International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Effective Prognosis of Diabetes Using Machine Learning Techniques

Dr. Atheeq C[1], Dr. G. Shree Devi[2], Rajeswari. P[3], Dr. Layak Ali[4], Dr. M. A Rabbani[5]
[1]Assistant Professor, Computer Science and Engineering Department
GITAM University
Hyderabad, India.
atheeq.prof@gmail.com
[2]Assistant Professor, Department of Computer Applications
B. S. Abdur Rahman Cresent Institute of science and technology
Chennai, India
shreedevi@crescent.education
[3]Assistant professor
R.M.D Engineering College
Chennai, India.
prajimtech@gmail.com
[4]Assistant Professor, Electronics and Communication Engineering Department
Central University of Karnataka, Karnataka
layakali@cuk.ac.in
[5]Professor, Department of Computer Applications
B. S. Abdur Rahman Cresent Institute of science and technology
Chennai, India
marabbani@crescent.education

Abstract—Diabetes has emerged as a primary disease in most of the nations like China, India and United States. unchecked Diabetes leads to serious health issues which can cause damage to human tissues and organs. The presence of high sugar quantity in the blood stream is the specific cause of diabetes. However, this disease isn't curable and can solely be controlled. If it is not medicated, it will cause many difficulties. This difficulty might end up in death. Severe difficulties lead to foot sores, cardiovascular disease and eye blurriness. The aim of this project is to build a system which could predict the patient's diabetic also called sugar risk level with a higher accuracy. Model development is based on categorization algorithm like random forest, logistic regression and k-nearest neighbour algorithms. The performance of each algorithm is analysed and the model with highest accuracy is chosen for prediction of diabetes.

Keywords- Random Forest, Machine Learning, Analysis and prediction, Logistic regression-nearest neighbour

I. INTRODUCTION

Diabetes has become a common chronic disease in the world. It results due to increased level of glucose in the blood than the normal value. The increase in the glucose level is a result of defective insulin (a hormone emitted from the pancreas that help cells of the muscles, fat to absorb glucose present in the blood) secretion or other biological effects. Diabetes leads to many other health problems like disfunction and damaging of various tissues, particularly heart, kidneys, eyes, nerves and heart.

It is broadly divided into two types, first is Type 1 Diabetes and second is Type 2 Diabetes. Immune system of the Patients with first type diabetes destroys the cell of the pancreas which forms insulin. It is generally tested in kids and young adults, however it can emerge at any age. Excessive thirst and frequent urination, as well as high levels of blood glucose, are the most prevalent indications and symptoms of diabetes. Insulin treatment is essential for survival in patients with this kind of diabetes because medication alone is not sufficient to reverse the disease. The second form of diabetes develops when a patient develops insulin resistance, leading to an increase in blood sugar levels. This happens when the patient's body becomes resistant to insulin. Hereditary diabetes refers to a condition that occurs in a person's family and is passed down from generation to generation. It is more common in middle-aged and elderly adults, who frequently have health conditions such as obesity, hypertension, cholesterol, arteriosclerosis, and other disorders [1].

Diabetes has become an increasingly common ailment in people's day-to-day lives as a result of an increase in the number of developments that have taken place inside the living requirements of humans. Therefore, a method that is both quick and accurate in assessing and diagnosing diabetes is a topic worthy of further investigation. Sugar checking out is done in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

accordance with random blood glucose stage, fasting blood glucose, and glucose tolerance testing in the majority of hospitals in today's modern world. Patients are required to travel to a diagnostic center, talk with their primary care physician, and then relax for at least one day in order for their test results to become definitive [2]. This is the conventional method for distinguishing patients. Moreover, whenever they need to induce their diagnosing report, they need to waste their cash vainly. At the same time every person who is not suffering diabetes should go for a check-up. As the number of victims is more, quick treatment is required. Diabetes can be easily controlled if the diagnosis is obtained as early as possible. So, fast testing is required as the population is more [3].

When a blood glucose test is done, it usually takes a few days to obtain the result. Whereas, by using an algorithm we can get the result of not only one patient but many within no time. This makes a clear way for checking our health status. Many patients wait in queue in front of hospitals, which is difficult for them as well the medical staff to maintain such huge number of people. AI can serve as a source of perspective for professionals, and it can also assist people in making a preliminary determination about diabetes based on the results of their regular actual assessment data [4].

Currently, a number of different machine learning algorithms have been developed to predict diabetes. These algorithms integrate standard methods of machine learning, such as Naive Bayesian, support vector machine (SVM), decision tree (DT), and so on. By utilizing these various approaches to machine learning, preferable results can be acquired for the purpose of diabetes prediction. The decision tree technique is one of the most extensively used machine learning algorithms, and it possesses grateful type power. This technique is employed in the therapeutic field. In this work, we determine the sugar level by utilizing the Random Forest algorithm, the Logistic regression algorithm, and the okay closest neighbor algorithm. The accuracy that was achieved with the use of Random Forest is above 90 percent. When contrasted with other machine learning algorithms for diabetes prediction, this demonstrates a higher level of accuracy.

II. LITERATURE REVIEW

In [5] proposed the SVM system mastery algorithm has been developed in this paper in order to forecast diabetes. Python is used as the programming language for the implementation, and an information set is used for testing the SVM method. A training component and a trying out component have been separated from the dataset. Next, the SVM version is educated as a direct consequence of the outcome. The performance of SVM kernels is analyzed and compared in this research. 0020 The model is trained on the four distinct kernels that are available for use with SVM, and its prediction accuracies are determined with the help of the testing set [6]. The SVM is tested using four different kernels, which might be either a linear kernel, a polynomial kernel, a sigmoid kernel, or an RBF kernel. For the purpose of diabetes prediction, the first-rate SVM kernel was chosen and utilized.

The Levenberg-Marquardt education set of rules was utilized by to grade overall performance. The community model underwent intensive training on multiple occasions in order to achieve an accuracy rate of more than 80 percent. There were 768 people included in the dataset; 268 of them had diabetes, whereas the other 500 did not have diabetes. The proposed version produced satisfactory results, with an accuracy of 82 percent, when applied to this non-uniform dataset. In subsequent research, it will be necessary to broaden the scope of this investigation by enhancing the teaching methodology and modifying the activation function using a deep neural community in order to more accurately predict the onset of DM at an earlier stage [7].

A predictive model was proposed by [8] makes use of factors that are comparable to the uncertainty associated with type 2 diabetes. The data came from a database that was kept in Tabriz, Iran, and it was devoted to sugar control methods. All of the information pertaining to diabetic screening patients who were referred for care between the years 2009 and 2011 was compiled into a database. In order to strengthen the model, the "Choice Tree" technique and the "J48" computation were implemented into the WEKA (3.6.10 adaption) programming. Following the preprocessing and planning of the information, information mining was performed on a total of 22,398 records. The accuracy of the model to correctly identify patients was 0.717. The age component was moved to the root hub of the tree in order to maximize the amount of information gained [9].

The machine learning techniques developed for predicting a variety of medical data sets, including the diabetic disorders dataset (DDD). In this study, they forecast multiple medical datasets, including the diabetes dataset (DD), by using support vector machines (SVM), logistic regression, and naive bayes with 10 fold cross validation.

The researchers compared the accuracy of several algorithms with their performance, and based on their findings, they came to the conclusion that the SVM (support Vector Machine) algorithm offers the highest level of accuracy compared to the other algorithms that are mentioned on the website above [10 - 12].

Clustering calculation additionally was utilized (head part Analysis (PCA) and Expectation amplification (EM) for pre-handling and commotion eliminating prior to applying the standard. Different clinical dataset (MD) was utilized like bosom malignant growth, Heart, and Diabetes Develop choice help for various sicknesses including diabetes. The outcome was CART with noise removal can provide effective and better in health/diseases prediction and it is possible to safe human life from early death.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

III. PROPOSED SYSTEM

The suggested system employs algorithms from Determine 1. Random wooded area, logistic regression, and okay-nearest neighbor are base-type algorithms.

A. Dataset Description

Data is gathered for the purpose of training advanced machine learning algorithms from a variety of resources, such as the UCI repository or Kaggle. The dataset incorporates numerous reports from a variety of victims, each of which is comprised of the following characteristics.

The 8th attribute of every data point is the class variable. The output of the class variable indicates whether the person is diabetic or not. The outcome 0 indicate negative result whereas the outcome 1 indicate a positive result showing that the person is suffering from diabetes.

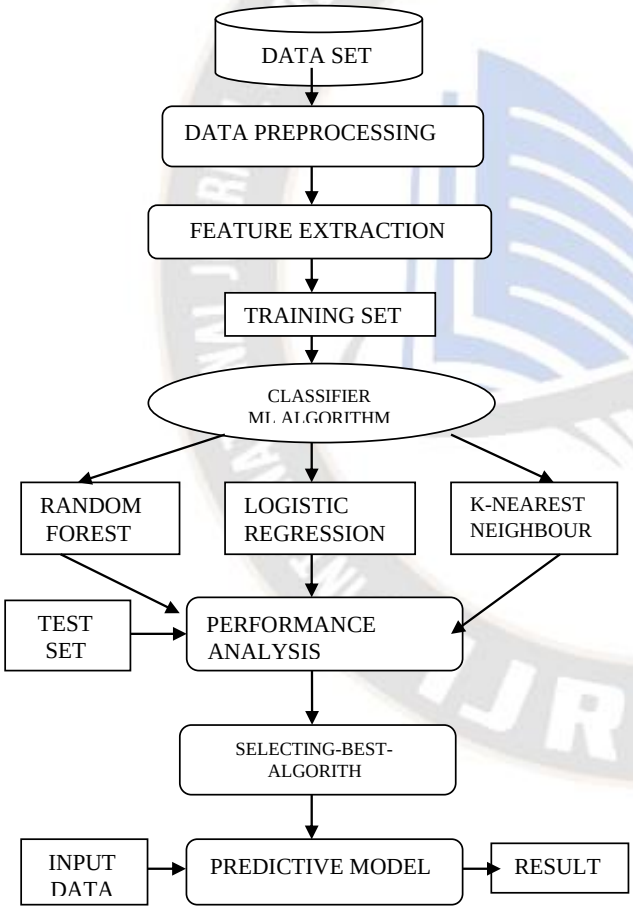

Figure. 1 Block diagram of diabetes prediction system

TABLE I. DATA ATTRIBUTES FOR PATIENT

S. No	Attribute
1	Age
2	Glucose
3	Insulin
4	Blood pressure
5	Skin thickness
6	Body Mass Index
7	Diabetes Pedigree Method

B. Data Preprocessing

Most of the times the raw data that have being obtained from different sources is not completely valuable and efficient to be used by the any machine learning algorithm. Therefore, the data should be processed first in order to increase data quality and effectiveness. It is very important step as it helps in successful prediction and getting accurate results.

The raw data obtained may contain many missing values or irrelevant data. Missing values are handled by replacing them with either the attribute's mean, median or the most frequent value. Irrelevant or inconsistent data are modified to make these records consistent with the order records in the dataset. If any redundant records are present in the dataset these records are simply eliminated. Noise or outliers are detected and they are removed by using either regression or clustering. If the eight attributed of the records contain categorial data like positive or negative they are converted to numeric values because categorical value might cause problem in algorithms that use mathematical equations. So positive is converted to 1 and negative to 0.

C. Feature Extraction

Feature extraction involves reducing the number of attributes that describe a dataset. This is done because a dataset with large number of attributes generally requires huge amount of computation power and also consume a lot of memory which might cause overfitting of training samples leading to less accurate results. Feature extraction creates new attributes from the original ones. This new reduced set of attributes summarizes most of the information contained in the original set of features. Machine learning mechanism is assign to this reduced set of attributes. This increases the efficiency of the classifier and helps in obtaining better and accurate results

D. Splitting of data

Next step is splitting of data. After cleaning, normalizing and feature extraction, data is divided in two parts one is the training set and other is the test set. 75 percent of the total data comes under training set and the rest 25 percent comes under test set. Training set is utilized to make learn various machine learning algorithms individually and the test part is utilized for obtaining the accuracy of the trained model.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

E. Apply Machine Learning

After the information is split the training set is applied to Machine Learning algorithms. The main objective to employ Machine learning strategy is to investigate the overall performance of the algorithms and discover accuracy amongst them, and moreover so that you might parent out the optimum accountable/crucial function that functions a primary duty in prediction of diabetes. The multiple device studying algorithms employed are as follows.

Random Forest

Random Forest was developed by Leo Berman. It is widely used ML algorithm that belongs to the class of supervised learning technique. It is used for classification problems and regression problems. It is the most popular algorithm because it can handle large datasets as well as dataset that contains large number of missing values, it takes less time to train itself and predicts with greater accuracy as compared to other classifiers.

Random forest builds a large no of decision tress using various subset of dataset. These decision tresses are build using the various subsets of the training dataset. Greater the number of decision tree a random forest classifier construct, greater is the accuracy. random forest make prediction by taking prediction from each decision tree constructed and based on the majority it gives the final result.

Algorithm

1. Step one is to select randomly k data records from the dataset.
2. Using the above selected data records construct a decision tree using the best split.
3. Choose the number of such decision trees that have to be constructed by the classifier.
4. Repeat the steps 1-3 for n decision trees
5. For predicting the output of a new data tuple, the prediction from each decision tree constructed by the above steps are found out
6. Select the majority voted output as the final result.

Logistic Regression

Logistic regression belongs to the class of supervised learning algorithm. It is used for only classification problems and predicts the output of categorial dependent variable which can be either discrete or categorial using independent variables. Instead of providing the exact values it provides deterministic values that lie between 0 and 1. In this the concept of threshold is used in which the values that are greater than the specified threshold are 1 and the values that are less than the specified threshold are 0. The sigmoidal mathematical function is used to map the predicted values to probabilities. The s shaped curve called the sigmoidal function of this model predicts the likelihood of something.

The primary objective of logistic regression is to find the model that provides the best explanation for the link that exists between the target variable and the predictor variable.
Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

K-Nearest Neighbour

KNN is one of the simple and easiest ML algorithm. This algorithm is easy to understand and implement. It belongs to the class of a supervised machine learning technique. KNN helps in solving classification problems as well as regression problems.it is a non-parametric algorithm because it doesn't make any assumption about the dataset used.

KNN is called a lazy learner algorithm because when a the training dataset is supplied to it, it doesn't train itself instead it simply memorize the data set. Each time when a prediction has to be made about data record KNN algorithm simply finds the k closest datapoints to the given record in the entire training set. The closest records are found using some distance formula like Euclidean distance. The records maximum class variable is given as the final output.

The Euclidean distance between two points a and b i.e. A (a1,a2, . An) and B (b1, b2,..bn) is defined by the following equation:-

$$\text{Sqrt}((a1-b1)2 +(a2-b2)2+..+(an-bn)2)$$

Algorithm

1. Load the training dataset
2. Choose the no of nearest neighbour to be considered that is choose a positive value k.
3. For each data record in the input set. find the distance of the input set with all the records present in the training set using Euclidean, Manhattan or hamming distance.
4. Sort the distances in increasing manner.
5. Choose the top most k rows from arranged patterns.
6. Depending on recent class of the topmost n rows output the result

F. Performance Analysis

When performing classification predictions, there are four types of outcomes that could occur.

1. True positives are when a model correctly classify a positive sample as positive.
2. True negatives are when a model correctly classify a negative sample as negative or if the system guess wrongly the negative output class of the sample.
3. False positives are when a model incorrectly classify a negative sample as positive or if the system guess wrongly the output class of the sample.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

4. False negatives are when a model incorrectly classify a positive sample as negative or when a model incorrectly predicts the output class of the sample.
These four outcomes are plotted on a confusion matrix.

Metrics

1. Accuracy is termed as the percentage of correct predictions for the test data. It is the ratio of correct predictions to number of total predictions.

Accuracy = correct predictions /all prediction

2. Precision is referred to as the percentage of relevant examples, also known as true positives, out of the total number of examples that were anticipated to belong to a particular class.

Precision = TP/(TP + FP)

3. Recall is the percentage of examples that have been expected to belong to a category in comparison to all of the examples that unquestionably belong inside the group.

Recall = TP/(TP + FN)

4. F1-Score: One of the qualities of Precision and Recall is denoted by the letter F1. It is important to have a high F1 Score since we are looking to strike a balance between precision and recall, which has an uneven distribution of its magnificence..

F1=2* Precision*Recall/(Precision*Recall)

IV. RESULTS AND DISCUSSIONS

In the wake of taking the information dataset the model will foresee the information by applying the ML calculations and give the best bring about the type of correlation between to anticipate the best precision to treat diabetes.

The results are analyzed for various algorithms and their performance is also calculated. The sample data collection for diabet[ps prediction is considered and is listed in table 2 and the correlation between the attributes of data record is given in figure 2.

TABLE II. FIRST 5 ROWS OF DATA COLLECTED

S. No	PG	GL	BP	SK	IN	BMI	BD	AG	OC
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1

PG- Pregnencies
GL- Glucose
BP- Blood Pressure
SK- Skin Thickness
In- Insulin
BMI- Body Mass Index
DP- Diabetics Prediction function
AG- Age
OC- Outcome

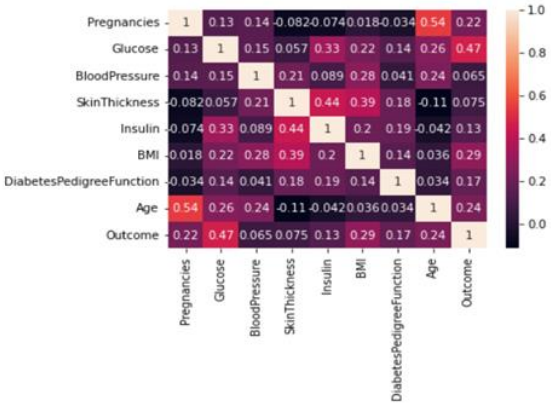

Figure 2. Correlation between the attributes of the data record

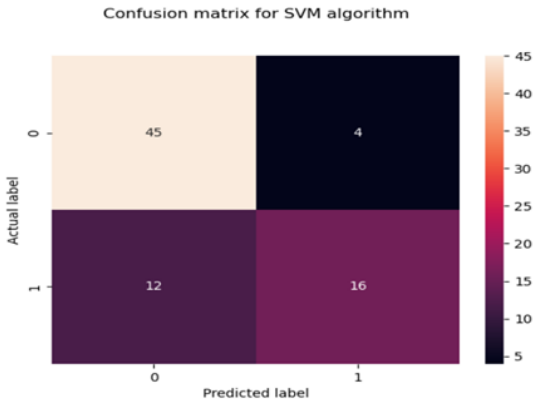

Figure 3. SVM Classifier Confusion Matrix

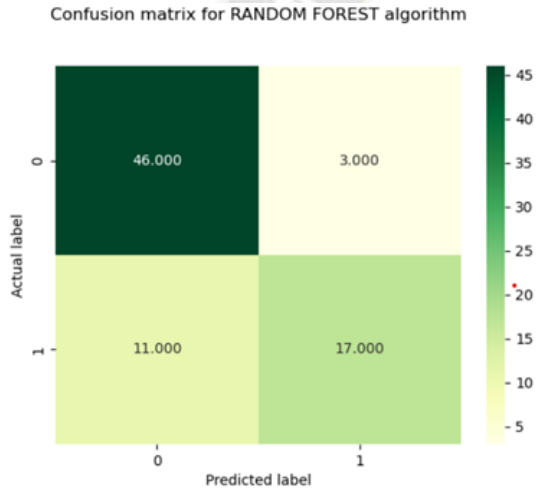

Figure 4. Random Forest Classifier Confusion Matrix

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

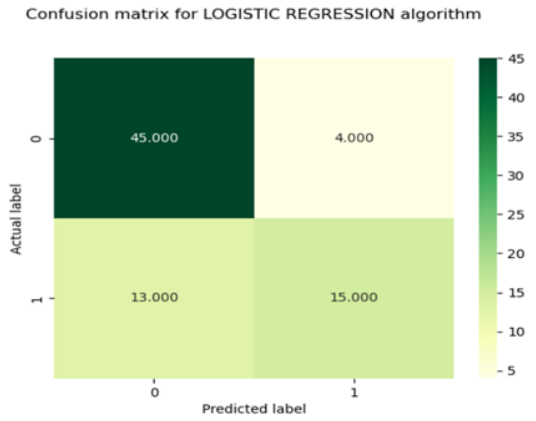

Figure 5. Logistic Regression Classifier Confusion Matrix

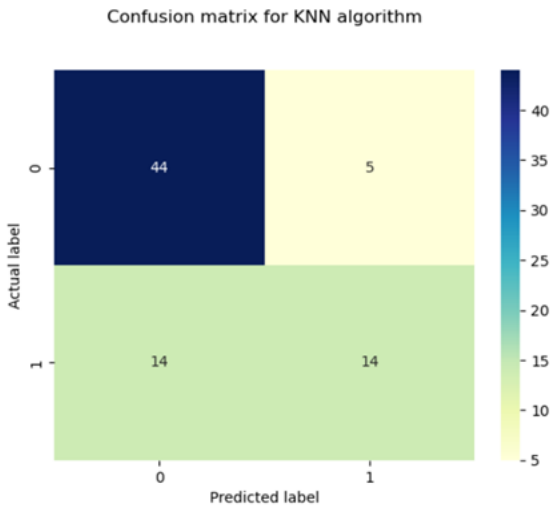

Figure 6. KNN Classifier Confusion Matrix

Figure 7. Naive Bayes Classifier Confusion Matrix

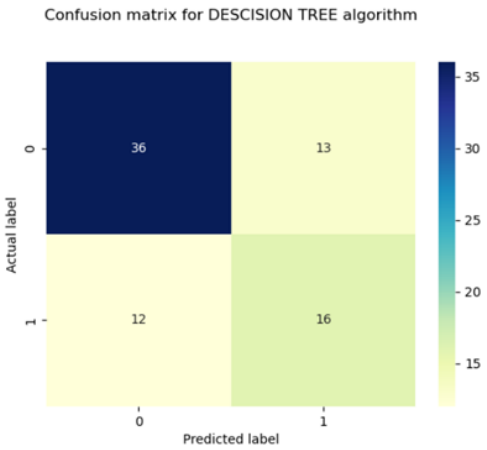

Figure 8. Decision Tree Classifier Confusion Matrix

In all the above figure from 3 to 8, various algorithms confusion matrix is represented having the actual and predicted labels. Ant the accuracy of all algorithms is given in table 3 and is also represented in figure 9.

TABLE III. ACCURACY OF DIFFERENT MACHINE LEARNING ALGORITHMS

S. No	Algo.	Class	Precision	Recall	F1 Score	Support	Accu.
1	SVM	0	0.78	0.90	0.83	155	0.76
		1	0.70	0.49	0.57	76	0.66
2	RF	0	0.86	0.81	0.83	53	0.77
		1	0.63	0.71	0.67	24	0.57
3	LR	0	0.85	0.77	0.81	53	0.75
		1	0.59	0.71	0.64	24	0.52
4	KNN	0	0.87	0.75	0.81	53	0.75
		1	0.58	0.75	0.65	24	0.51
5	NB	0	0.86	0.72	0.78	53	0.72
		1	0.55	0.75	0.63	24	0.48
6	DT	0	0.82	0.62	0.71	53	0.64
		1	0.46	0.71	0.56	24	0.41

SVM- Support Vector Machines
RF- Random Forest
LR- Logistic Regression
KNN- K-Nearest Neighbor
NB- Navie Bayes
DT- Decision Tree

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

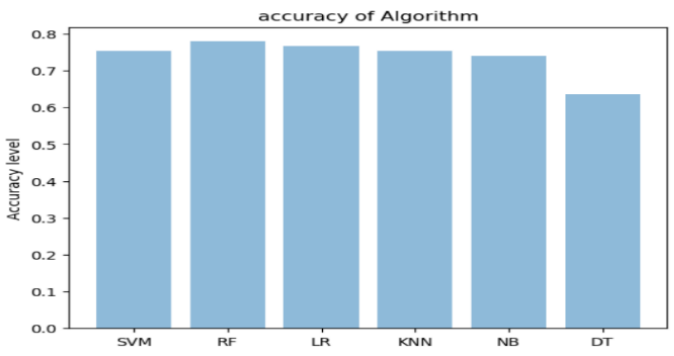

Figure 9. Comparison of Accuracy of all Algorithms

V. CONCLUSION

SVM: When the records are unknown. SVM calculation works well despite unstructured and semi-organized inputs like text, photos, and wood. To achieve remarkable order outcomes using SVM, a few important boundaries must be specified correctly. Choice tree is easy to identify and control. Choice tree instability must be obvious to small record layout adjustments in an acceptable desire tree. Often wrong. Naive Bayes manages missing attributes by ignoring chance estimation. Delicate with data sources. Inclined when preparing datasets. KNN's forecasts are accurate and easy to use. Managing large, complex records takes a long time. Logistic regression delivers viable results when compared to Naive Bayes in terms of accuracy, but it's better for producing widely conventional results. Random Forest has varied impacts with different rules and more accuracy than previous device learning algorithms.

References

[1] Veena Vijayan V. and Anjali C, 2015, 'Prediction and Diagnosis of Diabetes Mellitus A Machine Learning Approach', IEEE Recent Advances in Intelligent Computational Systems. pp 122-127. Prediction and diagnosis of diabetes mellitus — A machine learning approach | IEEE Conference Publication | IEEE Xplore

[2] P. Suresh Kumar and V. Umatejaswi, 2017, 'Diagnosing Diabetes using Data Mining Techniques', International Journal of Scientific and Research Publications, Volume 7, Issue 6, ISSN 2250-3153.

[3] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, 2017 'Application of Machine Learning Algorithms on Diabetic Retinopathy', International Conference On Recent Trends In Electronics Information & Communication Technology. pp 2046-2051. Application of machine learning algorithms on diabetic retinopathy | IEEE Conference Publication | IEEE Xplore

[4] Berina Alic, Lejla Gurbeta and Almir Badnjevic, 2017, 'Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases', Mediterranean Conference On Embeded Computing. pp 1-4. Machine learning techniques for classification of diabetes and cardiovascular diseases | IEEE Conference Publication | IEEE Xplore

[5] Dr. M. Renuka Devi and J. Maria Shyla, 2016, 'Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus', International Journal of Applied Engineering Research, Volume 11, Number 1 (2016) pp 727-730.

[6] Rahul Joshi and Minyechil Alehegn, 2017, 'Analysis and prediction of diabetes diseases using machine learning algorithm' Ensemble approach, International Research Journal of Engineering and Technology. Volume: 04 Issue: 10. irjet.net/archives/V4/i10/IRJET-V4I1077.pdf

[7] Zhilbert Tafa and Nerxhivan Pervetica, 2015, 'An Intelligent System for Diabetes Prediction', Mediterranean Conference on Embedded Computing, pp 378-382. An intelligent system for diabetes prediction | IEEE Conference Publication | IEEE Xplore

[8] Sumi Alice Saji and Balachandran K, 2015, 'Performance Analysis of Training Algorithms in Diabetes Prediction', International Conference on Advances in Computer Engineering and Applications, pp 201-206. Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction | IEEE Conference Publication | IEEE Xplore

[9] Aakansha Rathore and Simran Chauhan, 2017, 'Detecting and Predicting Diabetes Using Supervised Learning', International Journal of Advanced Research in Computer Science, Volume 08. https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09765697&AN=124636576&h=8ifOIZHxNyKo%2Brabloa98E5TjcBe%2FeeRJzV99wSbbPFtNqmtrglO4fPWk4SW4BqTPJ%2B0MRtRAYvn7A7mln%2F1Rw%3D%3D&crl=c

[10] April Morton, Eman Marzban and Ayush Patel, 2014, 'Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients', International Conference on Machine Learning and Applications, pp 428-431. A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients | IEEE Conference Publication | IEEE Xplore

[11] Prof. Dhomse Kanchan B. and Mr. Mahale Kishor M, 2016, 'Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis', International Conference on Global Trends in Signal Processing, Information Computing and Communication, pp 5-10. Study of machine learning algorithms for special disease prediction using principal of component analysis | IEEE Conference Publication | IEEE Xplore

[12] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, 2016, 'Diabetes Disease Prediction Using Data Mining', International Conference on Innovations in Information, Embedded and Communication Systems.