PREDICTIVE MODELLING AND ANALYTICS FOR DIABETES USING A MACHINE LEARNING APPROACH

Prateek Mishra[1], Dr.Anurag Sharma[2], Dr.Abhishek Badholia[3]
[1,2,3] *Computer Science and Engineering, MATS University, Raipur,India*
[1] *mishra19881221@gmail.com,* [2]*anusiraag@gmail.com*

Abstract: Adverse effects can be seen in the entire body due to the major disorders known as Diabetes. The risk of dangers like diabetic nephropathy, cardiac stroke and other disorders can increase severally because of the undiagnosed diabetes. Around the globe the people are suffering from this disease. For a healthy life early detection of this disease is very curtail. As the causes of the diabetes is increasing rapidly this disease might turn up as a reason for worldwide concern. Increasing the chances for a more accurate predictions and form experiences automatic learning by computational method may be provided by Machine Learning (ML). With the help of R data manipulation tool for trends development and with risk factor patterns detection in Pima Indian diabetes technique of machine learning is been used in the current researches. With the use of R data manipulation tool analysis and development five different predictive models is done for the categorization of patients into diabetic and non- diabetic. supervised machine learning algorithms namely multifactor dimensionality reduction (MDR), k-nearest neighbor (k-NN), artificial neural network (ANN) radial basis function (RBF) kernel support vector machine and linear kernel support vector machine (SVM-linear) are used for this purpose.

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

Introduction

As a common metabolic diseases Diabetes can be seen. In general Diabetes type 2 onsets can be seen in a life time and in adulthood sometimes. But in present generation even in the childhood the report of these incidences can be seen. Development of diabetes can happen due to various factors such as sedentary lifestyle, food habits, weight and susceptibility. Complications such as foot ulcer, diabetic cardiac stroke, neuropathy, nephropathy and retinopathy might be caused due to the very high blood glucose level that is known as hyperglycemia because of the undiagnosed diabetes end. For the enhancement of the patient's anticipation and quality of lifetime it is extremely important to detect diabetes at the early stages. The event of techniques and algorithms are taken care by the Machine learning that allows the computer for gaining and finding the intelligence that supports the experiences of past. It is related closely to the statistics and it's an AI branch. Understanding and spotting the input file is the ability of the system in such a way that makes predictions and decisions that can be supported by it, and all this is meant by the learning.

Form different resources and various means gathering knowledge is the start point of the learning process. Organizing the info is the constitutive step, so info related challenges repairing and by removing data that is irrelevant or choosing the info of interest only scaling back the space dimensionality. Since there is a large quantity of knowledge that is used for the purpose of learning, forming a decision by the system is a difficult task, hence using few control theories, statistics, probability, logic, etc. Designing of the algorithm is done for past experiences knowledge retrieving and the info research. To calculate performance and accuracy of the system the model testing is the next step and simultaneously the system Optimization or say by using data set or new rules the Nobel improvisation. For pattern recognition, prediction and classification use of the machine learning technique is done. In different fields such as traffic management, disease prediction, robotics, gaming, character recognition, related advertisements, face tagging and recognizing, email filtering, website ranking and program. For development of the predictive model, it is an essential learning process. In today's scenario high dimensional biomedical data automatic analysis use of machine learning algorithm is done. ML biomedical application number of the samples is analysis of genetic and genomic data, risk assessment for disorder, cancer classification, skin lesions and diagnosis of disease. SVM algorithm is successfully used for diagnosis of the diseases. Classification models use such as Naïve Bayesian (NB), logistic regression (LR) and support vector machine (SVM) for diagnoses of major clinical depression (MDD) supported EEG dataset. With the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

help of supervised machine learning techniques implementation of our novel model is done in R for Pima Indian diabetes dataset for the pattern that is known for the process of knowledge discovery in diabetes. History regarding the diabetes onset of the Pima Indian population's medical is discussed in this dataset. Various independent variables are included in this along with the diabetes one variable class value in terms of 0 and 1. Five different models performance is studied during this work that is used to detect diabetes in the female patients on the bases of the multifactor dimensionality reduction (MDR) algorithms, artificial neural network (ANN), k-nearest neighbor (k-NN), radial basis kernel support vector machine (SVM-RBF) and linear kernel support vector machine (SVM-linear).

II. Related Material and Method

From UCI machine learning repository collection of minimum age of twenty-one year of Pima Indian population of female patients dataset is done. The National institute of diabetes and digestive and kidney diseases originally owns this dataset. Total 768 instances are there in the duration of this dataset that into two classes are categorized as: non-diabetic and diabetic. With risk factors of eight different types like age, diabetes pedigree function, body mass index, two-hour serum insulin, triceps skin fold thickness, diastolic vital sign, plasma glucose concentration of two hours in an oral glucose tolerance test and number of times pregnant. With the help of the powerful R data manipulation tool we investigate this dataset of diabetes, in machine learning process application crucial step is Feature engineering. For building practical machine learning model with many attributes description of the modern data set is done. In general to the classification of the supervised machine learning most of the attributes are irrelevant. Data processing phase includes k-NN imputation, removal of outliers and feature selection for missing value prediction. For inconsistent and irrelevant data handling there are several methods. The attributes that includes data that is highly correlated are chosen in this work. By feature selection method implementation of this step is done that either by Boruta wrapper algorithm or 'manual method' might be done. From a system of data essential features unbiased selection and stable is provided by the Boruta package whereas it is prone to errors when using the manual method. With R package Boruta assistance, feature selection has been through. As R package tactic is open. For machine learning algorithms convenient interface is provided by this package. Within the R implemented classification algorithm of random forest built around wrapper is called as Boruta package. With all attributes and important yielded four attributes on the Pima Indian dataset implementation of or run of Boruta wrapper is done. Calculation of parameters like the recall, precision, accuracy, attributes and other parameters are done.

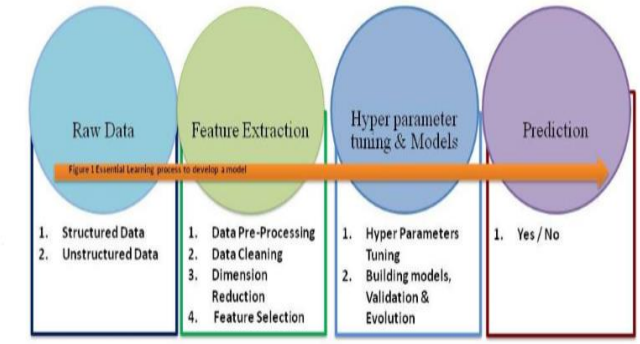

Figure1: Essential Learning process to develop a predictive model.

For the machine learning process implementation will be wanted by the couple of machine learning techniques present there. Most frequently used learning techniques like unsupervised learning and supervised learning. when the historical data is out employment of the Supervised learning technique is done for specific problems. With the respective responses and inputs the training of the system is done latest data response prediction. The Naïve Bayes classifier, support vector machines, decision tree, back propagation and artificial neural network are included in the general supervised approaches. When unlabeling of the available training data is done employment of the unsupervised learning technique is done. Any training or prior information is not provided to the system. From the available data identification and exploring of the pattern is done by the algorithm for the formation of the prediction or the decision. Generally hidden-Markov model and principle component analysis, hierarchical clustering and k-means clustering are included in the unsupervised approaches. Diabetes dataset of Pima Indians categorization binary classification is performed by the chosen supervised machine learning algorithms. Five different algorithms are use to predict if the patient has diabetes or doesn't have diabetes. Those five algorithms are as multifactor dimensionality reduction (MDR), artificial neural network (ANN), k-nearest neighbour (k-NN), kernel support vector machine (SVM) and linear kernel and radial basis function (RBF).

Algorithms used in our machine learning predictive models details are as follows:

Support Vector Machine

In both of the regression and classification employment of the Support Vector Machines (SVM) is done.

Representation of the info point is done of the space and into group's categorization and hence, in the same group the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

points with similar properties fall in the SVM model.

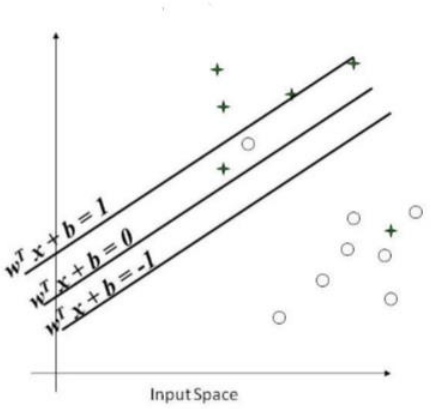

Figure 2: Representation of Support Vector Machine

In linear SVM, the provided data set as p-dimension vector is taken into the account that by the maximum of p-1 planes will be separated also known as hyper-plane. As in the figure 2 among the info groups for regression or classification problems, these planes set the boundaries or separate the info space. The amount of hyper plane on the thought of distance between the 2 classes it separates, the selection of the simple hyper-plane is done. Namely maximum-margin hyper-plane is between 2 classes the plan that has the utmost margined.

Definition of n data points:

(X_1,Y_1)…….,(X_n,Y_n)……1

Here real vector is represented by X1 and 1 or -1 is Y1, the class to which X1 belongs is represented.

Construction of hyper-plane can be done so to the distance between to classes minimization y=1 and y=-1, is defined as follows -

W. X- b = 0……2

Here Normal vector is represented by W and offset of hyper-plane is b.

.

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

On the linear and non linear data prove of the efficiency of the Support vector machine is shown. To classify nonlinear data implementation of the Radial base function is done.

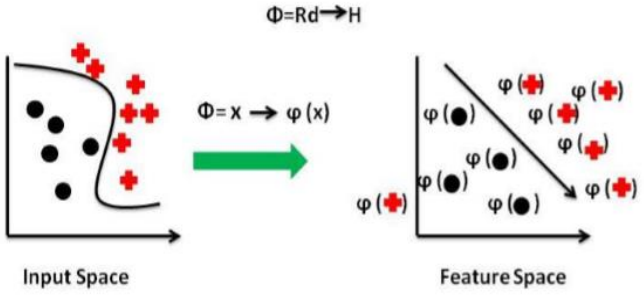

Figure 2: Representation of Radical Basis Function (RBF) Kernel support vector machine.

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

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

Yielding excellent results although it is a single algorithm by the k- Nearest neighbour. It's an instance-based, nonparametric and lazy learning algorithm. In both the regression as well as the classification problem utilization of this problem is often done. To the seek out out the class k-NN which the belongs to the unlabeled object, in the classification. Setting a 'k' for this where k is number of neighbours to be considered that in general is odd and hence nearest to the object info points distance is calculated by the methods such as Minkowski distance, Manhattan distance or Hamming distance, Euclidean's distance. The 'k' nearest neighbour after the space calculation are selected as the resultant class of new object that is calculated or analyzed on the thought of the votes of the neighbors. The result with the highest accuracy is predicted by the k-NN.

D. Artificial neural network (ANN)

The functionality of the human brain is mimicked by the artificial neural network. In generally as a set of nodes it is seen called as artificial neurons. At least to one another all of that node can transmit information. By some state 0 or 1 neurons are represented. To every node assignment of some weight might be done to define within the system the importance or strength. Into the layers of multiple nodes; it reaches the output layer by the info travels from first layer input layer and after passing through middle layers hidden

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

layers, into some important information the info is transformed by every layer and the specified output is given eventually. In neurons functioning important role is played by activation and transfer function. All the weighted input is summed up by transfer function:

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

……6

To specific range the output of the transfer function is flattened by the activation function. In can be linear as well as nonlinear. Simple activation function is:

$$f(z) = z$$

……7

As no limitations are applied on the data by the function, use of sigmoid function is done that can be expressed as follows:

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

……8

D. Multifactor Dimensionality Reduction (MDR)

The approach for representation and location of the independent variables consolidation that will in some way influence the variable that are dependent is known as Multifactor Dimensionality Reduction (MDR). In general it's designed to get the interaction in-between the variables that will affect the system output. It is not dependent on the sort of used model or parameters that as compared to the tradition system will make it better. The two or more attributes are take and combined into one attribute. The knowledge space representation is changed by this conversion. The category variable prediction by system performance improvement is lead by this. In Machine Learning utilization of the various MDR extensions is done. Few of them are as follows- covariates, risk scores, odds ratio, fuzzy methods and many more.

III. Predictive Model

Figure 4 shows predictive model proposed by us, in this model data pre-processing is done by us and to receive better results various feature engineering techniques. For the missing values prediction k-NN imputation and removal of outliers is involved in pre-processing. For choosing feature employment of Boruta wrapper algorithm is done as it gives unbiased choice from a data system of unimportant features and important features. In supervised learning featuring a significant role by feature engineering then training of data. For an improved outcome highly correlated variables are used by us. Checking data used for confusion and prediction matrix is indicated here by the input file.

For anticipation enhancement and for improving the standards of the patient lifetime can be very helpful for diabetes early diagnosis. For diabetes detection different models are wanted to be developed by the supervised algorithm, this provides a view of the different machine learning models with tuning parameters that are optimized, on Pima Indian diabetes dataset are trained. In "R" programming studio examination of the All techniques of classification were done. Into two parts the info set are classified as training and testing. Our model is trained with the training data to be 70% and 30% remaining data for testing. For testing whether the patient is diabetic or non diabetic

. For this purpose, linear kernel support vector machine (SVM-linear), radial basis

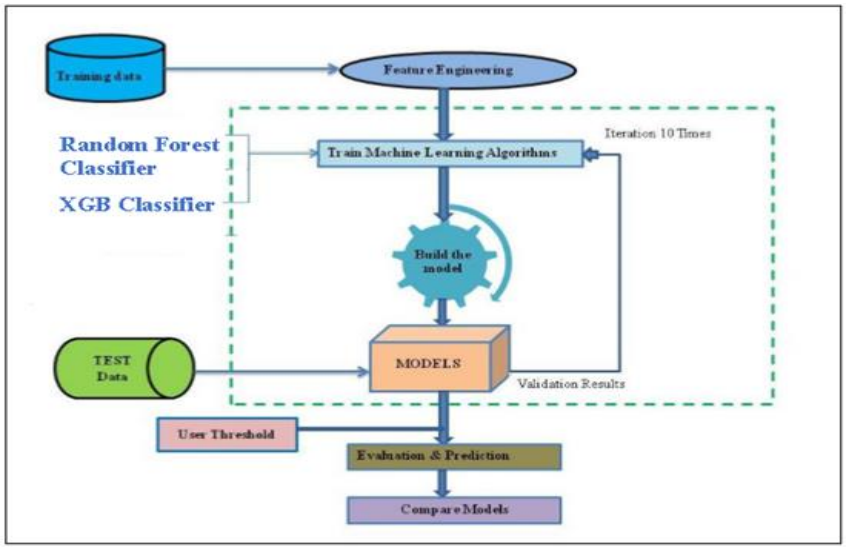

Figure 4: Framework for evaluating Predictive Model.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

using supervised learning development of five different model is done. Use of MDR algorithm, ANN, k-NN, kernel support vector machine, radial basis function (RBF) and linear kernel support vector machine (SVM-linear) was done for this purpose. To diabetes diagnosis for Pima Indian population, evaluation of the performance of all five differentiating models on the characteristics like F1 score, area under curve (AUC), recall and precision are evaluated. So to avoid the issue of under fitting and over fitting, completion of the tenfold cross validation. how frequently correct is the classifier in diagnosis of whether patient is non diabetic or diabetic is indicated by the accuracy. Classifier's ability to give diabetes correct positive predictions is determined by the precision. In our work to find out the diabetes cases proportion of actual positive cases by classifier being correctly identified uses of sensitivity or recall. Ability of determining diabetes negative cases by computer classifier is getting wanted by specificity, due to the F1 score is given by the recall and weighted average of precision hence in the account of both this scores are considered. Termed as best one is the F1 score classifiers near 1. For a binary classifier algorithm to ascertain performance as a tool of documentation should have been a Receiver operating characteristic (ROC) curve. As to a selected class the edge for assigning observations are varied, it's plot of true positive rate against false positive rate. In-between 0.5 to1 classifier Area under curve (AUC) value might be. Indicating a gaggle of random data that cannot differentiate between true and false is the value below 0.5. Area under the curve (AUC) near 1.0 has value of Optimal classifier. Like random guessing is the value if it is near 0.5. For all the models evaluation various parameters are represented, it is calculated the 0.89 is the linear kernel SVM model accuracy. 0.84 is the radial basis function kernel SVM model accuracy. 0.88 is the k-NN model accuracy. Whereas 0.86 is the ANN accuracy. 0.83 is the MDR based model accuracy. Out of the actual positives diabetic cases the correctly identified proportion is indicated by the sensitivity or recall, 0.87 is for SVM-linear case model, 0.83 is for SVM-RBF case model, 0.90 is for k-NN model, 0.88 is for ANN model, 0.87 is for MDR model. Precision for of MDR, ANN, k-NN, SVM-RBF and SVM-linear models is found to be 0.82, 0.85, 0.87, 0.85 and 0.88 respectively. F1 score of MDR, ANN, k-NN, SVM-RBF and SVM-linear models is found to be 0.84, 0.86, 0.88, 0.83 and 0.87 respectively. Our models performance measurement by area under the curve (AUC) is calculated by us it's found that AUC of MDR, ANN, k-NN, SVM-RBF and SVM-linear model the values are respectively 0.89, 0.88, 0.92 0.85 and 0.90. So from the above observations and numbers we can conclude that the two best models on the thought of all the parameters are k-NN and SVM-linear for detecting if the patient have diabetes or does not have diabetes. As compared to the k-NN the SVM-linear precision and accuracy is high. But when SVM-linear is compared with the k-NN, the k-NN has higher F1 score and recall. When we carefully study the diabetes dataset it comes out as the class that is imbalanced with the 500 instances that are negative and 268 instances that are positive. That gives 1.87 imbalanced ratio. just for imbalanced class case binary classifier performance as a good indicator accuracy alone is not sufficient. Just in case of uneven class distribution better insight to the performance of the classifier is provided by the F1 score as it gives a balance between recall and precision. So care is need in the case of the duration of the F1 score. Additionally it is frequently seen that the AUC value 0.90 and 0.92 is often shown by the SVM-linear and k-NN.

IV. Patient demographics

From This dataset consisting of 768 female patients this dataset has been taken, a minimum of 21 years old, diabetes diagnoses (diabetic or control) Pima Indian heritage, gor the diabetic patients there are 268 cases and of control patients there are 500 cases. Nine variables are contained in this dataset as : count of pregnancies numbers (1), concentration of plasma glucose -a two hour in an oral glucose tolerance test(2), diastolic vital sign (mm Hg)(3), triceps skin fold thickness (mm)(4), 2-hours serum insulin (mu U/ml)(5), body mass index (weight in kg/ (height in m)2)(6), diabetes pedigree function(7), age in years(8), class variable (diabetic or control)(9). At the duration of this dataset zero blood sugar level is found in 5 patients, for 35 patients zero diastolic vital sign, body mass index os zero in 27 patients, skin fold thickness is zero in 227 patients, serum insulin level is zero in 74 patients. How so ever meaningless was all these zero values.

.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

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:

For anticipation enhancement and for improving the standards of the patient lifetime can be very helpful for diabetes early diagnosis. For diabetes detection different models are wanted to be developed by the supervised algorithm, this provides a view of the different machine learning models with tuning parameters that are optimized, on Pima Indian diabetes dataset are trained. In "R" programming studio examination of the All techniques of classification were done. Into two parts the info set are classified as training and testing. Our model is trained with the training data to be 70% and 30% remaining data for testing. For testing whether the patient is diabetic or non diabetic using supervised learning development of five different model is done. To fulfill this purpose, radial basis upon parameters like F1 score, area under curve (AUC), recall and precision, linear kernel support vector machine (SVM-linear). So if issue of under fitting and over fitting is avoided , completion of validation of tenfold cross is done area under the curve value in optimal classifier is around 1.0. This value is such as a random guessing if it is around 0.5. How the classifier is indicated by the accuracy to judge our classifier and may lie in-between 0.5-1. To a specific class is varied is the value is below. On the bases of the parameters like F1 score, area under curve (AUC), recall and precision. So if issue of under fitting and over fitting is avoided, completion of validation of tenfold cross is done. In patient diagnosis if the patient is diabetic or non diabetic, it is indicated by the accuracy that how frequently the proposed classifier is correct in diagnosis. For the ability of the classifier to give a correct diabetes prediction of positive result for determination we need precision. By the use of classifier correct identification the actual positive cases correctly of diabetes out of the proportion for seeking out in our work employment of sensitivity or recall is done. For diabetes negative cases determination the capability of our classifier is getting used by Specificity. In our work to find out the diabetes cases proportion of actual positive cases by classifier being correctly identified uses of sensitivity or recall. Ability of determining diabetes negative cases by computer classifier is getting wanted by specificity, due to the F1 score is given by the recall and weighted average of precision hence in the account of both this scores are considered. Termed as best one is the F1 score classifiers near 1. For a binary classifier algorithm to ascertain performance as a tool of documentation should have been a Receiver operating characteristic (ROC) curve.

As to a selected class the edge for assigning observations are varied, it's plot of true positive rate against false positive rate. In-between 0.5 to1 classifier Area under curve (AUC) value might be. Indicating a gaggle of random data that cannot differentiate between true and false is the value below 0.5. Area under the curve (AUC) near 1.0 has value of Optimal classifier. Like random guessing is the value if it is near 0.5. Using the approach of median for solving missing value problem is adapted by us and in the duration of our classification paradigm within the process simplicity is offered. For this problem approaching there are various methods this should be noted and inside the paper scope present, with the help of the approach that is median-based simplification is done by us. It should be noted that on the type of info it is also depended and hence the info density also. Since the simple data is used by us here, with the prevailing approach comparable results are yielded here by our strategy whereas for the system the novelty is the comprehensive analysis. Info variables few statistical information from the table that shows various parameters for all this method evaluations, it's found that 0.89 is the accuracy of the kernel SVM model. 0.84 is the accuracy of the radial basis function kernel SVM. 0.88 is the accuracy of the k-NN model. 0.86 is the accuracy of the ANN model. 0.88 is the accuracy of the MDR model. Actual positives diabetic cases portion correctly identified is indicated by the sensitivity or recall, 0.87 is for SVM-linear model, 0.87 is for SVM-linear model,0.90 is for k-NN model,0.88 is for ANN model and 0.87 is for MDR model. Precision of MDR, ANN, k-NN, SVM-RBF and SVM-linear model is found to

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

be 0.82, 0.85, 0.87, 0.85 and 0.88 respectively. F1 score for MDR, ANN, k-NN, SVM-RBF and SVM-linear model is found to be 0.84, 0.86, 0.88, 0.83 and 0.87 respectively. Measurement of the area under the curve (AUC) is calculated to the live performance of our model of MDR, ANN, k-NN, SVM-RBF and SVM-linear model the values are respectively 0.89, 0.88, 0.92, 0.85 and 0.90 respectively. So from the above observations and numbers we can conclude that the two best models on the thought of all the parameters are k-NN and SVM-linear for detecting if the patient have diabetes or does not have diabetes. As compared to the k-NN the SVM-linear precision and accuracy is high. But when SVM-linear is compared with the k-NN, the k-NN has higher F1 score and recall. An example for imbalanced class is with the 500 instances that are negative and 268 instances that are positive. That gives 1.87 imbalanced ratios. Just for imbalanced class case binary classifier performance as a good indicator accuracy alone is not sufficient. Just in case of uneven class distribution better insight to the performance of the classifier is provided by the F1 score as it gives a balance between recall and precision. So care is need in the case of the duration of the F1 score. Additionally it is frequently seen that the AUC value 0.90 and 0.92 is often shown by the SVM-linear and k-NN.

Table 2(a): ExperimentPredictive 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

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

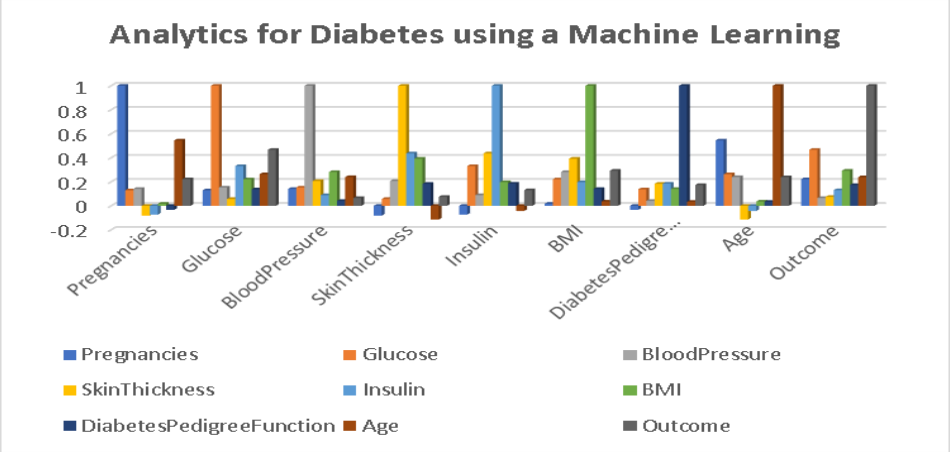

Figure 5: Predictive Modelling and Analytics for Diabetes

VI. CONCLUSION AND FUTURE WORK

To detect diabetes we have developed five different models using the algorithms such as MDR algorithms, ANN, k-NN, support vector machine (SVM-RBF), radial basis kernel and linear kernel support vector machine (SVM-linear). By using Boruta wrapper algorithm selection of feature of dataset is done that gives unbiased selection of important features. on the basis of various parameters evaluation of all the models is done and these different parameters are: AUC, F1 score, precision, recall and accuracy. All the models shows good results can be suggested by the experimental results, giving 0.89 best accuracy by SVM-linear model and 0.88 precision for the diabetes prediction in comparison to the other models. Best F1 score and recall of 0.88 and 0.90 is shown by the k-NN model. As an example of imbalanced class is provided by our dataset. Into performance of our model better sight may be given by the F1 score. Between recall and precision a balance provided by the F1 score. Additionally we can see that 0.92 and 0.90 is the AUC value of the k-NN model and SVM-linear model. We can say that as an optical classifier for diabetes is k-NN and SVM-linear model because of such high value of the AUC. So from all the above statistics and results we can conclude that to find if the patient is have diabetes or not having. The k-NN and diabetes linear kernel support vector machine (SVM-linear) are two best models on the basics of all the parameters. This entire paper also gives a suggestion that for feature selection use of that Boruta wrapper algorithm can be done. Compared with the less medical domain knowledge choosing the attributes manually using the Boruta wrapper features selection algorithm is far better as suggested from all the above studies. Hence with the parameters being in the limited numbers still we have received higher precision and accuracy in the Boruta feature selection algorithm.

References

[1]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey "Multifactor Authentication Methods: A Framework for Their Selection and Comparison" accepted for publication in International Journal of Future Generation Communication and Networking Vol. 13, No. 3, (2020), pp. 2522–2538, ISSN: 2233-7857 (Web of Science).

[2]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey "A Novel Hybrid Approach for Cyber Security in IoT network Using Deep Learning Techniques" accepted for publication in International Journal of Advanced Science and Technology ISSN:2394-5125, ISSN: 2005-4238 (Scopus indexed Journal).

[3]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey entitled "Development of Real Time Automated Security System for Internet of Things (IoT)" accepted for publication in International Journal of Advanced Science and Technology Vol. 29, No. 6s, (2020), pp. 4180 – 4195, ISSN: 2005-4238 (Scopus indexed Journal).

[4]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey "A Proposed Security Framework for Internet of Things: An Overview" presented in international Conference held on 20-22 December,2019, MTMI, Inc. USA in Collaboration with at amity Institute of Higher Education, Mauritius.

[5]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey "Control of Public Services for Public Safety through Cloud Computing Environment" presented in international Conference held on 04-05 January,2020, Organized by Atal Bihari Vajpayee University, Bilaspur in association with MTMI, USA and sponsored by CGCOST, Raipur (C.G), India.

[6]. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey "A Study on Security and Data Privacy

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

issues of IoT based Application in Modern Society" presented in international Conference held on 04-05 January,2020, Organized by Atal Bihari Vajpayee University, Bilaspur in association with MTMI, USA and sponsored by CGCOST, Raipur (C.G), India.

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

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

[9]. 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.

[10]. 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.

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

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

[13]. 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.

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

[15]. 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.

[16]. 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.

[17]. 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.

[18]. D. K. Chaturvedi, Soft Computing Techniques and Their Applications, In Mathematical Models, Methods and Applications, 31-40. Springer Singapore, 2015.

[19]. A. Tettamanzi, M. Tomassini. Soft computing: integrating evolutionary, neural, and fuzzy systems. Springer Science & Business Media, 2013.

[20]. 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.

[21]. G. B. Huang, Q. Y. Zhu, C. K. Siew, Extreme learning machine: theory and applications. Neurocomput. 70 (1) (2006), 489-501.

[22]. S. A. Dudani, The Distance-Weighted k-Nearest-Neighbor Rule, IEEE Trans. on Syst., Man, and Cybernet. SMC-6 (4) (1976) 325-327,

[23]. T. Kohonen, An introduction to neural computing. Neural networks 1, no. 1 (1988): 3-16.

[24]. 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.

[25]. 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 [25] 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.

[26]. 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.

[27]. 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.

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

[29]. 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.

[30]. [30] 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.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731