International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 11 Issue 5, May-2022, Impact Factor: 7.957

Comparative Study of Machine Learning Algorithms for Diabetes

Kasturi Pal[1], Namrata Khadse[2], Rashika Raut[3], Renu Iyer[4], Prof. Vivek Deshmukh[5]

[1,2,3,4] Student, Dept. of Electronics and Telecommunication Engineering, S. B. Jain Institute of Technology, Management and Research, Nagpur
[5]Associate Professor, Dept. of Electronics and Telecommunication Engineering, S. B. Jain Institute of Technology, Management and Research, Nagpur

ABSTRACT

Machine Learning algorithms are applied in many applications as a standard procedure for extracting useful information and knowledge to reinforce the major decision-making processes, analysing the large volume of available data. Diabetes mellitus is a deadly syndrome which happens when our body isn't able to take up sugar into its cell and use it for energy resulting in high blood glucose. Early symptoms are related to hyperglycemia and include nerve damage, excessive thirst or urination, weight loss, blurred vision or weight loss. In this project work, we have put forward a diabetes prediction model for better classification of diabetes which includes few external as well as regular factors like IBM, Glucose, Pedigree Glucose Function, Age, etc. responsible for diabetes. Several algorithms are used which classifies Diabetes mellitus data efficaciously. The advantages and limitations of machine learning algorithms are analysed thoroughly. The assessment of performance of algorithms is carried out to determine the best one of them.

Keywords: Diabetes mellitus, machine learning, classification, regression, hyperglycemia

INTRODUCTION

A group of metabolic diseases where a person experiences high blood glucose levels either because the body produces inadequate insulin or the body cells do not respond properly to the insulin produced by the body is known as Diabetes mellitus. Diabetic patients often experience frequent urination, increased thirst and increased hunger. There are 3 types of Diabetes:

a. Type 1 Diabetes: For this type of diabetes, the insulin is not produced enough by the body. Insulin-dependent diabetes, early-onset diabetes or juvenile diabetes is also referred as type 1 of diabetes. Before a person is 40-years-old i.e., in early adulthood or teenage this Type 1 diabetes usually starts develop. Patients with this type of diabetes have to take insulin injections for the rest of their life. Proper blood-glucose levels must also ensure by them by carrying out regular blood tests and following a special diet.

b. Type 2 Diabetes: For this type of diabetes, enough insulin is not produced by the or insulin resistance is displayed by the cells in the body. May be some people be able to control their type 2 diabetes symptoms by following a healthy diet, losing weight, monitoring their blood glucose levels and doing plenty of exercise. However, type 2 diabetes is typically a progressive disease – it gradually gets worse – and the patient have to probably end up taking insulin, usually in tablet form- Being physically inactive, overweight and eating the wrong. A Comparative Analysis in the Prediction (RatnaPatil) on the Evaluation of Classification Algorithms 3967 foods all contribute to our risk of developing type 2 diabetes. With increase in age the risk of developing Type 2 diabetes also increases.

c. Gestational Diabetes: Gestational Diabetes affects females during pregnancy. Some women have very high levels of glucose in their blood, and the insulin is unable to produce by their bodies to transport all of the glucose into their cells, which results in progressively rising levels of glucose. For gestational diabetes the majority of the patients can control their diabetes with diet and exercise. Between 10% to 20% of them will need to take some kind of blood glucose-controlling medications. Uncontrolled or undiagnosed gestational diabetes can raise the risk of complications during child birth.

MODEL CONSTRUCTION

Model Construction will take place using Random Forest, SVM, KNN, Logistic Regression, Naïve Bayes, Decision Tree, Ada Boost and their performance will be evaluated.

Random Forest

Number Of decision trees is contain by the random forest classifier on different subsets of the given data. It also take the average to improve the accuracy of the data. Random forest gets the prediction of the trees instead of depending on one decision tree based on the majority votes and from that it predict the final output. **The greater the no. Of trees than it higher the accuracy and also prevent from over fitting problem.**

The below diagram explains the working of the Random Forest algorithm:

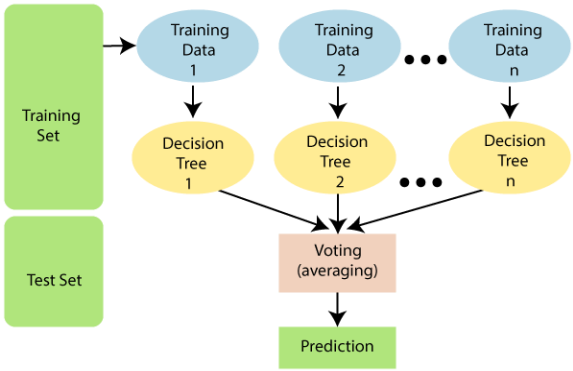


Figure 1: Random Forest Algorithm

Support Vector Machine (SVM)

It is one of the most popular supervised learning algorithm. It is used for classification as well as regression problems, but it is mainly used for classification problem in machine learning.

SVM algorithm is used to create decision boundary. The best decision boundary is called as a hyper plane.
The hyper plane is help to create by the SVM chooses the vector. The. Extreme cases are called as SVM.

Consider the below diagram of hyper plane:

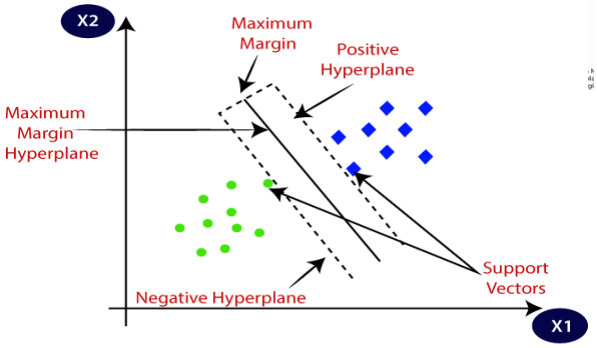


Figure 2: Support Vector Machine Algorithm

K-Nearest Neighbor (KNN)

K-NN algorithm assumes the similarity between the data and available cases. Based on the similarity of all the available data and classifies of K-NN algorithm. Using K-NN algorithm new data appears then it can be easily classified.

K-NN is a **multivariate algorithm**, it means it does not make any assumption on data set. It is also called a **lazy learner** algorithm. Instead it stores the dataset and at the time of classification, on dataset it performs an action.

K-NN algorithm stores the dataset and gets the data to classify into a category that is much similar to the new dataset.

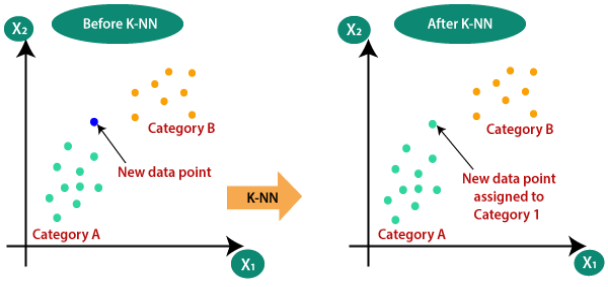


Figure 3: K-nearest neighbour Algorithm

Logistic Regression

Logistic Regression is a supervised classification technique. The variable which is dependent should be categorical. The independent variable must not have multiple co linearity. It is used for predicting the categorical dependent variable using the given set of independent variables. Logistic Regression is similar to Linear Regression except how they are used. Logistic Regression predicts the output of a categorical dependent variable. Therefore the outcome must be discrete value. It can be True or False, 0 or 1, etc. It gives the probabilistic values which gives True or False.

In this algorithm, we fit the "S" shaped logistic function, which help to anticipate two maximum values. The S-curve is called the sigmoid function. The curve indicates the likelihood of something such as whether the patient is cancerous or not, whether the patient is diabetic or not, etc. It is a significant algorithm because it has the ability to provide probability and classify new data using the datasets. In this algorithm, we use the concept of threshold value, which defines the probability. The values above the threshold tends to 1 and value below the threshold tends to 0.

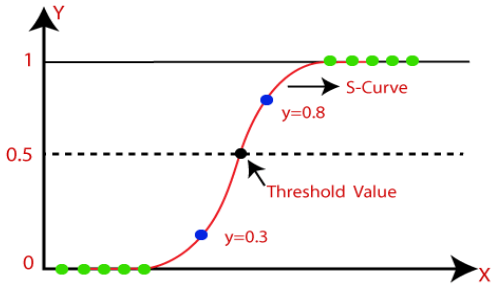


Figure 4: Logistic Regression Algorithm

Naïve Bayes

It predicts on the basis of the probability and hence called as probabilistic classifier. When we are working with data that has millions of records, the recommended algorithm is Naïve Bayes. It is used for large volumes of data, it gives adequate results when it comes to sentimental analysis. It is based on Bayes theorem.

Bayes Theorem: The theorem works on conditional probability. Conditional Probability is that something happen when something else is already occurred. It gives us the probability of an event using the prior knowledge.

Naïve Bayes assumption is that each feature makes contribution which is independent and equal to the outcome.

- Convert the given dataset into tabular form.
- Generate likelihood table by determining the probabilities.
- Use Bayes theorem to calculate the probabilities.

Decision Tree

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub-trees. Below diagram explains the general structure of a decision tree:

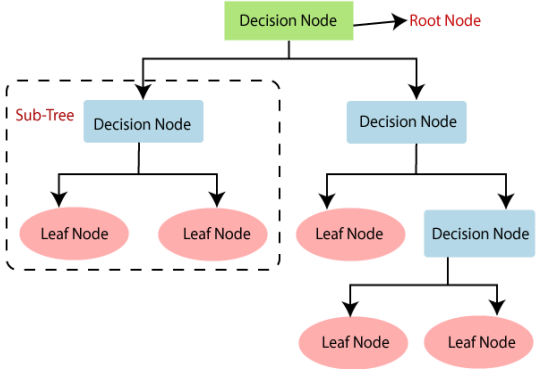


Figure 5: Decision Tree Algorithm

Ada Boost

Adaptive Boosting is a meta-learning method which was built to increase the efficiency of binary classifiers. The principle behind the algorithm is that we built model on the training dataset, then another model is built to rectify the errors present in the first model and until the errors are minimized this process is continued. It is a decision tree with only one level i.e. with only one split.

It builds a model and all the data points are given equal weights. The wrongly classified points are assigned with the higher weights. Now the higher weights points are given more importance in the next stage and the process continues unless the error received is lower. It is used to fit the sequence of weak stumps on modified versions of data repeatedly. It combines multiple weak classifiers into single strong classifier. The nodes are called as base learner or decision stumps. It is sequential process and each sub sequence tries to correct the error of previous model.

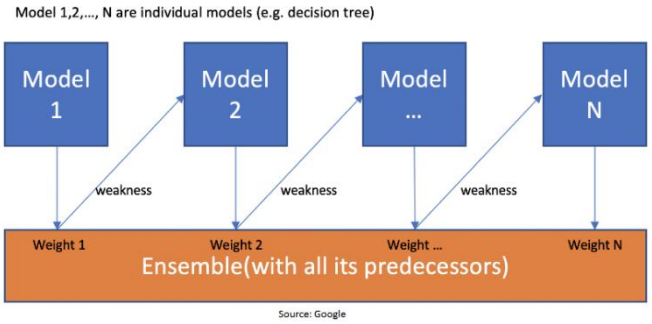


Figure 6: AdaBoost Algorithm

WORK FLOW

Data, feature, and software tool

In our research, the dataset is collected from Kaggle, which is originated from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). In the dataset, all the patients are female, and at least 21 years old. The data set contains information about 17768 patients and their nine unique attributes. In Table1thedescriptionofthe attributes of this dataset is shown. The nine attributes that are used for the prediction of diabetes are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Diabetic.

FIG 7: PROPOSED MODEL DIAGRAM

Table 1 The attributes of Dataset.

Attribute	Description	Type	Average/Mean
Preg	Number of times pregnant.	Numeric	3.305
Glucose	Plasmaglucoseconcentration2hinanoralglucosetolerancetest.	Numeric	109.920
BP	Diastolicbloodpressure(mmHg).	Numeric	70.895
Skin Thickness	Tricepsskinfoldthickness(mm).	Numeric	27.569
Insulin	2-hourseruminsulin(µlU/mL).	Numeric	128.859
BMI	Bodymassindex(kg/m^2).	Numeric	31.607
DPF	Diabetespedigreefunction.	Numeric	0.410
Age	Age(years).	Numeric	30.604
Diabetic	Diabetes diagnose results (tested_positive:1, tested_negative:0)	Nominal	0.334

The attribute 'diabetic' is taken as a dependent or target variable, and the remaining eight attributes are taken as independent/feature variables. The attribute' diabetic' consists of binary value whereas 0meansnon-diabetes, and1impliesdiabetes. In our research, we have used data mining and machine learning algorithms for prediction whether a patient has diabetes or not with enhanced accuracy. This obesity dramatically increases people' srisk of developing Type 2 diabetes. Table1shows that theaveragebodymassindexis32forthe17768patients.Thedataset is for the Type 2 diabetes patients, as the people with a BMI of 30 or greater are considered obese.

We have used We ka, an open-source machine learning, and data mining software tool for the diabetes dataset's performance analysis. It contains tools for data preprocessing, clustering, classification, regression, visualization, and feature selection. And also this steps are implemented in the Jupiter Notebook, and the Python programming language is used for coding.

Data preprocessing

Preprocessing helps to transform the data so that a better model of machine learning can be built, which will be providing higher accuracy. The preprocessing performs various functions like outlier rejection, missing values filling, normalization of data, feature selection to improve the quality of data. In the dataset we are using, 5952 samples are class ified as diabetic, and 11816 were non-diabetics.

Missing value identification

Using the excel and by using python, we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.

Null values and removal

Using the Weka tool and by using python in jupyter notebook, we filtered the dataset for detecting the missing or null values. The number of missing values are shown in Table 2.

Table2

Thenumberofmissingvaluesindataset.

Attributes	No.ofmissingvalues
Preg	0
Glucose	18
BP	125
Skin Thickness	800
Insulin	1330
BMI	39
DPF	0
Age	0

Table3

Thecorrelationbetweeninputandoutputattributes.

Attributes	Correlationcoefficient
Preg	0.38
Glucose	0.10
BP	0.088
Skin Thickness	0.14
Insulin	0.23
BMI	0.22
DPF	0.17
Age	0.33

Feature selection

Person's correlation method, to find the most relevant attributes/features is a popular method. The correlation coefficient is calculated in this method, which will correlate with output and input attributes. The coefficient value remains in the range be-tween 1 and 1. The value above 0.5 and below 0.5 will indicate a notable correlation, and the zero value means no correlation. In the Weka tool, correlation filter is used to find the correlation coefficient, and the results are shown in Table3.

Normalization

We have performed feature scaling by normalizing the data from 0 to 1 range, which boosted the calculation speed of algorithms. The mean and standard deviation is resulting for all the attributes after normalization are shown in Table 4.

In **Fig. 8**, we can see that, after completing preprocessing, we have 17738 samples/instances where 11816 patients have no diabetes, and 5952 patients have diabetes. After the preprocessing, the correlation between input and output attributes is shown in Fig. 9

Table4 Describing Standard deviation and Minimum- Maximum values

Attributes	Std deviation	Min	Max
Preg	3.385	0.000	17.000
Glucose	32.348	0.000	199.00
BP	17.183	0.000	122.00
Skin Thickness	15.080	0.000	110.00
Insulin	131.722	0.000	846.00
BMI	9.518	0.000	80.600
Diabetes Pedigree	0.371	0.078	2.420
Age	12.090	21.00	81.00
Outcome	0.471	0.000	1.000

Data settra in and test method
After cleaning of data and preprocessing, the dataset becomes ready to train and test. We 70% train/tests plitting method separately to test the different machine learning model's performance. In this train/split method, we will split the dataset randomly into training and testing set and then performance will be evaluated.

Design and implementation of classification model
For this research work, comprehensive studies are done on the PIDD applying different ML classification techniques like DT, KNN, RF, NB, LR, AB, SVM, and neural network (NN).We used Kth value 7 for the KNN algorithm. The modeling ram proposed is shown in Fig.7

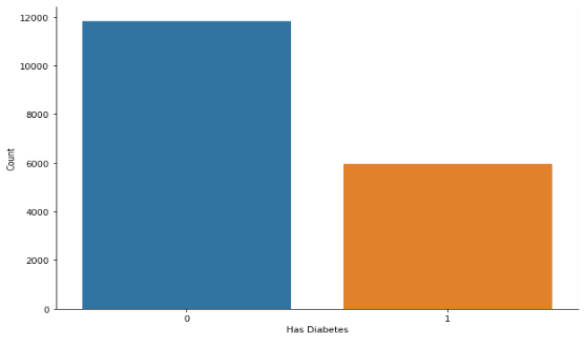


Figure 8: After pre-processing the number of diabetes and non diabetes patients.

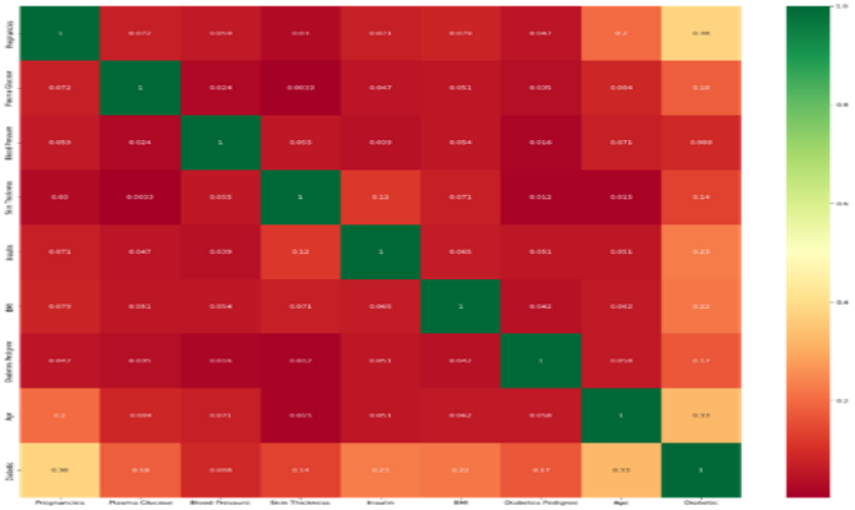


Figure 9: After pre-processing correlation between input and output attributes.

Table5 Confusion matrices for DT, KNN, RF, NB, AB, LR, SVM classifier.

Test method	LR			KNN			SVM			NB			DT			RF			AB		
		0	1		0	1		0	1		0	1		0	1		0	1		0	1
Train/test splitting	0		413	0		221	0	312	941	0	305	849	0		235	0		128	0		292
		311			330									335			340			326	
	5			7									0			0			2		
	1	764	1039	1	587	1216	1	772	1014	1	747	1036	1	272	1474	1	196	1607	1	281	1496

RESULT AND DISCUSSION

Results for ML method DT, KNN, RF, NB, AB, LR, SVM

The machine learning algorithm's accuracy can be calculated from the confusion matrix. In the abstract term, the confusion matrix is given below:

	PredictedNo(0)	Predicted Yes
(1)ActualNo(0)	TN	FP
ActualYes(1)	FN	TP

Here, FP = False Positive, FN = False Negative, TN = True Negative, and TP = True Positive. These are used to calculate classification method's performance measurement.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FN + FP)}$$

Recall**:** Recall is the number of correct positive results divided by the number of *a* relevant samples. In mathematical form it is given as-

$$\text{Recall} = \frac{(TP)}{(TP+FP)}$$

Precision: Precision is the number of correct positive results upon the number of positive results predicted by the classifier. It is expressed as-

$$\text{Precision} = \frac{(TP)}{(TP+FP)}$$

F1 score- F1- score is used to measure a test's accuracy. It's Harmonic Mean between precision and recall. The range is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$\text{F - measure} = \frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})}$$

Confusion matrix of DT, KNN, RF, NB, AB, LR, SVM classifier for Train/Test splitting is shown in Table 5. The performance measure value of all the classification algorithm used on the dataset is shown in Table 6.In Table6, we can see that the all classification methods have accuracy above 75%. Moreover, RF and DT both methods are showing that the accuracy is better for both testing methods.

All classifier's performance based on the differentmeasures with train/test splitting methods is plotted via a graph in Fig 10. In Fig12 the ratio of Diabetic and Non Diabetic patients is shown in the form of pie chart. The below figure is the graphical representation of all the attributes present in the dataset plotted individually Fig 13.

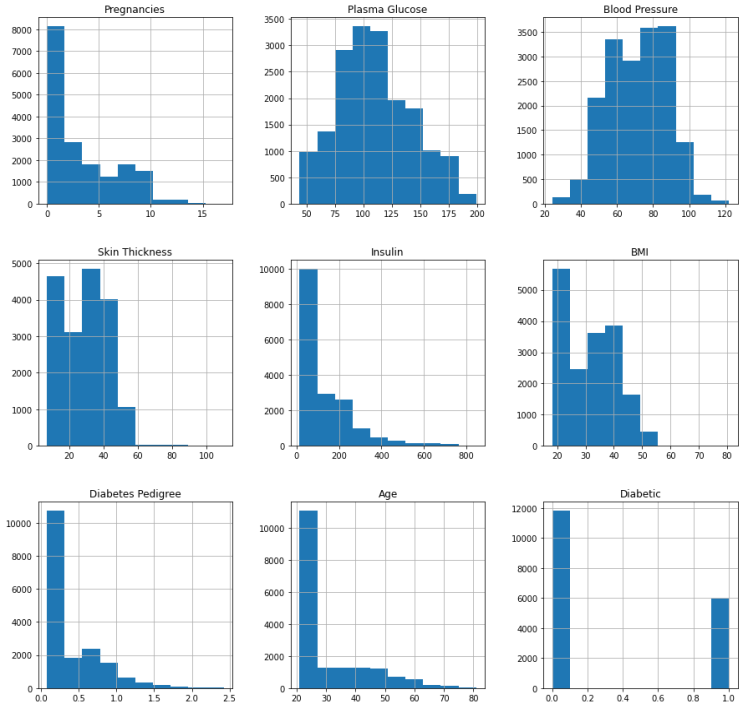
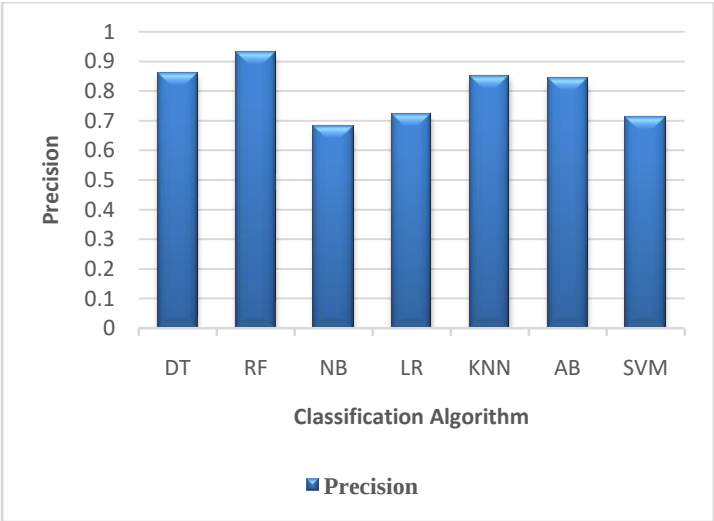


Figure 11: Graphical Representation of the Attributes in Dataset

Table 6 The performance measure of all classification methods for Train/Tests plitting method.

Classification	Precision	Recall	F-measure	Accuracy
DT(0)	0.92	0.93	0.93	90%
(1)	0.86	0.84	0.85	
RF(0)	0.95	0.96	0.95	94%
(1)	0.93	0.89	0.91	
NB(0)	0.80	0.86	0.83	77%
(1)	0.68	0.58	0.63	
LR(0)	0.80	0.88	0.84	78%
(1)	0.72	0.58	0.64	
KNN(0)	0.85	0.94	0.89	85%
(1)	0.85	0.67	0.75	
AB(0)	0.92	0.92	0.92	89%
(1)	0.84	0.84	0.84	
SVM(0)	0.80	0.88	0.84	78%
(1)	0.71	0.57	0.63	



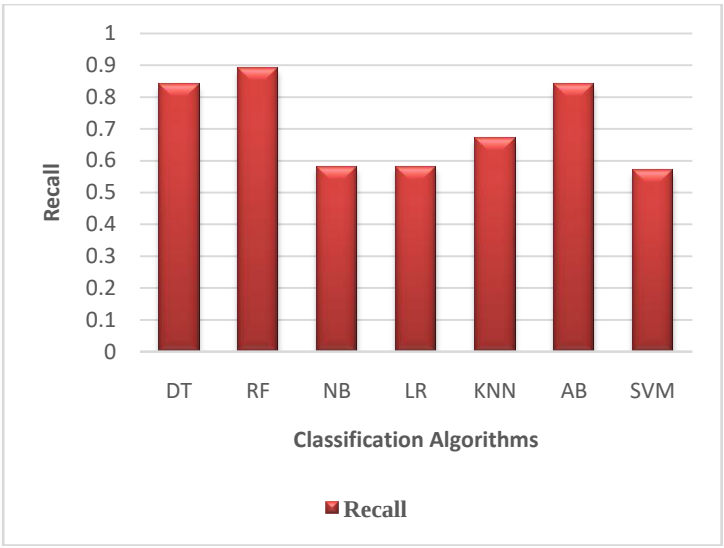
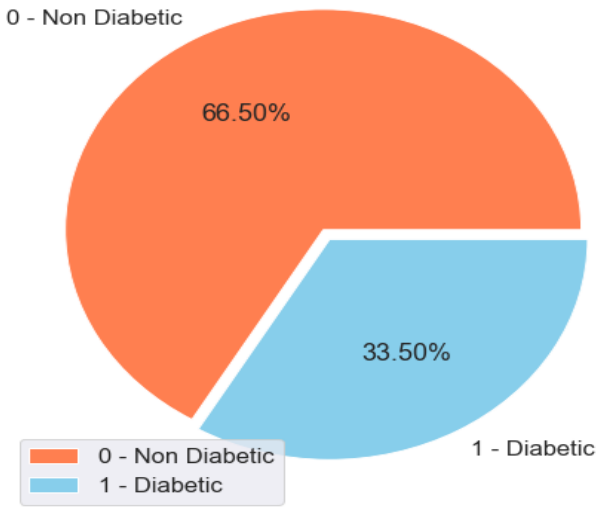
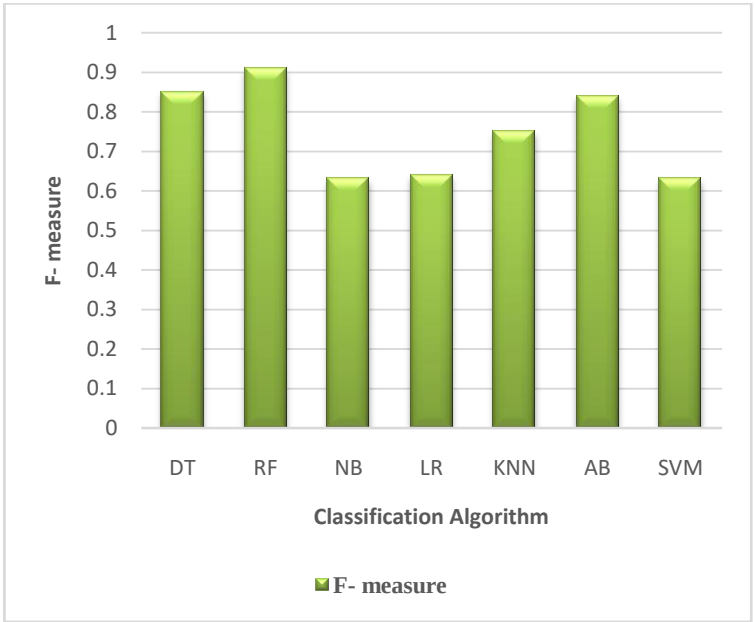


FIGURE 12: RATIO OF NO. OF DIABETIC AND NON- DIABETIC PATIENTS





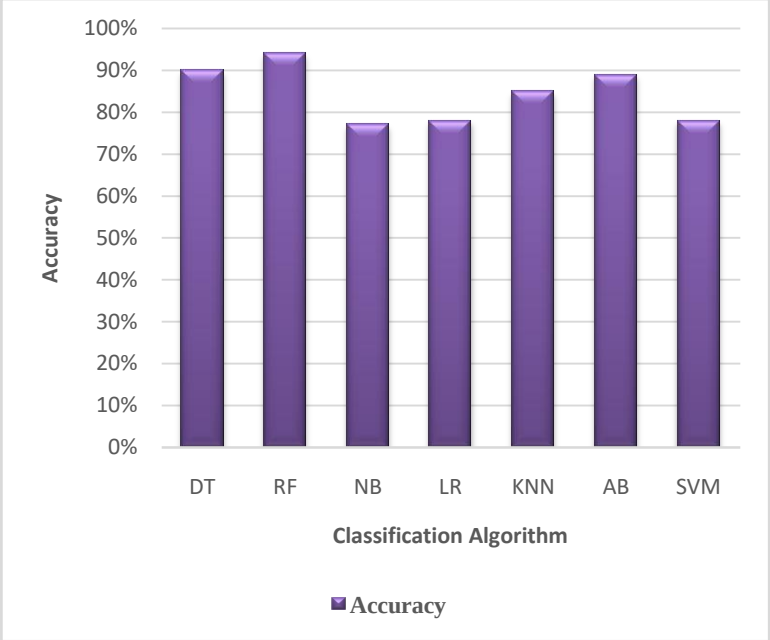


Fig 13: Graphical presentation of the performance of classifier with train/test splitting method.

CONCLUSION

In this paper, we have inspected the execution of seven machine learning algorithms which are namely SVM, Logistic Regression, Ad boost, Random Forest, Naïve Bayes, KNN and Decision Tree. The performance measurement is compared in terms of Accuracy, Precision and Recall. Here the study conclude that the Random Forest achieves the higher test accuracy of 93.30% than other classifiers. All the models shows accuracy greater than 75%. . This study can be used to select best classifier for predicting diabetes.

The accuracy found are SVM (77.7%), Logistic Regression (77.9%), Ad aboost (89%), Random Forest (93.30%), Naïve Bayes (76.79%), KNN (84.84%) and Decision Tree (90.26%).

REFERENCES

[1]. Stoklasa, R.; Majtner, T.; Svoboda, D. Efficient k-NN based HEp-2 cells classifier. Pattern Recognit. 2014, 47, 2409–2418.
[2]. https://www.sciencedirect.com/science/article/pii/S1877050921015350
[3]. https://www.mdpi.com/2227-7390/9/15/1817/htm
[4]. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
[5]. AiswaryaIyer, S. Jeyalatha and RonakSumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[6]. Marco-Antonio Moreno-Ibarra, "Classification of Diseases Using Machine Learning Algorithms: A Comparative Study" , International Journal of Pure and Applied Mathematics, 2021
[7]. Swati Chauhan, Sanjeev Kumar Prasad, "A Comparative Study of Early Detection of Diabetes Risk by Machine Learning" , Springer, Singapore 2021
[8]. Daag Singh, E JebamalarLeavline, "Diabetes prediction using medical data" Journal of Computational Intelligence in Bioinformatics 10 , 2017
[9]. Md. Aminul Islam, NusratJahan, " Prediction of Onset Diabetes using Machine Learning Techniques", International Journal of Computer Applications, 2017
[10]. Hasan Temurtas, NejatYumusak, Feyzullah Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with applications 36, 2009
[11]. Vincent Sigillito, "Pima Indians Diabetes Database", National Institute of Diabetes and Digestive and Kidney Diseases, 1990
[12]. Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood, Random Forests and Decision Trees, International Journal of Computer Science, 2012

[13]. Yao,H.,Hamilton, H.J., Buzz, C.J, A foundational Approach to mining itemset utilities from databases, In 4th SIAM Inter-national Conference on Data Mining, Florida USA,2004
[14]. N Yuvaraj, K R Sripreetha Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster, 2019
[15]. Amardip Kumar Singh, "A Comparative Study on Disease Classification using Machine Learning Algorithms", Jawaharlal Nehru University, March 11, 2019
[16]. https://www.niddk.nih.gov/healthinformation/diabetes/overview/symptoms-causes.
[17]. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-earning-ai/.