IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Disease Prediction By Using Machine Learning Algorithms

V.Yamuna[1],D.Ushanthi[2,],B.Krishna chaitanya[3],Y.Divya sri[4],T.Jagadish[5]

Assistant professor[1],B.Tech Students [2,3,4,5]

Dept.Of Computer Science and Engineering

Aditya Institute of Technology and management, Tekkali

Srikakulam, Andhra pradesh,532201,India

Abstract: This paper deals with the prediction of Diabetes Disease by using SVM, KNN, Decision Tree, LR, Random Forest Classifiers. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 90%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

Keywords: Diabetes disease, Machine learning(ML), SVM, KNN, Decision Tree, LR, Random Forest Classifiers.

1. INTRODUCTION

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers. A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using

unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

1.1 How does Machine Learning Work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the learning and inference. First of all, the machine learns through the discovery of patter507ns. This discovery is made thanks to the data. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a featurevector**.** You can think of a feature vector as a subset of data that is used to tackle a problem.

The machine uses some fancy algorithms to simplify the reality and transform this discovery into a model. Therefore, the learning stage is used to describe the data and summarize it into a model.

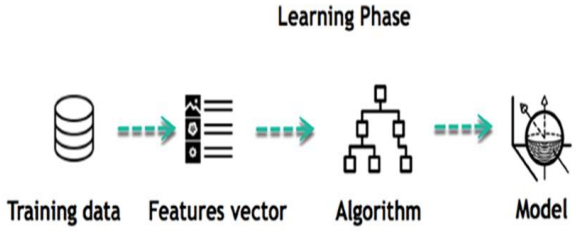


Fig.1.1 Machine Learning Working

For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model

1.1.1 Inferring

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

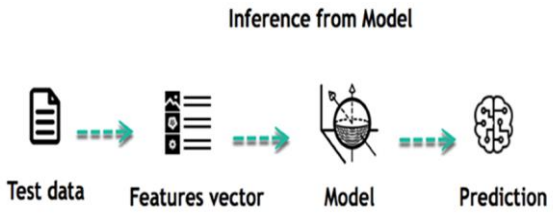


Fig. 1.1.1 Inference from Model

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

1.2 Machine Learning Algorithms

In this paper, we have used five different classification algorithms, they are KNN, SVM, Decision Tree, Random Forest, Logistic Regression.

1.2.1 K-Nearest Neighbor

The k-closest neighbor's algorithm (k-NN) is a non-parametric strategy utilized for arrangement and regression. In both cases, the information comprises of the k nearest preparing models in the component space. The yield relies upon whether k-NN is utilized for grouping or relapse. In k-NN grouping, the yield is class participation. An article is characterized by a majority vote of its neighbors, with the item being relegated to the class most basic among its k closest neighbors (k is a positive number, ordinarily little). In the event that k = 1, at that point the article is essentially appointed to the class of that solitary closest neighbor. In k-NN relapse, the yield is the property estimation for the item. This esteem is the normal of the estimations of its k closest neighbors. K-NN is a type of instance based learning or lazylearning, where the capacity is just approximated locally and all calculation is conceded until order. The k-NN calculation is among the least complex of all AI algorithm. Both for order and relapse, a valuable procedure can be utilized to allot weight to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more removed ones. For instance, a typical weighting plan comprises in giving each neighbor a weight of 1/d, where d is the separation to the neighbors. The neighbors are taken from a lot of articles for which the class (for k-NN arrangement) or the item property estimation (for k-NN relapse) is known. This can be thought of as the preparation set for the calculation, however no unequivocal preparing step is required. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

1.2.2 Support Vector Machine

In support-vector machines (SVMs, likewise support-vector networks) are managed to learn models with related learning calculations that break down information utilized for grouping and relapse examination. Given a lot of preparing precedents, each set apart as having a place with either of two classes, a SVM preparing calculation constructs a model that allocates new guides to one classification or the other, making it a non-probabilistic paired straight classifier (despite the fact that techniques, for example, Platt scaling exist to utilize SVM in a probabilistic characterization setting). An SVM display is a portrayal of the models as focuses in space, mapped with the goal that the instances of the different classifications are separated by an unmistakable hole that is as wide as could be expected under the circumstances. New models are then mapped into that equivalent space and anticipated to have a place with a class dependent on which side of the hole they fall.

In addition to performing straight characterization, SVMs can effectively play out a non-direct grouping utilizing is known as the kernel trick, certainly mapping their contributions to high-dimensional element spaces [13]. At the point when information is unlabelled, regulated learning is absurd, and an unsupervised learning approach is required, which endeavors to discover characteristic bunching of the information to gatherings, and afterward map new information to these framed gatherings.

1.2.3 Decision Tree

Decision tree is a decision support tool that uses a tree-like model of decision and their conceivable results, including chance occasion results, asset expenses, and utility. It is one approach to show a calculation that just contains contingent control articulations. Decision trees are generally utilized in tasks look into, explicitly in choice examination, to help distinguish a system destined to achieve an objective, but at the same time, a well known instrument in machine learning [6].

Although a good number of traditional classification methods for breast cancer are proposed by many researchers, for the first time, Zhang et al. (2000) [12] realized that artificial neural network (ANN) models are alternative to various conventional classification methods which are based on statistics. ANNs are capable of generating complex mapping between input and the output space and thus they can form arbitrarily complex nonlinear decision boundaries. Along the way, there are already several artificial neural networks, each utilizing a different form of learning or hybridization. As compared to higher order neural network, classical neural networks (Example: MLP) are suffering from slow convergence and unable to automatically decide the optimal model of prediction for classification. In the last few years, to overcome the limitations of conventional ANNs, some researchers have focused on higher order neural network (HONN) models used for better performance.

1.2.4 Random Forest

Random forest or random decision are a ensemble learning technique for arrangement, relapse and different errands that works by building a huge number of choice trees at preparing time and yielding the class that is the method of the classes (grouping) or mean expectation (relapse) of the individual trees [8][9]. Random choice backwoods right for choice trees' propensity for over fitting to their preparation.

The primary calculation for arbitrary choice timberlands was made by Tin Kam Ho utilizing the irregular subspace method, which, in Ho's detailing, is an approach to execute the "stochastic separation" way to deal with order proposed by Eugene Kleinberg. An augmentation of the calculation was created by Leo Bremen and Adele Cutler, who registered [10] "Irregular Forests" as a trademark (starting at 2019, possessed by Minitab, Inc.). The expansion joins Breiman's "stowing" thought and arbitrary choice of highlights, presented first by Ho and later

autonomously by Amit and Geman [11] so as to build an accumulation of choice trees with controlled change.

1.2.5 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Logistic regression is a classification model in machine learning, extensively used in clinical analysis. It uses probabilistic estimations which helps in understanding the relationship between the dependent variable and one or more independent variables. Diabetes, being one of the most common diseases around the world, when detected early, may prevent the progression of the disease and avoid other complications. In this work, we design a prediction model, that predicts whether a patient has diabetes, based on certain diagnostic measurements included in the dataset, and explore various techniques to boost the performance and accuracy.

2. RELATED WORKS

Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and SoumenMoulik[1]- This paper deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e. K-Nearest Neighbors, Naive Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction. Index Terms—Diabetes disease, Machine Learning (ML), Disease risk analysis, Confusion Matrix,Scikit-learn, Body mass Index (BMI), Precision, Recall, F1-Score, Pandas, NumPy and Python.

Abbas, Hasan [2]- In this paper, we revisit the data of the San Antonio Heart Study, and employ machine learning to predict the future development of type-2 diabetes. To build the prediction model, we use the support vector machines and ten features that are well known in the literature as strong predictors of future diabetes. Due to the unbalanced nature of the dataset in terms of the class labels, we use 10-fold cross-validation to train the model and a hold-out set to validate it. The results of this study show a validation accuracy of 84.1% with a recall rate of 81.1% averaged over 100 iterations. The outcomes of this study can help in identifying the population that is at high risk of developing type-2 diabetes in the future. Index Terms—Disease Prediction, support vector machine, type 2 diabetes.

Kaur, Harleen, and Vinita Kumari[3]- Diabetes is a major metabolic disorder which can affect entire body system adversely. Undiagnosed diabetes can increase the risk of cardiac stroke, diabetic nephropathy and other disorders. All over the world millions of people are affected by this disease. Early detection of diabetes is very important to maintain a healthy life. This disease is a reason of global concern as the cases of diabetes are rising rapidly. Machine learning (ML) is a computational method for automatic learning from experience and improves the performance to make more accurate predictions. In the current research we have 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 have 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 neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR).

Sisodia, Deepti, and Dilip Singh Sisodia[4]- Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

Islam, MdAminul, and NusratJahan[5]- Machine learning algorithms can help us to detect the onset diabetes. Early detection of diabetes can reduce patient's health risk. Physicians, patients, and patient's relatives can be benefited from the prediction's outcomes. In low resource clinical settings, it is necessary to predict the patient's condition after the admission to allocate resources appropriately. Several articles have been published analyzing Prima Indian data set applying on various machine learning algorithms. Shankar applied neural networks to predict the onset of diabetes mellitus on Prima Indian Diabetes dataset and showed that his approach for such classification is reliable [4, 5 and 6]. Machine learning techniques increase medical

diagnosis accuracy and reduce medical cost [2, 3]. In this study, the main focus is to investigate different types of machine learning classification algorithms and show their comparative analysis. The purpose of this study is to detect the diabetic patient's onset from the outcomes generated by machine learning classification algorithms.

P. S. Kohli and S. Arora[6]- The application of machine learning in the field of medical diagnosis is increasing gradually. This can be contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early detection of fatal diseases and therefore, increase the survival rate of patients significantly. In this paper, we apply different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases.

J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes[7]- Neural networks or connectionist models for parallel processing are not new. However, a resurgence of interest in the past half decade has occurred. In part, this is related to a better understanding of what are now referred to as hidden nodes. These algorithms are considered to be of marked value in pattern recognition problems. Because of that, we tested the ability of an early neural network model, ADAP, to forecast the onset of diabetes mellitus in a high risk population of Pima Indians. The algorithm's performance was analyzed using standard measures for clinical tests: sensitivity, specificity, and a receiver operating characteristic curve. The crossover point for sensitivity and specificity is 0.76. We are currently further examining these methods by comparing the ADAP results with those obtained from logistic regression and linear perceptron models using precisely the same training and forecasting sets. A description of the algorithm is included.

A. Mir and S. N. Dhage[8]- Healthcare domain is a very prominent research field with rapid technological advancement and increasing data day by day. In order to deal with large volume of healthcare data we need Big Data Analytics which is an emerging approach in Healthcare domain. Millions of patients seek treatments around the globe with various procedure. Analyzing the trends in treatment of patients for diagnosis of a particular disease will help in making informed and efficient decisions to improve the overall quality of healthcare. Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. This paper aims at building a classifier model using WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. The research hopes to recommend the best algorithm based on efficient performance result for the prediction of diabetes disease. Experimental results of each algorithm used on the dataset was evaluated. It is observed

that Support Vector Machine performed best in prediction of the disease having maximum accuracy.

Alam, TalhaMahboob[9]- Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. Our findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the Apriori method. Artificial neural network (ANN), random forest (RF) and K-means clustering techniques were implemented for the prediction of diabetes. The ANN technique provided a best accuracy of 75.7%, and may be useful to assist medical professionals with treatment decisions.

Selvakumar, S., K. SenthamaraiKannan, and S. GothaiNachiyar[10]- Data mining is the search of large datasets to extract hidden and previously unknown patterns. Classification is the one of the task in data mining. Health care data are often huge, Complex and heterogeneous because it contains different variable types. Nowadays, knowledge from such data is a necessity. Data mining can be utilized to extract knowledge by constructing models from healthcare data such as diabetic patient sets. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using data mining methods to aid people to predict diabetes has gain major popularity. In this paper, to predict the persons whether diabetic or not. In this paper classification techniques such as Binary Logistic Regression, Multilayer Perceptron and K-Nearest Neighbor are classified for diabetes data and classification accuracy were compared for classifying data.

3. METHODOLOGY

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. In this project we aim to develop a prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Random Forest Classifier.Classification is one of the most important decision making techniques in many real world problem. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problem, the higher number of samples chosen but it doesn't leads to higher classification accuracy. In many cases, the performance of algorithm is high in the

context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Support Vector Machine, Logistic Regression, and Artificial Neural Network are most suitable for implementing the Diabetes prediction system.

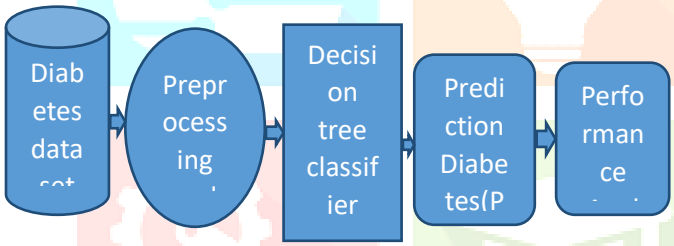


Fig 3.1: Architecture of Proposed System

A.Data Collection

Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided “logical time” slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00). Thus paper records have fictitious uniform recording times whereas electronic records have more realistic time stamps. Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

Future	Description
Pregancies	Number of Pregnancies patients had earlier.
Glucose	Glucose level present in the patient.
Blood Pressure	Recorded blood pressure level at that particular time.
Skin Thickness	Skin thickness level of the patient.
Insulin	Amount of Insulin present in the body.
BMI	Body Mass Index of the individual.
Diabetes Pedigree Function	Family history of Diabetes disease.
Age	Age of an individual.

• Pregnancies:

Those who develop gestational diabetes are at higher risk of developing type 2 diabetes later in life. The subjects with more number of pregnancies have a higher risk of developing diabetes.

•Glucose:

The subjects were given an oral glucose test, whereby, they were administered glucose and a reading of their plasma glucose concentration was taken after 2 hours. The subjects with higher levelsof glucose concentration after 2 hours have a higher risk of developing diabetes.

• Blood pressure:

Having blood pressure over 140/90 mmHg of Mercury are linked to having increased risk of developing diabetes. Although, certain subjects having diastolic blood pressure 70 mmHg may develop diabetes.

• Skin Thickness:

Skin thickness is primarily determined by collagen content and is increased in the case of insulin dependent diabetic patients. The subjects' tricep skin fold were measured and results showed that having a skin thickness of 30mm of greater are at a higher risk.

• Insulin:

Normal insulin levels after 2 hours of glucose administration is 16-166 mlU/L. Subjects having lower or higher levels than said value are at a higher risk.

• Body Mass Index (BMI):

Subjects having a BMI over 25 have a relatively high risk in having diabetes.

B.Data Pre-processing

The dataset, which is quoted above, has lapsed and have shed data. To make the dataset serviceable and obtain the knowledge from it, we have performed data preprocessing. In order to handle erroneous data, we have analyzed the dataset for the unusual entries and fixed them manually. Missing values are handled with the help of calculating the standard deviation of that particular feature and allotting it to the missing spaces. To make the dataset useful, we have used Pandas [6] and NumPy [7] library for handling the dataset efficiently and easy data handling throughout the experiment.

C. Setting Classification Metrics

To classify disease and get a prediction result, we need to set a few metrics which will help us in predicting the Diabetes disease. Since we are using scikit-learn (Sklearn) machine learning library [8] for our experiment, we have used confusion matrix as the classification measure metrics. All the used metrics, i.e. Precision, Recall, F1-Score and Accuracy in our analysis, are listed below

• Precision (P) is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp). Mathematically,

$$P = TP/(TP+FP)$$

• Recall (R) is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

$$R = TP/(TP+FN)$$

A ROC curve (receiver operating characteristic curve) is a diagram explaining the completion of a classification model at all classification thresholds.

This curve plots two parameters:

- TPR=True Positive Rate
- FPR=False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows: TPR=TP/(TP+FN)

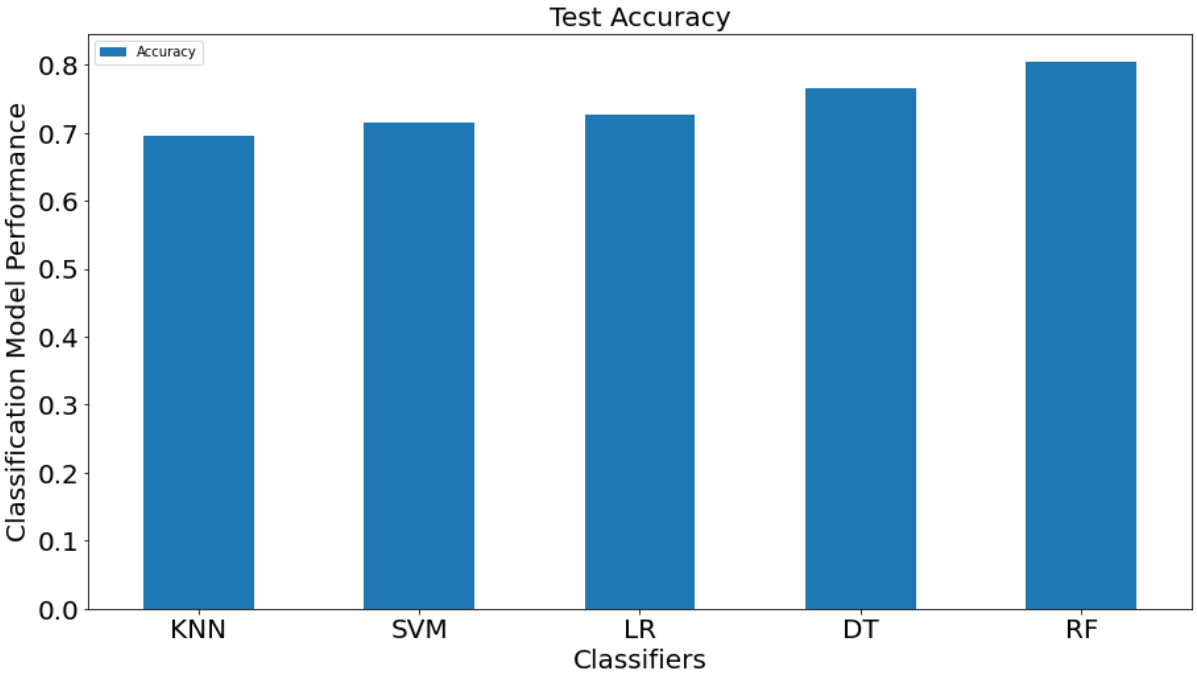
False Positive Rate (FPR) is defined as follows: FPR=FP/(FP+T

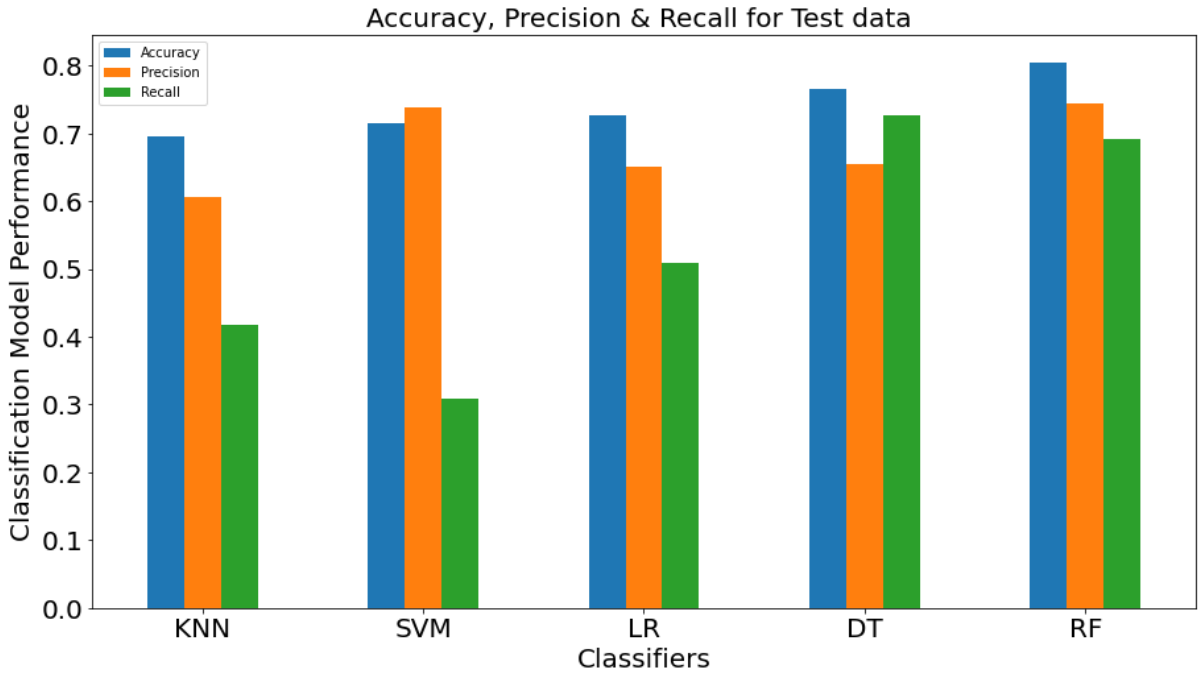
4. EXPERIMENTAL RESULTS

The developing environment for the proposed method is Spyder(python)- 3.9 version on a system with the Intel or AMD x86-64 proccessor, 1.8GHz, 8GB Ram and Microsoft window Family.In Experimental setup we are explaining about the Data Set i:e Diabetes Disease Prediction Using Machine Learning Algorithms, Spyder.

Classifier	**Accuracy**	**Precision**	**Recall**
Random Forest	80%	74%	50%
SVM	71%	73%	30%
DeisionTree	76%	65%	72%
LR	72%	65%	50%
KNN	69%	60%	41%

In this project, we have used some of the algorithms like SVM, KNN, Random Forest, Decision Tree, LR. From these algorithms, we got the Random Forest accuracy is highest 80%. In this project, we have used accuracy, precision, recall as performance matrics.





5. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage.The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. However, in this study, systematic

efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. The proposed approach uses various classification and ensemble learning method in which SVM, KNN, Random Forest, Decision Tree, Logistic Regression are used. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 80% using the Random Forest Classifier. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well. This project results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future.

REFERENCES

[1]- Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and SoumenMoulik. "Diabetes Disease Prediction Using Machine Learning Algorithms." *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2021.

[2]- Abbas, Hasan, et al. "Predicting diabetes in healthy population through machine learning." *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019.

[3]- Yahyaoui, Amani, et al. "A decision support system for diabetes prediction using machine learning and deep learning techniques." *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE, 2019.

[4]- Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science* 132 (2018): 1578-1585.

[5]- Islam, MdAminul, and NusratJahan. "Prediction of onset diabetes using machine learning techniques." *International Journal of Computer Applications* 180.5 (2017): 7-11.

[6] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[7] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, “Using the adap learning algorithm to forcast the onset of diabetes mellitus,” Proceedings - Annual Symposium on Computer Applications in Medical Care, vol. 10, 11 1988.

[8] A. Mir and S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.

[9]Alam, TalhaMahboob, et al. "A model for early prediction of diabetes." *Informatics in Medicine Unlocked* 16 (2019): 100204.

[10]Selvakumar, S., K. SenthamaraiKannan, and S. GothaiNachiyar. "Prediction of diabetes diagnosis using classification based data mining techniques." *International Journal of Statistics and Systems* 12.2 (2017): 183-188.