Journal of Information Systems Engineering and Management
2025, 10(28s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The Role of Machine Learning Algorithms for Diagnosing Diabetes Mellitus Based on Different Datasets with Different Attributes

Ahmad Hussain AlBayati [1], Shnoo Abdual Aziz Zangana [2]
[1, 2] *Computer Science Department, Computer Science and Information Technology, University of Kirkuk, Iraq*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Diagnosing diabetes type 1 and type 2 early can avoid a variety of complications, including nephropathy, retinopathy, neuropathy, and multiple diseases such as renal diseases, visual impairments, and cardiovascular diseases. This paper employs machine learning algorithms using data mining techniques to predict diabetes effectively. We focus on three different datasets with varying attributes that complement each other. Even the Pima Indian dataset and the Healthcare Diabetes dataset are commonly used in machine learning research. Still, they lack essential attributes, such as HbA1c level, crucial for diabetes research. In contrast, the Iraq dataset includes HbA1c levels and risk factors, such as hyperlipidemia tests measuring cholesterol, triglycerides, high-density lipoproteins (HDL), and low-density lipoproteins (LDL). Hence, using different data sets provides a more comprehensive evaluation of type 2 diabetes. Our data mining process involves data cleaning and ensuring data integrity. For more integrity, we compare machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, Gaussian Naive Bayes, Decision Tree, and K-Neighbors, to identify the most effective method for diabetes prediction. The new methodology relies on the predictive accuracy of robust machine learning algorithms, where the evaluation of the algorithms is achieved through multiple metrics such as precision, recall, and F1 score. However, we utilised k-fold cross-validation and train-test split techniques to assess the models. The results indicate that Gradient Boosting performed best in predicting diabetes within the Pima Indian dataset, while the K-Neighbors algorithm demonstrated superior performance in the Healthcare Diabetes dataset. Moreover, the Decision Tree method showed greater efficiency in the Iraq dataset than the other algorithms.

Keywords: Diabetes Prediction, Machine Learning Algorithms, Data Mining, Logistic Regression LR, Random Forest RF, Gradient Boosting GB, Gaussian NB GNB, Decision Tree DT, K Neighbors KNN.

INTRODUCTION

Today, the world is grappling with a growing epidemic of chronic diseases, including heart disease, cancer, diabetes, and tuberculosis. The importance of early detection cannot be overstated, as individuals often endure these conditions for long periods. Despite ongoing research efforts to manage these illnesses, their incidence has risen alarmingly. Therefore, more comprehensive studies are needed to combat these health threats effectively. This paper aims to shed light on diabetes mellitus (DM), one of the most pervasive chronic diseases. Diabetes is an order that results in elevated blood sugar levels. In a healthy body, insulin facilitates the movement of sugar from the bloodstream into cells to convert it into energy. However, in diabetes, the body either produces inadequate insulin or ceases to produce it altogether, leading to significant health complications. Chronic DM takes a heavy toll on individuals, making awareness vital. The most common forms of diabetes include type 1, type 2, Prediabetic, and gestational diabetes. By understanding these conditions, we can take meaningful action to reduce their impact on society. Type 1 diabetes is a serious, chronic condition where the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. In contrast, type 2 diabetes arises when the body can no longer produce

enough insulin, resulting in dangerously high blood sugar levels. Remarkably, research shows that early identification and intervention can prevent up to 80% of type 2 diabetes cases, underscoring the importance of regular screenings. Pre-diabetes, characterized by elevated blood sugar levels that fall short of a type 2 diagnosis, serves as a critical warning sign. Taking proactive steps now can dramatically change your health future.

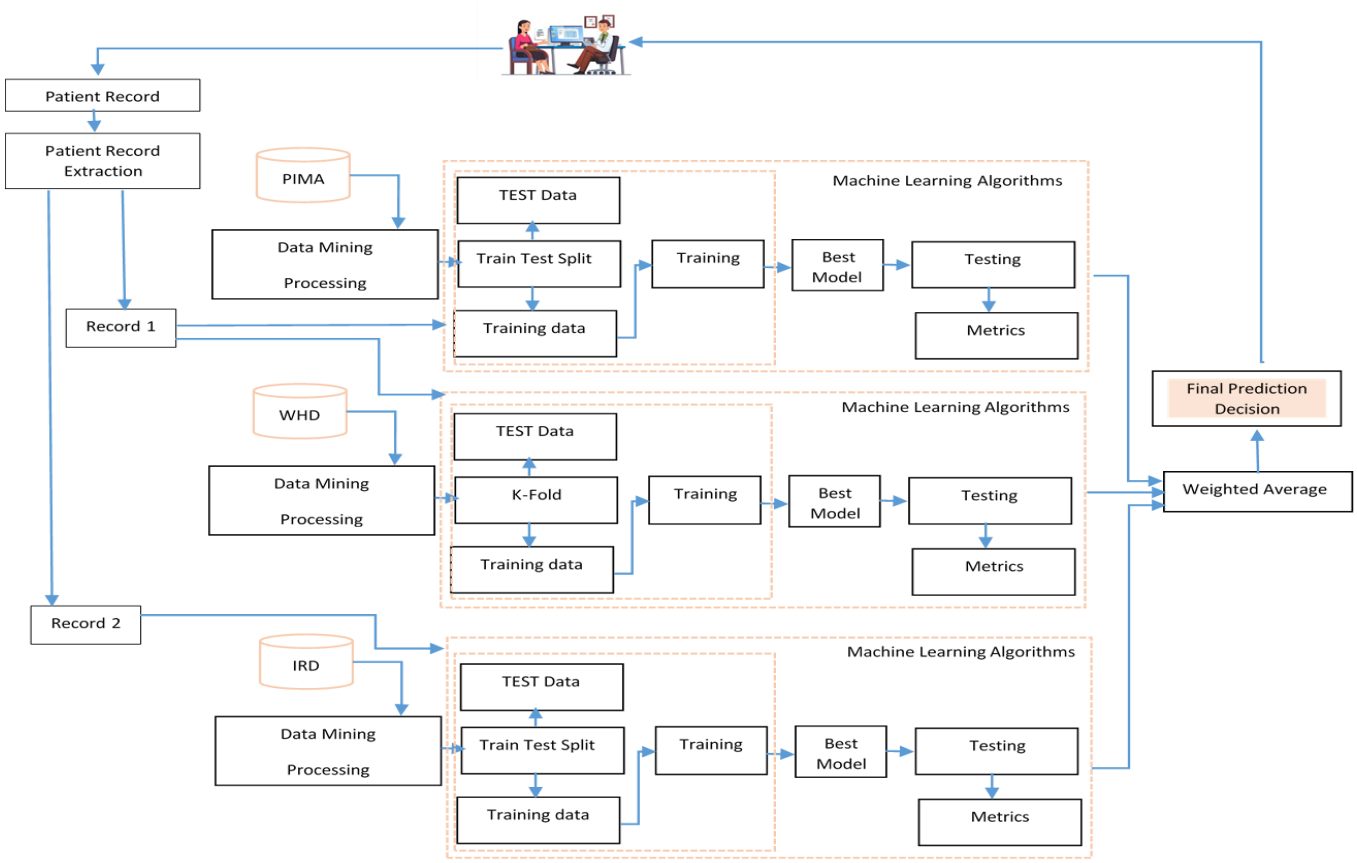

Figure 1 Structure of new methology for Diabetes Mellitus predication

Gestational diabetes is a serious condition affecting pregnant women, marked by elevated blood sugar levels[1]. This form of diabetes not only poses risks to the mother's health but also affects the developing baby[2]. On the other hand, Type 2 diabetes mellitus (T2DM) is a chronic disease characterised abnormal blood glucose levels due to insulin deficiency or resistance. This condition arises from various controllable factors, including insufficient physical activity, poor dietary choices, tobacco use, excessive alcohol consumption, and obesity. Importantly, T2DM significantly increases the risk of developing cardiovascular diseases such as coronary heart disease, stroke, peripheral arterial disease, and aortic diseases—all linked to high blood pressure resulting from diabetes. Understanding these risks is crucial for preventing and maintaining health[3]**.** Imagine a future where doctors can accurately diagnose diabetes where doctors can accurately diagnose diabetes in patients before symptoms even appear. By utilising a robust model that considers eight critical factors—the number of pregnancies, plasma glucose levels, diastolic blood pressure, skin thickness, insulin levels, body mass index, diabetes pedigree function, and age. to accurately identify at-risk individuals. This proactive approach enhances early detection and transforms diabetes management[4].

With the rapid advancements in machine learning and artificial intelligence, we now have access to various robust classifiers and clustering algorithms, including Logistic Regression, Random Forest, Gradient Boosting, Gaussian Naive Bayes, Decision Trees, and K-Neighbors. These tools are practical and essential for tackling pressing challenges today. Classification is the premier machine learning technique in medicine, addressing real-life problems that impact our daily lives[5]. Furthermore, we strategically utilise feature selection methods to significantly boost the accuracy of our results, ensuring the best possible outcomes.

METHODOLOGY

This section will delve into the powerful classifiers that machine learning employs to predict diabetes with remarkable accuracy. The proposed methodology, as depicted in Figure 1, is designed to meet the rigorous standards set by modern medical studies. The accompanying block diagram outlines a clear and systematic approach to conducting this research, focusing on data analysis through advanced machine learning and data mining techniques. The primary objective is to harness various datasets, including PIMA, WHD, and IRD, to tackle specific challenges and successfully meet the study's goals. The process commences with meticulous data collection and processing. Initially, the data undergoes thorough cleaning to eliminate errors and invalid values. For data sourced from diverse locations, integration follows to ensure uniformity. Subsequently, the dataset is divided into two key groups: Training and Test datasets. Utilizing K-fold cross-validation, the dataset is segmented into k homogeneous folds—one fold is allocated for testing, while the remaining folds are amalgamated to form the training subset. This strategic method significantly reduces errors that may arise from data partitioning. During the training phase, six leading machine learning algorithms are employed to develop robust models. The most effective model is chosen based on each algorithm's performance and evaluated through precise criteria and metrics such as accuracy, sensitivity, and other indicators of quality. A weighted average technique ultimately synthesises predictions from these distinct models, providing a final prediction based on the aggregated results. This rigorous process is executed for each dataset (e.g., Record 1 and Record 2), ensuring uniformity and thoroughness in methodology to achieve precise and insightful results. This redundancy empowers the system to manage diverse data types adeptly, achieving exceptional accuracy in prediction and classification.

DATA MINING TECHNIQUES

Data mining is an essential and innovative technique for uncovering valuable insights from vast data volumes. By leveraging existing data, businesses can gain meaningful conclusions that drive decision-making. This powerful process involves analysing intricate models within large datasets through machine learning, statistical analysis, and robust database systems. As a result, organisations can effectively identify patterns, classify data through clustering, detect anomalies, and reveal crucial relationships or dependencies. Classification, a vital data mining function, categorises database objects into distinct target groups. This approach is an essential pre-processing step before implementing data into the classification framework. By aggregating similar elements, we can define a comprehensive dataset that includes all relevant attributes and combine them based on their similarities. Clustering, akin to segmentation, is a strategic method for organising unprocessed instances according to various characteristics, ultimately leading organisations to make well-informed, data-driven decisions[2].

A. Data Collection

To guarantee the robustness of our model, we assembled three datasets, each comprising a distinct number of attributes.

B. Data Pre-Processing

Preprocessing is the preliminary phase that converts input data into an appropriate format for processing. This phase encompasses several actions, such as merging, restructuring, and manipulating data to cleanse, integrate, reduce, and discretise it.

C. Data Cleaning

This step entails identifying incomplete, erroneous, or inaccurate data and absent values. While this dataset lacks missing values, it has undergone a cleaning process to eliminate duplicates and null entries[6].

D. Data Integration

Data integration is a strategy for retrieving and consolidating disparate information into a cohesive format and structure[2]. Transforming the original data measurement scale into a comprehensible format for the analytic tool is essential.

E. Data Reduction

This phase entails a data reduction procedure to enhance storage efficiency while reducing time and expenses. Nevertheless, the author opts not to employ this stage, as the diabetic dataset is already structured appropriately for the analysis phase.

F. Data Transformation

This step entails converting the data into a format appropriate for the chosen analytical approach. Data transformation is essential to change the original measurement data into a format compatible with the analyser to analyse the diabetes dataset. In this instance, the author omits this step as the diabetes dataset is already structured in a manner conducive to study.

MACHINE LEARNING ALGORITHMS

Machine learning (ML) is an essential branch of artificial intelligence that derives valuable insights from data patterns. The learning process occurs in two main stages[7]: first, it analyses the provided dataset to identify the system's unknown dependencies; second, it generates predictions based on these identified dependencies to assess the system's likely new outputs. Additionally, K-fold cross-validation is a key technique for improving the training and evaluation of machine learning models without requiring the duplication of datasets[8]. This method offers a thorough understanding of model performance across different data subsets and is also known as Rotation Estimation. Its versatility and effectiveness make it a valuable tool for achieving reliable model evaluations[9].

A. K Nearest Neighbor Algorithm (KNN)

The k-NN algorithm is frequently regarded as one of the most straightforward machine learning techniques. Model construction solely entails retaining the training dataset. The algorithm determines the nearest neighbours from the training set to predict a new data point.

B. Naive Bayes algorithm

Naïve Bayes is a prevalent model in machine learning applications because it is simple, permitting all qualities to contribute equally to the ultimate conclusion. This simplicity improves computing efficiency, rendering the Naïve Bayes method attractive and appropriate for diverse fields. The fundamental elements of the Naïve Bayes Classifier consist of three principal aspects: prior probability, posterior probability, and class conditional probability[10]. Prominent uses of this technique encompass spam email filtration and document classification[5].

C. Decision Tree Algorithm

The Decision Tree (DT) algorithm stands out in supervised machine learning for effectively addressing regression and classification challenges. Systematically partitioning data based on chosen variables creates a clear structure of nodes, with the leaves representing final decisions. The essence of a decision tree is to construct a model that accurately predicts the target variable, leveraging straightforward decision rules derived from training data[5]. With their intuitive approach, decision trees empower us to make informed choices from established options while uncovering additional potential solutions. Embrace this powerful tool to enhance your predictive capabilities and achieve better decision-making outcomes[7].

D. Logistic Regression

Logistic regression is a supervised learning algorithm predominantly employed for binary classification tasks. It utilises a mathematical model with the logistic function to ascertain the probability that a particular data point is classified into a specific category. Numerous intricate extensions of logistic regression exist. It operates as a regression model that forecasts the likelihood of an item belonging to a particular class.

E. Random Forest Classification

The random forest is an ensemble model that can serve as a nearest-neighbor predictor. The fundamental concept of ensemble methods is that multiple models can collaborate to produce a more robust predictor.

OA represents the overall accuracy, and *IE* denotes a metric that assesses the concurrence between the model predictions and the actual class values as if occurring randomly. The random forest enhances the conventional

machine learning methodology of decision trees by employing the ensemble idea. A key advantage of using a random forest classifier is its efficacy in addressing multiple challenges, such as efficient execution times, imbalanced datasets, and absent values[11].

F. Gradient Boosting classification

The novel dataset or test data is distributed across all the constructed sub-trees in a random forest. Each decision subtree in the forest can determine the class of the dataset. The model then determines the most appropriate class via majority voting. The ensembles are developed sequentially, with each subsequent ensemble rectifying the flaws of its predecessor[12].

PERFORMANCE CRITERIA

This is the final phase of the predictive model. The prediction outcomes are evaluated using various assessment indicators. Metrics encompass classification accuracy, confusion matrix, and F1-score. Classification accuracy denotes the ratio of valid predictions to the total number of input samples.

- Root mean square error (RMSE): It is a prominent methodology to assess the error within the model for forecasting statistical data. The Root Mean Square Error (RMSE) scores fall within the range of 0.0 to 0. 5, indicating that the model can accurately forecast the data.

$$\sqrt{R}MSE = \frac{1}{m}\sum_{i=0}^{m}(xi - yi)^2 \qquad (1)$$

- Absolute error (MAE): MAE may be applicable in cases where outliers indicate damaged parts within the data set. The absolute error (MAE) does not overly penalise outliers in the training data as (the L1 criterion effectively mitigates the impact of any outliers), thus providing a comprehensive and constrained performance assessment. For the model. Conversely, if the test set contains many outliers, the model performance will likely be unsatisfactory.

$$MAE = \frac{1}{m}\sum_{i=0}^{m}|xi - yi| \qquad (2)$$

- Kappa value: The kappa value quantifies the statistical efficacy of a machine learning classifier in categorising events and assigning labels based on established truths while considering the anticipated accuracy of a random classifier. It assesses the effectiveness of a classifier for a specific dataset.

$$Kappa = \frac{OA - i}{1 - IE} \qquad (3)$$

- The Receiver Operating Characteristic (ROC) curve is a powerful tool that visually represents the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across various thresholds. This visualisation is essential for distinguishing between true and error in data. Furthermore, the mean absolute error provides a precise measure of accuracy by calculating the average absolute differences between actual and predicted values, offering valuable insights into the model's performance across the entire dataset[1].
- The Confusion Matrix is an essential tool that effectively highlights the model's overall performance. It uses the terms TP, FP, FN, and TN—short for True Positive, False Positive, False Negative, and True Negative—to provide valuable insights into the model's accuracy.
- - Accuracy: You can easily gauge the model's reliability by calculating the mean of the values along the primary diagonal of the matrix. Remember, 'N' indicates the total number of samples, making this assessment informative and straightforward.

$$Accuracy = \frac{TP+FN}{N} \qquad (4)$$

- The F1 score is a critical metric for evaluating the effectiveness of a test, as it accurately measures precision. By representing the harmonic mean of precision and recall, the F1 score ranges from zero to one, providing a clear

indication of your classifier's accuracy and reliability. It is essential for achieving a balance between precision and recall, ensuring that your model performs optimally. Its mathematical formulation underscores its importance in assessing performance in a nuanced way:

$$F1 = 2 \frac{1}{\left(\frac{1}{Precision}\right)+\left(\frac{1}{Recall}\right)} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

As a result, after applying many machine learning algorithms to three separate datasets—Pima Indian Data, Healthcare data, and Iraq data—the resulting accuracies are detailed below.

RESULTS

This part relates to the findings and examination. The proposed technique is implemented using Python and Google Collab notebook. Python is an open-source programming language. The execution is accelerated. It supply integrated library files for executing Python program which is most suited for data mining applications [13]. After applying many machine learning algorithms to three separate datasets—Pima Indian data, Healthcare data, and Iraq data where the resulting accuracies are detailed as follows

A. Data mining techniques Processes

a) Data Collection

To guarantee the robustness of our model, we assembled three datasets, each comprising a distinct number of attributes. The first dataset was obtained from the Indian Pima Diabetes database accessible on Kaggle. The dataset comprises 768 entries categorized into eight attributes and two classes: Class 1 contains 268 entries, whereas Class 0 encompasses 500 instances. The second dataset comprises 2,768 entries, categorized into eight attributes and two classifications within the healthcare domain. Category 1 comprises 718 items, whilst Category 0 contains 1,581 entries

- (https://www.kaggle.com/uciml/pimaindians-diabetes-database**).**

The study also utilized data from Iraq, accessible from the following site:

- (https://data.mendeley.com/datasets/wj9rwkp9c2/1/files/2eb60cac-96b8-46ea-b971-6415e972afc9)

Where this dataset, accessible to the public, was uploaded on the Mendeley website in July 2020 by the University of (Chol), Fasting Lipid Profile (LDL, VLDL), Triglycerides (TG), HDL Cholesterol, and HbA1c.

It similarly includes the classifications for diabetes (Y), non-diabetic (N), and pre-diabetic (P). In this study, we categorize the pre-diabetic (P) class as a subset of the diabetic (Y) class. There are 897 occurrences for the positive class and 103 for the negative class**.** The dataset has 1,000 instances with 12 attributes: Patient Number, Age, Gender, Creatinine Ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting Lipid Profile (LDL, VLDL), Triglycerides (TG), HDL Cholesterol, and HbA1c.

Table 1. Samples of diabetes data with attribute values

Pregnancies	*Glucose*	*Blood Pressur*	*Skin Thickness*	*Insulin*	*BMI*	*Diabetes*	*Pedigree Function*	*Age*	*Outcome*
1	*6*	*148*	*72*	*35*	*0*	*33.6*	*0.627*	*50*	*1*
2	*1*	*85*	*66*	*29*	*0*	*26.6*	*0.351*	*31*	*0*
3	*8*	*183*	*64*	*0*	*0*	*23.3*	*0.672*	*32*	*1*
4	*1*	*89*	*66*	*23*	*94*	*28.1*	*0.167*	*21*	*0*

b) Data Pre-Processing

Preprocessing is the preliminary phase that converts input data into an appropriate format for processing. This phase encompasses several actions, such as merging, restructuring, and manipulating data to cleanse, integrate, reduce, and discretize it. The preprocessing methods employed may differ according on the objectives to be attained.

Choosing the correct procedures is essential, as efficient data preprocessing can greatly improve classification results. To guarantee high-quality data for analysis, preprocessing must encompass essential procedures including Data Integration, Data Cleaning, Data Reduction, and Data Transformation. The author has already conducted these preparation actions to generate high-quality data.

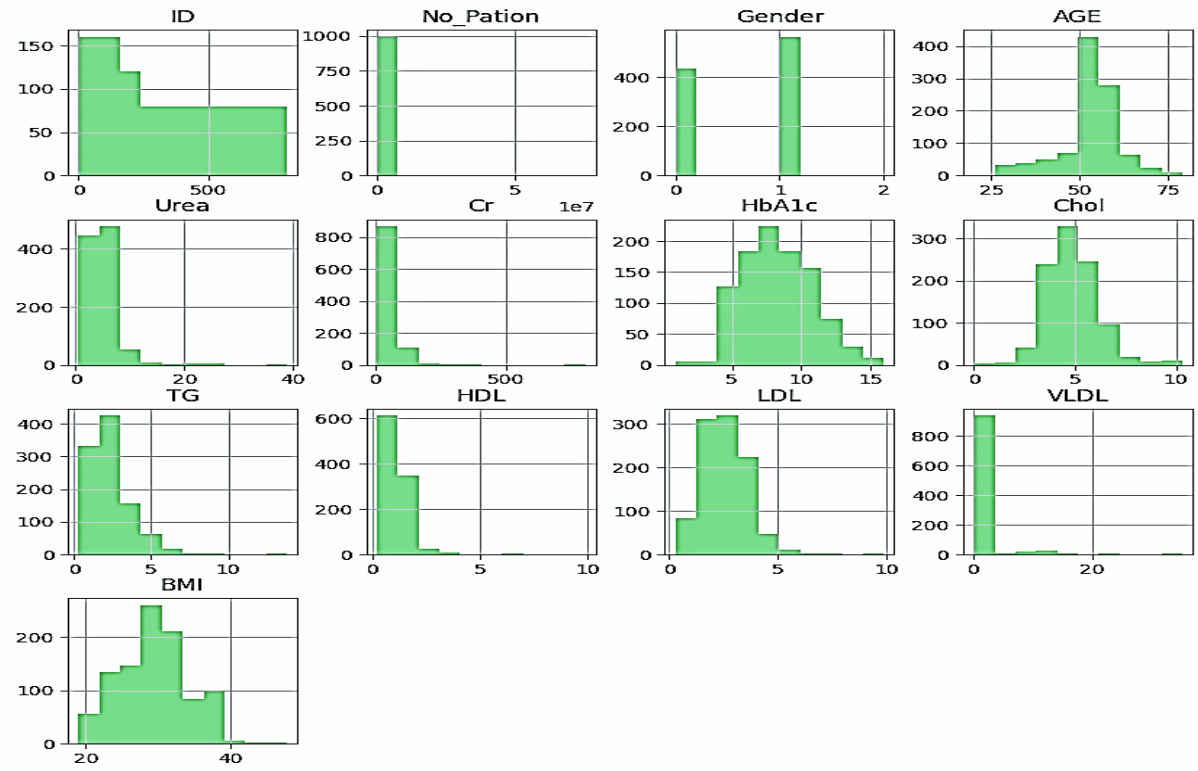

Figure 2. Data Distribution (Diabetes (Iraq))

c) Data Cleaning

Data cleansing is conducted intuitively, utilising information from 25 distinct instruments or through a structured deployment. Data cleansing is often called data scrubbing or data purification[2]**.**

Table 1 Description of the dataset

Properties	***PIDD***
Attribute	9
Instances	768
Missing Values	0
Outcome	2(0,1)

LITERATURE SURVEY

In [1], researchers showcased an array of machine learning predictive algorithms aimed at effectively predicting diabetes, including K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Logistic Regression (LR). Notably, SVM achieved a commendable accuracy of 74%. However, the KNN and RF algorithms excelled on the German database, reaching a remarkable accuracy of 98.7%,

demonstrating their strong potential in real-world applications. To enhance diabetes diagnosis and prevention, the authors advocated for leveraging data extraction techniques [2]. Their application of Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machines was meticulously evaluated using accuracy, confusion matrices, and sensitivity metrics, revealing the highest accuracy at 82.64% compared to other models. In [14], the potential of machine learning to save lives is underscored by the development of an artificial intelligence model designed to predict diabetes. This study employed robust supervised machine learning techniques by comparing KNN and Naïve Bayes algorithms through various health features in the dataset. The Naïve Bayes algorithm emerged as the superior choice, with outstanding average results of 76.07% accuracy, 73.37% precision, and 71.37% recall, solidifying its place as a vital tool in diabetes prediction.

The two-stage model selection methodology[6] is a pivotal advancement in predictive analytics, particularly healthcare. In the first stage, various powerful algorithms—including logistic regression, support vector machines, nearest neighbours, gradient boosting, Naive Bayes, and random forest were rigorously evaluated based on their ability to predict outcomes from patients' preconditions. Impressively, the Random Forest model emerged as the leader, achieving a notable accuracy of 80.7% after implementing the Synthetic Minority Over-sampling Technique (SMOTE). The widely recognised CRISP-DM framework ensured systematic data management and model development[15]. The analysis was conducted using the versatile R programming language. Remarkably, the Random Forest model further enhanced its credibility by attaining an accuracy of 90.43%.

Moreover, feature selection was meticulous, derived from a secondary dataset comprising seventeen well-defined attributes devoid of irrelevant data or missing values[16]. The dataset was expertly processed using the Ada Boost algorithm, with Decision Trees as the foundational model, in conjunction with support vector machines (SVM) and an innovative ensemble model rooted in machine learning principles.

However, [17] aims to effectively predict diabetes in patients by utilising diagnostic measurements from a comprehensive dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases. Out of 768 records analysed, a significant 500 (65.1%) were healthy, whereas 268 (34.9%) were diagnosed with diabetes. Notably, the accuracy of various algorithms employed in this research ranged impressively from 81% to 89%, showcasing the potential for reliable diabetes detection and early intervention.

The study in [18] accurately predicts diabetes in patients by utilising diagnostic measurements from a comprehensive dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases. Out of 768 analysed records, 500 (65.1%) were classified as healthy, while 268 (34.9%) were diagnosed with diabetes. Notably, the accuracy of various algorithms used in this research ranged impressively from 81% to 89%, highlighting their potential for reliable diabetes detection and early intervention. In contrast, the random forest model is more effective when additional features are included.

Finally, [19] proposes a robust framework for developing a diabetes prediction model to assist in the clinical diagnosis of diabetes with different supervised machine learning models, including the Random Forest (RF) model, the Support Vector Machine (SVM) model, and Twice-Growth Deep Neural Network (2GDNN) model for classification. The results conducted on the PIMA Indian and LMCH diabetes datasets yielded impressive results with the proposed 2GDNN model, achieving precision, sensitivity, F1-score, training accuracy, and testing accuracy scores of 97.34%, 97.24%, 97.26%, 99.01%, and 97.25% for training, as well as 97.28%, 97.33%, 97.27%, and 99.57%, 97.33% for testing, respectively.

DISCUSSION

Through rigorous study and simulation, we have conducted a comprehensive examination of cutting-edge predictive algorithms for diabetes, a vital area within healthcare. To predict diabetes onset effectively, we have hand-picked a selection of powerful machine learning algorithms: logistic regression (LR), random forest (RF), gradient boosting, Gaussian naive Bayes (NB), decision tree (DT), and k-nearest neighbors (KNN). Our innovative framework is purpose-built to determine the most effective algorithm for diabetes datasets that share similar characteristics, ensuring optimal performance. Hence, we source our datasets from Kaggle, where they are meticulously curated to eliminate missing values. During the preprocessing phase, we utilized a class balancing filter to guarantee that all instances are fairly represented. In addition, our analysis uses a comprehensive range of evaluation metrics, including precision, recall, ROC area, kappa value, mean absolute error (MAE), root mean square error (RMSE), and relative

absolute error (RAE) as shown in Tables 3-7. This multifaceted strategy not only strengthens the reliability of our predictions but also enables healthcare professionals to make well-informed decisions in managing diabetes effectively. By leveraging our findings, we can transform how diabetes is predicted and treated, ultimately improving patient outcomes.

Moreover, TP, or true positive, represents the number of cases accurately classified as sick, highlighting the effectiveness of our classification system. TN, or true negatives, indicates cases that were misleadingly classified as sick. False positives (FP) represent instances where healthy conditions are incorrectly identified as being unwell, while false negatives (FN) denote cases where sick conditions are mistakenly deemed healthy. Therefore, in our research, we harnessed the power of feature importance, as illustrated by the Iraq dataset Figure 3. Our analysis using a Decision Tree revealed that the most critical features were HbA1c, AGE, and BMI, resulting in an impressive accuracy of 0.9916. Furthermore, the decision tree classifier demonstrated a remarkable accuracy of 0.9463 on the global health data when all features were considered, as depicted in Figure 4. This evidence underscores the reliability of our approach and the potential for significant contributions to health outcomes.

Table 2 PID dataset analysis

NO.	*Algorithm*	*Accuracies*	*Recall*	*F_Score*	*Precision*	*cohen_kappa*
1.	*LR*	*0.947*	*0.935*	*0.945*	*0.95*	*0.885*
2.	*RF*	*0.869*	*87*	*86.5*	*87*	*0.704*
3.	*GB*	*0.982*	*0.975*	*0.985*	*0.985*	*0.962*
4.	*GNB*	*0.964*	*95.5*	*0.96*	*0.975*	*0.923*
5.	*DT*	*0.965*	*0.9571*	*0.964*	*0.964*	*0.839*
6.	*KNN*	*0.929*	*0.93*	*0.925*	*0.925*	*0.852*

Table 3 Healthcare dataset analysis

NO.	*Algorithm*	*Accuracies*	*Recall*	*F_Score*	*Precision*	*cohen_kappa*
1.	*LR*	*0.7648*	*0.7493*	*0.7491*	*0.75006*	*0.5282*
2.	*RF*	*0.9526*	*0.94464*	*0.94462*	*9453*	*0.9051*
3.	*GB*	*0.9937*	*0.977460*	*0.977461*	*0.9767*	*0.9874*
4.	*GNB*	*0.7510*	*0.7383*	*0.7372*	*0.7436*	*0.4997*
5.	*DT*	*0.9463*	*0.9086*	*0.9078*	*0.9092*	*0.8925*
6.	*KNN*	*0.9968*	*0.983394*	*0.983393*	*98360*	*0.9937*

Table 4 IRD analysis

NO.	*Algorithm*	*Accuracies*	*Recall*	*F_Score*	*Precision*	*cohen_kappa*
1.	*LR*	*0.935*	*0.933*	*0.933*	*0.936*	*0.9030*
2.	*RF*	*0.9763*	*0.9767*	*0.973*	*0.9767*	*0.964*
3.	*GB*	*0.989*	*0.99*	*0.986*	*0.99*	*0.983*
4.	*GNB*	*0.988*	*0.99*	*0.986*	*0.99*	*0.982*
5.	*DT*	*0.991*	*0.993*	*0.990*	*0.993*	*0.987*
6.	*KNN*	*0.9968*	*0.983394*	*0.98339*	*98360*	*0.993*

Table 5 IRD error rate analysis

No.	*Alg.*	MAE	*RMSE*	*AVE*	*Time*
1.	***LR***	*0.071*	*0.29*	*0.081*	***0.8ms***
2.	***RF***	*0.031*	*0.21*	*0.046*	***33ms***
3.	***GB***	*0.021*	*0.20*	*0.042*	***16ms***
4.	***GNB***	*0.023*	*0.21*	*0.047*	***0.89ms***
5.	***DT***	*0.016*	*0.18*	*0.033*	***1.6ms***
6.	***KNN***	*0.022*	*0.19*	*0.039*	***23ms***

Table 6 Healthcare dataset error rate analysis

No.	*Alg.*	MAE	*RMSE*	*AVE*	*Time*
1.	***LR***	*0.23*	*0.48*	*0.17*	***3.12ms***
2.	***RF***	*0.04*	*0.21*	*0.04*	***22.3ms***
3.	***GB***	*0.006*	*0.07*	*0.006*	***24.4ms***
4.	***GNB***	*0.24*	*0.49*	*0.18*	***0.7ms***
5.	***DT***	*0.05*	*0.23*	*0.05*	***0.4ms***
6.	***KNN***	*0.0031*	*0.05*	*0.0032*	***8.9ms***

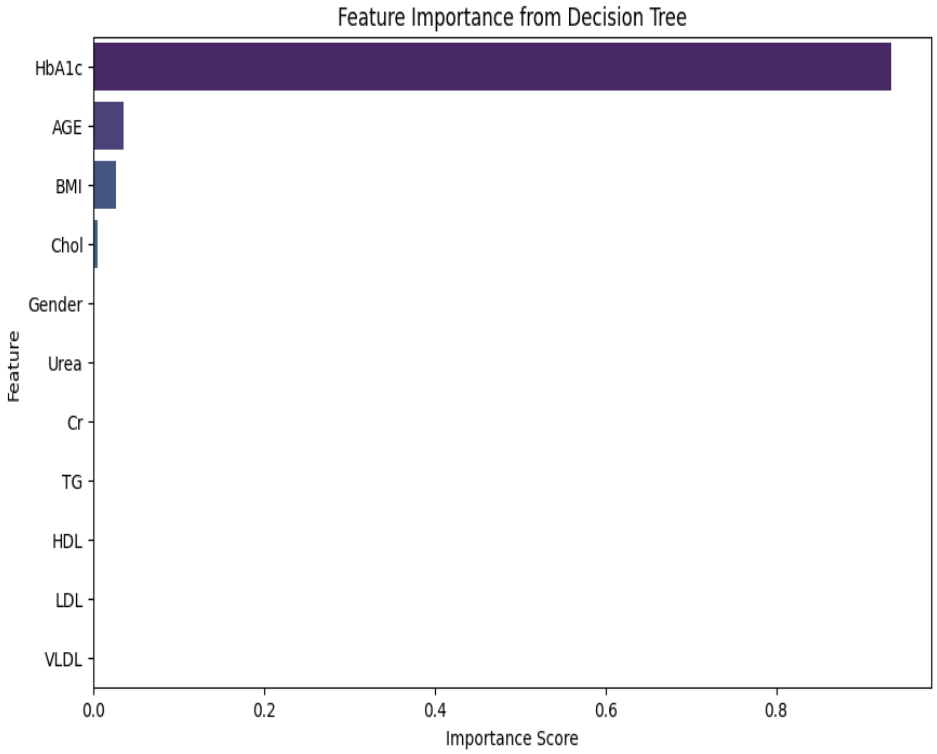

Figure 3 Feature Importance in Decision Tree (Iraq)

Feature importance evaluates each feature's significance in a tree's decision-making process. It is a value ranging from 0 to 1 for each attribute, where zero signifies "not utilised at all", and one indicates "perfectly predicts the target", as shown in Figure 3. The three most essential features in the decision tree are HbA1c, Age, and BMI.

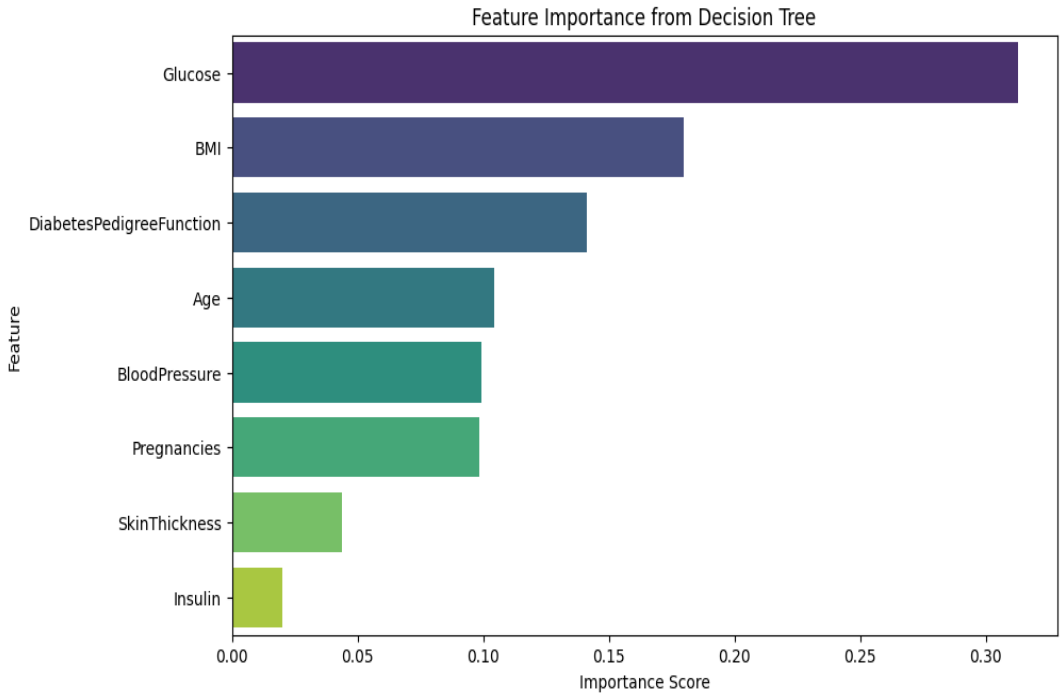

Figure 4 Feature Importance in Decision Tree

The importance of attributes in the decision tree model shows that "Glucose" is the most significant variable, followed by "BMI" as the second most important attribute.

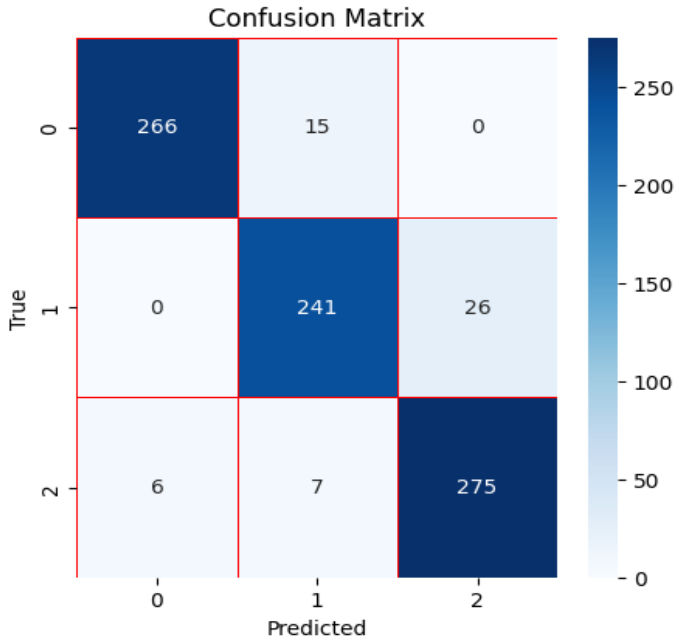

Figure 5 Confusion matrix of model Logistic regression

CONCLUSION

Diabetes mellitus (DM) is a significant global health concern, with its prevalence rapidly increasing. Undiagnosed diabetes can lead to numerous complications, including retinopathy, nephropathy, neuropathy, and various vascular disorders. Both type 1 and type 2 diabetes are among the leading causes of mortality worldwide and are associated with renal disease, visual impairment, and cardiovascular conditions. Data mining techniques can enhance healthcare decision-making by facilitating accurate disease diagnosis and treatment, thereby reducing the burden on

healthcare specialists. We can leverage machine learning to reduce mortality by developing an artificial intelligence model capable of predicting diabetes. Our methodology involves comparing various algorithms Logistic Regression, Random Forest, Gradient Boosting, Gaussian Naive Bayes, Decision trees, and K-neighbors to identify the most effective one for predicting diabetes. This research presents a model that predicts whether a patient will develop diabetes. Our approach focuses on the predictive accuracy of robust machine learning algorithms, evaluated using multiple metrics, including precision, recall, and the F1 score. We implemented k-fold cross-validation and utilized a train-test split based on different metrics. Three datasets were used: the Pima Indian Diabetes Dataset, a healthcare dataset sourced from Kaggle, and the Iraq dataset, which was used to forecast the onset of diabetes via diagnostic methods. However, the Pima Indian Diabetes Dataset is not comprehensive for diagnosing type 2 diabetes, as it omits critical data such as HbA1c levels and does not account for essential risk factors like hyperlipidemia and metabolic syndrome. In contrast, the other datasets incorporated additional risk factors, including assessments for hyperlipidemia through cholesterol, triglyceride, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) measurements, providing a more thorough assessment of type 2 diabetes and its associated risk factors. Results from our analysis indicated that the Gradient Boosting algorithm was the most effective for predicting diabetes using the Pima Indian dataset, while the K-Neighbors algorithm performed best with the Healthcare Diabetes dataset. The Decision Tree method showed greater efficiency in predicting diabetes for the Iraq dataset compared to the other algorithms. Additionally, we evaluated the efficacy of our proposed model through the confusion matrix, sensitivity, and accuracy metrics. As a result, the paper focuses on significant features rather than utilizing all available features, and we conducted data cleansing with a focus on potential predictive factors.

REFERENCES

[1] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics,* vol. 12, no. 3, pp. 1728-1737, 2023.

[2] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors,* vol. 25, p. 100605, 2023.

[3] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications,* vol. 83, no. 8, pp. 24153-24185, 2024.

[4] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine,* vol. 13, no. 3, p. 406, 2023.

[5] I. M. Ibrahim and A. M. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *learning,* vol. 4, no. 5, p. 6, 2021.

[6] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," *Mobile Information Systems,* vol. 2022, no. 1, p. 6521532, 2022.

[7] A. Urgiriye and R. Bhartiya, "Review of machine learning algorithm on cancer data set," *International Journal of Scientific Research & Engineering Trends,* vol. 6, no. 6, pp. 3259-3267, 2020.

[8] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data,* vol. 11, no. 1, p. 44, 2024.

[9] A. Ram and H. Vishwakarma, "Diabetes prediction using machine learning and data mining methods," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1116, no. 1: IOP Publishing, p. 012135.

[10] R. P. A. Murti, S. M. Putra, S. A. Kurniawan, and Y. R. Nugraha, "Naïve Bayes classifier for journal quartile classification," 2019.

[11] S. V. Lakshmi, M. Meena, and N. Kiruthika, "Diagnosis of chronic kidney disease using random forest algorithms," *International Journal of Research in Engineering, Science and Management,* vol. 2, no. 3, pp. 559-562, 2019.

[12] A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, and S. Dutta, "Emerging technologies in data mining and information security," *Proceedings of IEMIS-2018,* 2018.

[13] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science,* vol. 165, pp. 292-299, 2019.

[14] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science,* vol. 216, pp. 21-30, 2023.

[15] V. Garcia-Rios, M. Marres-Salhuana, F. Sierra-Liñan, and M. Cabanillas-Carbonell, "Predictive machine learning applying cross industry standard process for data mining for diagnosing diabetes mellitus type 2," 2023.
[16] O. Y. Hang, W. Virgiyanti, and R. Rosaida, "Diabetes Prediction using Machine Learning Ensemble Model," *Journal of Advanced Research in Applied Sciences and Engineering Technology,* vol. 37, no. 1, pp. 82-98, 2024.
[17] Ö. F. Akmeşe, "Diagnosing Diabetes with Machine Learning Techiques," *Hittite Journal of Science and Engineering,* vol. 9, no. 1, pp. 9-18, 2022.
[18] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications,* vol. 35, no. 22, pp. 16157-16173, 2023.
[19] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine,* vol. 220, p. 106773, 2022.