iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.64988

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 12 Issue XI Nov 2024- Available at www.ijraset.com

Diabetes Prediction Using Supervised Learning Models

Riza Atik
Indira Gandhi Delhi Technical University for Women, Delhi, India

Abstract: ***About 30 million people in India suffer from diabetes. These patients can be provided with proper treatment if the signs of diabetes are identified early on. This study aims to assess the risk of diabetes among individuals based on parameters such as age, body mass index (BMI), blood glucose level, blood pressure etc. The possibility of an individual suffering from diabetes is predicted using three different machine learning models. The accuracy and F1 score for the predictions of logistic regression, K-nearest neighbors and support vector machine model are calculated. These scores are then compared and support vector machine (SVM) model is found to be the most accurate among the three chosen models.***

I. INTRODUCTION

Diabetes is a common disease in which either the pancreas doesn't produce enough insulin or the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. If not managed in a timely manner, it can lead to serious health issues such as damage to the heart, nerves, tissues, eyes, kidneys etc.

There are three different types of diabetes :

1) *Type 1 Diabetes:* This type of diabetes is an autoimmune condition. As the name suggests, the immune system of the patient mistakenly destroys the cells of the pancreas, which is responsible for producing insulin. Both genetic and environmental reasons are considered responsible for this type of diabetes.
2) *Type 2 Diabetes:* In this type of diabetes, the body of the patient slowly becomes resistant to insulin. Due to this, the pancreas has to increase the insulin production to meet the body's demands. This in turn decreases the efficiency of the pancreas, thus leading to an overall decrease in insulin production. This causes the sugar level in the blood to increase. Factors such as genetics, obesity etc. contribute to a person suffering from type 2 diabetes.
3) *Gestational Diabetes:* This type of diabetes happens only during pregnancies due to the insulin blocking hormones produced in the pregnancy stage. It is often seen in people with a family history of diabetes.

It is also found that people suffering from gestational diabetes have a higher chance of developing type 2 diabetes in the future.

A. Symptoms of Diabetes

The symptoms of diabetes may occur suddenly. Although in type 2 diabetes, the symptoms might start out mild and may take many years to be noticed. Following are some common symptoms of diabetes :

- Frequent urge of Urination
- Increased thirst
- Tiredness/Sleepiness
- Weight loss
- Blurriness of vision
- Mood swings
- Confusion and difficulty in concentrating
- Frequent infections

Diabetes can, over time, damage blood vessels in the heart, eyes, kidneys and nerves. People suffering from it have a higher risk of health problems including heart attack, stroke and kidney failure. It can also cause permanent vision loss by damaging blood vessels in the eyes. Many people with diabetes often develop problems with their feet from nerve damage and poor blood flow. This can cause foot ulcers and may lead to amputation.

B. Causes of diabetes

The primary cause of diabetes is genetics. It is brought on by at least two mutated genes in chromosome 6, which affects the body's reaction to diverse antigens. Additionally, viral infection may affect the likelihood of diabetes types 1 and 2 occurring in a particular person.

Research has indicated that infection with viruses like Coxsackievirus, rubella, hepatitis B virus, cytomegalovirus, and mumps raise the chance of getting diabetes.

II. RESEARCH PROBLEM

Diabetes is regarded as one of the deadliest chronic diseases that raises blood sugar. Many issues arise when diabetes is left untreated and undiagnosed. The laborious identification process ends with a patient visiting a diagnostic facility and seeing a doctor. However, the development of machine learning techniques resolves this significant issue. The goal of this research is to create a model that can accurately predict a patient's likelihood of having diabetes. We have chosen three supervised learning models, namely logistic regression, support vector machine (SVM) and k-nearest neighbors (KNN), to predict the occurrence of diabetes in a patient using factors such as age, glucose, blood pressure, skin thickness, BMI, insulin etc. The accuracy and F1 scores for all three models are calculated and then compared to find the model that is best at predicting the occurrence of diabetes.

III. PREDICTION USING SUPERVISED LEARNING MODELS

Supervised learning is a type of machine learning in which machines are trained using well labelled training data. The supervised learning algorithms then predict the output, on the basis of the data it has been trained on. Labelled data implies that the inputs values are already tagged with the correct output values.

As data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of a cross validation process. Supervised learning is extremely useful in solving a variety of real world problems at scale. It can be used to build highly accurate machine learning algorithms.

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning problems can be further classified into regression and classification problems.

1) Classification: *In a classification problem, the output variable is a category, such as 'red' or 'blue', 'disease' or 'no disease', 'true' or 'false' etc.*
2) Regression: In a regression problem, the output variable is a real continuous value, such as 'dollars' or 'weight'.

Nowadays, due to current environment and living habits, humans face various diseases. There is an urgent need for the identification and prediction of such diseases at their earlier stages, to prevent the extremity of it. It is difficult for doctors to manually identify the diseases accurately most of the time. Using cutting-edge machine learning techniques can prove immensely helpful in this process. The process can be simplified down to a few steps of putting in patient medical data and history and identifying their risk for dangerous diseases such as diabetes.

In an effort to implement the prediction of diabetes using supervised machine learning models, three appropriate models are chosen :

A. Logistic Regression

Logistic regression is a supervised machine learning algorithm that is mainly used for classification tasks. It is used in cases where the goal in to predict the probability that an instance belongs to a particular class or not. It is a statistical algorithm that analyzes the relationship between two data factors.

B. K- Nearest Neighbors

KNN algorithm works by determining the 'K' number of nearest neighbors to a given data point based on a distance metric, such as euclidean distance. The class or value of the data point is then determined by the average of the k neighbors or the majority vote. The approach of this algorithm allows it to adapt to different patterns and make predictions based on the local structure of the data.

C. Support Vector Machine

Support vector machine (SVM) is supervised machine learning algorithm that is used for linear and non-linear classification, regression and even outlier detection. It is best suited for classification problems. The main objective of SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be maximum. The dimension of the hyperplane depends on the number of features.

IV. LITERATURE REVIEW

Results from related research that analyzed various healthcare datasets and made predictions using a variety of methods and strategies are presented. Researchers have created and used a variety of prediction models utilizing different data mining techniques, machine learning algorithms, or even a mix of these techniques.

The research in paper [1], proposes a robust framework for diabetes prediction where outlier rejection, filling the missing value, data standardization, feature selection, k-fold cross-validation were used to preprocess data. The data was then used for training algorithms such as k-nearest neighbor, decision tree, random forest, AdaBoost, naive bayes, XGBoost and multilayer perceptron. AUC is chosen as the performance metric which is maximized using grid search technique during the process of hyperparameter tuning.

In research paper [2], machine learning classification and ensemble techniques are used on a dataset to predict diabetes. The models used are k-nearest neighbors, logistic regression, decision tree, support vector machine, gradient boost and random forest. The accuracy of each model is compared and it is concluded that random forest technique achieves a higher accuracy than the other models employed.

This paper by Ayan Mir et al. [3] is concentrated on diabetes prediction. Diabetes databases for PIMA Indians are used. On the Weka interface, the classification methods Naive Bayes, SVM, Random Forest, and Simple CART are employed. As a result, SVM offers greater accuracy than the competition. Aakansha Rathore et al. [4] made use of the Diabetes dataset for PIMA Indians experimentally, and R Studio was used to assess the performance measurements. SVM and Decision Tree are two machine learning techniques that were employed. In [5], the authors drew a comparison between logistic regression, artificial neural networks and decision tree model for predicting diabetes or pre-diabetes. The participants of the study came from two communities in Guangzhou, China. 735 patients confirmed to have diabetes or pre-diabetes and 752 didn't suffer from either. The decision tree model had best performance followed by logistic regression while ANN gave the lowest accuracy.

In [6], collection of disease symptoms was performed for preparing the dataset along with a person's living habits, and related doctor consultations. This data was used to predict diabetes in patients. The performance of various algorithms such as naive bayes, decision tree and logistic regression was also compared.

V. METHODOLOGY

Supervised machine learning algorithms derive insights, patterns, and relationships from a labeled training dataset. It means that the dataset already contains a known value for the target variable for each record. It is called supervised learning because the process of an algorithm learning from the training dataset is like an instructor supervising the learning process. The correct answers are known, the algorithm iteratively makes predictions on the training data and the instructor corrects it. Learning ends when the algorithm achieves the desired level of performance and accuracy. In this study, three different supervised machine learning models are chosen namely logistic regression, support vector machine and k-nearest neighbor. The 'PIMA Indians Diabetes Database' dataset is first pre-processed and then split into 'train' and 'test'. The 'train' part of the dataset is used to train the models. Once the training is completed, the performance of the models is evaluated using the 'test' part of the dataset. The accuracy and F1 score for each model is calculated and compared to find out which model is the best at predicting diabetes.

A. Implementation steps

The methodology involved collecting and pre-processing the data. We split the data into training and test set. After which the models are trained and evaluated. They are discussed in detail below. These steps are also summarized in Fig. 1.

1) *Data collection and pre-processing:* The process begins by selecting and loading the dataset, where each sample consists of input features and a target output label. Pre-processing is performed to clean the data, handle missing values and scale features where necessary. This step ensures that the data is of high quality, which is essential for training effective supervised learning algorithms.

2) *Splitting the Data:* The data is divided into two sets, the training set (80%) and the test set (20%). The training set is used to teach the model, while the test set is used to evaluate the model's performance on unfamiliar data. This split ensures that the model can perform well on new, unseen data, a key concept in supervised learning.
3) *Training the Models:* The chosen models are then trained by feeding them the input data and output labels. Patterns in the data are learnt by the model by adjusting internal parameters. This training is guided by a loss function which helps the model minimize prediction errors.
4) *Evaluating the Model:* After the training of the selected models is completed, the models are then tested using the unseen test set. Each model's performance is assessed using metrics such as accuracy and F1 score.

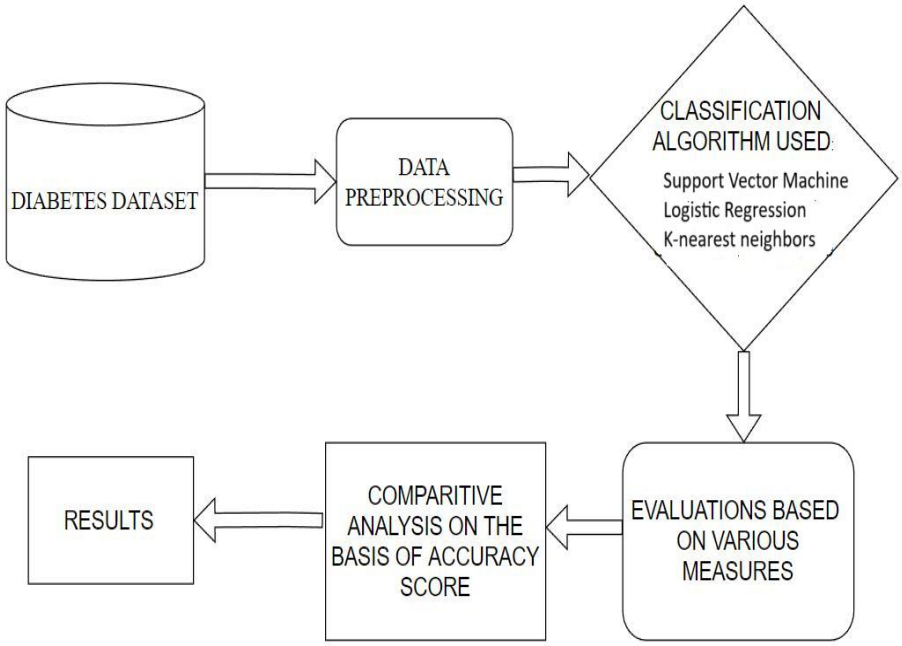


Fig. 1. Proposed Methodology

B. Brief description of the algorithms used

1) *Logistic Regression:* This kind of statistical model, commonly referred to as a logit model, is frequently used in categorization and predictive analysis. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. A logit transformation is performed to the odds in logistic regression, which is the probability of success divided by the probability of failure. The natural logarithm of odds or the log odds are other names for this.
2) *K- Nearest Neighbour:* The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another. A class label is chosen for classification problems based on a majority vote, meaning that the label that is most commonly expressed around a particular data point is adopted. Despite the fact that this is officially "plurality voting," literature more frequently refers to "majority vote."
3) *Support Vector Machine:* Support Vector Machine (SVM) is an effective classification and regression method that increases a model's predicted accuracy without overfitting the training set. SVM is particularly well suited for data analysis with a very large number of predictor fields (could be in thousands). SVM is used in a variety of fields, including CRM, bio-informatics, text mining concept extraction, intrusion detection, protein structure prediction, voice and speech recognition, facial and other image recognition, and CRM. SVM categorizes data points even when they are not otherwise linearly separable by mapping the data to a high-dimensional feature space. Once a separator between the categories is identified, the data are converted to enable the hyperplane representation of the separator. The group to which a new record should belong can therefore be predicted using the features of new data.

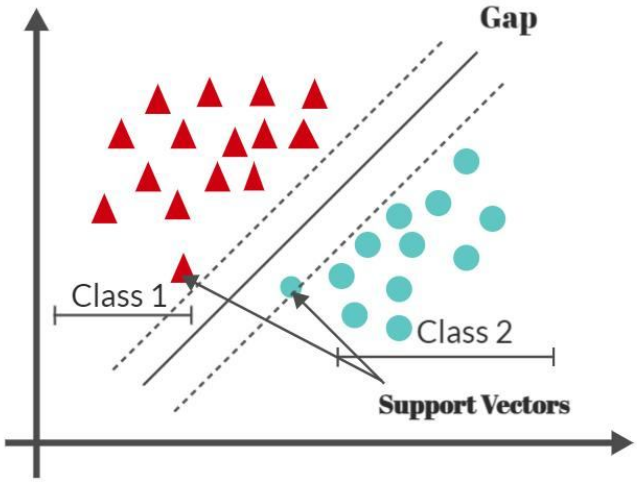
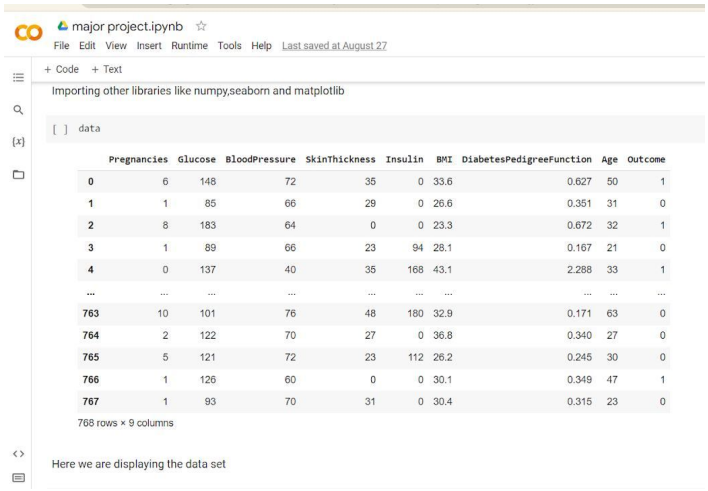


Fig. 2. Supply vector machine algorithm implementation

C. Dataset Description

We have chosen the PIMA Indians Diabetes Database available on Kaggle [7]. The chosen dataset contains 9 columns representing various factors that could possibly contribute to diabetes as shown in Fig. 3. The data was collected for 768 individuals. There are no null values present in the dataset. The outcome column represents whether the person suffers from diabetes (1 represents disease) or doesn't (0 represents disease).



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Fig. 3. PIMA Indians diabetes database

D. Model performance evaluation

This is the last stage in the prediction model. Here, we use measures like classification accuracy and F1-score to evaluate the prediction results.

Accuracy: The percentage of correct predictions to all the input samples is known as the classification accuracy.

F1 score: The F1 score is used to assess the correctness of a test. The Harmonic Mean of memory and precision is the F1 Score. F1 Score has a range of [0, 1]. It informs you of the robustness and precision of your classifier.

VI. RESULT

Three supervised machine learning models were trained on the 'PIMA Indians Diabetes Dataset'. The models were then tested using the 'test set' and their F1 score and accuracy were measured for comparative analysis. Out of the three chosen models, logistic regression, k-nearest neighbors and support vector machine, support vector machine (SVM) gives the highest accuracy of 77.272727 and highest F1 Score of 0.615385.

Table I. represents the Accuracy and F1-Score performance metrics.

TABLE I. PERFORMANCE MATRIX

Model Name	Accuracy of Model	F1 Score
Logistic Regression	75.324675	0.577778
K-Nearest Neighbours	72.727273	0.533333
Support Vector Machine	77.272727	0.615385

Visualizing these scores helps comprehend the difference in model performance of each algorithm clearly. Fig. 4. depict the accuracy of each model and Fig. 5. represents the F1 score of each model.

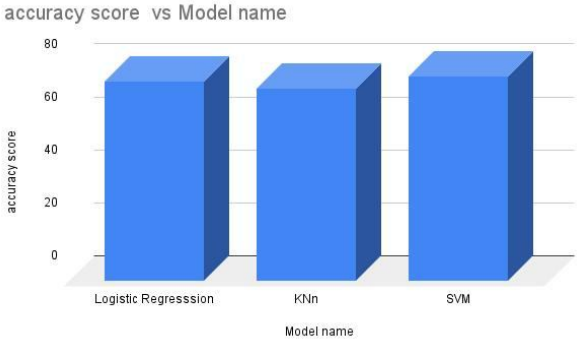


Fig. 4. Accuracy score

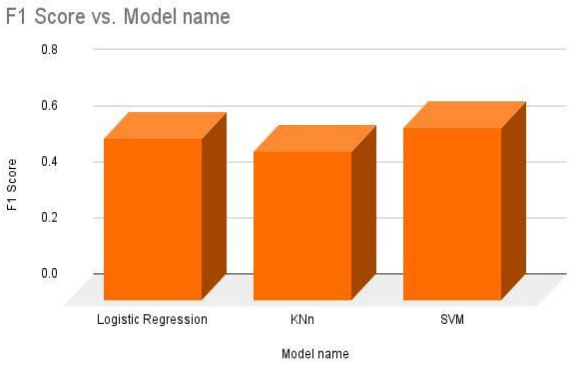


Fig. 5. F1 Score

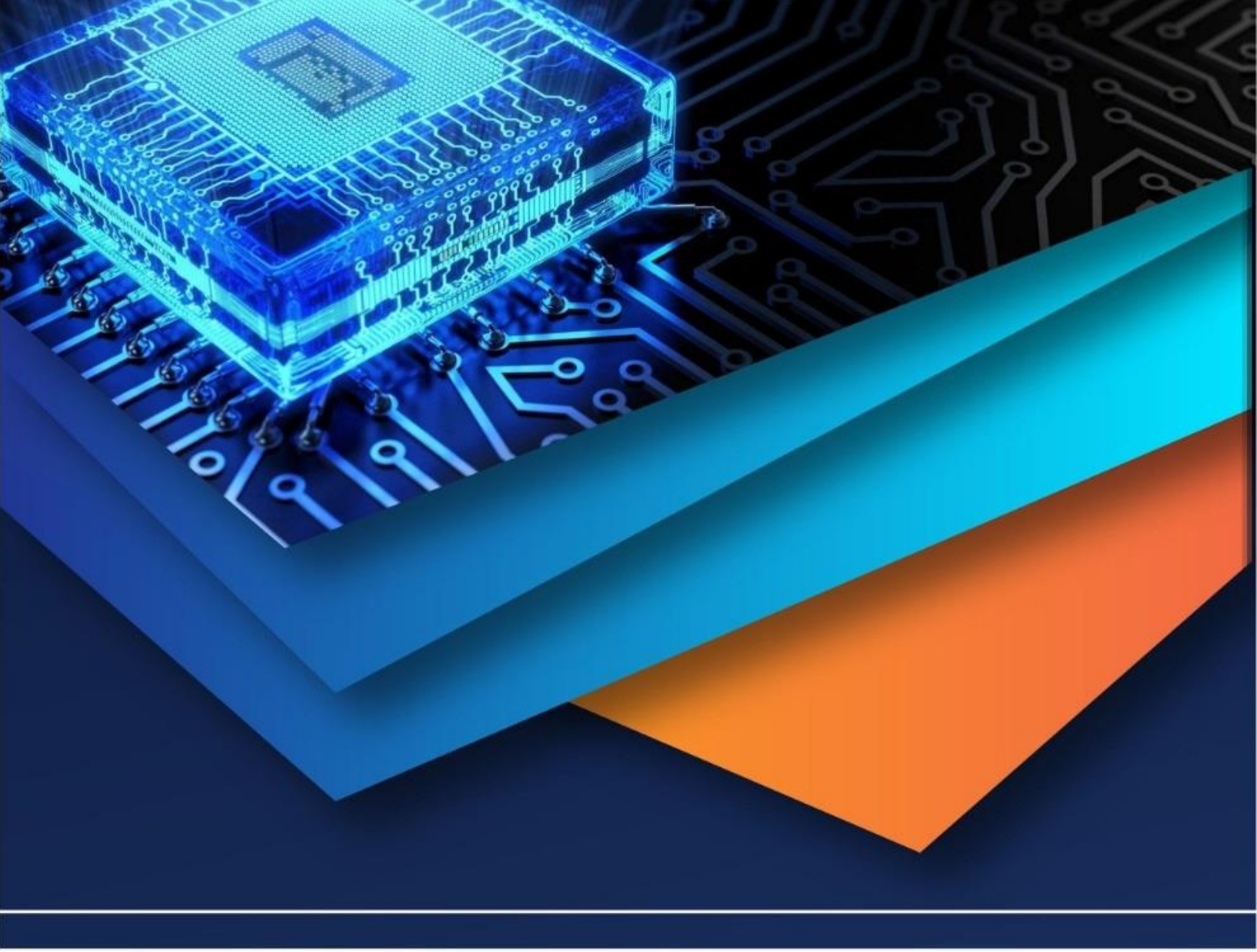
VII. CONCLUSION

The early identification of diabetes is one of the major real-world medical issues. This study makes methodical attempts to design a system that forecasts the development of diabetes. In this paper, three machine learning algorithms namely logistic regression, K-nearest neighbors and supply vector machine were used to predict diabetes. These algorithms were applied on the 'PIMA Indians Diabetes Dataset'. Experimental results determine the adequacy of the designed system with an achieved accuracy of 78 percent using Supply Vector Machine algorithm.

In the future various other machine learning models as well as deep learning models can be used for diabetes prediction. Ensemble techniques can be used to improve the accuracy of predictions. There is also scope in using similar approaches to predict other diseases such as cancer, heart disease etc. This kind of work can be extended and improved such that it can be used to automate the process of diabetes analysis.

REFERENCES

[1] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.
[2] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(09), 2278-0181.
[3] Mir, A., & Dhage, S. N. (2018, August). Diabetes disease prediction using machine learning on big data of healthcare. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.
[4] Rathore, A., Chauhan, S., & Gujral, S. (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women. International Journal of Advanced Research in Computer Science, 8(5).
[5] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.
[6] Alanazi, R. (2022). Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022(1), 2826127.
[7] Kaggle dataset description : https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY