International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Prediction of Diabetes Using Machine Learning Algorithms in Healthcare

Sharath Kumar S R[1], Bharathi Hegade S R[2], Bhavana G T[3], Deepti R Bhat[4], Nikitha K V[5]

***Mr* .Sharath Kumar S R** Professor, Department of ISE JNN College of Engineering,

Bharathi Hegade S R, Bhavana G T, Deepti R Bhat, Nikitha K V, Department of ISE JNN College of Engineering,

Abstract - This study introduces a new approach to prognosticating the onset of detection of diabetes via using machine learning ways. By assaying a dataset containing different demographic, clinical, and life factors, the study identifies crucial predictors for assessing diabetes threat. Several ML algorithms, similar as logistic regression, KNN, Ada-boost, and support vector machine, are utilized and estimated for their prophetic performance.

The results demonstrate that the ML- grounded models effectively identify individualities at high threat of developing diabetes. These models give precious decision making aids for healthcare interpreters, easing early intervention and substantiated operational strategies. Overall, this approach has the implicit to significantly reduce the burden of diabetes on public health systems.

***Keywords*: :** prognosticating, individualities, operational strategies, Public health systems.

1. INTRODUCTION

Diabetes is a fleetly growing complaint affecting individualities of all periods, including youthful people. Understanding how the body functions without diabetes is pivotal for comprehending its development. Glucose, deduced from carbohydrate-rich foods, serves as the body's primary energy source. When consumed, these foods are split into glucose, which circulates in the blood. Insulin, is developed by the pancreas, facilitates the movement of glucose from the bloodstream, these foods are transported into cells for energy. operation. Insulin insufficiency or resistance leads to elevated glucose situations in the bloodstream, performing in diabetes mellitus. In order to complete particular jobs, computers use data to learn. Simple tasks can be programmed directly, but more complicated activities could need that robots create their own algorithms from the data they are given. By examining datasets, Machine learning techniques such as ensemble methods and classification can be employed to accurately predict diabetes

The major goal is to construct a system capable of accurately predicting diabetes early in life by combining the results of diverse machine learning approaches. When predicting diabetes, each algorithm's model accuracy is evaluated and the most accurate model is chosen. The advancement of AI (ML) and related calculations has taken out a significant number of the significant boundaries to sickness recognition and gave direct, exact, and careful outcomes. AI has shown to find lasting success and corresponding to the clinical business in the present day.

Machine learning (ML) facilitates early disease identification and may even stop diseases before they reach critical stages by evaluating individual features. The major goal is to forecast diabetes utilizing machine learning algorithms to offer individuals with early treatment and intervention. To further develop diabetes expectation precision, a multimodal technique utilizing different AI strategies is utilized.

This involves notable methods like AdaBoost, Calculated Relapse, K Closest Neighbor (KNN), and Backing Vector Machine (SVM). KNN utilizes vicinity based order to find potential models by uncovering complex connections between pieces of information. Famous for its ability to perform well on paired grouping issues, strategic relapse reveals understanding of the likelihood climate around diabetes. SVM, recognises for its efficiency in managing high-dimensional feature spaces, Support Vector Machines. Attempt to pin-point the best hyperplanes for class separation in the data given. In the meantime, AdaBoost, an ensemble learning method, builds a strong prediction model resistant to noise and overfitting by utilizing the combined knowledge of weak classifiers.

The predictive analytics platform aims to provide healthcare professionals with a powerful tool for early diabetes detection and risk stratification by combining these many approaches. By utilizing the revolutionary

potential of machine learning, this integrated method transcends conventional diagnostic modalities and signifies a paradigm shift in the healthcare industry. It provides a thorough and sophisticated understanding during the initial stages of diabetes and is distinguished by improved accuracy.

This model under offers a thorough and innovative method for managing and forecasting illnesses. Through the use of intelligent computer programmes and the integration of data from multiple sources, it offers a powerful tool for pre-emptive health care. There are numerous approaches to improving people's health as we continue to develop this paradigm. Future generations now have hope for a better and more robust future.

2. LITERATURE SURVEY

The paper introduces a pioneering method to enhance the Diabetes is a serious and long-lasting medical condition caused by the body's incapacity to either make or use insulin as needed. To stop its progression, early detection is essential. A comprehensive documenting of diabetes cases is made possible by the utilisation of healthcare informatics, opening the door for predictive analysis through the machine learning deployment and data mining techniques. The categorization of diabetes datasets using K-nearest neighbor, Adaptive Boosting, and Decision Tree algorithms is the major goal of this work. In order to Increase healthcare outcomes by prompt intervention, the objective is to develop an effective prediction model for the timely identification of diabetes that remains undiagnosed[1].

Healthcare professionals have a great chance to use analytics of the big data to make informed decisions regarding the health and treatment of their patients. the health and treatment of their patients. Using a dataset of patient medical records, six different machine learning (ML) algorithms are applied in this paper to investigate predictive analytics in healthcare. The best method for predicting diabetes is found by carefully examining and contrasting the accuracy and performance of these algorithms. Through data-driven insights and a clear understanding of the advantages and disadvantages of various machine learning approaches ,this seeks to help medical professionals detect diabetes early and improve patient outcomes[5].

to predict diabetes, this approach develops a model that combines the Random Forest and Support Vector Machine(SVM) learning methods. With a stated highest accuracy of 98% and ROC of 99%, the suggested model achieves impressive performance using a real dataset from Security Force Primary Health Care[11].

3. METHODOLOGY

To predict the Diabetes we use these steps: first data collection Secondly pre-processing the data, data split, model selection and training the model. Finally evaluating the model to see its which model is best among all(1) .

3.1 Dataset

The dataset previously owned for our experimental. It consists of 8 attributes. It contains 768 entries. Each entry contains Features like Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigree Function, Age.The taken dataset below contains eight features and are given in table 1.

Features	Description
Pregnancies	Number of Pregnancies of patients .
Glucose	Glucose concentration is in the patient.
Blood Pressure	The documented blood pressure.
Skin Thickness	Skin thickness of the skin layer
Insulin	Concentration of Insulin in the body.
BMI	Individuals BMI
Diabetes Pedigree Function	The familial background of Diabetes.
Age	Individuals age.

Table 1. Description of Dataset.

3.2. Preprocessing

Following data collection, a preprocessing stage ensues to validate its accuracy and suitability for analysis. This phase encompasses tasks such as data cleaning, managing missing values, outlier detection and removal, and formatting the information into an appropriate structure for analysis. Moreover, techniques like data normalization or scaling may be utilized to standardize features onto a comparable scale.

3.3. Data Split

The gathered and preprocessed data is commonly split into two subsets training set and test set. Training set is utilized to train the prediction model, and the other is employed to assess its performance. This partitioning of data is typically done

randomly, guaranteeing that both subsets accurately represent the overall data distribution.

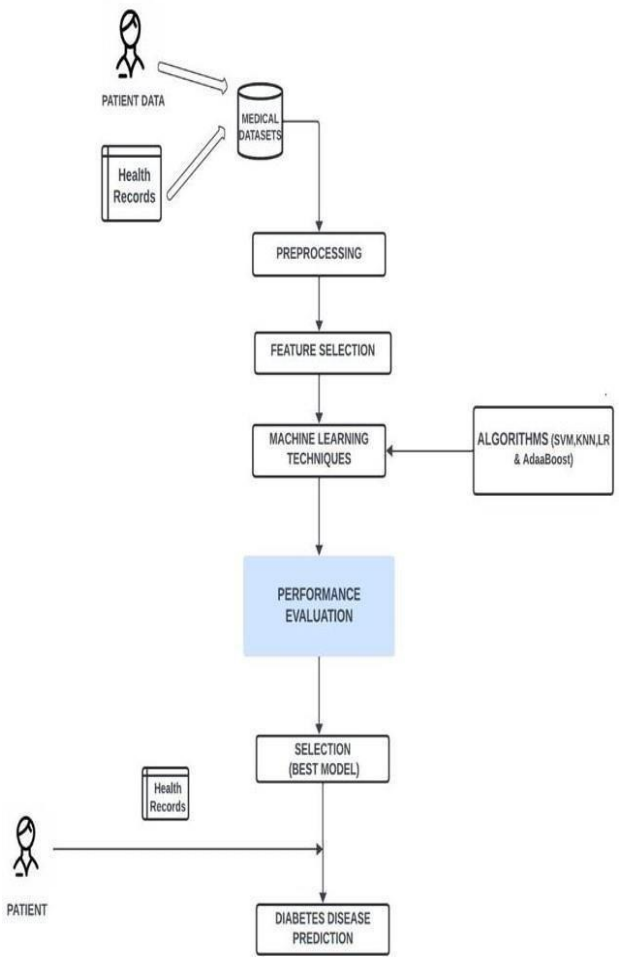

Fig 1 : Proposed Diabetes Prediction Model

3.4 Algorithms used:

Support Vector Machine (SVM)

SVM is used for classification and regression tasks. It operates by identifying the hyperplane that splits a dataset into classes most effectively. The support vectors, or data points nearest to the hyperplane, are significant because they establishes the margin. Using kernel functions to convert the input space into a more complex dimensional space where a hyperplane can be utilised to divide classes, SVM can handle both linear and nonlinear data.

K-Nearest Neighbors (KNN)

KNN is straightforward, non-parametric, lazy learning technique that can be used for regression and classification. A data point is classed according to classification of its neighbours. KNN finds the K-nearest instances to the query by calculating the distance (e.g., Euclidean) between the query instance and the labelled samples in the dataset. It then conducts a majority vote to select the class label for the query instance.

Logistic Regression

A statistical technique for forecasting binary outcomes from a dataset is called logistic regression, To determine the likelihood that a particular input point belongs to a certain class, it employs the logistic function. One uses a dichotomous variable (such as yes/no or 0/1) to measure the result.

Adaptive Boosting (AdaBoost)

AdaBoost is a Troupe learning strategy that forms a strong classifier by consolidating a few powerless students. It constructs a last model with loads changed in light of the precision of every classifier by efficiently utilizing a frail classifier on iteratively adjusted duplicates of the information, underlining the misclassified models more.

Calculation of Performance Metrics

Accuracy: the percentage of real results true positives and true negatives among all the cases that were looked at.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: The percentage of actual favourable outcomes for every optimistic prediction. It displays how well the classifier identified good examples.

Recall (Sensitivity): The percentage of genuine positive outcomes in all real positive cases. It assesses the classifier's capacity to recognise every pertinent event.

F1 Score: The precision and recall harmonic mean, which offers a balance between the two. It comes in especially handy when there is an unequal distribution of classes.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

All classification models, including SVM, KNN, Logistic Regression, and AdaBoost, can be used with these formulas. They come from a table called the confusion matrix that describes how well a classification model performs on a set of test data that has known true values.

3.5 Prediction model:

In machine learning, a prediction model is a computational programme that has been trained on past data to forecast future, unknown data. It entails picking pertinent characteristics, deciding on a suitable method, and using a labelled dataset to train the model.

Name	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7878	0.70	0.58	0.64
SVM	0.7922	0.72	0.58	0.64
Adaptive Boosting	0.753	0.63	0.54	0.58
KNN	0.770	0.68	0.54	0.60

Tabel 2: Comparison among the models

3.6 Results Discussion

Above Table 1, illustrates the performance attributes of four distinct machine learning models that were implemented in the investigation. The metrics include accuracy, precision, recall, and F1-score for each model. When the results are compared, the SVM method performs better than the others, with an astounding 80% accuracy rate.

Figure 1 shows the models that were used and their respective comparative accuracies. The figure shows that the KNN model records the lowest accuracy at 77%, while the SVM method obtains the maximum accuracy at almost 80%, and the findings indicates that the Support Vector Machine Adaptive Boosting model accomplishes a 75% accuracy, Logistic Regression model achieves a 78% accuracy rate.

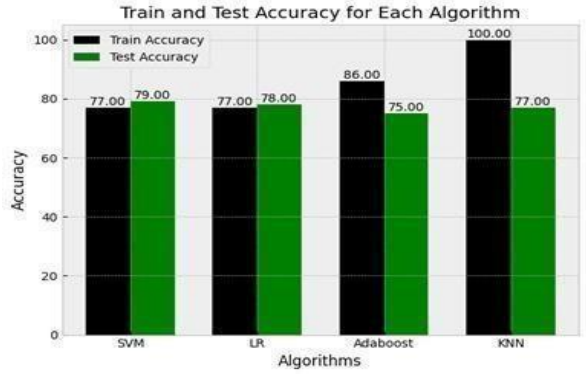

Fig 2: models Accuracy Comparison

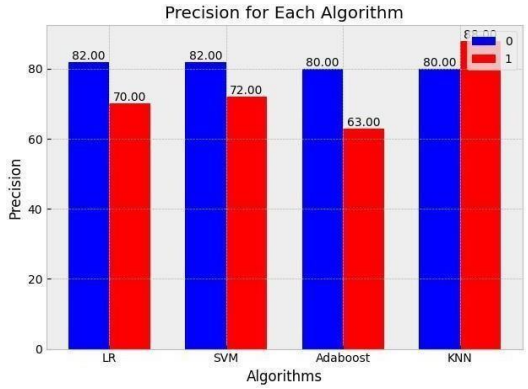

Fig 3: Precision comparison

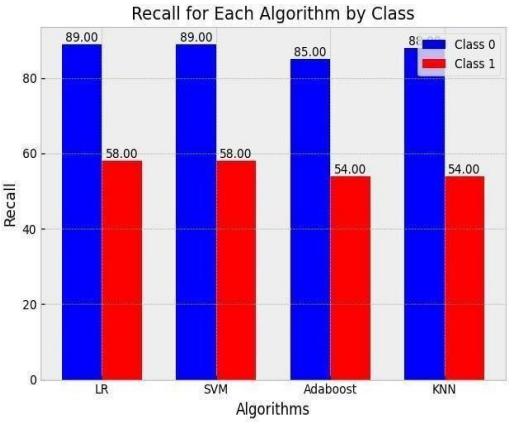

Fig 4:Recall comparison

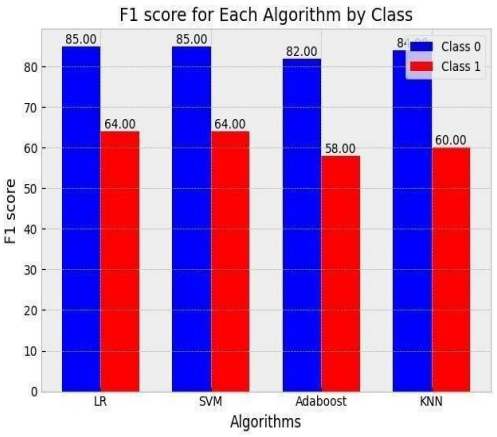

Fig 5: F1 score compassion

The improved performance of the Support Vector Machine algorithm with regards to precision, recall, and F1-score demonstrates how robust it is when processing high-dimensional, complicated datasets with distinct class margins. The kernel trick adds to its effectiveness by enabling it to handle non-linear interactions with ease. Due to this, SVM is especially well-suited for datasets where minimising missed positives (recall) and accurately recognising positives (accuracy) are critical balances. An optimal balance between both metrics is indicated by the significant F1-score.

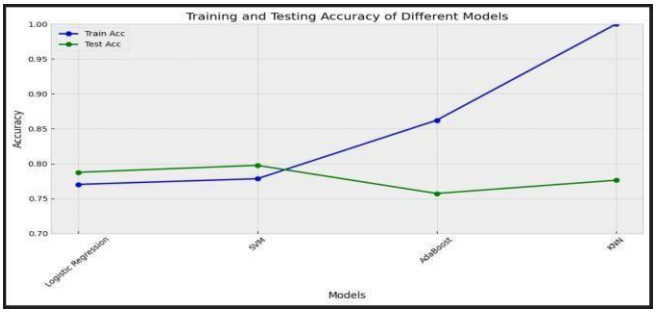

Fig 6: Complete Accuracy of Different models

4. CONCLUSION

By demonstrating the capability of different machine learning (ML) approaches to forecast the probability of diabetes onset, this study offers significant perspectives for individualized healthcare strategies and early intervention. As shown by their exactness, AI models display potential forecast execution by using an extensive variety of segment, clinical, and way of life boundaries. SVM has an accuracy of 89%, logistic regression has an accuracy of 78%, adaptive boosting has an accuracy of 75%, and KNN has an accuracy of 77% for diabetes prediction. In addition, our research emphasizes the significance of collaborating with medical facilities to obtain larger and more diverse datasets. More information makes it conceivable to prepare AI models all the more actually, which could build their precision and capacity to sum up to genuine circumstances.

REFERENCES

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

[2]."A Model-Based Approach for an Early Diabetes Prediction Using Machine Learning Algorithms" Abrar M. Alajlan presents, IEEE Access, Vol.12, No.3, pp. 3957–3965, 2021.

[3]."Diabetes Disease Prediction Using Machine Learning Algorithms," Prashant Kumbharkar, Deepak Mane, Santosh Borde, and Sunil Sangve, ProQuest, vol. 20, pp. 2225-2231, 2022.

[4]."Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah, Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, 2018.

[5]."Supervised machine learning for diabetes prediction" Gustian Paul Sendani Febriana, Muhammad Exell Febriana, Fransiskus Xaverius Ferdinana, Ferdinana,7th International Conference on Computer Science and Computational Intelligence in 2022.

[6]. "Diabetes Prediction using Machine Learning Algorithms," Aishwarya Mujumdar, Dr. Vaidehi V., Proceedings of the International Conference on Recent Trends in Advanced Computing, pp.292-299,2019.

[7]. “Diabetes Prediction Through Machine Learning”, KM Jyoti Rani, Science Direct, Vol. 6, No. 4, pp. 294– 305,2020.

[8]. "An Optimised Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression," V.K. Daliya, T.K. Ramesh, IEEE Acceses, 2021.

[9]. "Prediction of diabetes disease using machine learning algorithms," Monalisa Panda1, Debani Prashad Mishra1, Sopa Mousumi Patro1, Surender Reddy Salkuti2, IEEE Access, Vol.11, No.1, pp. 284-290, 2022.

[10]. "On The Analysis of some machine learning algorithms for the prediction of diabetes," Bello A. Bodinga, Mukhtar A. Abdulsalam, Bello A. Buhari, and Muzzammil Mansur, IEEE Access, Vol.14, No.1, pp. 5294-5299, 2022.

[11]. "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus,"Aeshah Saad Alanazi and Mohd A. Mezher, University of Tabuk, Kingdom of Saudi Arabia,Proceedings of the International Conference on Computing and Information Technology, pp. 55–57,2020.