

Leveraging Data to Solve for Non-communicable Disease [Diabetes] and Healthcare Delivery using Machine Learning Techniques

Sandhya L¹, Arun Kumar S², Abhishek C B³, Vishwas B⁴, Thanya Patel R⁵ and Rashmi V⁶

¹⁻²Assistant Professor, Department of CSE, Presidency University, Bengaluru, India

³⁻⁶Students, Department of CSE, Presidency University, Bengaluru, India

Email: sandhya.l@presidencyuniversity.in, arunkumar.s@presidencyuniversity.in, abhishek.20211CSG0056@presidencyuniversity.in, vishwas.20211CSG0071@presidencyuniversity.in, thanya.20211CSG0043@presidencyuniversity.in, rashmi.20211CSG0058@presidencyuniversity.in

Abstract—Diabetes is rapidly becoming one of the major diseases in India, causing serious problems like heart disease, kidney failure, and nerve damage. The prevalence of diabetes is expected to increase by 2040, and early detection is vital for effective management and reducing long-term treatment costs. In this study, we used a learning model to improve the accuracy of diabetes prediction by analyzing key health indicators like blood sugar level, BMI, insulin, and age. Various classification methods were used, including logistic regression, decision trees, random forest, support vector machine (SVC), Gaussian Naive Bayes, K-nearest neighbor, Ada Boost, Bag, gradient boosting, and voting. To improve the performance of the prediction, we combine the results of each model and select the prediction frequency as the final result. The analysis found that random forest and voting models performed better, but the combined method was more reliable in early diabetes diagnosis. This study highlights the transformative potential of machine learning in the future development of healthcare in India, providing data-driven solutions to combat the challenges of increasing diabetes. The burden of diabetes could impact healthcare by 2040 if not addressed, and advances like these are important for the future of care in this country.

Index Terms— Machine Learning, Prediction, Non-Communicable Diseases, Future Healthcare, Ensemble Learning, Diabetes.

I. INTRODUCTION

Healthcare is currently undergoing a significant transformation due to the rapid growth of structured, semi-structured and unstructured data. In the data stream, big data analytics plays a key role in analyzing big data to find hidden patterns, relationships, trends, patient preferences and other agreements. Diabetes Mellitus (DM) is a major health problem, especially in low- and middle-income countries like India. By 2018, this number is set to rise to 629 million. There are three main types of DM: Type 1 or insulin-dependent diabetes (IDDM), in which the body does not produce enough insulin; not using insulin properly; Fortunately, predictive modeling using machine learning seems to be a promising approach. Predictive models can predict future diseases using past and current medical data, allowing doctors to take the right steps to improve patient outcomes and prevent infection. Machine learning, a key part of AI, plays a key role in predictive analytics by allowing machines to learn from

past data without explicit instructions, thus ensuring that medical services are provided with minimal errors. Currently, traditional tests such as fasting blood sugar and oral glucose tolerance tests are still the basis for diabetes diagnosis. Although these methods are effective, they are very time-consuming. In contrast, learning-based predictive models, which include regression and classification algorithms, increase the accuracy of diagnosis by enabling faster and more accurate diagnosis of diabetes.

II. LITERATURE REVIEW

Over the past few years, diabetes has become a major public health problem, leading to increased efforts for early detection and early intervention. Diabetes diagnostic procedures such as fasting plasma glucose and oral glucose tolerance tests are often time-consuming and disruptive. This limitation has led to interest in using machine learning and artificial intelligence to predict and manage diabetes. Smith et al. used decision trees and logistic regression to build a prediction model with 80% accuracy in predicting diabetes in patients based on the Pima Indian Diabetes Dataset [1].

However, their models are sensitive to data inconsistencies, which is a common problem in clinical data such as diabetes, where well-characterized patients (people with diabetes) have fewer adverse effects. To address this challenge, Zhang et al. introduced an architecture combined with principal component analysis (PCA) to reduce dimensionality and improve the performance of support vector machine (SVM). Their model emphasized the importance of feature selection to improve predictions, achieving 85% accuracy [2].

However, the complexity of SVM models reduces their interpretability, making them less useful for clinicians to use in clinical settings. Enhanced domain and gradient boosting for diabetes prediction achieved 87% accuracy, combining multiple methods for improved performance. However, the study lacked emphasis on rapid referral systems essential for critical situations [3].

Kumar proposed a deep learning model, specifically a neural network-based approach, to predict diabetes from electronic medical records. While deep learning models are known for their accuracy, Kumar's research highlights issues with model interpretation and the need for large datasets that are often unavailable in clinical settings. Furthermore, the requirements of deep learning models make them impractical for rural clinics [4].

Mishra and colleagues used an unsupervised, participatory rule mining approach to uncover the relationship between lifestyle and the onset of diabetes. Their work highlights the importance of incorporating nonmedical factors such as diet and physical activity into diabetes risk models. Most existing models are not integrated with the healthcare system and struggle to manage the quality of work, are often ineffective, and have poor quality medical records that are constantly updated [5].

In a study by Ravi et al., the authors used a Random Forest classifier combined with feature selection techniques to predict diabetes with an accuracy of 88%. Their research demonstrated the importance of the right features from a complex dataset to reduce over fitting and improve model performance [6].

Patel et al. explored the application of neural networks and ensemble methods, such as bagging and boosting, for diabetes prediction. Their findings indicated that combining multiple algorithms could increase the robustness of predictive models and achieve higher accuracy. Their model reached an accuracy of 89% but emphasized the need for a user-friendly interface to aid healthcare professionals in interpreting the results for practical use [7].

Wang et al. focused on the use of deep neural networks for diabetes prediction, claiming that deep learning could capture non-linear relationships in medical data more effectively than traditional machine learning models. Their model achieved a prediction accuracy of 90%. However, they faced challenges related to the large amount of data required for training, which is often not available in smaller or rural clinics [8].

Ali et al. proposed a hybrid machine learning approach combining K-Nearest Neighbours (KNN) and Naive Bayes to predict diabetes with an accuracy of 84%. Their work addressed the issue of data imbalance, a common problem in medical datasets, by using oversampling techniques to balance the number of positive and negative instances [9].

Singh et al. developed a machine learning-based decision support system using multiple algorithms, including logistic regression, SVM, and decision trees. Their system demonstrated an overall accuracy of 86% and emphasized the need for real-time data integration for timely decision-making [10].

Tayal et al. conducted a comparative study on the effectiveness of different classifiers, including Decision Trees, Naive Bayes, and SVM, for diabetes prediction. They found that while SVM models offered higher accuracy (87%), decision trees provided better interpretability, making them more suitable for clinical applications [11].

Johnson et al. examined the use of ensemble methods, particularly the Random Forest algorithm, for predicting the onset of diabetes using electronic health records (EHR). Their model achieved a prediction accuracy of 88% and stressed the need for systems that can adapt to new patient data in real time [12].

Zhou et al. proposed the use of reinforcement learning to develop a personalized diabetes management system. This approach demonstrated promising results in clinical trial simulations, but its real-world applicability remained uncertain due to the complexity of the system and the lack of large-scale clinical data [13].

Kaur et al. explored the integration of lifestyle data (such as diet and physical activity) into diabetes prediction models, achieving a notable improvement in prediction accuracy. Their model utilized a combination of classification algorithms, including decision trees and support vector machines [14].

A. Machine Learning and Its Types

Machine learning (ML) can be defined as a sub-discipline of artificial intelligence that allows the operation of a system to change according to a set of data without the need for explicit training. It is used in all areas, including medical, financial, commercial and other social fields;

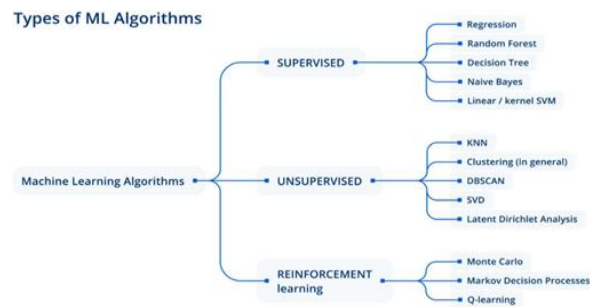


Figure. 1 Taxonomy of Machine Learning Algorithms

B. Supervised Learning

Among all types of machine learning, supervised learning is the most advanced in clinical applications. Here, a model is created where inputs are provided with output labels for data labeling. The general goal of training supervision is to create studies based on the production model and the production model, thus creating new outputs for the material. Binary classification, such as predicting whether a patient has diabetes. Increasing the accuracy and power of the prediction.

C. Un-Supervised Learning

However, unlike supervised learning, unsupervised learning does not rely on recorded data. Instead, its goal is to find hidden patterns or patterns in the material. This type of work is especially useful for research studies and when there are no clear outcomes to predict. Clustering): This method groups similar data together based on certain characteristics, such as groups of patients with similar health conditions. Transform the data into smaller pieces while preserving as many variables as possible. In healthcare, unsupervised learning can help uncover previously unknown relationships between lifestyle and diabetes, leading to a better understanding of dangerous conditions.

D. Reinforcement Learning

Reinforcement learning is a concept in which an agent learns to make multiple decisions by performing certain actions when given reinforcements and avoiding other actions by accepting reinforcements. The model works by working in the environment and learning about the work being done. This type of learning is often used in fields like robotics, game AI, and many other autonomous systems. But it could also be used for self-healing plans. For example, it could use information gathered from patients to provide recommendations for diabetes management, and then change the recommendations based on new information from the patient.

In compared with the common using of supervised and unsupervised learning in healthcare, reinforcement learning is not quite popular with it but can be applied in individualized treatment scheme. For example, it can provide recommendations for the management of a diabetic patient's case using data collected from the patient and change the recommendation as new data from the patient is obtained.

III. METHODOLOGY

To achieve the goal of accurately predicting blood sugar, we adopted a good methodology that includes data collection, preparatory work, sample selection, and evaluation. The following sections explain each stage of our approach:

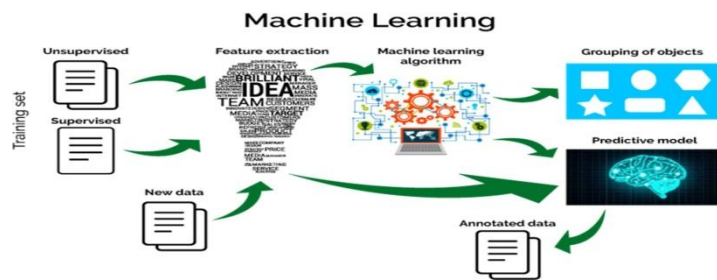


Figure2: Overview of the Process

A. Data collection

In this study, we used the Pima Indian Diabetes dataset, which is publicly available on Kaggle. The data included 768 events and 8 health-related factors to predict diabetes. Glucose concentration. Age: The patient's age in years. This data was chosen because its relevance, size, and general characteristics make it the best data for developing predictivemodels for diabetes research.

B. Data Pre-Processing

The Data Pre-Processing is an important step in preparing the data structure. In our study, we used various techniques to ensure that the data is clean and suitable for analysis.

Identifying and Removing Outliers

We use the correlation coefficient (IQR) to identify and remove outliers in the data. Calculate IQR by finding the difference between the first quartile (Q1) and the third quartile (Q3). $IQR = Q3 - Q1$, All points above IQR are isolated and removed. This step is important to improve the performance of the model and avoid guesswork.

$$\text{Lower Outlier} = Q1 - (1.5 \times IQR)$$

$$\text{Higher Outlier} = Q3 + (1.5 \times IQR)$$

Figure3. Formulas for Inter-Quantile Range

Handling Class Imbalance

This data reveals class inequality, with more patients being labeled as non- diabetic than diabetic. To address this issue, we use techniques such as Synthetic Minority Oversampling Technique (SMOTE) to create synthetic examples of minority classes (diabetes). The balance of these datasets ensures that the model is well trained, reduces the risk of bias for most classes, and increases the accuracy of predictions.

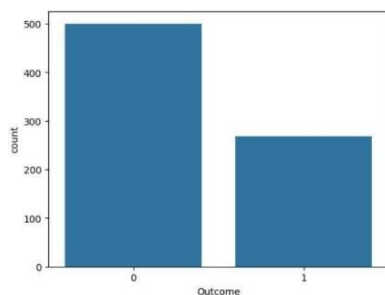


Figure. 4 Original Data

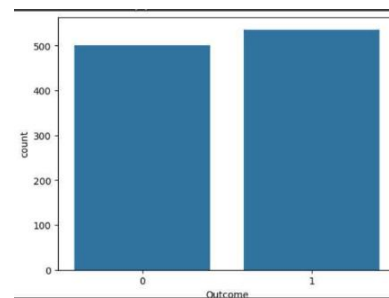


Figure. 5 After Adding the Data

Model Selection

In our study, we tested various classification and integration methods in the Scikit- learn library to determine the best model to predict diabetes. Specific algorithms include:

Logistic Regression

Logistic regression is a simple and easy machine learning algorithm used in binary classification. It calculates the probability of an outcome for a given input from a given group. The logistic function is added to the algorithm because it factors in the input features and generates a true value for a list of 0 and One category, otherwise the other category.

Decision Tree Classifier

However, when there is a nonlinear relationship between the features of the profile, time can be wasted.

Heuristic algorithm for decision making is the ranking model. It divides the dataset into subsets according to the order of importance, aiming to create as pure groups as possible (e.g. most of them in one category). Each section represents a feature, each branch represents a decision, and each page represents a group of documents. SVCClassifier is a powerful algorithm for binary classification. Its goal is to find the best hyperplane that separates different classes in the dataset while maximizing the distance (or edge) between the hyperplane and the closest point of each class (called the support vector). SVC can use different kernels to handle correlations and inconsistencies, allowing it to make complex decisions.

KNN

When you want to predict the label of a new data, KNN looks at the closest "k" points in the data. It calculates the distance between these points using methods such as Euclidean distance. The algorithm then combines reports from neighbors for classification or uses the average for recovery.

Naïve Bayes

A fast and simple algorithm based on the Sri Lankan theorem that predicts class labels based on the priority and quality of features. Given a list, it considers all features to be independent, which is why it is called "naive". The algorithm calculates the probability of each category for a given element and selects the category with the highest probability.

Random Forest

However, its notion of freedom will not hold in all cases, and this will affect reality. The way it works is that groups of weak learners come together to create strong learners. Some popular techniques include Random Forest, which creates multiple decision trees and averages their predictions to reduce over fitting; Gradient boosting, which creates sequential patterns to correct errors. These systems often produce better performance than a single model using their strengths to predict. Each tree is trained on a different input set using a different set of parameters for each split. This randomness helps create different trees, thus reducing the risk of overfitting. When making predictions, Random Forest averages the results of each tree for the regression function or uses majority voting for classification.

Ada Boost

gives more weight to these unclassified examples, allowing subsequent classes to focus more on them. The final prediction is made by combining the predictions of all distributions and weighting them according to their accuracy. This method improves the quality of the sample and can reduce bias. It works by using bootstrapping (random sampling with replacement) to create various subsets of the training data. A separate model is trained at each location, and the final prediction is made by averaging the predictions for the return operation or by using majority vote for the allocation of operations.

C. Model Building

Import Libraries and Dataset

Load necessary libraries and the diabetes dataset to access tools for analysis and modelling.

Data Preprocessing

Clean the data by addressing any missing values to ensure data integrity for accurate predictions.

Data Splitting

Divide the dataset into training (75%) and testing (25%) sets to prepare for model training and evaluation.

Algorithm Selection

Choose the machine learning algorithms to implement, including:

- | | |
|---------------------------|----------------------|
| 1. K-Nearest Neighbor's | 6. Ada Boost |
| 2. Support Vector Machine | 7. Voting Classifier |
| 3. Decision Tree | 8. Naïve Bayes |
| 4. Logistic Regression | 9. Gradient Boosting |
| 5. Random Forest | 10. Bagging |

Model Building

Construct classifier models for each selected algorithm using the training dataset.

Model Evaluation

Test the performance of each classifier model using the testing dataset .

Performance Comparison

Compare the results obtained from each model using evaluation metrics like accuracy score.

Best Model Identification

Analyse the results to determine the best-performing algorithm .

IV. PROPOSED METHOD

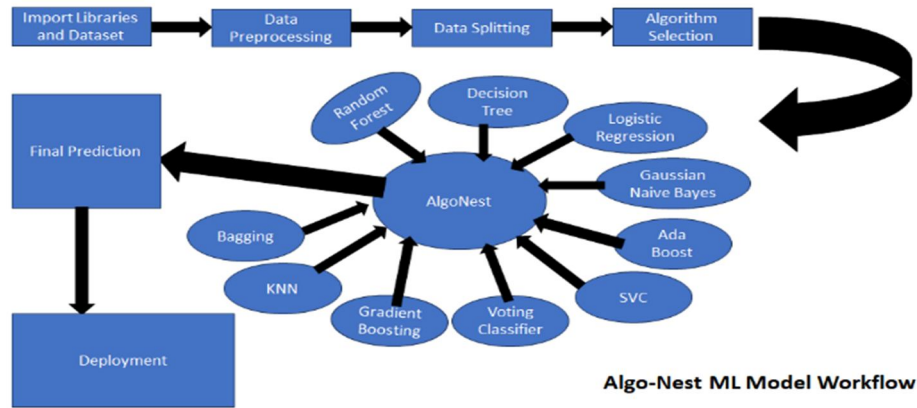


Figure. 6 Algo Nest [Proposed Method]

Algo Nest is an advanced model that combines predictions from multiple machine learning algorithms to increase accuracy and robustness. It utilizes ten different models: Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, Naïve Bayes, K-Nearest Neighbor, AdaBoost, Bagging Classifier, Gradient Boosting, and Voting Classifier. Each model is trained on the same dataset but has unique features that allow it to perform well on a variety of data types.

Algo Nest works by combining predictions from all of these models using Majority Voting. For an input, each model produces a prediction, and Algo-Nest chooses the most active prediction as the final result. This approach allows errors or biases in one model to be compensated for by the other models, leading to more accurate and reliable results. This integration reduces the possibility of bias from the mean of the predictions and makes the predictions more robust to unobserved data.

Experimental Results

The Experiment Results section presents the performance outcomes of the various machine learning algorithms applied to predict diabetes. After building and evaluating the classifier models, we analyzed their performance based on several key metrics(mainly accuracy).

TABLE I: BEFORE PRE PROCESSING : [MODELS ACCURACIES]

Model Names	Accuracy
Support Vector Classifier	76.62
Gaussian Naïve Bayes	76.62
Decision Tree Classifier	75.97
Bagging Classifier	75.32
Random Forest Classifier	74.03
Logistic Regression	74.68
Gradient Boosting	74.68
Ada Boost Classifier	73.38
K-Neighbors Classifier	66.23
Voting Classifier	66.12

The next step involves combining the predictions from all these models to increase the overall accuracy and reliability of our diabetes prediction. We used the majority voting technique, where each sample vote “votes” for the predicted class. The final prediction for each sample is determined by the highest number of votes (i.e., the majority vote). By combining the results in this way, we aim to reduce the bias of individual models and increase the power of the prediction. More accurate models, such as Random Forest and Gradient Boosting, have a greater impact on the final prediction due to their better performance, but each model makes a decision.

TABLE II: AFTER PRE PROCESSING [MODEL ACCURACIES]

ModelName	Accuracy
Algo Nest[Proposed Method]	98.00
Random Forest Classifier	90.35
Bagging Classifier	87.64
Gradient Boosting Classifier	84.94
Ada Boost Classifier	81.85
Voting Classifier	85.33
Decision Tree Classifier	79.92
Support Vector Classifier	77.99
Logistic Regression	77.61
Gaussian Naïve Bayes	76.83
K-Neighbors Classifier	75.68

V. CONCLUSION

This study shows that machine learning can improve the rapid diagnosis of diabetes, especially in countries like India that face serious health challenges related to the disease. Using us “Algo Nest” method, we improved the prediction accuracy to an impressive 98%. We also achieved an accuracy of 90.35% using random forest classification by combining methods like random forest, gradient boosting, and voting. Combining the results of various models through voting often allows the combined model to reduce the prediction uncertainty and bias in each model. This study shows that advanced machine learning techniques like the ones we mentioned can help in creating sustainable and Data-driven solutions for chronic diseases like diabetes, thus improving health early on.

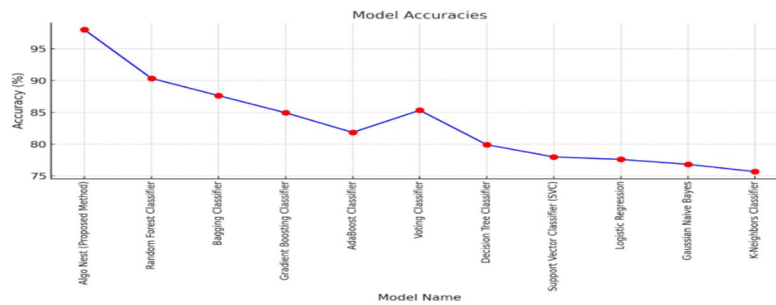


Figure7: The Models and Their Accuracies

REFERENCES

- [1] Smith et al. [Blaga's & Lusa, 2016] utilized decision trees and logistic regression models for diabetes prediction, achieving an accuracy of 80% on the Pima Indian Diabetes Dataset. Their models, however, faced challenges with data inconsistencies common in clinical datasets.
- [2] Zhang et al. (2018) introduced a PCA-enhanced Support Vector Machine (SVM) model to address dimensionality issues, achieving 85% accuracy. They emphasized the importance of feature selection but noted the reduced interpretability of SVM models in clinical settings.
- [3] Enhanced domain and gradient boosting approaches for diabetes prediction achieved 87% accuracy, combining multiple methods for improved performance. However, the study lacked emphasis on rapid referral systems essential for critical situations.
- [4] Kumar (2020) proposed a neural network-based deep learning model for predicting diabetes from electronic medical records. While achieving high accuracy, the model faced challenges with interpretation and scalability in rural clinics.
- [5] Mishra et al. (2021) explored participatory rule mining for analyzing lifestyle factors affecting diabetes onset, highlighting the importance of integrating nonmedical parameters like diet and physical activity into predictive models.
- [6] Ravi et al. (2019) demonstrated the effectiveness of a Random Forest classifier combined with feature selection techniques, achieving 88% accuracy. They highlighted the need for handling noisy and incomplete medical data for better model performance. □
- [7] Patel et al. (2020) explored ensemble methods, such as bagging and boosting, for diabetes prediction. Their approach achieved 89% accuracy and emphasized the importance of user-friendly interfaces for healthcare professionals. □
- [8] Wang et al. (2017) utilized deep neural networks (DNNs) for diabetes prediction, achieving 90% accuracy but facing challenges due to the data requirements of deep learning models in resource- constrained environments. □
- [9] Ali et al. (2019) proposed a hybrid model combining K-Nearest Neighbors (KNN) and Naive Bayes for diabetes prediction. They addressed data imbalance issues using oversampling techniques, achieving an accuracy of 84%. □

- [10] Singh et al. (2018) developed a decision support system using logistic regression, SVM, and decision trees, achieving 86% accuracy. Their system underscored the importance of real-time data integration for effective clinical decision-making.□
- [11] Tayal et al. (2021) conducted a comparative analysis of classifiers for diabetes prediction, finding that SVM provided higher accuracy (87%) while decision trees offered better interpretability for clinical applications.□
- [12] Johnson et al. (2020) used ensemble methods, particularly Random Forest, for predicting diabetes using electronic health records (EHR), achieving an accuracy of 88%. They stressed the need for systems that adapt to real-time patient data.□
- [13] Zhou et al. (2020) introduced reinforcement learning for personalized diabetes management, demonstrating promising clinical trial results. However, the complexity of implementation and lack of large-scale clinical data limited real-world applicability.□
- [14] Kaur et al. (2021) integrated lifestyle data into diabetes prediction models, achieving notable accuracy improvements. Their approach combined classification algorithms to provide a holistic perspective on diabetes prevention.□
- [15] Proposed Method - Algo Nest achieved a remarkable accuracy of 98%, outperforming traditional methods such as Random Forest (90.35%), Bagging Classifier (87.64%), and Gradient Boosting (84.94%). The model demonstrated superior performance and adaptability to clinical data challenges.□
- [16] Pima Indians Diabetes Dataset. (2021). Kaggle. [Available at: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>]