Measurement: Sensors 25 (2023) 100605

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Diabetes prediction model using data mining techniques

Rashi Rastogi [a,b,*], Mamta Bansal [a]

[a] Department of CSE, Shobhit Institute of Engineering and Technology (Deemed University), Meerut, India
[b] Sir Chottu Ram Institute of Engineering & Technology, Ch. Charan Singh University, Meerut, India

ARTICLE INFO

Keywords:
Diabetes prediction
Disease
Data mining
SVM
Naïve bayes
Regression and random forest

ABSTRACT

Diabetes is the leading cause of death in the world, and it also affects kidney disease, loss of vision, and heart disease. Data mining techniques contribute to health care decisions for accurate disease diagnosis and treatment, reducing the workload of experts. Diabetes prediction is a rapidly expanding field of research. Early diabetes prediction will result in improved treatment. Diabetes causes a variety of health issues. Therefore, it is critical to prevent, monitor, and raise awareness about it. Type 1 and Type 2 diabetes can cause heart disease, renal problems, and eye difficulties. In this paper, we propose a diabetes prediction model using data mining techniques. We apply four data mining techniques such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. The proposed mechanism is trained using Python and analysed with a real dataset, which is collected from Kaggle. Furthermore, the performance of the proposed mechanism is analysed using the confusion matrix, sensitivity and accuracy performance metrices. In logistic regression, the accuracy is high, i.e., 82.46%, in comparison to other data mining techniques.

1. Introduction

Diabetes is a chronic condition defined by an elevated blood glucose level. Diabetes causes progressive kidney, eye, and heart damage over time [1]. Early diabetes detection is a challenging task. Diabetes disease can be divided into three categories: Type 1 diabetes, also known as juvenile diabetes or insulin-dependent diabetes, occurs when the body's immune system damages insulin-releasing cells, halting insulin production [2]. Type 2 diabetes, also known as insulin-independent diabetes, occurs when the body develops insulin resistance or stops producing insulin [3]. It can occur at any age. Gestational diabetes is a form of diabetes that affects pregnant women [4] (see Table 1).

Data mining is the extraction of previously unknown or previously concealed patterns from a massive database or data warehouse [5]. It used to play a significant role in a variety of industries, including finance, education, healthcare, etc. [6]. Numerous organisations utilise data mining to analyse enormous datasets, to enhance the decision-making process, and to obtain better long-term results. Fig. 1 depicts the diabetes prediction model. In this model, the diabetes dataset is processed by a trained model. The data is collected from sensor devices. The trained model analyses data and predicts the outcome of diabetes for the clients (see Fig. 2).

1.1. Types of diabetes

In this section types of diabetes are discussed. There are broadly four types of diabetes. These are:

- **Type 1 DM**: Historically, the terms "insulin-induced diabetes mellitus" (IDDM) and "juvenile diabetes" were employed. The cause is unknown. Diabetes affects youth and those under the age of 20. Type 1 will harm pancreatic cells, rendering them dysfunctional [7]. Due to a lack of insulin secretion, type 1 diabetes patients have been afflicted throughout their entire lives and are insulin dependent. Patients with type 1 diabetes should constantly engage in physical activity and have a balanced diet.
- **Type 2 DM**: Insulin resistance, which occurs when cells do not respond appropriately to insulin, is the underlying cause of type 2 diabetes. Insulin insufficiency may develop as the condition advances. The phrase 'non-insulin-based diabetes mellitus' or 'adult-induced diabetes' has been used previously. Obesity and lack of exercise are the leading causes [8]. Typically, it occurs at the age of four.
- **Gestational Diabetes**: This is the third basic kind of gestational diabetes, which occurs when pregnant women develop elevated blood sugar levels without a history of diabetes. Approximately 18%

* Corresponding author. Department of CSE, Shobhit Institute of Engineering and Technology (Deemed University), Meerut, India.
E-mail addresses: rastogi.rashi4@gmail.com (R. Rastogi), mamta.bansal@shobhituniversity.ac.in (M. Bansal).

https://doi.org/10.1016/j.measen.2022.100605
Received 10 October 2022; Received in revised form 2 November 2022; Accepted 23 November 2022
Available online 5 December 2022

Table 1
Attributes used in dataset.

Attributes	Description
Glucose	Plasma glucose concentration over 2 h in an oral glucose tolerance test.
Pregnancies	It shows how many times patient is pregnant.
Blood Pressure	It indicates BP of patient.
Skin Thickness	It shows skin fold thickness.
Diabetes Pedigree Function	It shows family history of patient.
BMI	It indicates Body mass index.
Insulin	2-Hour serum insulin (mu U/ml)
Age	It shows age of patient. The age group to be used is 21–81 for analysis.
Outcome	1 for diabetes and 0 for non-diabetes.

of pregnant women have diabetes, according to the most recent study on diabetes. Possibility of increased gestational diabetes in elder women. The third major type is frequently induced by excessive blood sugar levels in pregnant women.

- **Pregestational Diabetes**: Pregestational diabetes occurs prior to the onset of insulin-dependent diabetes during pregnancy. A guy with prediabetes is more likely to receive a score of 2 under such settings or measurements.

1.2. Effects of diabetes

Diabetes can affect human body such as loss of vision, kidney neuropathy, liver and on heart [9]. In this section effects of diabetes have been discussed. These are:

- **Loss of vision**: It is a condition that affects the retina and optic nerve of the eye. Due to night-time vision issues and swelling in the retina, mental contact may be diminished. A diabetic's eye vision can be restored with a few diagnostic procedures or pharmaceuticals.
- **Kidney Neuropathy**: Higher amounts of blood sugar harm the renal arteries, resulting in chronic kidney disease or diabetic neuropathy. The kidney is effective in transporting waste and large quantities of water into the blood.
- **Liver Problems**: Through the breakdown of starch via glucogenesis or glycogenolysis, the liver plays a crucial role in regulating the quantity of blood glucose in the circulation. Diabetes type 2 increases the chance of developing liver problems. A fatty liver is responsible for the development of a liver tumour.
- **Heart Disorders Cardiovascular Ailments**: It's continuous damage to blood vessels and neurons leads to the deception of the circulatory system or organ frame. Cardiovascular disease risk factors include hypertension, abnormally high cholesterol and triglyceride levels, obesity, and lack of physical activity. Multiple clinical characteristics, such as poor glycaemic management and insulin resistance in diabetes, influence cardiac issues.

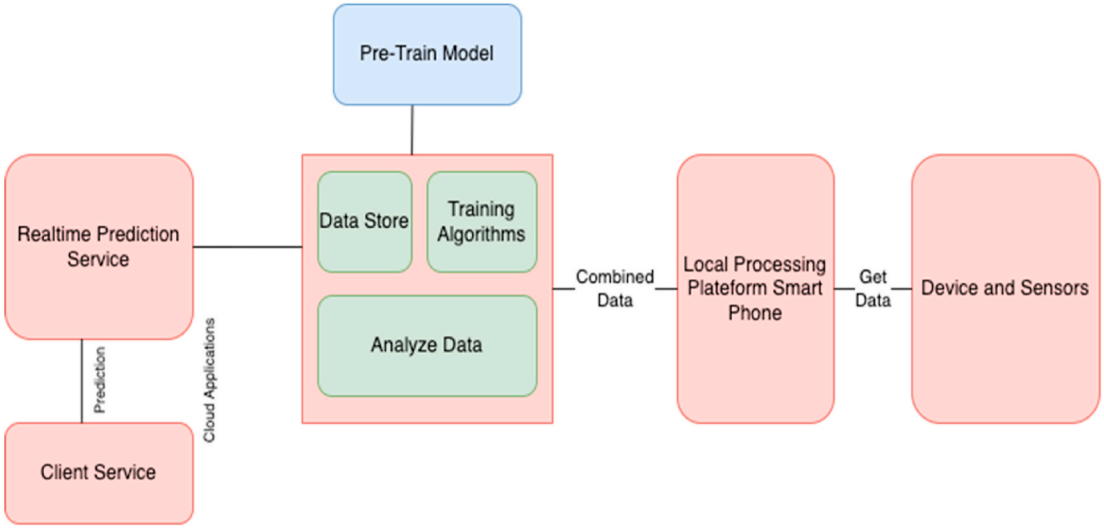

Fig. 1. Diabetes prediction model.

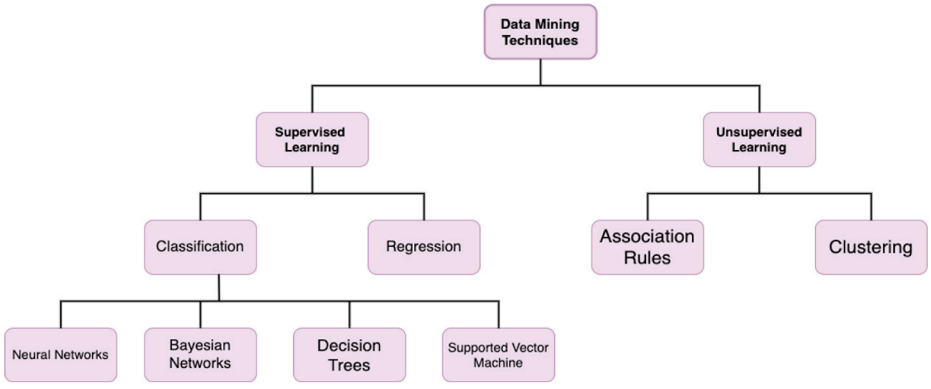

Fig. 2. Data mining techniques.

Organization: The paper is organised into VI sections. Section I provides an introduction to diabetes prediction, including types of diabetes and the various effects of diabetes; Section II discusses data mining techniques; Section III discusses literature reviews and survey inferences; and Section IV presents a proposed mechanism in detail. Section V presents results and analysis of the proposed mechanism. In section VI, the conclusion and future of the manuscript have been presented.

1.2.1. Contribution

- We present a model for early prediction of diabetes disease.
- We discuss different data mining techniques.
- The proposed mechanism is trained using Python for Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes, and analysed with a real dataset, which is collected from Kaggle.

2. Data mining techniques

Data mining is a relatively new technique for extracting knowledge from enormous quantities of data. Mining involves using and processing accessible data to make judgments [10]. It involves the investigation of models in big data sets using techniques at the intersection of machine learning, statistics, and database systems [11]. It facilitates the examination of patterns, such as the categorization of data through cluster studies, the recognition of abnormal records, also known as anomaly detection, and associated rules or dependencies [12].

•**Classification**: Classification is a data mining function that can be assigned to target categories for database objects. This approach can be used as a pre-processing step before to storing data in the classification model. Assemblages of identical components We must cluster to locate a collection of data whose properties capture everything and combine them according to their similarities. Clustering is comparable to segmentation. It is used to classify groups of unprocessed cases based on a range of criteria [13].

- **Decision Tree (DT)**: Decision Tree (DT) is mostly used classification model. As a statistical model for data processing, DT leverages a decision tree [14]. It is used to estimate diseases and classification techniques using patient data. It is possible to design and translate DT quickly.
- **Knowledge Discovery Dictionary (KDD)**: It is the process of data and information collection discovery. It entails data processing, data selection, data preparation, information establishment on data sets, and interpretation of the most effective ways based on observed findings [15]. It consists of an iterative data integration sequence and recognition of DM patterns.
- **Regression**: Regression is a frequently employed function. Regression receives identical grading. It is used to forecast innumerable numbers, age, weight, temperature, and diseases [16]. All of these can be predicted using regression techniques. Regression tasks can resolve several difficulties in the medical industry. Linear regression and logistic regression are the most prevalent types of regression.
- **Association**: Association is the most important DM characteristic that determines the probability of things. As rules of the association, the relationships between working together and the items are illustrated.
- **K- means clustering**: It is the process for grouping identical objects. There are a variety of classifications, including partitions and hierarchical clusters, however the K-means clustering technique was chosen for this investigation. K-Means clustering, which employs numerical data and describes K as a cluster core, is very straightforward or understandable.
- **Artificial Neural Network**: ANN is an information processing device modelled on the capability of the human brain. NNs are typically formed of numerous interconnected nodes with an activation function arranged in layers. The input layers provide the network with patterns that connect to one or more hidden layers, where a system of weighted ties performs the actual processing. After that, a secret layer is connected to an output layer for results.
- **Random Forest**: RF is a tree-based collection classifier in which the randomly assigned vectors are distributed completely differently, with the input X serving as the unit for each tree [17]. The random forest method is a scalable, effortless, and straightforward algorithm that mixes tree predictors. Its performance is difficult to improve, and it can process numerous types of numeric, binary, and nominal data.
- **Naive Bayes Classification**: Bayesian networks frequently contain classification tasks. It consists of direct acyclic graphics with one parent and only numerous children (relative to the observed nodes), with a crucial premise of autonomy between children in the context of their parents [18].
- **Bayesian Networks**: BN are visual models intended to explain relationships between events and concepts and evaluate their likelihood or degree of uncertainty. As a probabilistic model, a Bayesian network describes the variables as well as their relationships.
- **Support Vector Machines (SVMs)**: SVMs are a class of supervised learning techniques used to detect, identify, or regress shapes [19].

2.1. Application of data mining in smart healthcare

In this section applications of data mining in healthcare system such as disease prediction, better treatment etc. has been discussed [20]. There are as follows:

- **Disease Diagnosis and Prediction**: In terms of social security firms, we must infer and predict suffering, which is one of the most essential reasons for using knowledge in social insurance.
- **Classification of Various Hospitals**: The data mining strategy focuses on each of the various health centres' areas of concern, considering their ultimate classification objective. Different health clinics are identified by associations as having the ability to treat patients with real illnesses, i.e., higher-ranking treatment clinics are well equipped to treat high-risk patients during times of higher need.
- **Successful Treatments**: In contrast, medical knowledge mining is used to divide the efficacy of pharmaceuticals into components such as causes, signs, symptoms, and cost.
- **Infection Prevention in Hospital sites**: Information mining is used to investigate contamination to find such unforeseen situations in disease control data. A trained individual investigates these instances further for contamination management.
- **Identifying Patients at High-Risk**: American Health techniques enable diabetic clinics to increase efficiency and reduce administrative costs for diabetic patients [21].

3. Literature survey

For the purpose of diagnosing and preventing T2DM, Yang et al. [22] suggested a novel approach using data mining techniques. Enhancing the prediction model's precision and generalising the model's predictions beyond a single data set are primary goals. The model is split into two sub-parts: an enhanced K-means algorithm and a logistic regression method, both of which are based on a collection of pre-processing steps. In order to forecast diabetes risk, Islam et al. [23] need a dataset that includes information about people who are newly diagnosed with diabetes or who are at risk of developing the condition. They used a 520-item dataset that was acquired from surveys administered to Sylhet Diabetes Hospital patients in Bangladesh. Furthermore, they applied Naive Bayes, Logistic Regression, and Random Forest algorithms to examine the data. Similarly, Woldemichael and Menaria

[24] presented a mechanism using data mining methods to foresee cases of diabetes. predicting whether or not a person has diabetes using a back propagation algorithm. Predictions of diabetes were made using a number of methods, including J48, naive bayes, and support vector machine. Similarly, a prediction model was developed by Fiarni et al. [25] to forecast the occurrence of three major complications of diabetes in Indonesia, and key factors associated with these complications are identified. The seven risk factors for diabetes were identified as age, gender, BMI, family history of diabetes, blood pressure, length of time diabetic, and blood glucose level. Therefore, k-means clustering and the Naive Bayes Tree classification methods were utilized to examine this data set. Furthermore, application software developed by Aldallal et al. [26] is used by doctors and other medical professionals to foresee the onset or recurrence of chronic diseases (NCDs). In this endeavour, they utilized a data-mining technique that could foresee future outcomes. Information about patients from Bahrain Defense Force Hospital was used to test the programme. Moreover, Khan et al. [27] go into the area of glycemic management for diabetes and research data mining-based diagnosis and prediction solutions. In a similar manner, Kavakiotis et al. [28] carried out a systematic analysis of diabetes research using machine learning, data mining approaches, and tools Using three distinct data mining organization methods: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree. Kumar et al. [29] evaluated and analysed the prospective approaches to forecast the chance of heart disease for diabetic patients based on their predictive accuracy. Moreover, the goal of Mahesh et al. [30]'s blended ensemble learning (EL)-based forecasting system was to identify the best classifier for evaluating clinical outcomes through a standardised set of metrics. In this article, we suggest an EL that uses Bayesian networks and radial basis functions. Standard predictive methods, such as K-nearest neighbour (KNN) and logistic regression, are used in Oza and Bokhare [31]. By comparing the different ways that machine learning can be used, a model is proposed to improve performance and measure accuracy. Prediction of diabetes mellitus using data mining is well reviewed by Anil et al. [32]. The goal is to examine and compare the predicted accuracy of the various analytical approaches currently employed in this sector through a study of studies that have used these approaches. Logistic regression analysis with data mining techniques such as decision trees using the J48 and LMT algorithms, Naive Bayes, and Artificial Neural Networks (ANN) was proposed by Paisanwarakiat et al. [33]. The risk of developing diabetes was predicted using Random Forest, KNN, and Support Vector Machine by Arumugam et al. [34]. As seen in the outcomes, the support vector methodology is very trustworthy. A variety of classifiers have been presented, and their structures make it possible to choose the most appropriate one for future data analysis and interpretation. Abdollahi and Moghaddam [35] employed a genetic algorithm-based ensemble training methodology to effectively identify and predict the consequences of diabetes mellitus. Experimental data and genuine data on Indian diabetes from the University of California's website are used in this study.

3.1. Inferences drawn from literature review

- Data mining techniques are utilized by authors to predict diabetes diseases.
- Accuracy of different algorithms are compared to find more efficient algorithm to predict diabetes dataset.
- Different type of diabetes and their effects are studied in detail.
- Data mining algorithms are very useful to get more accurate information at early stages are studied.

It is noticed in literature review that the Data mining algorithms can help doctors to predict diabetes at early stage in more accurate manner so that they can diagnose the patient.

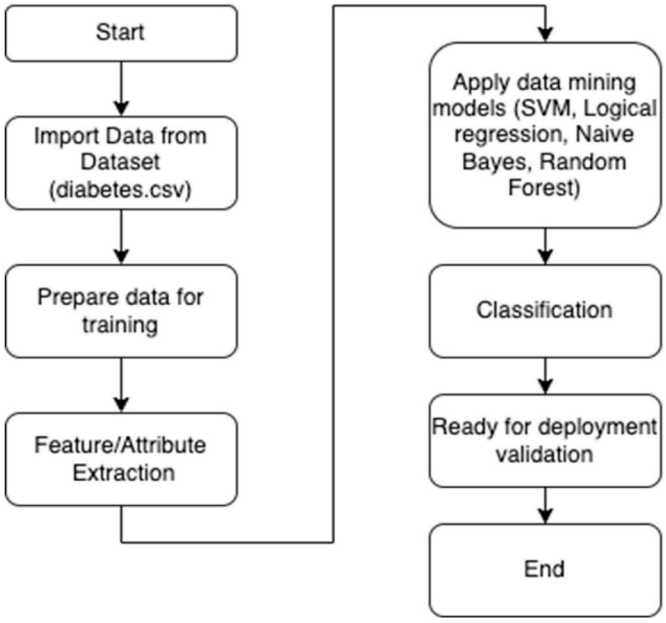

Fig. 3. Proposed framework.

4. Proposed model

4.1. Problem statement

Doctors depend on normal information for treatment. When standard information is insufficient, examines are summed up after a certain number of cases have been considered. However, this interaction sets aside time, though if AI is utilized, the examples can be recognised before. To utilise AI, a gigantic amount of information is required. There is an exceptionally restricted amount of information accessible relying upon the infection. Additionally, the quantity of tests having no illnesses is high compared with the number of tests actually having the infection. Various researchers have proposed different techniques to predict diabetes by applying various ML classification techniques, but each one has pros and cons. Some techniques have less accuracy while some have a high accuracy rate, but elapsed time also increases. So, to overcome these kinds of issues, it is intended to propose a mechanism to predict diabetes outcomes with a high accuracy rate. In the proposed mechanism, we use RF, SVM, Linear Regression, and Naive Bayes algorithms to predict diabetes.

4.2. Proposed Framework

In this section proposed mechanism has been presented. Fig. 3 presents proposed model. Diabetes disease may occur when immune system of human body can't work in proper manner due to variation in the sugar level of body or less immunity of body. In proposed mechanism diabetes dataset is taken for processing. The dataset contains diabetes attributes such as pregnancy, sugar level, BMI, skin thickness etc. In next step extract features from dataset using trained model. After that apply data mining models Naïve Bayes, SVM Logistic regression and Random Forest for classification of diabetes into two states 1 or 0. Furthermore once model is trained then validate model to retrieve performance of models with accuracy, sensitivity, and confusion matrix.

Proposed Algorithm.

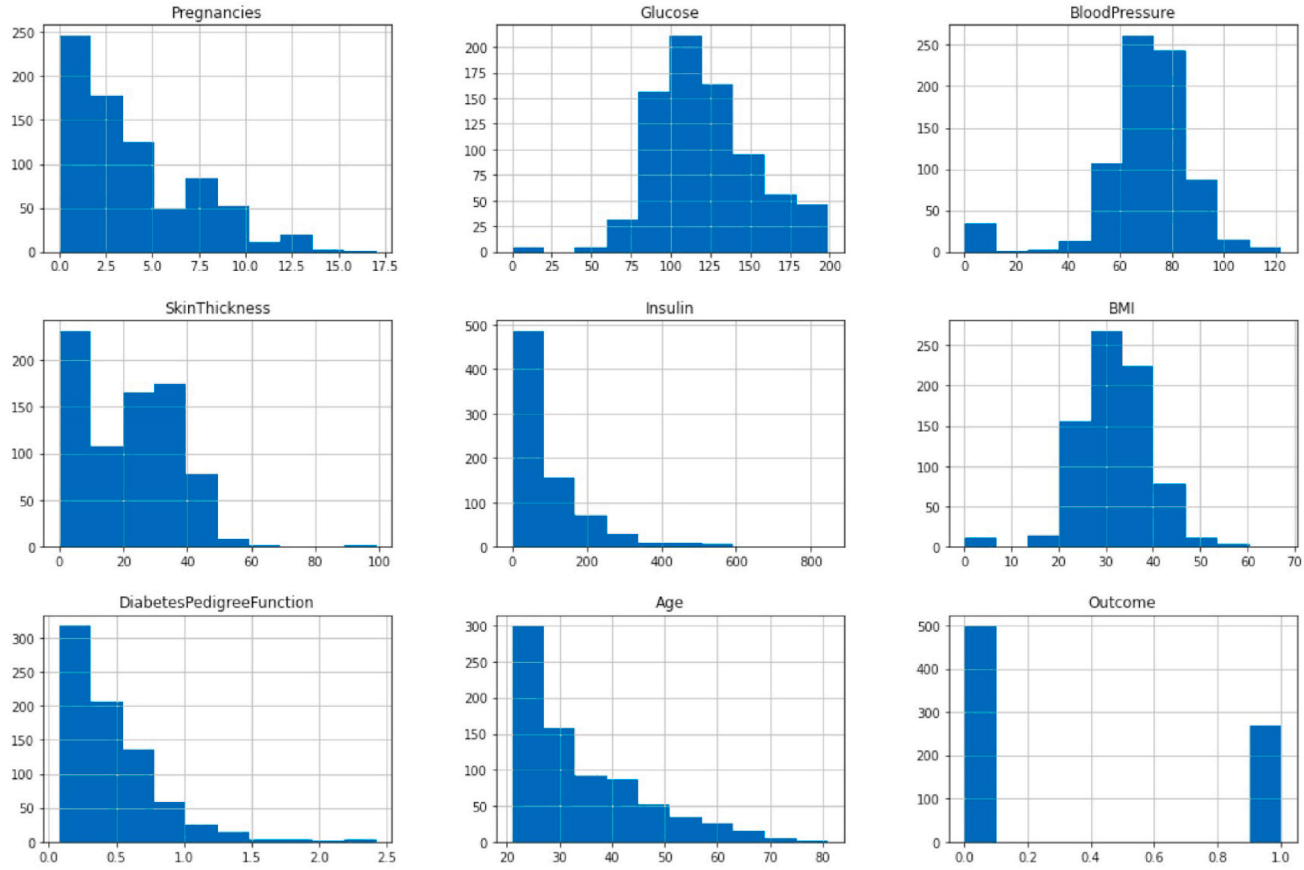

Fig. 4. Histogram representation of dataset Attributes.

1. INITIALIZE: x, dataset, data_mining_model,
2. IMPORT: dataset = diabetes.csv
3. PREPARE: training dataset: training_data
4. EXTRACT: features -> from training_data
5. APPLY: data mining model ->new_data = data_mining_model(training_data)
6. CLASSIFY: new_data
7. DEPLOY new_data -> dataset
8. END

4.3. Implementation strategy

The datasets are gathered from the information base. In stage two, the information will be pre-handled, which will incorporate information cleaning, mixing, and changing. By utilising RF calculation, we can achieve better precision when compared with other calculations.

The information was acquired from Kaggle. It was gathered and, along these lines, the data was inputted as instructing tests and back-to-back investigated to supply a decent model. Data variety is the arrangement of accommodating pertinent data that is assembled through an exploitation question measure. The information is separated in a very specific way, with a number of very specific categories.

- Data Pre-processing: Information pre-processing is a critical advancement in the philosophy of information disclosure. The majority of medical service data contains missing value, wheezy, and irregularity information.
- Data cleaning is the strategy of discovering and revising (or eliminating) bad or erroneous records from a record set, table, or information and refers to recognising insufficient, incorrect, mistaken, or distracting portions of the information, some replacement, adjusting, or erasing the messy or coarse information [8]. Data purifying is also performed intuitively with information from 25 different instruments, or as execution through prearranging. Data cleansing is also known as data cleaning or data purging.
- Data integration: Data integration could be a strategy in which heterogeneous information is recovered and joined as a fused kind and design. Information reconciliation enables clients, associations, and applications to incorporate completely and potentially extraordinary

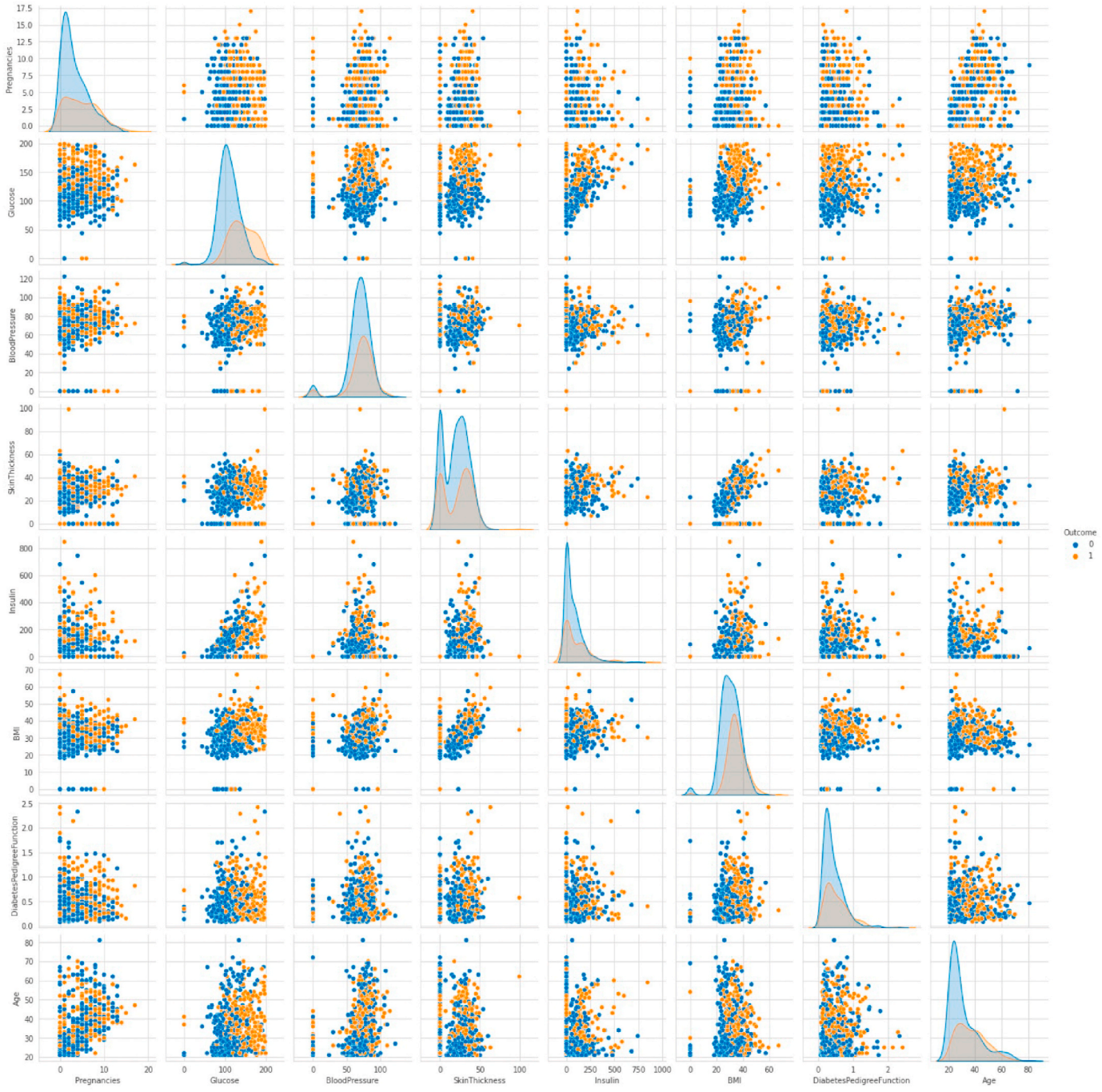

Fig. 5. Pair plot representation of Diabetes prediction outcomes.

data types (for example, data sets, reports, and tables) to be used as close to home or business measures as possible.

- Data reduction is the change of mathematical or sequential advanced information inferred through exact perception or by experimentation into an adjusted, requested, and improved type. The central development is the decrease of undeterminable measures of information in every one of the methods directly down to the deliberate segments.

5. Results

In this section, results and analysis have been discussed. The proposed mechanism is implemented using python and jupyter notebook. It is an open-source programming language. The execution of a programme is fast in Python. It can provide inbuilt library files to run users' programs. Python is best suited for data mining programs.

5.1. Metrics used

- **Sensitivity**

It depicts total only number of positive outcomes after processing of dataset. The formula for sensitivity calculation is as follows:

$$Senitivity = TP/(TP + FN)$$

- **Accuracy**

It shows the overall number of positive (+ve) outcomes in

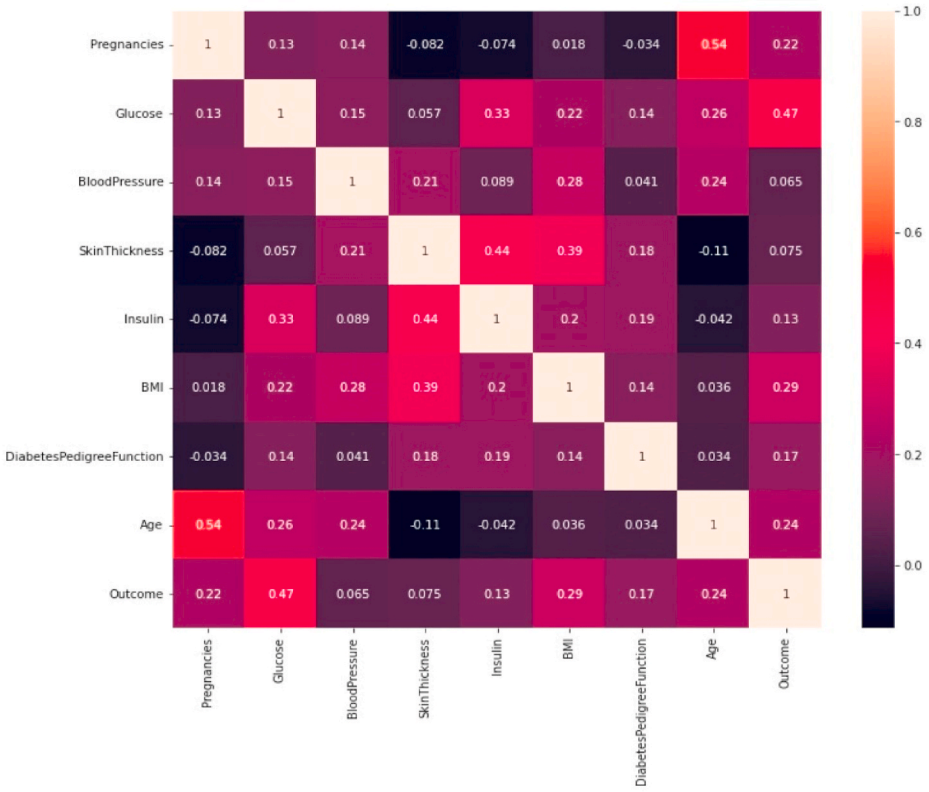

Fig. 6. Confusion matrix of proposed model.

Table 2
Record of used patient dataset (Age group (21–81)).

Male	Female	Total
358	256	614

Table 3
Performance comparison using Confusion Matrices.

Techniques	Confusion Matrix
Logistic Regression	$\begin{bmatrix} 98 & 9 \\ 18 & 29 \end{bmatrix}$
SVM	$\begin{bmatrix} 98 & 9 \\ 23 & 24 \end{bmatrix}$
Naïve Bayes	$\begin{bmatrix} 93 & 14 \\ 18 & 29 \end{bmatrix}$
Random Forest Classifier	$\begin{bmatrix} 95 & 12 \\ 16 & 31 \end{bmatrix}$

comparison to the total number of negative (-ve) outcomes in the entire dataset.

$Accuracy = TP + TN/ (TP + FP + FN + TN)$

- **Confusion Matrix**

It depicts the prediction values of data in terms of TP, TN, FN, FP i.e., true + ve, true –ve, false + ve and false -ve. Based on these parameters the sensitivity and accuracy of techniques has been computed.

5.2. Dataset used

To analyse proposed mechanism, we use diabetes dataset downloaded from Kaggle. The dataset contains data values of patients having age more than 20. The predicted outcome i.e., Positive or negative has been calculated on the basis of various parameters such as pregnancy, sugar level, BMI etc. Fig. 4 depicts the histogram representation of diabetes dataset attributes. Histograms are visual representations of data points clustered into intervals determined by the viewer. The histogram, which resembles a bar chart, summarises a sequence of numbers into a manageable visual by clustering them into discrete categories called "bins."

So here we have 9 criteria that can be used to estimate and predict the diabetes. These criteria are the pregnancy (that is condition of

Table 4
Accuracy and Sensitivity comparison.

Techniques	Accuracy (%)	Sensitivity (%)
Logistic Regression	82.46	68.23
SVM	79.22	59.99
Naïve Bayes	79.22	64.44
Random Forest Classifier	81.81	68.88

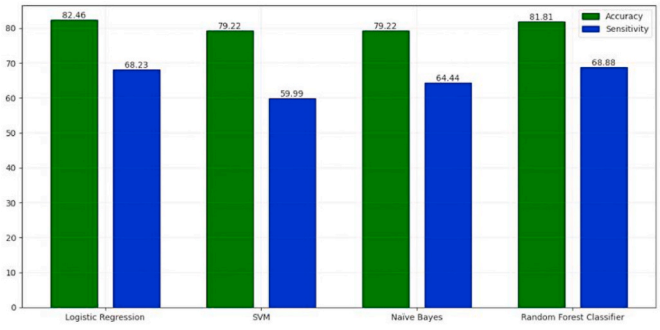

Fig. 7. Accuracy and sensitivity comparison.

women when give birth to child), quantity of glucose, the number or value of blood pressure, the thickness of human skin, Body mass Ratio also called BMI, the function of diabetes pedigree and the year to which human body has gone through with disease & the age, these all collectively going to predict the diabetes. The collective dataset of feature are going to give crucial information about the estimation of diabetes in a human being. All the features of our dataset play a significant task in the prediction of diabetes. Calculating importance of each feature will help us to find that how much each feature is relevant in finding the output of our model.

Fig. 5 represents the pair plot of the outcome of the diabetes datasets (see Table 2). It shows the relationship between attributes of data. Fig. 6 depicts the confusion matrix of the proposed model. The confusion matrix represents the values of each attribute used in the dataset. Table 3 shows the comparison of SVM, RF, logistic regression, and naive bayes models in terms of their confusion matrix. The rate of computation of positive outcomes is high in logistic regression as compared to other techniques. Furthermore, Table 4 depicts the comparison between data mining techniques in terms of accuracy and sensitivity. In logistic regression, the accuracy is high in comparison to other models.

It is crucial to provide a prompt and correct diagnosis while dealing with disorders like diabetes. If proper care is not taken, a delay in diabetes diagnosis can have devastating effects on health. Therefore, the accuracy of data mining algorithms used to anticipate a patient's status must be maximised. A false negative has a much greater cost impact than a false positive. If the subject is given a false diagnosis, they may try to relax despite the significance of their situation. Fig. 7 depicts the sensitivity and accuracy comparison of the proposed model. In the logistic regression model, the accuracy is high, i.e., 82.46% as compared to other models. whereas in SVM the accuracy is low, i.e., 79.22% as compared to other models. Furthermore, the sensitivity is slightly higher in RF, 68.88% in comparison to other models, and SVM is lower, 59.99% in comparison to other models such as naive bayes logistic regression and random forest.

6. Conclusion and future work

Nowadays, data mining plays a crucial role in diabetes prediction in the healthcare system. Diabetes is a major health challenge in the world. Early prediction of diabetes will result in improved results. This paper presents a diabetes prediction model with the help of data mining techniques. We apply Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine techniques to predict diabetes disease. The proposed mechanism is implemented using Python. To analyse the proposed mechanism, a real dataset is collected from Kaggle. Accuracy, confusion, and sensitivity matrices are used to assess performance. In the logistic regression model, the accuracy is high, i.e., 82.46% as compared to other models. whereas in SVM the accuracy is low, i.e., 79.22% as compared to other models. In the future, it is intended to continue working on it and apply more classification algorithms to predict diabetes datasets. It is also meant to suggest a new way to make predictions about diabetes outcomes more accurate.

CRediT authorship contribution statement

Rashi Rastogi: Methodology, Writing – original draft, preparation. **Mamta Bansal:** Conceptualization, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.
[2] M.K. Hasan, M.A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531.
[3] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, T. Saba, Current techniques for diabetes prediction: review and case study, Appl. Sci. 9 (21) (2019) 4604.
[4] S.I. Ayon, M.M. Islam, Diabetes prediction: a deep learning approach, Int. J. Inf. Eng. Electron. Bus. 12 (2) (2019) 21.
[5] Z.S. Ageed, S.R. Zeebaree, M.M. Sadeeq, S.F. Kak, H.S. Yahia, M.R. Mahmood, I. M. Ibrahim, Comprehensive survey of big data mining approaches in cloud systems, Qubahan Acad. J. 1 (2) (2021) 29–38.
[6] W. Haoxiang, S. Smys, Big data analysis and perturbation using data mining algorithm, J. Soft Comput. Paradigm (JSCP) 3 (2021) 19–28, 01.
[7] M. Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, W. Albaker, A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM, Comput. Biol. Med. 147 (2022), 105757.
[8] H. Ikegami, Y. Hiromine, S. Noso, Insulin-dependent diabetes mellitus in older adults: current status and future prospects, Geriatr. Gerontol. Int. 22 (8) (2022) 549–553.
[9] Y. Liu, Q. Wang, K. Wu, Z. Sun, Z. Tang, X. Li, B. Zhang, Anthocyanins' effects on diabetes mellitus and islet transplantation, Crit. Rev. Food Sci. Nutr. (2022) 1–24.
[10] K.G. Al-Hashedi, P. Magalingam, Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019, Comput. Sci. Rev. 40 (2021), 100402.
[11] J.R. Saura, D. Palacios-Marqués, D. Ribeiro-Soriano, Using data mining techniques to explore security issues in smart living environments in Twitter, Comput. Commun. 179 (2021) 285–295.
[12] Y.S. Su, S.Y. Wu, Applying data mining techniques to explore user behaviors and watching video patterns in converged IT environments, J. Ambient Intell. Hum. Comput. (2021) 1–8.
[13] N.P. Jayasri, R. Aruna, Big data analytics in health care by data mining and classification techniques, ICT Express 8 (2) (2022) 250–257.
[14] R. Krishnamoorthi, S. Joshi, H.Z. Almarzouki, P.K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari, A novel diabetes healthcare disease prediction framework using machine learning techniques, J. Healthc. Eng. 2022 (2022).
[15] A. Mavrogiorgou, A. Kiourtis, G. Manias, D. Kyriazis, An optimized KDD process for collecting and processing ingested and streaming healthcare data, in: 2021 12th International C*onference on Information And Communication Systems (ICICS)*, IEEE, 2021, May, pp. 49–56.
[16] E. Pérez-Montalvo, M.E. Zapata-Velásquez, L.M. Benitez-Vazquez, J.M. Cermeno-Gonzalez, J. Alejandro-Miranda, M.A. Martinez-Cabero, Á. de la Puente-Gil, Model of monthly electricity consumption of healthcare buildings based on climatological variables using PCA and linear regression, Energy Rep. 8 (2022) 250–258.
[17] V. Simic, A. Ebadi Torkayesh, A. Ijadi Maghsoodi, Locating a disinfection facility for hazardous healthcare waste in the COVID-19 era: a novel approach based on Fermatean fuzzy ITARA-MARCOS and random forest recursive feature elimination algorithm, Ann. Oper. Res. (2022) 1–46.
[18] H. Yoshikawa, Can naive Bayes classifier predict infection in a close contact of COVID-19? A comparative test for predictability of the predictive model and healthcare workers in Japan, J. Infect. Chemother. 28 (6) (2022) 774–779.
[19] S. Chidambaranathan, A. Radhika, V.V. Priya, S.K. Mohan, M.G. Gireeshan, Optimal SVM based brain tumor MRI image classification in cloud internet of medical things, in: Cognitive Internet of Medical Things for Smart Healthcare, Springer, Cham, 2021, pp. 87–103.
[20] J. Santos-Pereira, L. Gruenwald, J. Bernardino, Top data mining tools for the healthcare industry, J. King Saud Univ. Comput. Inf. Sci. (2021).
[21] A. Guzzo, A. Rullo, E. Vocaturo, Process mining applications in the healthcare domain: a comprehensive review, Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 12 (2) (2022) e1442.
[22] H. Wu, S. Yang, Z. Huang, J. He, X. Wang, Type 2 diabetes mellitus prediction model based on data mining, Inform. Med. Unlocked 10 (2018) 100–107.
[23] M.M. Islam, R. Ferdousi, S. Rahman, H.Y. Bushra, Likelihood prediction of diabetes at early stage using data mining techniques, in: Computer Vision and Machine Intelligence in Medical Image Analysis, Springer, Singapore, 2020, pp. 113–125.
[24] F.G. Woldemichael, S. Menaria, Prediction of diabetes using data mining techniques, in: 2018 2nd International C*onference on Trends in Electronics and Informatics (ICOEI)*, IEEE, 2018, May, pp. 414–418.
[25] C. Fiarni, E.M. Sipayung, S. Maemunah, Analysis and prediction of diabetes complication disease using data mining algorithm, Procedia Comput. Sci. 161 (2019) 449–457.
[26] A. Aldallal, A.A.A. Al-Moosa, Using data mining techniques to predict diabetes and heart diseases, in: 2018 4th International Conference *on Frontiers Of Signal Processing (ICFSP)*, IEEE, 2018, September, pp. 150–154.
[27] F.A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, S.A.C. Bukhari, Detection and prediction of diabetes using data mining: a comprehensive review, IEEE Access 9 (2021) 43711–43735.

[28] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017) 104–116.
[29] A. Kumar, P. Kumar, A. Srivastava, V.D. Ambeth Kumar, K. Vengatesan, A. Singhal, Comparative analysis of data mining techniques to predict heart disease for diabetic patients, in: International Conference on *Advances In Computing And Data Sciences*, Springer, *Si*ngapore, 2020, April, pp. 507–518.
[30] T.R. Mahesh, D. Kumar, V. Vinoth Kumar, J. Asghar, B. Mekcha Bazezew, R. Natarajan, V. Vivek, Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease, vol. 2022, Computational Intelligence and Neuroscience, 2022.
[31] A. Oza, A. Bokhare, Diabetes prediction using logistic regression and K-nearest neighbor, in: Congress on Intelligent Systems, Springer, Singapore, 2022, pp. 407–418.
[32] K.S. Anil, R. Jain, Data mining techniques in diabetes prediction and diagnosis: a review, in: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, 2022, April, pp. 1696–1701.
[33] R. Paisanwarakiat, A. Na-udom, J. Rungrattanaubol, Combining logistic regression analysis with data mining techniques to predict diabetes, in: International Conference on Computing and Information Technology, Springer, Cham, 2022, pp. 88–98.
[34] S.S. Arumugam, V. Kuppan, V. Chakravarthi, K. Palaniappan, An accurate diagnosis of diabetes using data mining, in: AIP Conference Proceedings, vol. 2405, AIP Publishing LLC, 2022, April, 1, p. 020017.
[35] J. Abdollahi, B. Nouri-Moghaddam, Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction, Iran J. Comput. Sci. (2022) 1–16.