random guessing [20].

From which represents different parameter for evaluating all the models, it's found that accuracy of linear kernel SVM model is 0.89. For radial basis function kernel SVM, accuracy is 0.84. For k-NN model accuracy is found to 0.88, while for ANN it's 0.86. Accuracy of MDR based model is found to be 0.83. Recall or sensitivity which indicates correctly identified proportion of actual positives diabetic cases for SVM-linear model is 0.87 and for SVM-RBF it's 0.83. For k-NN, ANN and MDR based models recall values are found to be 0.90, 0.88 and 0.87 respectively. Precision of SVM-linear, SVM-RBF, k-NN, ANN and MDR models is found to be 0.88, 0.85, 0.87, 0.85 and 0.82 respectively. F1 score of SVM-linear, SVM-RBF, k-NN ANN and MDR models is found to be 0.87, 0.83, 0.88, 0.86 and 0.84 respectively. we've calculated area under the curve (AUC) to measure performance of our models. it's found that AUC of SVM linear model is 0.90 while for SVM-RBF, k-NN, ANN and MDR model the values are respectively 0.85, 0.92 0.88 and 0.89. So, from above studies, it is often said that on the thought of all the parameters SVM-linear and k-NN are two best models to hunt out that whether patient is diabetic or not. Further it is often seen that accuracy and precision of SVM- linear model are higher as compared to k-NN model. But recall and F1 score of k-NN model are above SVM- linear model. If we examine our diabetic dataset carefully, it's found to be an example of imbalanced class with 500 negative instances and 268 positive instances giving an imbalance ratio of 1.87. Accuracy alone won't provide a very good indication of performance of a binary classifier just in case of imbalanced class. F1 score provides better insight into classifier performance just in case of uneven class distribution because it. provides balance between precision and recall [21, 25]. So, during this case F1 score should even be taken care of. Further it is often seen that AUC value of SVM-linear and k-NN model are 0.90 and 0.92 respectively.

IV. Patient demographics

The dataset has been taken from This dataset consisted of 768 female patients, a minimum of 21 years old of Pima Indian heritage, diabetes diagnoses (diabetic or control). there have been 268 cases of diabetic patients and 500 cases of control patients. This dataset contain 9 variables: (1) number of times pregnant, (2) plasma glucose concentration-a two hour in an oral glucose tolerance test, (3) diastolic vital sign (mm Hg), (4) triceps skin fold thickness (mm), (5) 2-hours serum insulin (mu U/ml), (6) body mass index (weight in kg/ (height in m)2), (7) diabetes pedigree function, (8) age (in years), (9) class variable (diabetic or control). during this dataset five patient have zero blood sugar level, diastolic vital sign is zero for 35 patients, 27 patients have zero body mass index, 227 patients have zero skin fold thickness and 374 patients have zero serum insulin level. However, these zero values were meaningless.

and testing). We trained our model with 70% training data and tested with 30% remaining data. Five different models are developed using supervised learning to detect whether the patient is diabetic or nondiabetic. For this purpose, linear kernel support vector machine (SVM-linear), radial basis upon parameters like precision, recall, area under curve (AUC) and F1 score. so on avoid problem of over fitting and under fitting, tenfold cross validation is completed optimal classifier has value of area under the curve near 1.0. If it's near 0.5 then this value is like random guessing [20]. Accuracy indicates our classifier is how of a classifier may lie between 0.5 to1. Values below varied to a specific class. Area under

upon parameters like precision, recall, area under curve (AUC) and F1 score. so as to avoid problem of over fitting and under fitting, tenfold cross validation is completed Accuracy indicates our classifier is how often correct in diagnosis of whether patient is diabetic or not. Precision has been wonted to determine classifier's ability provides correct positive predictions of diabetes. Recall or sensitivity is employed in our work to seek out the proportion of actual positive cases of diabetes correctly identified by the classifier used. Specificity is getting used to work out classifier's capability of determining

Attribute No.	Attribute	Variable Type	
A1	Pregnancy	Integer	0-17
A2	glucose	Real	0-199
A3	blood pressure	Real	0-122
A4	skin Thickness	Real	0-99
A5	insulin	Real	0-846
A6	Body mass index (BMI)	Real	0-67.1
A7	Diabetes pedigree Function	Real	0.078-2.42
A8	Age	integer	21-81
Class		binary	1=Tested positive for diabetes
			0=Tested Negative for diabetes

Table 1: parameter of different Dataset

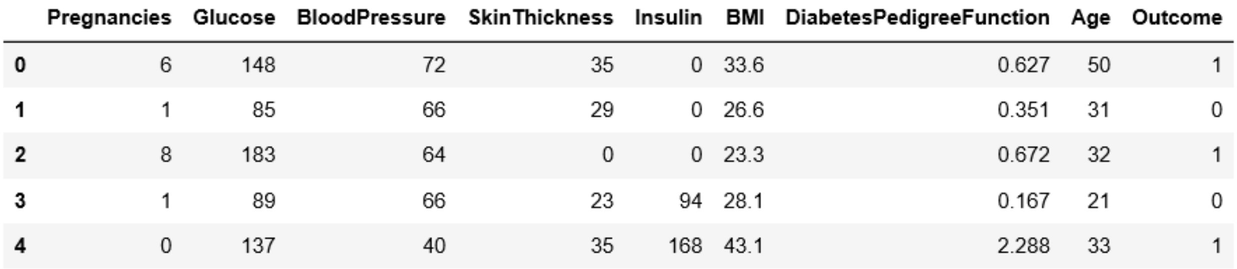
V.RESULT: Early diagnosis of diabetes are often helpful to enhance the standard of lifetime of patients and enhancement of their anticipation. Supervised algorithms are wont to develop different models for diabetes detection. Table 2 gives a view of the various machine learning models trained on Pima Indian diabetes dataset with optimized tuning parameters. All techniques of classification were experimented in "R" programming studio. the info set are partitioned into two parts (training

negative cases of diabetes. because the weighted average of precision and recall provides F1 score so this score takes under consideration of both. The classifiers of F1 score near 1 are termed as best one [18]. Receiver operating characteristic (ROC) curve may be a documented tool to see performance of a binary classifier algorithm [19]. it's plot of true positive rate against false positive rate because the threshold for assigning observations are

curve (AUC) value 0.50 indicated for a group of random data which couldn't distinguish between true and false. An often correct in diagnosis of whether patient is diabetic or not. Precision has been wonted to work out classifier’s ability provides correct positive predictions of diabetes. Recall or sensitivity is used in our work to hunt out the proportion of actual positive cases of diabetes correctly identified by the classifier used. Specificity is getting want to compute classifier’s capability of determining negative cases of diabetes. because the weighted average of precision and recall provides F1 score so this score takes into account of both. The classifiers of F1 score near 1 are termed as best one [18]. Receiver operating characteristic (ROC) curve could also be a documented tool to ascertain performance of a binary classifier algorithm [19]. it's plot of true positive rate against false positive rate because the edge for assigning observations are varied to a selected class. Area under curve (AUC) value of a classifier may lie between 0.5 to1. Values below 0.50 indicated for a gaggle of random data which couldn't distinguish between true and false. An optimal classifier has value of area under the curve near 1.0. If it's near 0.5 then this value is like random guessing [20].

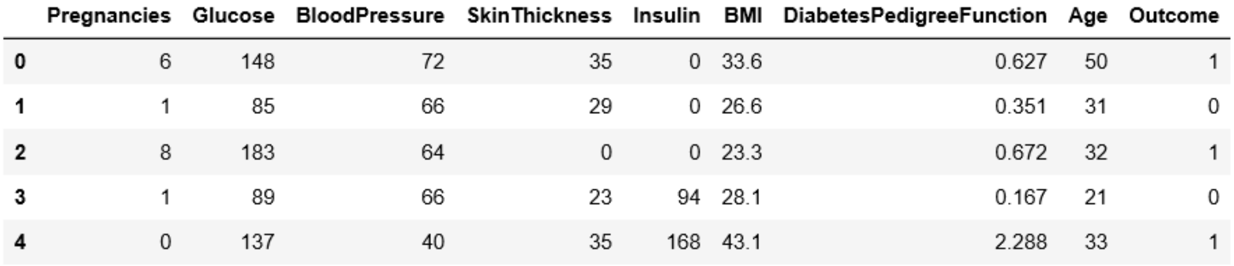
We adapted the missing value problem using the median approach and it offered the simplicity within the process during our classification paradigm. Note that, there a several methods for approaching this issue and within the present scope of this paper, we've simplified this using the present scope of this paper, we've simplified this using the median-based approach Note that it also depends upon the info types and therefore the density of the info. Since our data is simple, our strategy yields result which are comparable the prevailing approaches while comprehensive analysis is that the novelty of the system. Some statistical information of the variables of the info From Table which represents different parameter for evaluating all the models, it's found that accuracy of linear kernel SVM model is 0.89. For radial basis function kernel SVM, accuracy is 0.84. For k-NN model accuracy is found to 0.88, while for ANN it's 0.86. Accuracy of MDR based model is found to be 0.83. Recall or sensitivity which indicates correctly identified proportion of actual positives diabetic cases for SVM-linear model is 0.87 and for SVM-RBF it's 0.83. For k-NN, ANN and MDR based models recall values are found to be 0.90, 0.88 and 0.87 respectively. Precision of SVM-linear, SVM-RBF, k-NN, ANN and MDR models is found to be 0.88, 0.85, 0.87, 0.85 and 0.82 respectively. F1 score of SVM-linear, SVM-RBF, k-NN ANN and MDR models is found to be 0.87, 0.83, 0.88, 0.86 and 0.84 respectively. we've calculated area under the curve (AUC) to live performance of our models. it's found that AUC of SVM linear model is 0.90 while for SVM-RBF, k-NN, ANN and MDR model the values are respectively 0.85, 0.92 0.88 and 0.89. So, from above studies, it are often said that on the idea of all the parameters SVM-linear and k-NN are two best models to seek out that whether patient is diabetic or not. Further it are often seen that accuracy and precision of SVM- linear model are higher as compared to k-NN model. But recall and F1 score of k-NN model are above SVM- linear model.

Table 2(a): Experiment Predictive Modelling and Analytics for Diabetes



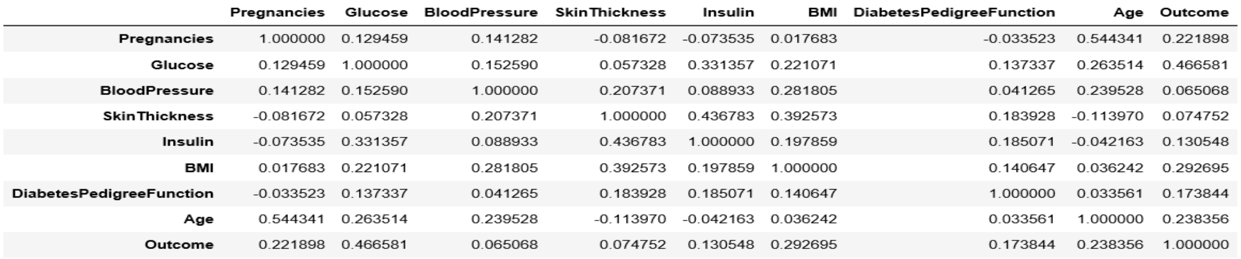
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Table 2(b): Predictive Modelling and Analytics for Diabetes



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Table 2(c): Predictive Modelling and Analytics for Diabetes



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

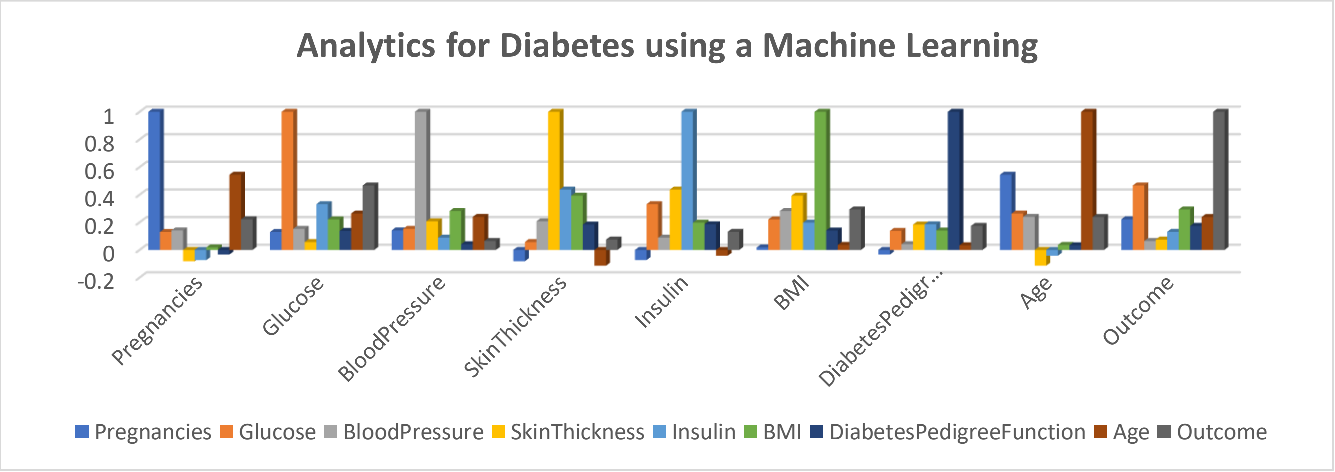


Figure 5: Predictive Modelling and Analytics for Diabetes.

an example of imbalanced class with 500 negative instances and 268 positive instances giving an imbalance ratio of 1.87. Accuracy alone may not provide a very good indication of performance of a binary classifier in case of imbalanced class. F1 score provides better insight into classifier performance in case of uneven class distribution as it provides balance between precision and recall [21, 25]. So, in this case F1 score should also be taken care of. Further it can be seen that AUC value of SVM-linear and k-NN model are 0.90 and 0.92 respectively

VI. CONCLUSION AND FUTURE WORK

We have developed five different models to detect diabetes using linear kernel support vector machine (SVM-linear), radial basis kernel, support vector machine (SVM-RBF), k-NN, ANN and MDR algorithms. Feature selection of dataset is done with the help of Boruta wrapper algorithm which provides unbiased selection of important features. All the models are evaluated on the basis of different parameters- accuracy, recall, precision, F1 score, and AUC. The experimental results suggested that all the models achieved good results; SVM-linear model provides best accuracy of 0.89 and precision of 0.88 for prediction of diabetes as compared to other models used. On the other hand, k-NN model provided best recall and F1 score of 0.90 and 0.88. As our dataset is an example of imbalanced class, F1 score may provide better insight into performance of our models. F1 score provides balance between precision and recall. Further it can be seen that AUC value of SVM- linear and k-NN model is 0.90 and 0.92 respectively. Such a high value of AUC indicates that both SVM- linear and k-NN are optimal classifiers for diabetic dataset. So, from above studies, it can be said that on the basis of all the parameters linear kernel support vector machine (SVM-linear) and k-NN are two best models to find that whether patient is diabetic or not. This work also suggests that Boruta wrapper algorithm can be used for feature selection. The experimental results indicated that using the Boruta wrapper features selection algorithm is better than choosing the attributes manually with less medical domain knowledge. Thus, with a limited number of parameters, through the Boruta feature selection algorithm we have achieved higher accuracy and precision.

References

[1] D. Soumya and B Srilatha, Late stage complications of diabetes and insulin resistance, J Diabetes Metab. 2(167) (2011) 2- 7.

[2] K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, D. Papazoglou, Complications of Diabetes, J. of Diabetes Res. 2015 (2015), 1-5.

[3] L. Mamykinaa, et al., Personal discovery in diabetes self-management: Discovering cause and effect using self-monitoring data, J. Biomd. Informat. 76 (2017) 1–8.

[4] A. Nather, C. S. Bee, C. Y. Huak, J. L.L. Chew, C. B. Lin, S. Neo, E. Y. Sim, Epidemiology of diabetic foot problems and predictive factors for limb loss, J. Diab. and its Complic. 22 (2) (2008) 77-82.

[5] Shiliang Sun, A survey of multi-view machine learning, Neural Comput. & Applic. 23 (7–8) (2013) 2031–2038.

[6] M. I. Jordan, M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science. 349 (6245) (2015) 255-260.

[7] P. Sattigeri, J. J. Thiagarajan, M. Shah, K.N. Ramamurthy, A. Spanias, A scalable feature learning and tag prediction framework for natural environment sounds , Signals Syst. and Computers 48th Asilomar Conference on Signals, Systems and Computers.(2014) 17791783.

[8] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics." Nature Reviews Genetics 16, no. 6 (2015): 321-332.

[9] K. Kourou, T. P.Exarchos, K. P.Exarchos, M. V.Karamouzis, D. I.Fotiadis, Machine learning applications in cancer prognosis and prediction, Computation. and Struct. Biotech. J. 13 (2015) 8-17.

[10]E. M. Hashem, M. S. Mabrouk, A study of support vector machine algorithm for liver disease diagnosis. Amer. J. of Intell. Sys. 4(1) (2014) 9-14. [11]W. Mumtaz, S. Saad Azhar Ali, M. Azhar, M. Yasin and A. Saeed Malik, A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD)." Medical & biological engineering & computing (2017): 114. [12]D. K. Chaturvedi, Soft Computing Techniques and Their Applications, In Mathematical Models, Methods and Applications, 31-40. Springer Singapore, 2015. [13]A. Tettamanzi, M. Tomassini. Soft computing: integrating evolutionary, neural, and fuzzy systems. Springer Science & Business Media, 2013. [14]M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, Support vector machines, IEEE Intell. Syst. and their Appl. 13 (4) (1998) 18-28. [15]G. B. Huang, Q. Y. Zhu, C. K. Siew, Extreme learning machine: theory and applications. Neurocomput. 70 (1) (2006), 489-501. [16]S. A. Dudani, The Distance-Weighted k-Nearest-Neighbor Rule, IEEE Trans. on Syst., Man, and Cybernet. SMC-6 (4) (1976) 325-327, [17]T. Kohonen, An introduction to neural computing.

Neural networks 1, no. 1 (1988): 3-16. [18]Z. C. Lipton, C. Elkan,B. Naryanaswamy, Optimal thresholding of classifiers to maximize F1 measure. in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg. (2014) 225-239. [19]L. B Ware, et al., Biomarkers of lung epithelial injury and inflammation distinguish severe sepsis patients with acute respiratory distress syndrome, Crit. Care. 17 (5) (2013) 1-7 [20] M. E. Rice, G. T. Harris, Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r, Law Hum Behav. 29 (5) (2005) 615-620.

[21]A. Ali, S. M. Shamsuddin, A. L. Ralescu, Classification with class imbalance problem: A Review, Int. J. Advan. Soft Compu. Appl . 5 (3) (2013) 176-204

[22] S. Park, D. Choi, M. Kim, W. Cha, C. Kim, I. C. Moon, Identifying prescription patterns with a topic model of diseases and medications, J. of Biomed. Informat. 75 (2017) 35-47.

[23] Kaur, H., Lechman, E. and Marszk, A. (2017), Catalyzing Development through ICT Adoption: The Developing World Experience, Springer Publishers, Switzerland.

[24] Kaur, H., Chauhan, R., and Ahmed, Z., Role of data mining in establishing strategic policies for the efficient management of healthcare system–a case study from Washington DC area using retrospective discharge data. BMC Health Services Research. 12(S1):P12, 2012.

[25] J. Li, O. Arandjelovic, Glycaemic index prediction: A pilot study of data linkage challenges and the application of machine learning, in: IEEE EMBS Int. Conf. on Biomed. & Health Informat. (BHI), Orlando, FL, (2017)357-360.