iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** V **Month of publication:** May 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.51696

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue V May 2023- Available at www.ijraset.com*

# Prediction of Diabetes using MachineLearning

Ramander Singh[1], Anas Sabbag[2], Anupriya Pal[3], Ayush Agarwal[4], Ayushmaan Sahu[5], Chirag Goel[6]
[1]*Assistant Professor,* [2, 3, 4, 5, 6]*Student B.Tech CSE Department, IMS Engineering College Ghaziabad, Uttar Pradesh India*

***Abstract:*** ***Diabetes is a chronic condition that could lead to a catastrophe in the world's healthcare system. 537 million people worldwide have diabetes, according to the International Diabetes Federation. By 2045, this number is anticipated to reach 783 million. A rise in blood glucose levels can result in diabetes. Numerous symptoms, which includes frequent urination, increased thirst, and increased appetite, are brought on by this raised blood glucose level. It is a significant contributor to heart failure, stroke, kidney failure, blindness, and amputations. The objective of the given study mainly is to develop the single collective system that combines result of many different types machine learning techniques, including Logistic Regression, Linear Regression, Support Vector Machine, and Random Forest, to more accurately predict diabetes in a patient. It collects the patient's records based on their pregnancy, blood sugar levels, blood pressure, insulin levels, body mass index, and many other factors. Each of the strategies will be used to determine the model's correctness, however, we found that the Support vector machine method had the highest accuracy (77%). The diabetes forecast is then made using the model with the highest accuracy.***
***Keywords: Machine Learning, SVM, Diagnosis, Diabetes.***

## I. INTRODUCTION

The healthcare industry is a very important study area with accelerating technology development and growing data every day. To handle the large amount of healthcare data, Big Data Analytics, an emerging strategy in the healthcare field, is necessary. Numerous procedures are used to treat millions of patients worldwide. Making informed and effective decisions to enhance the general standard of healthcare will be aided by analyzing the trends in patient treatment for the diagnosis of a specific condition. Machine learning is a promising tool for early illness diagnosis and may help practitioners make diagnosis judgements. Diabetes is a condition that results in a deficit because there is less insulin in the blood. Diabetes is a rapidly growing disease even among youngsters[1]. Understanding what occurs in the body is essential to understanding diabetes and how it develops.

A person without diabetes Sugar (glucose) is raised in the body of human from the intakes of several foods, mainly which are quite high in carbohydrates[1]. Everybody requires carbs, those also who have diabetes, as carbs are the body's primary source of energy. Bread, fruits, rice, dairy products,

and vegetables are all examples of foods that includes carbohydrates. Diabetes has the symptoms of Increased thirst, Tired/Sleepiness, Weight loss,mood swings, Confusion and difficulties faced in concentrating frequent infections.

The body converts these meals into glucose when we utilise them. Glucose flows in the overall via bloodstream. To think in effective manner and function in proper manner, brain also need glucose so some of the glucose transport to brain also. The remained amount of glucose is transferred to our body's cells so that it can be used as a fuel, and it is also stored as energy in our liver for prior functions of our body. Also the Insulin is required to our body for the utilization of glucose as fuel[2].

Pancreas having cells known as bets cells that generate the hormone insulin. Insulin functions as the key of any door. For the allowance of glucose to pass from bloodstream through the cell's doorways, insulin connects to them. The primary cause of diabetes is genetics. It is brought on by at least 2 defective genes on the 6 chromosomes, chromosomes are units that mainly influence how the body reacts to different antigens. Type 1 and type 2 development of diabetes may also be influenced by viral infections. According to the studies, having viruses such as hepatitis B, CMV, mumps, and rubella increases the chances to acquire diabetes.

### *A. The Technology Behind Diabetes prediction*

Beyond the integrity of data, Machine Learningtechniques mainly target, to allow software applications to become more accurate for the prediction of outcomes without being explicitly programmed. For input, Machine Learning mainly use the pre-used data for the output values prediction[4].

The reason is majorly that it can solve issues in effective speed and on the scale that cannot be achieved by the mind of human alone, ML has proved to be useful for this. Machines can be carefully trained to recognize patternsin and correlations between the incoming data by establishing relationships between the immense computational power behind the single activity or various other particular tasks. This allows machines to perform various process in repetitive manner.

The objective of the provided paper is to search that to what extent machine learning enhance the prediction result withaccuracy and information across the various different sources. For the conclusion of the prediction with higher accuracy, in this paper linear and lasso regression is used: which provides the higher accuracy as compared to other algorithm of machine learning[3]. Once a prediction rate of the diabetes has been achieved by the medical department, then further safetymeasures can be achieved.

### *B. Models Used*

#### *1) SVM*

Support Vector Machine, usually referred to as SVM, is the technique under supervised machine learning. Mostly used classification method is SVM. In the high-dimensional space, it generates the hyperplane or the group of hyperplanes. You may use these hyperplanes for the regression or classification as well. SVM differentiate between the samples in certain different classes and may categorize objects in absence of supporting data. For every class, the separation is carried out using a hyperplane, which performs the separation to the nearest point of training[5].

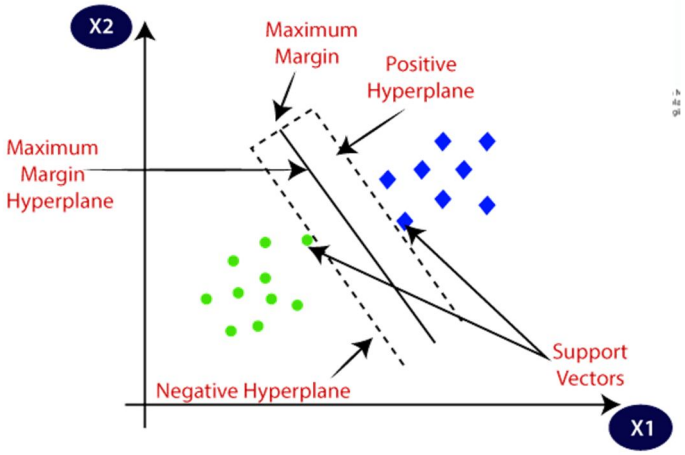



Fig 1: SVM

#### *2) Random Forest*

For the increment of the dataset's predictive accuracy, a classifier namely Random Forest makes use of several decision trees on different subsets of the provided input data. It is a very well-liked supervised machine learning technique used for the Classification and the Regression issues in machine learning[5].

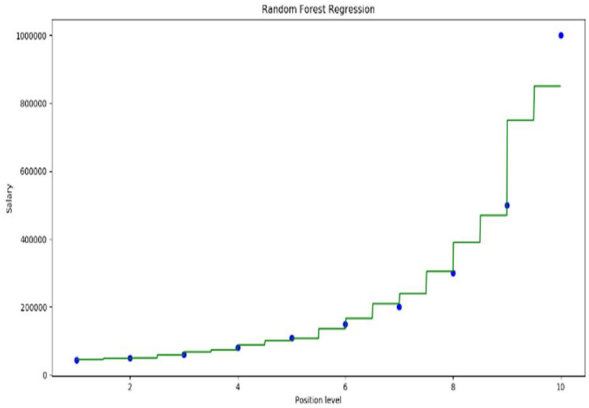



Fig 2 : Random Forest

#### *3) Logistic Regression Model*

This model is based on the various dependent variables, the machine learning classification process known as logistic regression is used for the forecast the likelihood of the provided classes. The logistic regression model mainly generates the logistic of the outcomes after computing the sum of the input features[5]. It forecasts the categorical dependent result of variables. So, the result must be very discrete.

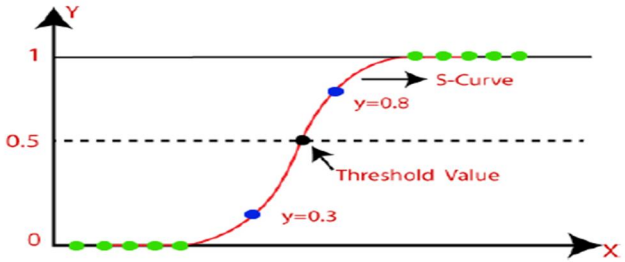


Fig 3 : Logistic Regression

*4) Software Requirement Proprieties*

Python 3.6.0 is a dynamic object-oriented programming language that may be applied to varieties of software development projects. It comes with extensive standard libraries, strong support for integration with other languages and technologies. Many Python developers claim to have seen considerable boosts in productivity and they have felt as if the language encourages the production of higher quality, more maintainable code[7].

Jupyter Notebook: A client-server program that mainly allow us to makes edit and run notebook papers by the web browser is known as the Jupyter Notebook App. The Jupyter Notebook App can also be deployed on the remote server and can be accessed through the help of internet, or it can be run on the desktop without the need for any kind of internet connection. The App contains a "Dashboard" , a "control panel" that exposes local files and allows to open or shutting down their kernel in addition to displaying, editing, and running notebook documents.

## II. LITERATURE REVIEW

In this literature, we propose the method random forest for the diabetes prediction in order to generate the machine learning system that made to be used in predicting accurately a patient's risk of developing diabetes early on. The results indicate that the generated prediction system can forecast the diabetes in very much effective and efficient manner and most significantly. The suggested model generate one of the best results for prediction of diabetes[6]. According to the International Diabetes Federation, there are 422 million people having diabeties worldwide. This will approx double to 783 million by 2045. High Blood glucose levels can cause diabetes.

In the world of medicine, ML algorithms are well-known for their ability to accurately predict diseases. In an effort to achieve one of the best and accurate findings, numerous different researchers have made the use of ML methods to predict the diabetes. Multiple classifiers, including SVM, KNN, and Random Forest, were used by the researcher Kandhasamy and the Balamurali[8]. Based on the accurate answer, sensitivity of the result, and specificity values, the outputs of the classifier were analysed. Using 5-fold cross validation, the classification process was carried out in two scenarios: with and without preprocessing of the dataset. The authors only kept in mind that the data had noise removed from it without going into detail about the pre-processing procedure that was used.

Negi and Jaiswal planned to make the use of SVM for the prediction of diabetes with some other researches[10]. The Diabetes 130-US and PIMA Indians datasets were both integrated to create a single dataset. Given that many other studies only used one dataset, the goal of this study was to confirm the accuracy of the findings. The dataset was pre-processed by converting the non-numbered values to numbered values, replacing missing values and out-of-range data with zero, and normalising it between 1 and 0. Before the model SVM was applied, various selection techniques for featuring were employed previously.

In addition, Tafa et al. provided an updated SVM and Naive Bayes model that is integrated and enhanced for predicting diabetes[12]. Using data gathered from three separate Kosovo locations, the model was used. Eight attributes and 402 patients were included in the sample, in which 80 had second type diabetes. The daily diets, physical activities, and genes history of diabetes are the few characteristics that were not previously studied. The pre- processing of the data was not mentioned by the authors. They divide the given dataset in half inorder to have the testing sets and training sets for the validation of test.

Authors were able to get a prediction accuracy of 80% using the configuration provided. Authors developed a prediction model that mainly based on the single machine learning algorithm in relevant studies displayed above.

It is evident that the single machine learning algorithm approach will not be able to produce such outstanding prediction results, but that the outcomes may be improved by combining various different machine learning techniques into an ensemble.

## III. PROPOSED METHODOLOGY

### A. Flow Chart

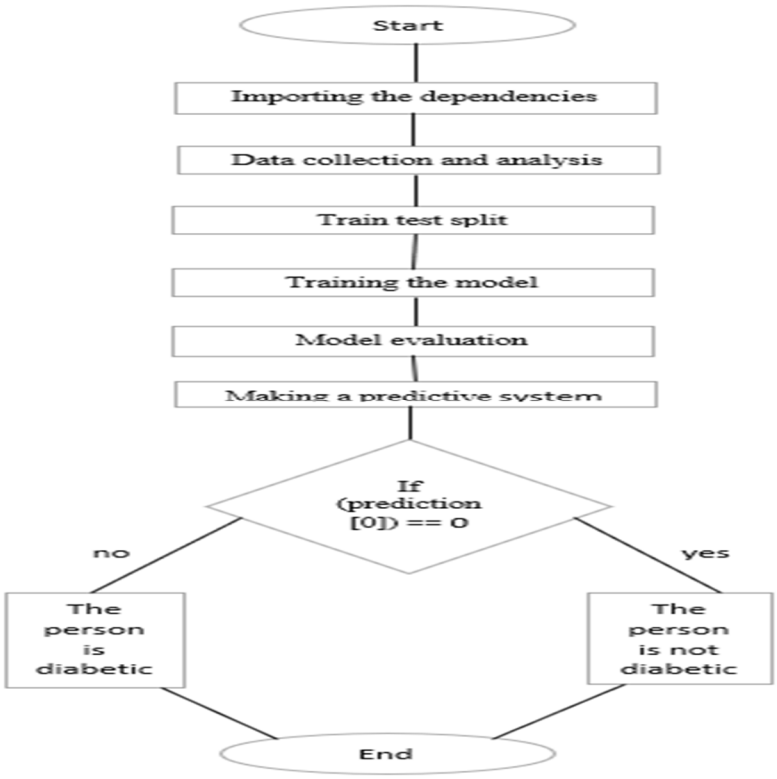



Fig 4 : Flow Chart of our Model

In this section, we go over the many algorithms that were used to develop this module as well as the necessary dataset. The model will be trained using a dataset of 92386 records[10]. The diabetes is analyzed by factors such as blood pressure , bmi , insulin , glucose and data processing was done to filter the data and remove some extraneous data from the Used Diabetes data set.. The Decision Tree Regressoion, Linear Regression ,Random Forest Regression, Logistic Regression, and Support Vector Machine (SVM) are five algorithms that we implemented.

### B. Data Collection

Initially the machine picks the knowledge from the given data. Obtaining effective data is important so that the machine learning models can find out the appropriate and effective patterns. The calibre of the data provided to the computers will state that how precise and concise your used model is. Inaccurate or old data will surely produce nonefficient outcomes or useless predictions**.** Make sure to get your data from some trusted source because it will have the direct impact on your outcome of the model. Good data is more relevant, as it contains duplicates minimum and information gaps, and reflects all the existing categories and the subcategories appropriately[11]. To contrast between the precisions of various ML models. The sample dataset that was 5235tilized had 9 attributes and nearly 786 records, and you will find it on Kaggle.

### C. Data Cleaning

Clean data, or data that is free of noise, is essential to the correctness of any ML model. The act of purging faultyrecords from a dataset is known as data cleansing.

Inaccuratedata could be duplicate data (use the drop duplicates()function to delete duplicates)

1) False data (simply delete the respective column for a largedataset).
2) Incorrectly formatted data.
3) A blank cell (put values calculated by mean, mode, median)[11]

*D. Choosing a Model*

Modeling the relationship between a scalar answer and one or more explanatory factors (often referred to as dependent and independent variables) using a linear approach is what linear regression does.

Model of logistic regression: A categorical dependent outcome of variable is predicted via the logistic regression. Consequently, the result can be the discrete value.

Random Forest: It is a kind of classifier that makes the uses many decision trees on different subsets of the provided input dataset andaverage the computed results for the increment of predicted accuracy of dataset. The most used classification method is SVM. In high-dimensional space, it produces 5236tilize hyperplanes or collection of hyperplanes. These hyperplanes can also be 5236tilized for the regression or the classification. SVM differentiatebetween various features in particular classes and has the ability to categorise items for entities that has no supporting data. Basically the usage of a hyperplanes, separation is carried out to the nearest training location for any class[11].

*E. Model Training*

Eighty percent of the dataset used for model training will be used in this step, while twenty percent will be used for model testing.

## IV. RESULT ANALYSIS

An algorithm i.e. "Support Vector Machine" (SVM) is a supervised machine learning technique that mainly applied for the classification or regression problems. It primarily 5236tilized, nevertheless,usually in classification issues.

PIMA Diabetes Dataset loading the diabetes dataset to a pandas DataFrame printing the first five dataset rows verifying the number of rows and columns obtaining the statistical measures of the data from the diabetes dataset we selected the computed outcome column and did the value count operation on it giving us the count as 0 or 1 where 0 represents as non diabetic and 1 represents as diabetic we perform group by function on outcome column to split it into 0 and 1 and on that we perform mean on the basis of outcome 0 and 1 we predict that glucose level is high for the patient who has diabetes and the age is also more who has diabetes. Seperating data and labels as we drop the outcome column and store the remaining table in a variable X and the dropped column into a new variable Y now we are splitting the data into trained data and tested data in the ratio 80:20 now time to train the model and model is linear so we are using support vector machine classifier now we load the svm model into classifier variable and now we will fit our training data to this classifier now comes model evaluation so basically evaluation is to check how many times our model is predicting correctly for model evaluation we have to check the accuracy score 1st we check accuracy score on training data so we will try to predict all these training data so we wont give the ml model these labels X train and Y train so we will try to predict the label for all these training data and we will compare the prediction of our model (x train prediction) to the original label (y train) which is white ring an then try to predict the accuracy score so accuracy score if it is above 75 its pretty good and in this case we are using very less number of data so there is a chance that we may get a kind of low accuracy score so if our accuracy score is greater than 75 than its pretty good because we can use other optimization techniques to increase that accuracy score so after printing the training data accuracy we see that our accuracy score is 78.6 which is almost 79 and it is pretty good so it means out of 100 predictions our model is predicting 79 times the correct predictions then we check accuracy score on the test data and it is the important step because the model has already seen the training data because we are training the model basically with the training data and it does not make sense if we only evaluate our model based on that so we need to use the model to predict some unknown data so it tells us how well our model is perfoming so its similar to exam case where the student is exposed to the questions to which they are not practiced so the accuracy score is 77 which is again pretty good for this small amt of data so its the good evidence that the model has not overtrained so overtraining represents the model just trains a lot on the training data that it cannot perform well on the test data so in that case training accuracy is very high and testing accuracy is very low so this concept is known as overfitting so now we have to make a predictive system that can predict whether a person has diabetes or not given all these data.

| Algorithms | Training Accuracy | Testing Accuracy |
|---|---|---|
| Linear Regression | 29 | 32 |
| Logistic Regression | 77 | 79 |
| Support Vector Machine | 78 | 77 |
| Random Forest | 1 | 74 |

Fig 5: Algorithms Accuracy

## V. FUTURE SCOPE

In data science, the proposed system uses the "SVM algorithm" to identify the diabetic disease. We have 5237tilized avariety of classification algorithms, including Linear Regression,Random Forest,Logistic Regression. In the future, we can be able to add additional algorithm inorder for the identification of outputs, and the methods can also be compared to determine which algorithm produces the model with the highest computed accuracy. We can also add-on a visitor enquiry module, where users may post questions to the administrators and the administrators can respond. We can even create a treatment section where physicians may upload patient related treatment information and patients can read that information[12]**.** An Android app can be created as a user interface for communicating with users. We intend to carefully craft deep learning network topologies, and employ adaptive learning rates, to train on data clusters rather than the entire dataset for better performance. We can also use more sophisticated machine learning methods, such as random forests, ensemble learning, which produces several decision/regression trees and significantly reducesoverfitting, to obtain even more accurate models.

## VI. CONCLUSION

Designing and implementation of a diabetes prediction system using machine learning techniques was one of a major goal of the research. That method's performance analysis was completed successfully. SVM, Random Forest, Logistic Regression, and Linear Regression classifiers are employed in the suggested methodology, which also incorporates ensemble learning techniques. Classification accuracy of 77% was acquired. The experimental findings can be used for early and quick prediction and decision-making to prevent diabetes and save lives.

The ability to foresee diabetes is crucial in the modern world because of the serious problems it might cause. Because diabetes is the leading cause of death globally. The System model focuses mostly on identifying diabetes using a fewcharacteristics. Physicians can even predict the diabetes early on and special thanks to their system with the help of which now the patients can receive the traditional treatments and the remedies. System did forecasting using some ML approaches in order to compute higher accurate results. Number of studies have been performed on the diabetic imprint. For the development of the hospitals and doctors, developing a diabetes prediction system is very much useful. This system helped doctors in treating the patients more effectively, a system can forecasts disease in its early stages. Suggested model is a real-time or an online programme that is designed for various hospitals and can make anticipation of disease more quickly and efficiently. We'll receive more precise and effective results when we'll utilise machine learning algorithms inorder to anticipate diseases.

## REFERENCES

[1] "Random Forest Algorithm for the Prediction of Diabetes," by K. Vijiya Kumar, B. Lavanya, I. Nirmala, and S. Sofia Caroline. International Conference on Systems,Computation, Automation, and Networking, 2019. Proceedings.

[2] Predicting Diabetes Onset: An Ensemble Supervised Learning Approach," Nonso Nnamoko, Abir Hussain, and David England. 2018 IEEE Congress on Evolutionary Computation

[3] "Diabetes Prediction Using MachineLearningTechniques," Tejas N. Joshi and Prof. Pramila M. Chawan. Internationa Journal o Engineering Researchand Application, Volume 8, Number 1, (Part II), January 2018,pp. 09–13.

[4] "Diabetes Disease Prediction Using Data Mining," byDeeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil. ICIIECS 2017, an international conference on innovationsin information, embedded, and communication systems Romania, October 29–30, 2020, Web Conference onMedicine and Pharmacy.

[5] Development of a Patient-Specific Model for Patients with Diabetes Type I Using Meal and Exercise Guidelines from Modern Schools of Diabetes" by Ghenadie Usic The 8th IEEE International Conference on E-Health and Bioengineering, or EHB 2020, will take place on October 29–30, 2020, at Grigore T. Popa University of Medicine and Pharmacy in Romania.

[6] "Diabetes Prediction Using Ensembling of DifferentMachine Learning Classifiers," by Mahmudul Hasan, MD. Kamrul Hasan, MD. Ashraful Alam, Dola Das, and Eklas Hossain (Senior Member, IEEE). Received on April 6, 2020, accepted on April 18, published on April 23, current as of May 7, 2020.

[7] Deepti Sisodia, "Prediction of Diabetes using Classification Algorithms," with Dilip Singh Sisodia international conference on data science and computational intelligence (ICCIDS 2018) The Writers. Elsevier Ltd. is the publisher.

[8] Diabetes Prediction Using Machine Learning Algorithms,International Conference on Recent Trends in Advanced Computing 2019, (8) Aishwarya Mujumdar and Dr. Vaidehi The Writers. Elsevier B.V. is the publisher.

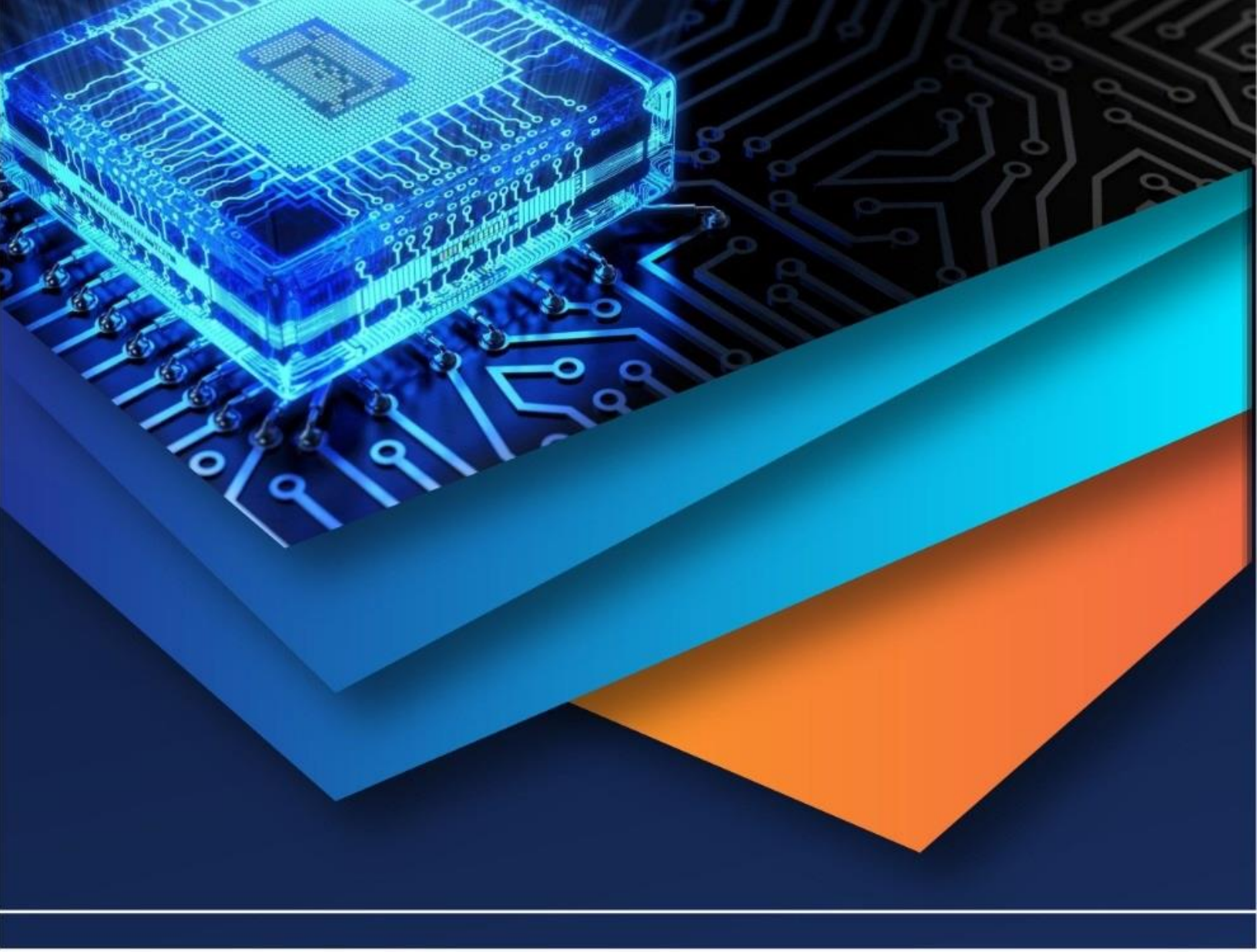
[9] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[10] A.K., Dewangan, and P., Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques, International Journal of Engineering and Applied Sciences, vol. 2, 2015. 48

[11] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[12] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY