Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Automated Machine Learning And Deep Learning Techniques For The Classification Of Diabetes Mellitus: A Model Development Study

Prof. Sanmati Kumar Jain [1], Dr. Sheetal Bawane [2]

[1] *Associate Professor & HOD, Department of Computer Science and Engineering, Vikrant Institute of Technology and Management, Indore, M.P. India*
sanmatijain2906@gmail.com
[2] *Associate Professor, Department of Electronics & Communication, Medi-Caps University, Indore, M.P., India*
sheetal.bawane@medicaps.ac.in

A dangerous disease in humans, diabetes mellitus is caused by elevated glucose levels. Untreated diabetes may cause a host of further serious health complications. This research aims to predict the occurrence of diabetes by analysing several human bodily characteristics. For the purpose of diabetes mellitus risk prediction, an ensemble method called En-RfRsK is suggested. This voting classifier combines three machine learning techniques: RF, R-SVM, and KNN. RF makes use of the results obtained from a multitude of models or trees that are not always evenly distributed. Using a function whose value varies as the distance from the origin increases, R-SVM is able to make predictions. By analysing diabetic data, KNN is able to understand the non-linear decision limits. This new method makes use of all the best features of various ML approaches. Since ensemble methods outperform single classifiers in terms of accuracy and flexibility, the suggested method is an amalgamation of preexisting ML techniques. Because of its superior prediction powers and accuracy, it provides the most suitable answers. Using the PIMA diabetes dataset, experiments were conducted. Experiments clearly show that the suggested method beats both the current baseline classifiers and the most cutting-edge ML diabetes mellitus prediction systems. An accuracy of 91% was achieved using the En-RfRsK method.

Keywords: Diabetes, Machine learning, Accuracy, Mellitus, Algorithm.

1. Introduction

Untreated diabetes poses an unfathomable threat to human prosperity and is among the most well-known chronic diseases in the world [1]. Diabetes may be classified into two main forms: type 1 and type 2. Type 1 diabetes occurs when the pancreas does not secrete enough insulin, despite the fact that the body searches for insulin-producing cells in that organ. A lack of proper cellular response to insulin—insulin resistance [2, 3]—is a hallmark of type 2 diabetes. Globally, 9.3% of the population would be unable to control their diabetes by 2019, with that

figure expected to rise to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045, according to surveys [4]. Along with that, half of all diabetics are unaware that they have the disease. The use of blood glucometers allows for the monitoring of diabetes over certain time intervals. Continuous glucose monitoring devices are used to assess a patient's current glucose level. These devices are minimally intrusive and allow for continuous measurement of diabetes [5]. Delay in diabetes diagnosis affects several bodily organs, including the kidneys, eyes, heart, and nerves [6]. Consequently, a prompt and precise diagnosis of diabetes is of the utmost importance.

To diagnose and understand diabetic data, proper data analysis is required when data management becomes a categorisation challenge [7]. Thus, it is highly appreciated when artificial intelligence is used to accurately anticipate diabetes. In today's world, some of the most cutting-edge technologies include ML, deep learning, AI, the internet of things, and big data [8]. Early on, patients may use ML to confirm their health, and it will help doctors with future research, too [9]. Regression and classification problems are both amenable to its use. Since diabetes prognosis is a classification issue, we may categorise individuals based on their diabetes status. Numerous ML methods are useful for looking at the data from different perspectives and combining it into useful knowledge. Learning include preparing the dataset, extracting and selecting features, doing training and testing, and evaluating outcomes [10]. Clinical information, text information, and sensor data are examples of the types of data generated by various wearable devices, and the majority of this data is in its raw form [11]. In order to make reliable predictions, pre-processing is required to transform this data into a usable format, which includes dealing with dataset missing values and imputed missing values. The key to making the best prediction is selecting the top attributes from the feature space, since there are several at-tributes in the dataset [12]. Predicting the likelihood of diabetes effectively may need the application of ML methods such RF, NN, DT, and SVM [13]. Training and testing the models on the test dataset allows one to ascertain whether or not the models are performing as expected. Researchers also find success in diabetes prediction by using ensemble algorithms, which sort the dataset using majority-voting-based bagging, boosting, and stacking approaches [14], [15], [16]. By casting ballots and selecting the model with the most votes, the best model results are integrated by majority voting [17], [18]. Despite researchers' best efforts, the outcomes for diabetes prediction remain inadequate [19], [20]. As a result, we need to propose new approaches to accurate and efficient forecasting. In order to determine whether a person has diabetes, we must first go through the data we have collected. As a fundamental portion of the adult population accounts for 9.3% of the global total, driving this evaluation is crucial. Themes and data plans related to diabetes should help us better anticipate if individuals have the disease [21]. This is a problem with directed AI's characterization [11]. In order to better categorize illnesses, researchers are constructing ensembles of classifiers, which consist of many individual systems [22], [23]. These systems include classification judgements at numerous levels, in contrast to traditional approaches that rely on a single classifier. To maximize their individual strengths, ensembles of classifiers with varying variances and biases are used. One objective of the research is to forecast the likelihood of diabetes in a patient based on exact estimations recorded in the dataset and an overview of different portions of it. Numerous researches have focused on diabetes mellitus [24], [13]. Excessively high blood glucose levels are associated with a host of complications

for diabetics, including blurred vision, shock, and even obviousness, according to the research. Significant difficulties or challenges related to a person's success might arise from an uneven distribution of blood sugar levels. In order to reap the advantages of any assumption or collecting system, it is essential to establish it. Machine learning (ML) is a powerful instrument for evaluating one-of-a-kind polluted setups, mandate data extraction from different wagers, build patient-genuinely stable organisations, and identify such potential clinical restrictions and their relationships. Afterwards, ML algorithms that are compatible with different dataset planning constraints are proposed for diabetes prediction. The proposed framework was built using the PIMA Indian diabetes dataset; it contains nine unique attributes; and even after pre-processing, there are 768 records or events. Several ML course of action calculations are proposed to verify the precision or correctness of the blood glucose assumption. For the purpose of evaluating the performance of ML algorithms, the logistic regression (LR) accuracy of 92.26% is used, which is more than all other existing organised estimates. Further, alternative convincing classifiers, such as artificial NN, show a significant increase in accuracy, which would typically enhance the decision support framework's precision [25]. Detectable diabetes at an early stage is a key concern in the clinical benefits field. Data pre-processing was facilitated by the WEKA mechanical assembly. This was followed by the component decline approach eliminating three characteristics. We utilised the PIMA dataset with a single outcome and five data characteristics: age, glucose, body mass index (BMI), insulin, pregnancy, and pregnancy status. A number of metrics were used to analyse the presentation, and seven alternative AI estimations—DT, KNN, RF, Naive Bayes (NB), AB, LR, and SVM—were used to predict diabetes. In terms of certain constraints, such as survey, F-measure, precision, and exactness, every model shows remarkable performance. The models' accuracy was higher than 70%. Applying the LR and SVM with train/test splits and K-cross validations yielded a 77% to 78% accuracy rate, respectively [26]. The authors sought to use three AI estimations—the key backslide, the support vector classifier, and the KNN—on the PIMA Indians Diabetes dataset in an effort to anticipate and study diabetes. Dataset analysis has shown that glucose, insulin, and body mass index (BMI) are the parameters with the most potential for improvement in association coefficient. A diabetes diagnosis is therefore more heavily influenced by these traits than by others. The findings show that SVM would differ the most from other estimations in terms of the correctness of the assumptions used, coming in at 85.06% [27]. One common clinical issue is determining the early warning signs of diabetes. In order to speculate on the risk related to type 2 diabetes mellitus, the researchers in [28] set out to build a framework. The outcomes of the six AI representation methods were differentiated and measured using many credible measures. The dataset, which was constructed using online and detached overviews, consisted of 18 diabetes-related enquiries, and tests were conducted on this. The exploratory findings show that out of all the models, RF has the best accuracy at 94.10%. To current paper propose a new ensemble machine learning method in this work.

2. Methodology

Whether diabetes is present or not can be predicted using the ensemble classifier En-RfRsK. Figure 1 shows the block diagram of the En-RfRsK method that has been suggested. Diabetes

data analysis, diabetes data modelling, model selection, and ensemble-based voting categorization are the following elements that make up this technique.

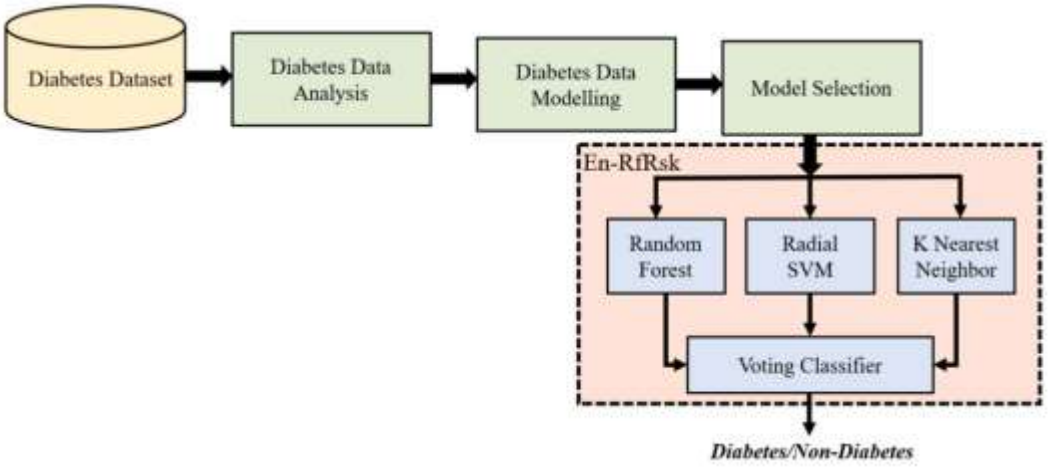

Fig. 1 Proposed machine learning techniques

2.1. Analyzing diabetes data

This study's perceptual evaluation was based on the PIMA Indian Diabetes dataset. Prior to being used as a foundation for initial evaluations of different request estimates, the dataset under consideration is cleansed using data pre-processing and data cleaning techniques. Important patient information is used to classify and predict the overall diabetes prevalence. The dataset for diabetes mellitus includes 768 records. It includes patient information such as the number of pregnancies, plasma glucose levels measured in an oral glucose strength test every 2 hours, diastolic blood pressure in millimetres of mercury, triceps skinfold thickness in millimetres, 2-hour serum insulin in micro-L/ml, body mass index (weight in kg/(level in m)2), age in years, and a class variable that can be either 0 or 1. During the data pre-processing and cleaning step, any values that are missing or otherwise abnormal are deleted from the dataset. Following pre-processing, the initial information is condensed to 722 entries that include the three crucial patient attributes. Out of the 722 patient peculiarities in the dataset, 474 are classified as non-diabetic, 248 as diabetes, and 46 as missing essential core features. The data credits consist of six numerical variables drawn from the dataset, and one component is utilized to determine the result quality.

2.2. Diabetes information modelling

Machine learning algorithms are an evolution of traditional algorithms. System intelligence is enhanced by the use of algorithms, which enable systems to autonomously learn from input data. Due to their learning and classification capabilities, these algorithms are being used to address a broad variety of problems in almost every industry. The two main types of these algorithms are supervised and unsupervised. The section that follows has described a few useful machine learning algorithms for diabetes mellitus prediction.

2.3. Regression using logistical models

The purpose of a logistic regression analysis is to use a set of independent variables to determine the likelihood of an event occurring. The dependent variable may take on values between zero and one since the result is a probability. The lb, which divides the success probability by the failure probability, is used to convert probabilities in logistic regression.

2.4. Recurrent neural network

Neural networks are supervised learning machine learning techniques. Because of its exceptional accuracy and ability to handle massive volumes of data, it is mostly used for classification purposes. The current study applied linear regression, decision tree, SVM and KNN networks for the model prediction. The performance of the models is evaluated and considered throughout the ensemble building process based on an accuracy measure. The ensemble model is fine-tuned by trial and error to determine the selection threshold. The ensemble voting method of categorization makes use of the chosen models. Machine learning models that are conceptually similar or distinct may be combined for prediction using majority vote by employing this classifier, which acts as a meta classifier. The proposed model is a hybrid of the RF and SVM classifiers with the KNN network. One use of the predict probability attribute column in soft voting classifiers is to offer the likelihood of each target variable.

3. Results and discussion

The tests were run on a Windows 10 machine with an Intel(R) Core (TM) i7-6700 CPU and 16 GB of random-access memory. Python was used to implement the suggested method. To evaluate the efficacy of the suggested En-RfRsK ensemble method, experiments were carried out utilizing the PIMA Indian Diabetes dataset.

Table 1. Sampling of data for diabetes patients

Pregnancy	**Glucose**	**Blood pressure**	**Thickness of skin**	**Insulin**	**BMI**	**Diabetic probability**	**Age**	**Outcome**
7	147	70	34	0	32.6	0.618	51	1
2	84	64	28	0	24.3	0.349	31	0
8	183	67	0	0	22.5	0.689	32	1
1	88	62	22	89	28.4	0.193	22	1
0	132	39	31	137	39.2	1.928	34	0

To make the diabetes prediction system more effective, pre-processing is a must. The range of [0,1] has been used for the normalization. Data distribution for diabetes is shown in figure 2.

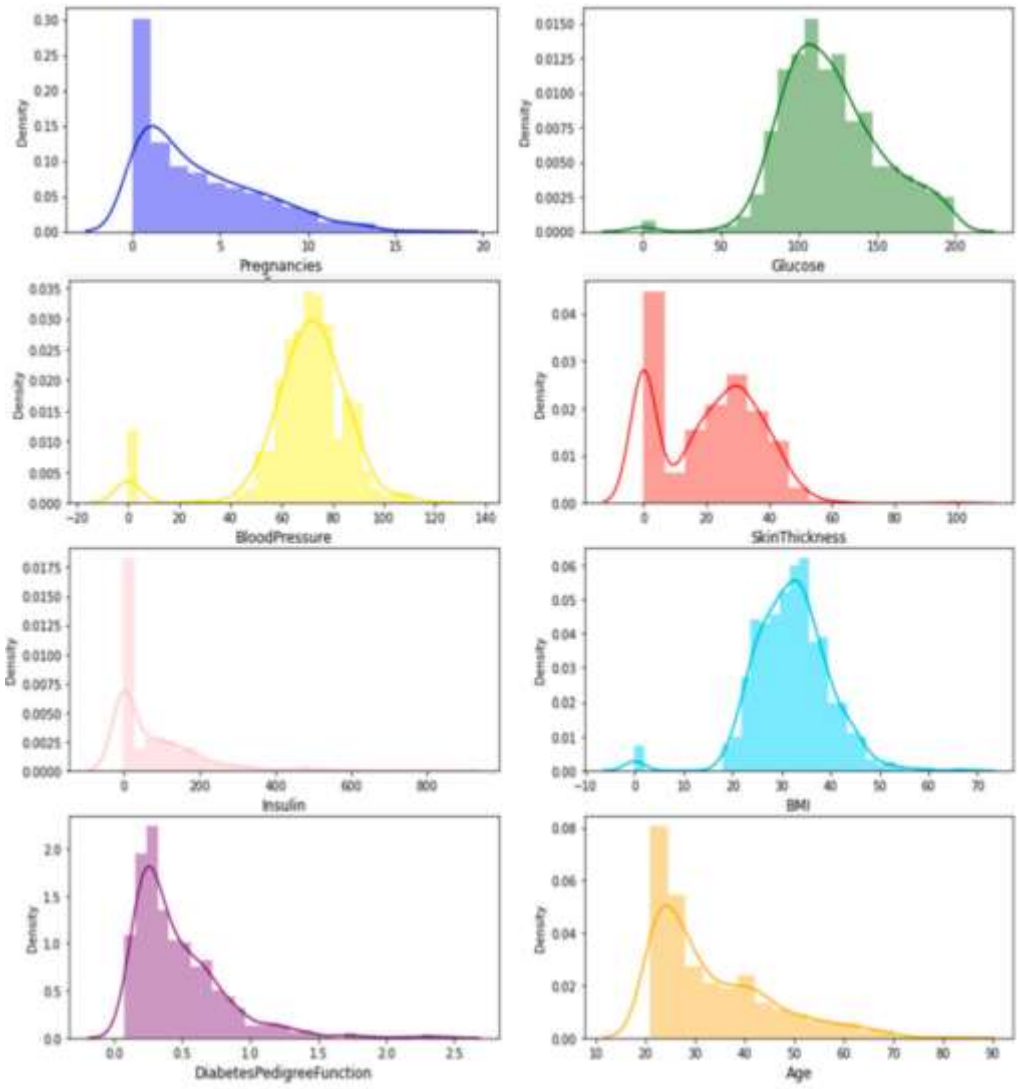

Fig. 2 Distribution of data under study

Finding important or potentially influencing factors, as well as establishing links between them, are both made easier using this distribution. A single variable's frequency of unique values is shown in each graphic. The figure suggests that insulin feature levels are lower than other diabetes characteristics. A visual representation of the values associated with each diabetes characteristic is provided by the distribution of those traits. The overall response patterns of a group may be shown in box plots. We have two outliers that are over the maximum, and the fact that the box plot for the pregnancy feature is skewed towards the bottom indicates that the majority of the values are closer to the lowest range. The graph shows how the characteristics are related to one another. Features have a high degree of correlation

when the values are near to one. There is little to no correlation between the traits if the values are around zero. The strong correlation between the characteristics is shown by the dark colors. Three sets of data were extracted from the diabetes dataset: one for testing, one for validation, and one for training. After experimenting with several splitting ratios to test the system's resilience, we found that the above-mentioned ratio produced the best results. A number of measures are used to assess the accuracy, precision, recall, false positive rate, accuracy, and error rate of the diabetes prognostication system.

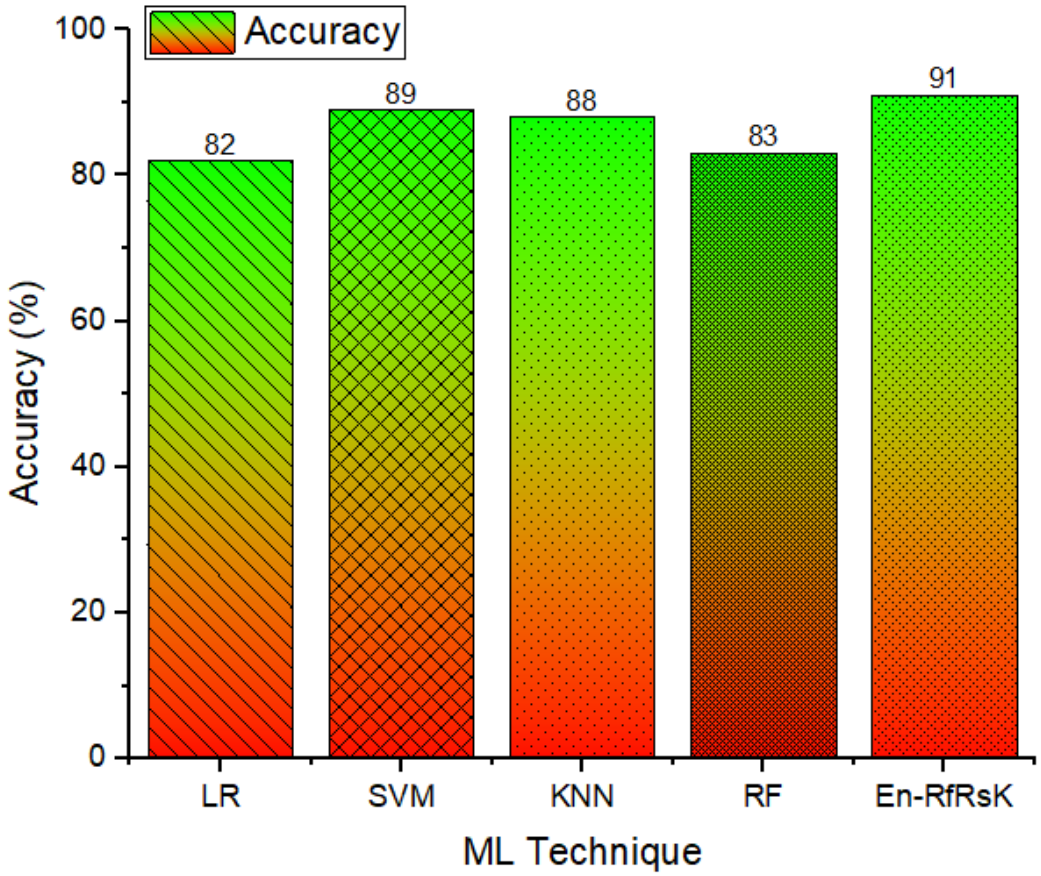

Fig. 3 Accuracy of machine learning algorithms

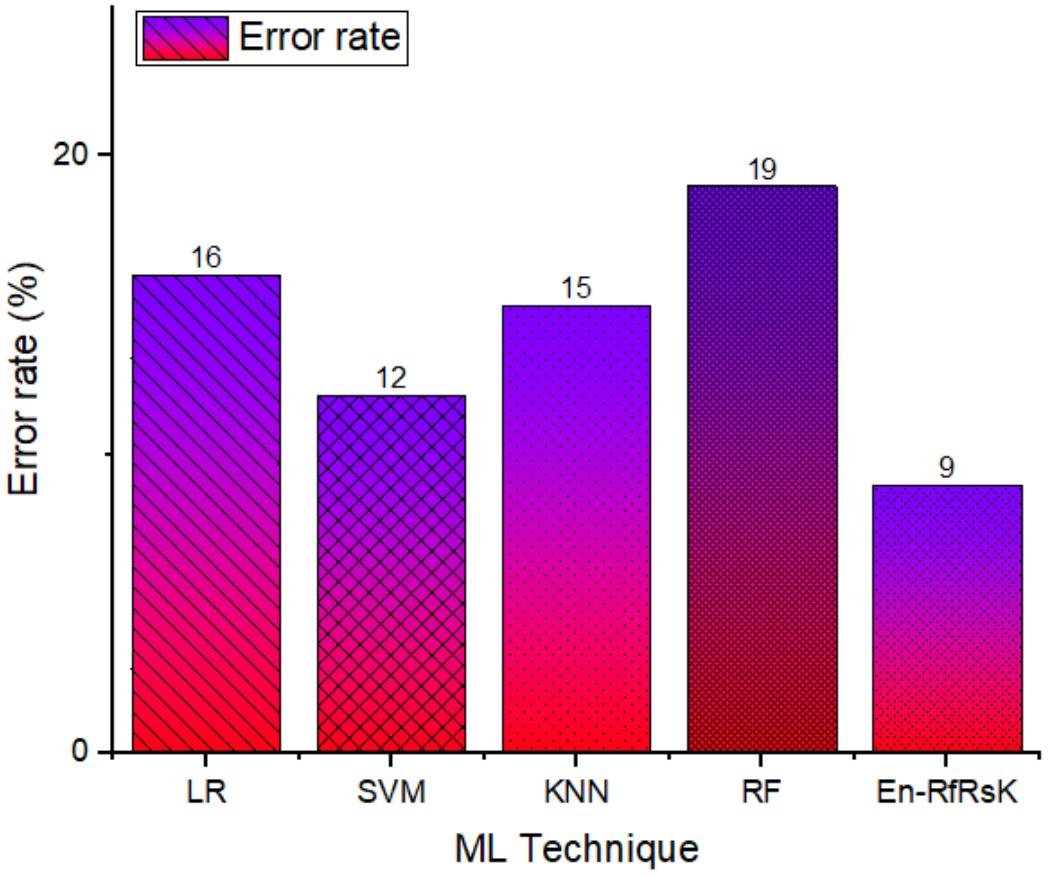

Fig. 4 Errors encountered in different machine learning algorithms

The accuracy and error rate metrics, as shown in Figure 3, and Figure 4, respectively, illustrate the performance of the suggested method. When compared to other machine learning methods, the suggested strategy clearly provides better results in terms of accuracy and lower error rates. When compared to other diabetes prediction systems and base classifiers, the suggested method clearly performs better in terms of accuracy, error rate, and performance metrics. This is especially true when considering the proposed En-RfRsK.

Table 2 Accuracy and error percentages of different machine learning classifiers

ML Technique	**Accuracy (%)**	**Error (%)**
Linear regression	82	16
SVM	89	12
KNN	88	15

Random Forest	83	19
En-RfRsK	91	9

4. Conclusion

Diabetes mellitus is a common condition among adults today. Hence, it is essential to provide early prognostication for this condition. The primary goal of this study is to develop a method that can accurately predict who will get diabetes. The veracity of previously used machine learning methods was investigated. Given this, we provide an updated approach called En-RfRsK. It combines three machine learning algorithms; linear regression, SVM, K-nearest neighbor, and decision tree. For this experiment, we will be using the PIMA Indians diabetes dataset. It was noted that diabetes poses a significant danger during pregnancy and requires proper treatment to ensure a healthy pregnancy. Keeping your body mass index (BMI) in a healthy range keeps diabetic complications to a minimum. An increasing number of individuals are being affected by diabetes as they age, starting around age 35. Outperforming both baseline classifiers and cutting-edge diabetes prediction systems, the ensemble method has produced findings with an accuracy of 88.89%. There is more room for performance improvement in terms of accuracy, and the ensemble classifier lacks interpretation, which are downsides of the suggested technique. Additional data sets with diverse features and risk variables for diabetes mellitus prediction may be explored in future studies. Finding a happy medium between accuracy, complexity, and interpretability is essential for diabetes prediction models. It is also possible to tweak the suggested method to investigate other approaches to outlier removal and missing value imputation, as well as to investigate alternative feature selection strategies for selecting useful characteristics and removing irrelevant ones. In addition, deep learning models have the potential to improve this method's accuracy down the road.

Reference

1. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. Appl Sci 2021;11(3):1173.
2. Lu H, Uddin S, Hajati F, Moni MA, Khushi M. A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus. Appl Intell 2022;52(3):2411–22.
3. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Proc Comput Sci 2020;167:706–16.
4. Ahmed U, Issa GF, Khan MA, Aftab S, Khan MF, Said RA, et al. Prediction of diabetes empowered with fused machine learning. IEEE Access 2022;10:8529–38.
5. Wadghiri M, Idri A, El Idrissi T, Hakkoum H. Ensemble blood glucose prediction in diabetes mellitus: a review. Comput Biol Med 2022:105674.
6. Islam M, Raihan M, Akash SRI, Farzana F, Aktar N, et al. Diabetes mellitus prediction using ensemble machine learning techniques. In: International conference on computational intelligence, security and Internet of things. Springer; 2019. p. 453–67.
7. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ml) algorithms. Neural Comput Appl 2022:1–17.

8. Zhang Y, Xu X. Machine learning tensile strength and impact toughness of wheat straw reinforced composites. Mach Learn Appl 2021;6:100188.
9. Tripathi G, Kumar R. Early prediction of diabetes mellitus using machine learning. In: 2020 8th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO). IEEE; 2020. p. 1009–14.
10. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In: 2017 international conference on intelligent sustainable systems (ICISS). IEEE; 2017. p. 533–8.
11. Singh A, Dhillon A, Kumar N, Hossain MS, Muhammad G, Kumar M. ediapredict: an ensemble-based framework for diabetes prediction. ACM Trans Multimed Comput Commun Appl 2021;17(2s):1–26.
12. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 2019;6(1):1–19.
13. Patil V, Ingle D. Comparative analysis of different ml classification algorithms with diabetes prediction through pima Indian diabetics dataset. In: 2021 international conference on intelligent technologies (CONIT). IEEE; 2021. p. 1–9.
14. Mahabub A. A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Appl Sci 2019;1(12):1–12.
15. Zhang Y, Xu X. Yttrium barium copper oxide superconducting transition temperature modeling through Gaussian process regression. Comput Mater Sci 2020;179:109583.
16. Zhang Y, Xu X. Disordered mgb2 superconductor critical temperature modeling through regression trees. Physica C, Supercond Appl 2022;597:1354062.
17. Laila Ue, Mahboob K, Khan AW, Khan F, Taekeun W. An ensemble approach to predict early-stage diabetes risk using machine learning: an empirical study. Sensors 2022;22(14):5247.
18. Warsi GG, Saini S, Khatri K. Ensemble learning on diabetes data set and early diabetes prediction. In: 2019 international conference on computing, power and communication technologies (GUCON). IEEE; 2019. p. 182–7.
19. Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-ii using hybrid machine learning based ensemble model. Int J Inf Technol 2020;12(2):419–28.
20. Mirshahvalad R, Zanjani NA. Diabetes prediction using ensemble perceptron algorithm. In: 2017 9th international conference on computational intelligence and communication networks (CICN). IEEE; 2017. p. 190–4.
21. Srivastava R, Dwivedi RK. Diabetes mellitus prediction using ensemble learning approach with hyperparameterization. In: ICT analysis and applications. Springer; 2022. p. 487–94.
22. Zhang Y, Xu X. Solubility predictions through lsboost for supercritical carbon dioxide in ionic liquids. New J Chem 2020;44(47):20544–67.
23. Zhang Y, Xu X. Modulus of elasticity predictions through lsboost for concrete of normal and high strength. Mater Chem Phys 2022;283:126007.
24. Srivastava R, Dwivedi RK. A survey on diabetes mellitus prediction using machine learning algorithms. In: ICT systems and sustainability. Springer; 2022. p. 473–80.
25. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng 2021;2:40–6.