iJRASET
International Journal For Research in
Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: IV Month of publication: April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41143

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

# Diabetes Prediction Using Machine Learning

Ashwini R[1], S M Aiesha Afshin[2], Kavya V[3], Prof. Deepthi Raj[4]
[1, 2, 3]Department of Telecommunication Engineering, Dayanandasagar College of Engineering Bengaluru, India
[4]Assistant Professor Department of Telecommunication, Engineering Dayanandasagar College of Engineering Bengaluru, India

***Abstract: The concept of machine learning has quickly become very attractive to the healthcare industry. Predictions and analyzes made by the research community on medical data sets help with appropriate care and precautions in the prevention of disease. of machine learning, the types of algorithms that can help make decisions and predictions. We also discuss various applications of machine learning in the medical field, with a focus on diabetes prediction through machine learning. Diabetes is one of the most increasing diseases in the world and it requires continuous monitoring. To check this, we explore various machine learning algorithms which will help in early prediction of this disease. This work explains various aspects of machine learning, the types of algorithm which can help in decision making and prediction. The predictions and analysis made by the research community for medical dataset support the people by taking proper care and precautions by preventing diseases. Discuss various applications of machine learning in the field of medicine focusing on the prediction of diabetes through machine learning. Diabetes is one of the fastest-growing diseases in the world and requires constant monitoring. To verify this, we are exploring different machine learning algorithms that will help with this baseline prediction.***

***Keywords: Decision Support Systems, Diabetes, Machine learning, Support vector Machine, Random Forest, K-Nearest Neighbor, Logistics Regression***

## I. INTRODUCTION

### *A. Overview*

Diabetes mellitus is a chronic, lifelong disease caused by excessively high blood sugar levels. Classification strategies are widely used in the medical field to classify data into different classes based on certain constraints against an individual classifier. Diabetes is a disease that affects the body's ability to produce the hormone insulin, thereby making carbohydrate metabolism abnormal and raising blood sugar level[4]. As reported by the World Health Organization report in 2019 reported 463 million are with diabetes, 1.5 million deaths, as the report indicates that is not difficult to guess how much diabetes is very serious and chronic.

Many researchers conduct experiments to diagnose diseases using different machine learning approach classification algorithms such as logistic regression, KNN, SVM, Random forest classifier because researchers have proven demonstrated that machine learning algorithms are more effective.

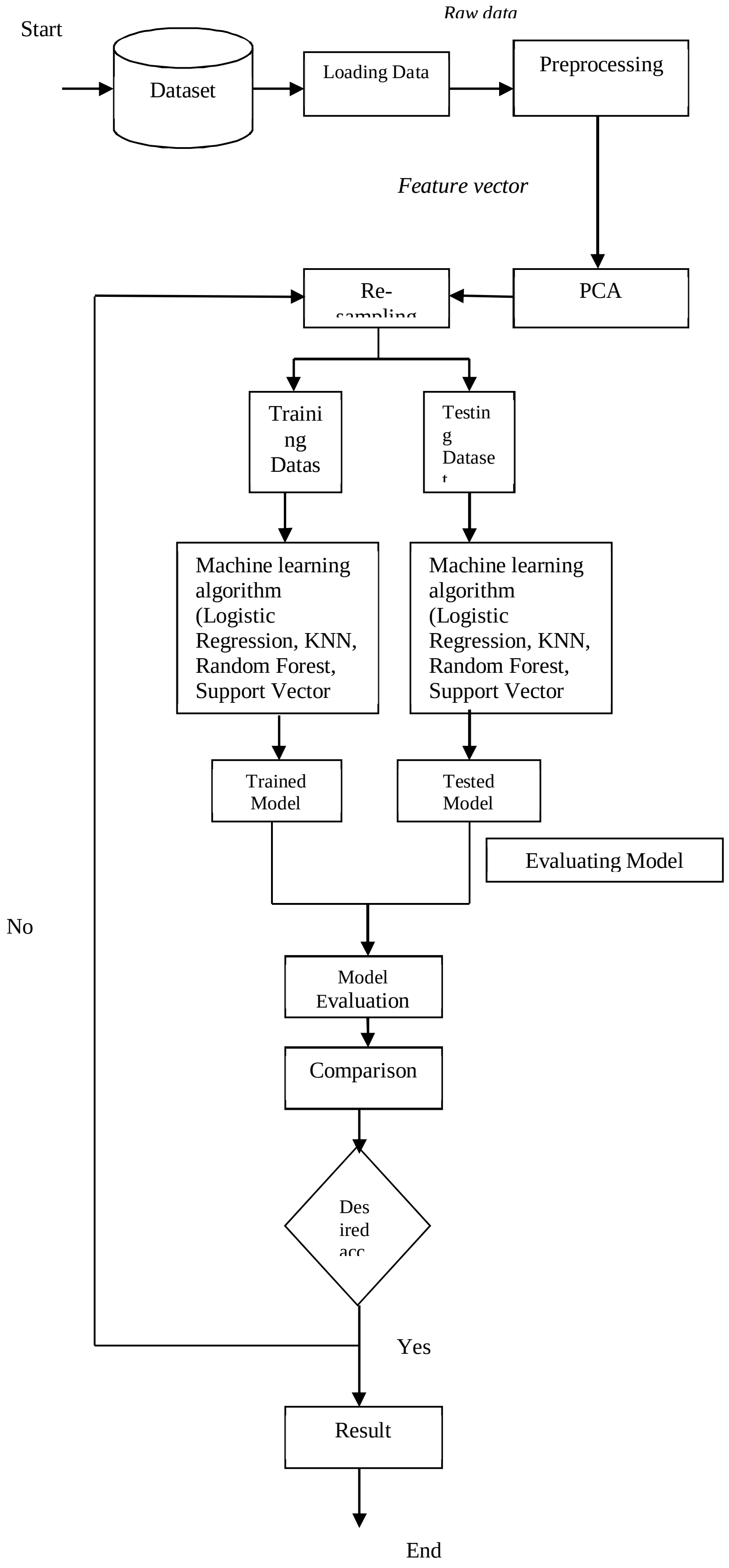
### *B. Data, feature, and software tool*

In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

1) Step 1: Import the required libraries, import the diabetes dataset.
2) Step 2: Preprocess the data to remove missing data.
3) Step 3: Perform 80% scaling to split the set data as a training set and 20% as a test set.
4) Step 4: Select the machine learning algorithm i.e.K-Nearest Neighbor machine learning, Support Vector Machine, Decision Tree, logistic regression, and Random Forest.
5) Step 5: Create a model classifier for the mentioned machine learning algorithm based on the training set.
6) Step 6: Test the classifier model for the mentioned machine learning algorithm based on the test set.
7) Step 7: execute a Comparative evaluation of test performance results obtained for each classifier.
8) Step 8: After analyzing based on various metrics, determine the best performing algorithm.

### *C. Proposed Diagram*

We propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different machine learning techniques are using like classification, regression and clustering. The major focus is to increase the accuracy by using resample technique on a benchmark well renowned PIMA diabetes dataset that was acquired from UCI machine learning repository, having eight attributes and one class label. The proposed framework is shown in Figure 1. The description of each phase is mentioned.


Start
Raw data
Dataset
Loading Data
Preprocessing
Feature vector
Re-
sampling
PCA
Traini
ng
Datas
Testin
g
Datase
t
Machine learning algorithm (Logistic Regression, KNN, Random Forest, Support Vector
Machine learning algorithm (Logistic Regression, KNN, Random Forest, Support Vector
Trained Model
Tested Model
Evaluating Model
No
Model Evaluation
Comparison
Des
ired
acc
Yes
Result
End

1) *Data Selection:* Data selection is a process in which the most relevant data is selected from a specific domain to derive values that are informative and facilitate learning. PIMA diabetes dataset having 8 attributes that are used to predict the diabetes at earlier stage. This dataset is obtained from UCI repository.
2) *Data pre-processing:* Data pre-processing is a Machine Learning technique that includes changing crude information into reasonable configuration. It includes Data cleaning, Data Integration, Data Transformation, and Data Discretization.
3) *Feature Extraction Through Principle Component Analysis:* Feature Extraction on the dataset to determine the most suitable set of attributes that can help achieve better classification. The set of attributes suggested by the PCA are termed as feature vector. Feature reduction or dimensionality reduction will be benefitted us by reducing the computation and space complexity.
4) *Re-sampling Filter:* The supervised Resample filter is applied to the pre-processed dataset. Re-sampling is a series of methods used to reconstruct your sample data sets, including training sets and validation sets. In this study, Boot strapping re-sampling technique to enhance the accuracy.

## II. MACHINE LEARNING ALGORITHM

### A. Logistic Regression

The predictive analysis which is used for the dependent variable is categorical called as Logistical Regression. Logistical Regression explains the relationship between one dependent variable and one or more independent variables. The various types of Logistic Regression are:

1) Multinomial Logistic Regression (many)
2) Binary Logistic Regression (two)
3) Ordinal Logistic Regression (1)

The categorical response has only two possible outcomes. Multinomial Logistic Regression hasthree or more outcomes without ordering whereas Ordinal Logistic Regression has three or more outcomes with ordering.

### B. K-nearest Neighbors

The supervised classifier which is a best choice for K-NN is called as k-Nearest Neighbor. It is a best choice for the classification of k-NN kind of problems.

In order to predict the target label of a test data,KNN which finds distance between nearest training data class labels and new test data point in the presenceof K value? KNN uses K variable value between 0 to 10 normally.

### C. Random Forest

The outfit learning technique used for the classification and regression that operates by constructing the multitude of decision trees at training time and outputting the class i.e mode of the classes or the regression of the individual trees. Irregular choice woods right for choice trees propensity which is used for over fitting on to their preparation set.

### D. Support Vector Machine (SVM)

SVM is a division of Supervised Learning Algorithm. The strategy used to perform regression, classification and outlier detection of data. SVM will be grouping the information dependent that on the hyperplane.

The hyper plane is used to totally isolate the two classes in the best way and the most extreme edge hyper plane ought to be picked as a best separator.

The two types SVM Classifiers that are been used areused are: Linear Classifier and Non-Linear Classifier.

## III. RESULTS AND ANALYSIS

Indian diabetes dataset named PIMA were used for analysis for this study. It consists of eight independent attributes and one independent class attribute. The study was implemented by R programming language using R Studio. Machine learning algorithms (Logistic regression, K-NN, Random Forest, Support vector machine) and clustering (k-means, hierarchical agglomerative) are used to predict the diabetics disease in early stages.

## IV. LITERATURESURVEY

A. *Sneha, N. and Gangil,T., Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), p.13.(2019):*

Authors have focus on selecting the attributes in early detection of Diabetes Miletus using predictive analysis and design a prediction algorithm using Machine learning techniques. The data is collectedfromUCImachinerepository.15attributeshavebeenused for the purpose of classification. Support Vector Machine, Random forest and Naïve Bayes are the classifiers used with an accuracy of 77.73 %, 75.39% and 73.48%.

B. *K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking,2019*

The author proposes a random forest algorithm for diabetes prediction to develop a system that can perform early prediction of diabetes for patients with higher accuracy by using the random forest algorithms. The proposed model gives the best results to predict diabetes and the results show that the prediction system can predict diabetes effectively, efficiently, and most importantly instantaneously. Nanos Nnamoko et al presented Prediction of diabetes onset: a group-supervised learning approach, they used five widely used classifiers for groups and one used Meta classifier. Results are presented and compared with similar studies that have used the same data sets in the literature. It is shown that by using the proposed method, prediction of the onset of diabetes can be made with greater accuracy.

C. *Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13*

Diabetes prediction is presented by machine learning techniques to predict diabetes through three different supervised machine learning methods including SVM, logistic regression, ANN. This project proposes an effective technique for the early detection of diabetes. Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patients diagnoses information. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient’s database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar proposed a study of diabetes prediction using machine learning algorithms in healthcare. They applied six different algorithms of machine learning. The performance and accuracy of the applied algorithms will be discussed and compared. Comparing different machine learning techniques used in this study shows which algorithm is best suited to predict diabetes. Diabetes prediction has become an area of interest for researchers to train programs to identify patients with diabetes by applying the appropriate classifier on the data set. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important, in computers, to handle the issues identified based on previous research.

D. *Sisodia, D. and Sisodia, DS, 2018. “Prediction of diabetes using classification algorithms. Procedia computer science”, 132, pp.1578-1585.(2018) .*

The authors designed a support system for estimating disease, including diabetes, using the Pima Indian Selected Diabetes Database (PIDD). In this study, three machine learning recognition algorithms, including Bayes Naive, SVM, and Decision Tree, were used to diagnose diabetes at an earlier stage with an accuracy of 76.3%, 65.1. % and 73.82%.

E. *Rahul Joshi and Minyechil Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm”: Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct - 2017.*

The authors have proposed the ML techniques which are used to guess the data set satan initial phase to save the life. Using KNN and Naïve Bayes algorithm. In this study they proposed method provide high accuracy with accuracy value of 90.36% and decision Stump provided less accuracy than other by providing 83.72% accuracy.Random Forest, Naive Bayes, and KNN, are the most widely employed predictive algorithms here. The single algorithm offered less precision than ensemble one. The decision tree was highly accurate in most of the tests. Java and Weka are the tools in this hybrid study for predicting diabetes data. They proposed a theory based on Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. To make this system as an ensemble hybrid model, the following algorithms are used: KNN, Naive Bayes, Random forest and J48 which is used to increase the performance and accuracy. J48 is one of the most popular as well as better accuracy. All these algorithms are used to enhance the accuracy and all these are advanced when compared to others.

The random forest provides better accuracy than J48 as well as Nave Bayes in 10 cross-validation splitting method. The fuzzy rule was developed to reduce the wrong treatment.

We can analyze the performance by using the result of this proposed theory

*F. Ridam Pal ,Dr. JayantaPoray, and MainakSen, , "Application of Machine Learning Algorithms on Diabetic Retinopathy", 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.*

The authors suggested diabetic retinopathy (DR) as one of the main causes of poor visual acuity in diabetics. In it, they tested the performance of a set of machine learning algorithms and verified their performance for a particular data set. They deployed machine learning techniques to predict whether an image contains signs of diabetic retinopathy.

*G. Dr. M. Renuka Devi and J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016).*

The authors propose to analyze different mining skills for diabetes prediction using Naive Bayes, Random Forest, Decision Tree and J48 algorithms. In this K-means and KNN are combined to overcome the computational complexity of large number of dataset. And the training set is verified with fuzzy structures and neural networks to supply higher results.

*H. Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)| 10- 12 December 2015 | Trivandrum.*

The authors suggest that diabetes is caused by increased blood sugar levels. Various computer information systems have been described using classifiers to predict and diagnose diabetes using decision trees, SVM, Naive Bayes, and ANN algorithms. They proposed a technique for the diagnosis of Diabetic patients such as Prediction and Diagnosis of Diabetes Mellitus – A Machine Learning Approach. They used this technique to provide high accuracy based on the AdaBoost algorithm. We can collect the local dataset by using the relating mean value as a part of the global dataset. As well as we can train and validate the dataset collection by using four base classifiers as a Decision tree, Support Vector Machine, Native Bayes and Decision stump. After that, we can also calculate Body Mass Index (BMI) by using the height and weight of a person. By analyzing all these techniques as well as by using the AdaBoost algorithm, we can easily get the performance accuracy, sensitivity, specificity and also error rate.

*I. Santhanam, T. and Padmavathi, M.S., 2015. Application of K- means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis. Procedia Computer Science, 47 ,pp.76-83.*

The authors propose a K-means method with the aim of removing noisy data and genetic algorithms to find the optimal feature set with Support Vector Machine (SVM) as the classifier.

The proposed model achieved an average accuracy of 98.79% for the Pima Indian diabetes dataset which was reduced from the UCI repository.

## V. CONCLUSION

Machine learning can help doctors identify and cure diabetes. We will conclude that improving the accuracy of the classification will help the machine learning models perform better.

The performance analysis is in terms of accuracy rate among all the classification techniques such as logistic regression, K-nearest neighbors, SVM, random forest.

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset.

The main aim is to design and implement diabetes prediction using machine learning methods and performance analysis of that method. The proposed method approach uses SVM, KNN, logistic regression, and random forest. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

We also find that the accuracy of the current system is less than 70%, that's why we suggest using a combination of classifiers called associative methods. The combined method takes advantage of the summation of the values of two or more techniques. We have hoped that our system provides us with more than 98% of accuracy.

## REFERENCES

[1] Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), p.13.(2019).

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[3] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-1

[4] Sisodia, D. and Sisodia, DS, 2018. Prediction of diabetes using classification algorithms .Procedia computer science, 132, pp.1578-1585.(2018) .

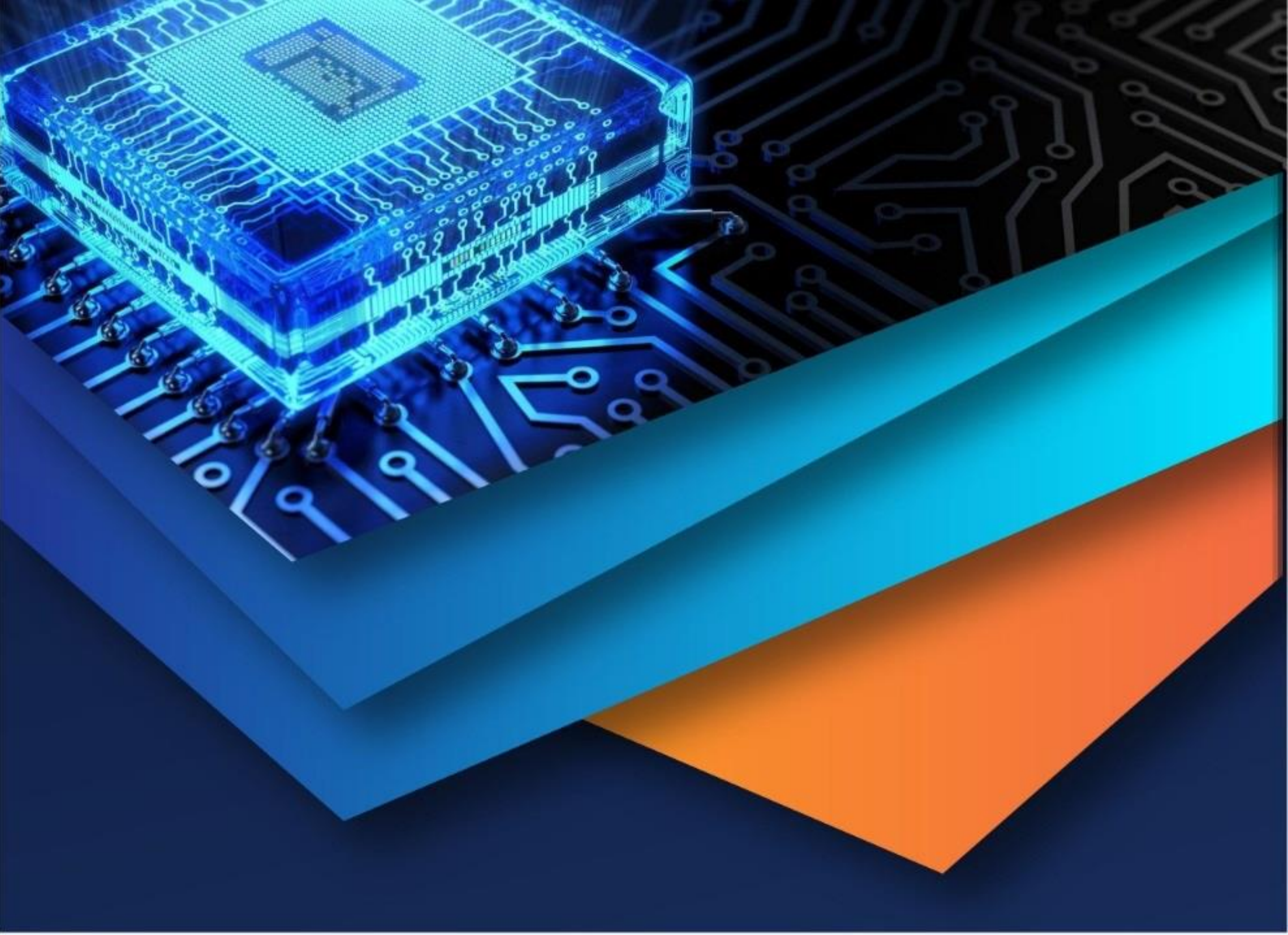
[5] Rahul Joshi and Minyechil Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm”: Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct - 2017.

[6] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, , “Application of Machine Learning Algorithms on Diabetic Retinopathy”, 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.

[7] Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016).

[8] VeenaVijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, “A Machine Learning Approach” ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)| 10- 12 December 2015 | Trivandrum.

[9] Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis .Procedia Computer Science, 47, pp.76-83.(2015).



doi crossref
10.22214/IJRASET

INDEX COPERNICUS
45.98

JISRAF
IMPACT FACTOR:
7.129

SJIFactor
TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY