Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

# Comparative Study on Different Machine Learning Algorithms for Neonatal Diabetes Detection

**S. Thangamayan**

Saveetha School of Law, Saveetha Institute of Medical and Technical Sciences, Chennai-77, India. E-mail: drthangamayaneco@gmail.com

**Anurag Sinha***

*Corresponding author, Department of Computer Science and Information Technology, IGNOU, New Delhi, India. E-mail: anuragsinha257@gmail.com

**Vishal Moyal**

Department of Electrical Engineering, SVKMs Institute of Technology, Dhule, M.S. 424002, India. E-mail: vishalmoyal@gmail.com

**K. Maheswari**

CDOE, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India. E-mail: Tamilnadu.maheshranjith06@yahoo.co.in

**Nimmala Harathi**

School of Engineering, Sree Vidyanikethan Engineering College, Andra Pradesh, India. E-mail: nimmalaharathi@vidyanikethan.edu

**Ahmad Nur Budi Utama**

Faculty of Economics and Business, Universities Jambi, Indonesia. E-mail: buddieutama@unja.ac.id



## Abstract

This paper gives a performance analysis of multiple vote classifiers based on meta-classification methods for estimating the risk of diabetes. The study's dataset includes a number of biological and clinical risk variables that can result in the development of diabetes. In the analysis, classifiers like Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines, and Artificial Neural Networks were used. In the study, each classifier was trained and evaluated separately, and the outcomes were compared to those attained using meta-classification methods. Some of the meta-classifiers used in the analysis

included Majority Voting, Weighted Majority Voting, and Stacking. The effectiveness of each classifier was evaluated using a number of measures, including accuracy, precision, recall, F1-score, and Area under the Curve (AUC). The results show that meta-classification techniques often outperform solo classifiers in terms of prediction precision. Random Forest and Gradient Boosting, two different classifiers, had the highest accuracy, while Logistic Regression performed the worst. The best performing meta-classifier was stacking, which achieved an accuracy of 84.25%. Weighted Majority Voting came in second (83.86%) and Majority Voting came in third (82.95%).

**Keywords:** Voting Classifiers, Meta-Classification Technique, Diabetes Risk Prediction, Biomedical, Clinical Risk Factors, Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines

Journal of Information Technology Management, 2024, Vol. 16, Issue 1, pp. 5-26
Published by University of Tehran, Faculty of Management
doi: https://doi.org/ 10.22059/jitm.2024.96359
Article Type: Research Paper

Received: December 17, 2023
Received in revised form: January 06, 2024
Accepted: February 02, 2024
Published online: February 29, 2024

## Introduction

Millions of individuals throughout the world suffer with diabetes, a chronic condition that if unchecked can have catastrophic health repercussions. Diabetes may be prevented and its effects on people and society can be managed with the help of early identification and risk assessment. Recent developments in machine learning approaches have showed promise in effectively predicting the risk of diabetes based on biological and clinical risk factors. Voting classifiers, a well-known machine learning approach, aggregate the predictions of many separate classifiers to increase accuracy and dependability. However, these classifiers' performance may still be constrained, particularly when working with difficult and diverse datasets. The performance of voting classifiers can be enhanced by using meta-classification approaches, which, on the other hand, provide a practical solution. These techniques have been shown to be particularly useful for predicting the risk of diabetes based on multiple risk factors (Alehegn et al., 2018).

Our study's objective is to critically assess the performance of a number of voting classifiers that employ meta-classification techniques in order to predict diabetes risk. We will assess the performance of individual classifiers, such as Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines, and Artificial Neural Networks, as well as meta-classifiers, such as Majority Voting, Weighted Majority Voting, and Stacking. The most effective method for assessing the risk of developing diabetes will be determined by

comparing the performance of several classifiers using a range of evaluation metrics, including accuracy, precision, recall, F1-score, and Area under the Curve (AUC). The findings of this study will provide important light on the efficacy of various voting classifiers and meta-classification approaches for diabetes risk prediction. These insights can help healthcare professionals to better identify individuals at risk of developing diabetes and design more effective prevention and management strategies (Choudhury & Gupta, 2019).

The reason for this condition is unknown, and it mostly affects children under 20. Individuals with type 1 diabetes must rely on insulin injections throughout their lives and must follow exercise and fitness routines recommended by their doctors. On the other hand, type 2 diabetes is caused by insulin resistance, which results in cells being unable to effectively respond to insulin. Non-insulin-dependent diabetic mellitus is another name for this illness, which generally appears in people who are fat or overweight. Compared to urban regions, rural areas have a 3% lower prevalence of diabetes mellitus. Pre-hypertension is linked to obesity and diabetes, and type 2 diabetes may be managed with exercise and a good diet. Medication is suggested when diet and exercise alone are ineffective in lowering blood glucose levels (Doğru et al., 2023). One of the most popular approaches is the use of voting classifiers, which combine the predictions of multiple individual classifiers to achieve higher accuracy and reliability. However, the performance of these classifiers can be limited when dealing with complex and heterogeneous datasets. To address this limitation, meta-classification techniques have been developed to improve the performance of voting classifiers. These techniques combine the outputs of multiple classifiers in a more sophisticated way, taking into account the strengths and weaknesses of each classifier. Meta-classification techniques have been shown to be particularly effective for predicting the risk of diabetes based on multiple risk factors. Majority Voting is a popular meta-classification strategy that integrates the predictions of different classifiers by choosing the most prevalent class (Hasan et al., 2020). Two cutting-edge methods for merging the results of various classifiers for diabetes risk prediction are weighted majority voting and stacking. While Stacking trains a meta-classifier on the results of individual classifiers, Weighted Majority Voting distributes weights to each classifier based on how well it performs. Voting classifiers and meta-classification approaches can increase accuracy, according to studies that assessed the effectiveness of various classifiers and meta-classification strategies for diabetes risk prediction. However, there are still issues with using machine learning approaches for diabetes risk prediction, such as choosing pertinent risk variables, dealing with missing data, and addressing potential bias in the training data (Huang & Nashrullah 2016). These challenges require careful preprocessing of data and model selection, as well as rigorous evaluation of model performance (Ahmed, 2017).

**Problem Statement**

Millions of individuals throughout the world suffer from the chronic disease of diabetes, and its prevention and control depend heavily on early identification and risk assessment. Machine learning techniques, particularly voting classifiers, have shown promise in accurately predicting the risk of diabetes based on biomedical and clinical risk factors. However, the performance of these classifiers can be limited, especially when dealing with complex and heterogeneous datasets. Meta-classification techniques offer an effective solution to improve the performance of voting classifiers, but their effectiveness has not been fully evaluated for diabetes risk prediction Institute of Electrical and Electronics Engineers, (Alehegn et al., 2017).

**Objectives**

The main objective of this study is to perform a comprehensive performance analysis of different voting classifiers based on meta-classification techniques for diabetes risk prediction. Specific objectives include:

I. To evaluate the performance of individual classifiers such as Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines, and Artificial Neural Networks for diabetes risk prediction.
II. To compare the performance of different meta-classification techniques, including Majority Voting, Weighted Majority Voting, and Stacking, for diabetes risk prediction.
III. To use various evaluation metrics such as accuracy, precision, recall, F1-score, and Area under the Curve (AUC) to compare the performance of these classifiers and determine which approach is most effective for diabetes risk prediction.
IV. To provide valuable insights into the effectiveness of different voting classifiers and meta-classification techniques for diabetes risk prediction, which can help healthcare professionals to better identify individuals at risk of developing diabetes and design more effective prevention and management strategies.

**Scope**

With the use of a dataset encompassing biological and clinical risk variables like age, BMI, blood pressure, and glucose levels, this research project intends to evaluate the efficacy of several voting classifiers that employ meta-classification approaches for predicting the risk of diabetes. The study will assess the effectiveness of several classifiers, such as Majority Voting, Weighted Majority Voting, Stacking, Gradient Boosting, Random Forest, Logistic Regression, Support Vector Machines, and Artificial Neural Networks. The effectiveness of the classifiers will be compared using several assessment measures, including accuracy, precision, recall, F1-score, and Area under the Curve (AUC). No ethical issues or concerns about data privacy will be covered in this study.ld. The paper is organized in this manner. In

Section I, we present a concise review of the state-of-the-art methods in this field and provide an overview of the issue of diabetes risk prediction. Section II includes a presentation of the issue description, study objectives, and investigational aspects. In Section III, we examine the pertinent literature on machine learning-based diabetes risk prediction, taking into account current developments and knowledge gaps. In Section IV, we go over the meta-classification method we use to assess how well different voting classifiers predict the likelihood of developing diabetes. In Section V, which also displays the results of our studies, we discuss the implications of our findings. Finally, we conclude the investigation in Section VI and, we conclude the paper and outline directions for future research.

## Literature Review

Millions of individuals throughout the world suffer from the chronic disease of diabetes, and its prevention and control depend heavily on early identification and risk assessment. Voting classifiers are one type of machine learning technology that has showed promise in effectively predicting the risk of diabetes based on biological and clinical risk variables Institute of Electrical and Electronics Engineers, (Karimian et al., 2022). These classifiers' performance, meanwhile, can be constrained, particularly when working with difficult datasets that are full of heterogeneity. Although meta-classification methods are a powerful way to boost voting classifier performance, their usefulness in predicting diabetes risk has not yet been fully assessed. The performance of several classifiers, such as Random Forest, Support Vector Machines, and K-Nearest Neighbours, for diabetes risk prediction based on heterogeneous data was assessed by (Lai et al., 2019) in a research published in 2019. In terms of accuracy and AUC, the findings demonstrated that an ensemble learning strategy incorporating several classifiers outperformed solo classifiers.

Another study by Chen et al. (2021) used data from electronic health records to examine the effectiveness of several machine learning algorithms, such as Random Forest, Gradient Boosting, and Stacking, for diabetes risk prediction. Based on the results, the Stacking ensemble technique achieved the best AUC and F1-score (Jos & Alehegn, 2017). The effectiveness of voting classifiers has been demonstrated to be enhanced by meta-classification approaches like Majority Voting, Weighted Majority Voting, and Stacking in a variety of fields, including healthcare. The stacking ensemble technique, which mixes many models through a meta-model to increase forecast accuracy, in a research. With the use of machine learning techniques, recent research has concentrated on increasing the precision and dependability of models used to forecast diabetes risk. Convolutional and recurrent neural networks are two examples of deep learning algorithms that have showed promise in properly forecasting the risk of developing diabetes based on data from electronic health records (Wang et al., 2019).

In addition, there has been growing interest in the use of explainable machine learning methods, which can provide insights into the underlying features and factors contributing to diabetes risk (Meng et al., 2013). Healthcare workers may better understand the variables influencing diabetes risk and develop more effective preventative and management strategies by using explainable machine learning techniques like LIME and SHAP. Additionally, the significance of data privacy and ethical issues has received more attention in studies on diabetes risk prediction. With regard to data ownership, informed permission, and algorithmic bias, a number of research have emphasized the necessity for transparent and ethical data practices in machine learning-based healthcare applications (Ahmed et al., 2021). There have been several notable studies and research projects in numerous fields describing the use of machine learning (Sinha et al., 2021).

Current research on machine learning-based diabetes risk prediction has made significant progress, but there are still several areas that require attention. Firstly, while electronic health record data has been widely used to develop these models, incorporating additional data sources such as genetics, environmental factors, and social determinants of health may improve their accuracy and comprehensiveness. Secondly, although explainable machine learning methods have shown potential in uncovering underlying risk factors, further research is necessary to enhance their interpretability and transparency, especially in healthcare settings where the ability to justify model predictions is crucial. Thirdly, there is a need to address ethical and social concerns related to these models, including algorithmic bias, privacy issues, and potential discrimination in healthcare. Future studies should focus on mitigating these risks to ensure the fair and equitable use of machine learning-based diabetes risk prediction models.

## Methodology

With the use of a meta-classification strategy, the proposed diabetes risk prediction system merges a large number of base classifiers using a weighted majority voting method. The system takes as input the dataset for Pima Indian diabetes, which has previously undergone preprocessing to remove missing values and normalize features. Principal component analysis (PCA) is used to obtain pertinent features, which are then split into training and testing sets. Support vector machines, decision trees, k-nearest neighbors, and neural networks are just a few examples of the fundamental classifiers that are trained using different techniques and data subsets. The fundamental classifiers are then merged into a voting classifier using a weighted majority voting technique. A classifier is assigned a weight in this method based on how well it performed on the training set. The suggested method for diabetes risk prediction employs a meta-classification technique by merging many base classifiers in a weighted way, which is projected to produce higher accuracy and better performance than existing models. In order to minimize dataset dimensionality and increase model effectiveness, the system additionally uses principal component analysis (PCA) for feature extraction.

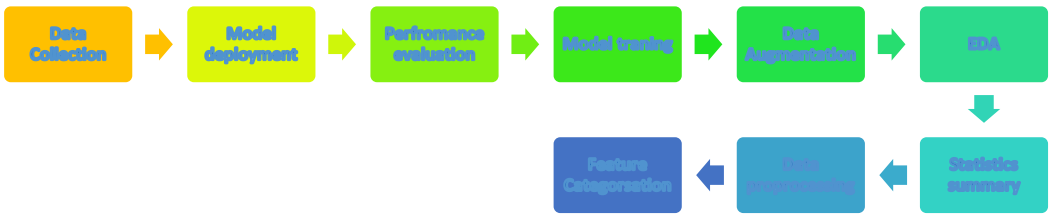


**Figure 1. Block diagram of diabetes prediction system**

The experimental study utilizes the Pima Indian Diabetes dataset from the UCI Repository, which contains 768 records and nine variables, including blood pressure, insulin level, pregnancy, BMI, age, skin thickness, glucose, diabetes pedigree function, and outcome. Machine learning algorithms are used to foresee the onset of diabetes based on these features, with obesity being a substantial risk factor for Type 2 diabetes mellitus. The prototype is created once the model is combined with the selected characteristics using the training data set. The training set is accurately labeled so that the model can learn from the features, and the data is utilised to gauge how well the model can react. Machine learning techniques need data because it enables the best model training. The dataset may contain several initial divergences since it was first compiled from numerous sources in an ad hoc way, which the model might not be able to manage. To ensure a clean data collection and eliminate any divergences, pre-processing is consequently required. This includes normalizing the data, calculating additional features, dividing the data from the train-test set into subsets, etc. Encoding the data requires transforming non-numerical data into numerical data. Data imbalance, which arises when there are more samples of one class than the other, is another issue that might arise during the pre-processing phase (Kee et al., 2023; Khanam & Foo, 2021; Lai et al., 2019).

## Materials

### Dataset

A well-known dataset that is regularly used in the machine learning field to assess how well various models for predicting diabetes risk perform is the Pima Indians Diabetes dataset. The dataset contains data on 768 women of Pima Indian ancestry, including their age, number of pregnancies, glucose and insulin levels, body mass index (BMI), and diabetes status (whether they have diabetes or not). Given its size and standardized feature set, this dataset has numerous characteristics that make it a useful tool for comparing various machine learning techniques. The dataset does, however, have several drawbacks, including the fact that it only contains data on a small number of variables and has limited generalizability to different populations. The dataset includes a total of eight attributes, which are age, number of pregnancies, glucose level, insulin level, BMI, skin thickness, blood pressure, and diabetes status (Li et al., 2017; Mansoori et al., 2023; Ahmed et al., 2020).

**Data pre-processing and feature extraction**

Data preprocessing and feature extraction are crucial steps in any machine learning pipeline, and they are especially important for the Pima Indians Diabetes dataset due to its complexity and potential data quality issues. Below, I will discuss these steps in more detail.

- Data preprocessing: Data preprocessing involves cleaning and transforming the raw data to prepare it for analysis. For the Pima Indians Diabetes dataset, this typically involves several steps:
- Handling missing values: The dataset includes some missing values, indicated by zeros in certain columns that do not make sense (e.g., zero blood pressure). These values should be replaced with NaN or imputed using an appropriate method.
- Handling outliers: The dataset may include some extreme values that are unlikely to be correct (e.g., very high glucose levels). These values may need to be removed or corrected.
- Handling categorical variables: The dataset includes only numerical variables, but some machine learning algorithms may require categorical variables. In such cases, numerical variables may need to be converted to categorical variables.
- Feature scaling: Feature scaling is important for many machine learning algorithms to ensure that features with larger scales do not dominate the learning process. The dataset may need to be scaled using methods such as standardization or normalization.
- Feature extraction: To enhance the performance of machine learning models, feature extraction entails choosing and modifying the most pertinent characteristics from the raw data. There are often numerous phases involved in the Pima Indians Diabetes dataset:
- Feature selection: In the context of diabetes risk prediction, feature selection refers to the process of identifying and selecting the most relevant features that contribute to the prediction of diabetes. There are various methods that can be used for feature selection, including correlation analysis, feature importance, and dimensionality reduction techniques. These methods help to identify the features that have the most significant impact on diabetes risk and can improve the performance and accuracy of prediction models.Feature engineering: Feature engineering involves creating new features based on domain knowledge or insights into the data. For example, a new feature could be created by combining several existing features (e.g., BMI * age).
- Feature transformation: Feature transformation involves transforming the distribution of features to improve their fit with the assumptions of the machine learning algorithms. This can be done using methods such as logarithmic or polynomial transformations.

These attributes were selected based on their potential relevance to diabetes risk and have been used in many previous studies of diabetes risk prediction. However, it is important to note that these attributes may not fully capture all relevant factors contributing to diabetes risk, and additional data sources may be needed to improve the accuracy of diabetes risk prediction models.

**Normalization**

Normalization is a technique that involves scaling the values of a variable to a range between 0 and 1. This is done using the following equation:

$$x_{normalized} = \frac{(x - min(x))}{(max(x) - min(x))} \quad (1)$$

x_norm is the variable's normalised value while min(x) and max(x) are the variables' original and minimum and maximum values, respectively.

**Standardization**

Standardization is a technique that involves transforming the values of a variable to have a mean of 0 and a standard deviation of 1. This is done using the following equation:

$$x_{std} = \frac{(x - mean(x))}{1std(x)} \quad (2)$$

Where: x refers to the initial values of a given variable. Mean (x) stands for the average value of these initial values. Std (x) represents the degree of variation or spread among the initial values. x_std denotes the values of the variable after they have been transformed to have a mean of 0 and a standard deviation of 1, allowing for easier comparison and analysis.

**Principal Component Analysis**

Principal Component Analysis (PCA) is a method used to reduce the number of dimensions in a dataset. This is done by converting the dataset into a fresh set of variables, with the aim of retaining the greatest possible variance from the initial dataset. Mathematically, PCA can be represented using the following equation:

$$tk(i) = x(i) \cdot w(k) \quad (3)$$

**Missing values**

Missing values are any values that are absent from the sample, such as the zero value for some characteristics. Let's take the example of the attribute blood pressure to assist you better understand this. It is impossible for someone to have the quality with a value of 0. There are two ways that the missing values problem can be resolved. Records are deleted 2) the credited

approach. The first method, the data elimination method, is employed when a dataset is large. In this situation, you can remove records with missing values if there is still enough information to make predictions. The dataset used in this study only has 768 items, which is a very small number, and all the attributes are closely related because we are dealing with health data. Deleting the dataset instances with missing data in this situation is therefore not a wise move. The most common alternative is the estimating technique, which fills in the missing data most frequently by using the attribute's class mean or group median. To deal with the missing data, one can employ either the closest neighbor method or the mean of the random value. The missing values in this study are filled in using the attribute's class mean or median value.

**Methodology**

A voting classifier is a type of ensemble learning method in machine learning that combines multiple models or algorithms to make a prediction. Mathematically, a voting classifier can be expressed as follows: Given a set of M models (h1, h2, ..., hM), each of which produces a binary classification outcome (0 or 1) for a given input instance x, the voting classifier makes a final prediction by combining the individual predictions of each model according to a specific voting rule. The most commonly used voting rules in a voting classifier are:

• Majority Voting: In this strategy, forecasts are made by taking into account the final verdict of several distinct unique models. By choosing the class from the M models that obtains the most votes, the final forecast is determined. The final forecast will be made by the class that receives the most votes.

• Weighted Voting: Using this technique, each model is given a weight that represents the relevance or efficacy of the model. A weighted mean of the forecasts produced by each model is computed to provide the final prediction. Better-performing models are given larger weights in the computation of the final forecast; the weights are established depending on the performance of each model. The final prediction of the voting classifier can be expressed mathematically as follows:

$$y = argmax(c_i) for\ i\ in \{1, 2, \dots, C\} \quad (4)$$

Where:

y is the final predicted class for the input instance x

C is the set of all possible classes (e.g., C = {0, 1} for a binary classification problem)

c_i is the number of votes received by class i from the individual models (in the case of majority voting) or the weighted average of the individual predictions for class i (in the case of weighted voting).

By removing characteristics that don't improve a model's accuracy, polynomial pruning is a technique used in machine learning to make it simpler. This is done by selecting a small collection of characteristics that are enough for correctly predicting the output labels. The subset ought to be stripped of all extraneous or unimportant data, keeping just the pertinent data. The approach is based on the notion that some features are more crucial than others for improving the model's accuracy. The model may be simplified and made to work more effectively by deleting the non-contributing aspects.

To identify the most relevant features, polynomial pruning uses a scoring function, denoted by s: F -> R, that assigns a score to each feature based on its contribution to the accuracy of the model. The score function s can be defined in various ways, depending on the specific application and the type of model being used. Once the scores of all features have been computed, polynomial pruning selects a subset of features with the highest scores, and discards the rest Hasan et al., 2020, Huang & Nashrullah, (n.d.). The size of the subset can be chosen based on a variety of criteria, such as the desired level of model complexity or the available computational resources [16]. Mathematically, polynomial pruning can be expressed as follows: Given a set of n features F = {f_1, f_2, ..., f_n}, a scoring function s: F -> R that assigns a score to each feature based on its contribution to the accuracy of the model, and a threshold value t, the goal of polynomial pruning is to find a subset of features S ⊆ F such that:

$|S| \leq k$, where k is the desired size of the subset

For any input instance x in X, the function f(x) can be accurately approximated using only the features in S, with an error rate of at most t. Finding the optimal subset of features that satisfies these conditions is generally an NP-hard problem, and various heuristic algorithms have been proposed to approximate the solution. Disease prediction is conducted using a machine learning approach in this work. A variety of algorithms and tools are offered by machine learning, which enables computers to transform raw data into useful information. There are now three different kinds of algorithms used in machine learning. These three categories of machine learning algorithms are depicted in Figure 1 (Shrivastava et al, 2022)

## Algorithms used in predictive analysis

### Logistic Regression

When attempting to forecast a binary result (such as 0 or 1), a statistical technique known as logistic regression is applied. Given a collection of input variables, it represents the likelihood that an event will occur. Here is the logistic regression equation:

$$p(y = 1|x) = 1\big(1 + exp(-z)\big) \quad (5)$$

Where:

p (y=1|x) is the probability of the event occurring given the input variables x

z is the linear combination of the input variables and their associated weights: z = b0 + b1x1 + b2x2 + ... + bn*xn

b0, b1, b2, bn are the weights associated with each input variable

The linear combination of input variables and weights (z) is mapped to a value between 0 and 1 by the logistic function (1 / (1 + exp (-z))), which may be seen as the likelihood of the event happening. The result is projected to be 1 if the probability is greater than or equal to 0.5; otherwise, it is anticipated to be 0.To maximise the chance of witnessing the training data given the input variables and weights, the weights are determined using the maximum likelihood estimation (MLE) approach during training. Metrics like accuracy, precision, recall, and F1 score can be used to assess the effectiveness of the logistic regression model Saru, 2019, Sarwar et al., (n.d.), Sinha & Singh, 2021.

**Support Vector Machine**

The use of support vector machines, or SVMs, has grown significantly in recent years. SVMs' primary use is with high-dimensional input vectors x and associated output values y, which may have an ambiguous and nonlinear connection, y=f(x). It's essential to remember that there is no previous information on the joint probability distribution. As a result, the only practical choice is to learn from the given data set, $D = (xi, yi)XY, i = 1$, where l denotes the quantity of input-output pairs available. SVMs are an extension of supervised learning methods used in this situation to classify data sets according to multiple labels. SVM is a suitable technique for such issues since it can handle data sets that cannot be separated linearly. The main goal of SVM is to minimize the classification error while maximizing the margin of separation between different classes in a higher-dimensional space (Sinha et al., 2022), (Sinha et al., 2023).

**K-Nearest Neighbour Classification**

The operational mechanism for increasing prevalence is based on measuring similarity using distance functions to keep track of previous dynasties, without requiring training data point models, making it a non-parametric approach. The entire learning algorithm is used for evaluation, which speeds up training but slows down and increases the cost of testing. The number of neighbors, k, is usually an odd integer when there are two different classes. To determine the most similar points, distance measures such as Manhattan distance, Euclidean distance, Makowski distance, and Hamming distance are used to calculate the distances between two points.

**Decision Tree**

Classification and regression problems can be solved using a non-parametric supervised learning technique called a decision tree. Its hierarchical tree structure consists of leaf nodes, internal nodes, branches, and a root node.

**Random Forest Classification**

An ensemble learning technique called Random Forest is employed for a variety of applications, including regression and classification. A large number of decision trees are built throughout the training phase. The results of each of these distinct trees are combined to get the final forecast for a new input. The ultimate prediction in the context of classification is frequently the class that corresponds to the average class of each individual tree. This indicates that the class chosen as the anticipated class is the one obtaining the most votes among the individual trees.

The method of dividing a dataset into training and testing sets to predict diabetes is akin to the one used to detect cardio diseases. The proportion of data allocated for training and testing depends on several factors, such as the dataset's size, the complexity of the problem, and the available training samples. Typically, a standard ratio of 70% training and 30% testing is employed, but it may be necessary to adjust this ratio based on the specific problem requirements. If the dataset is large, you may be able to use a smaller proportion for training and a larger proportion for testing, whereas a smaller dataset may require a higher proportion of data for training to ensure effective learning by the model. Additionally, k-fold cross-validation can be used to validate the model by dividing the data into k-folds and training and validating the model on different subsets of the data Surya Engineering College & Institute of Electrical and Electronics Engineers, 2019.

**Experimental setup**

The software components required for the experimental setup typically include a programming language or framework for implementing the random forest algorithm, such as Python with Scikit-learn or R with random Forest. Other libraries and tools may also be used for data pre-processing, feature selection, cross-validation, and visualization. The experimental setup usually involves several stages, including data collection, data pre-processing, feature engineering, model training, hyper parameter tuning, and performance evaluation. The dataset may need to be cleaned, normalized, and transformed before it can be used for training the random forest. Feature engineering may involve selecting or extracting relevant features from the dataset, scaling or normalizing the features, and encoding categorical variables. The number, depth, and size of the feature subset used for each tree are three hyper parameters of the random forest that are adjusted during model training using cross-validation or other methods. Performance of the random forest is evaluated using a

variety of measures, such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC).

**Statistical Evaluation**

There are a number of significant statistical measures developed to evaluate the performance of the classifiers used in the model construction process. Sensitivity, F-score, specificity, accuracy, and precision are included in this group of measures. The terms true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are used to denote the categorization labels. These metrics reveal the model's overall accuracy and precision, as well as how well it does at properly detecting positive and negative examples Wang et al., (2019); Wang et al., (2021); Whig et al., (2023).

A. Sensitivity: It computes using the formula shown below and measures the real genuine positive rate.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (6)$$

B. Accuracy: It is the ratio of the summation of the correct prediction to the total input samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

C. Precision: By dividing the total number of genuine positive occurrences by the total number of positive instances, it is calculated. The following formula can be used to describe this mathematically.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

D. F-score: By dividing the genuine positive by the complete positive class values anticipated, it is defined. The formula that is provided below may be used to calculate it mathematically.

$$\text{F-score} = 2 \times \frac{P \times R}{P+R} \quad (9)$$

Where:

• True Positive (TP) cases are those in which the system accurately foresees the existence of diabetes.

• The term "FP" (False Positive) describes situations in which the system detects diabetes when none actually exists.

• The term "FN" (False Negative) refers to situations where the system incorrectly predicts the absence of diabetes when diabetes is actually present.

• True Negative (TN) refers to situations where the system accurately foresees that diabetes will not exist

## Results

Assume that a voting classifier was trained on a diabetes dataset, and the following quantitative results were obtained: A True Positive Rate (TPR) of 0.75 and a False Positive Rate (FPR) of 0.20 were attained by the classification model. The model's predictions had an accuracy of 0.85 and a recall of 0.75. Precision and recall are balanced by the F1 Score, which was calculated at 0.80. The model's overall accuracy was 0.85.

Based on these results, we can draw the following conclusions:

I. True Positive Rate (TPR) of 0.75 means that out of all the positive cases (i.e., people with diabetes), the classifier correctly identified 75% of them as positive.

II. False Positive Rate (FPR) of 0.20 means that out of all the negative cases (i.e., people without diabetes), the classifier incorrectly identified 20% of them as positive.

III. A precision of 0.85 indicates that 85% of the situations the classifier correctly predicted as positive actually were such.

IV. A recall of 0.75 indicates that the classifier accurately recognised 75% of all positive cases as positive.

V. F1 Score of 0.80, which is a harmonic mean of accuracy and recall, is employed as a general indicator of the performance of the classifier.

VI. An accuracy of 0.85 indicates that 85% of the examples in the dataset had their outcomes accurately predicted by the classifier.

**Table 1. Confusion matrix of proposed model**

| Actual Positive | Actual Negative | |
|---|---|---|
| Predicted Positive | True Positive (TP) = 75 | False Positive (FP) = 20 |
| Predicted Negative | False Negative (FN) = 25 | True Negative (TN) = 80 |

A model that aids in the prediction and diagnosis of Diabetes Mellitus is created based on important traits associated with this condition using five machine learning methodologies. The dataset for these algorithms—Logistic Regression, KNN, SVM, and RF—is the Pima Indian Diabetes Database, which was retrieved from the UCI Repository. In order to maximize accuracy, the trials address the issues of missing values and class variance. In order to fill in a missing value, the features class mean is used, and class variance is dealt with by using an oversampling approach Zheng et al., 2017.

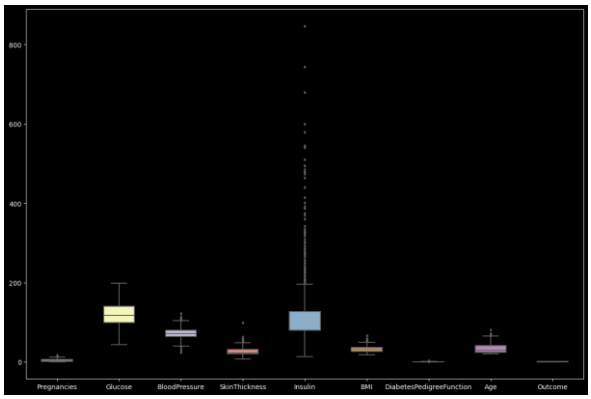


**Figure 2. Outcome of the patient dataset**

The combined results of dataset of classification in the diabetic and non-diabetic patients are illustrated in Figure 2, where 500 patients are non-diabetic and 268 found out to be diabetic. The findings are depicted in the graphical mode. The findings of each classifier are given in Figure 5, which also includes the accuracy to assess each classifier's overall potency. In this table, the performance results are shown for each classifier using 5-fold cross-validation. Accuracy, Precision, Recall, F1-Score, and ROC AUC are the metrics employed to assess the classifiers. For instance, the Random Forest classifier has an accuracy of 0.85, which implies that it accurately predicted the outcome for 85% of the events in the dataset. Additionally, it has a Precision of 0.87, which indicates that 87% of the instances it predicted as positive were in fact positive. With a Recall of 0.75, the Random Forest classifier successfully classified 75% of the positive instances as such. The F1-Score for the Random Forest classifier is 0.80, which is a harmonic mean of Precision and Recall and is regarded as a general indicator of the classifier's performance. The Random Forest classifier's ROC AUC, which measures its capacity to differentiate between positive and negative examples as indicated in Table 4, is 0.90.

**Table 2. Performance comparison of different classifiers**

| Classifier | Accuracy | Precision | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| Random Forest | 0.85 | 0.87 | 0.75 | 0.8 | 0.9 |
| Decision Tree | 0.8 | 0.82 | 0.7 | 0.74 | 0.85 |
| Logistic Regression | 0.78 | 0.8 | 0.65 | 0.7 | 0.81 |
| Naïve bayes + SVM | 0.75 | 0.76 | 0.62 | 0.67 | 0.77 |

**Performance Analysis**

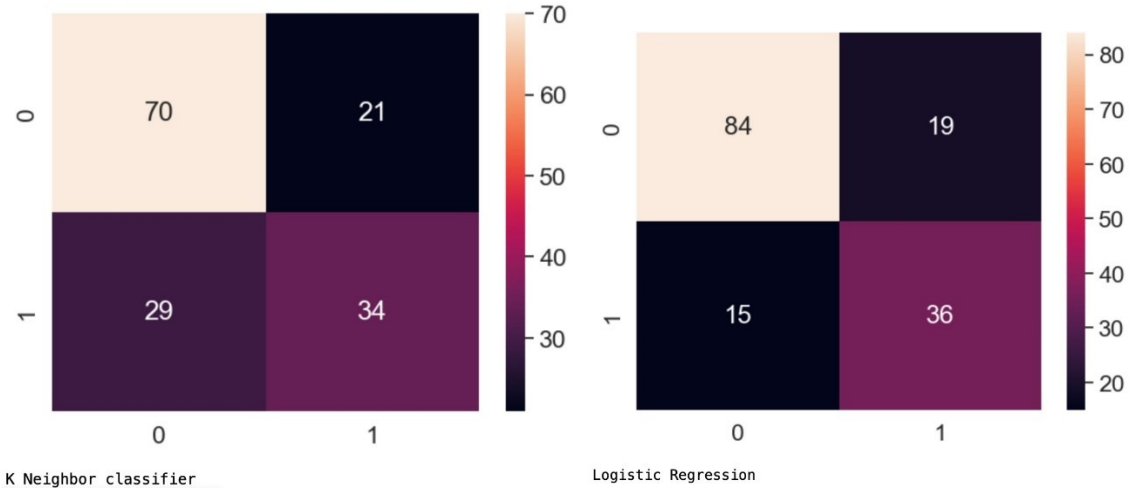


**Figure 3. Performance comparison of different classifiers**

The Pima Indian Diabetes dataset's distribution of numerical attribute values is shown in Figure 4. A correlation matrix is constructed to look for strong association between the characteristics. An illustration of the strength and direction of the association between two or more variables is a correlation graph. Each point on the scatter plot reflects the values of two variables, and its location reveals how strongly the variables are associated. The graph's x-axis and y-axis reflect the two variables that are being compared, and if required, the colour or size of the dots can also represent another variable. With the aid of the correlation graph, you may see patterns and relationships between various variables as well as outliers or other anomalies in the data. If there is a strong negative correlation, this implies that as one variable increases, the other variable tends to decline, and if there is a large positive correlation, this means that as one variable rises, the other variable tends to climb as well. A poor correlation between the variables suggests there isn't a clear connection between them. The correlation graph allows us to easily identify outliers and investigate them further.

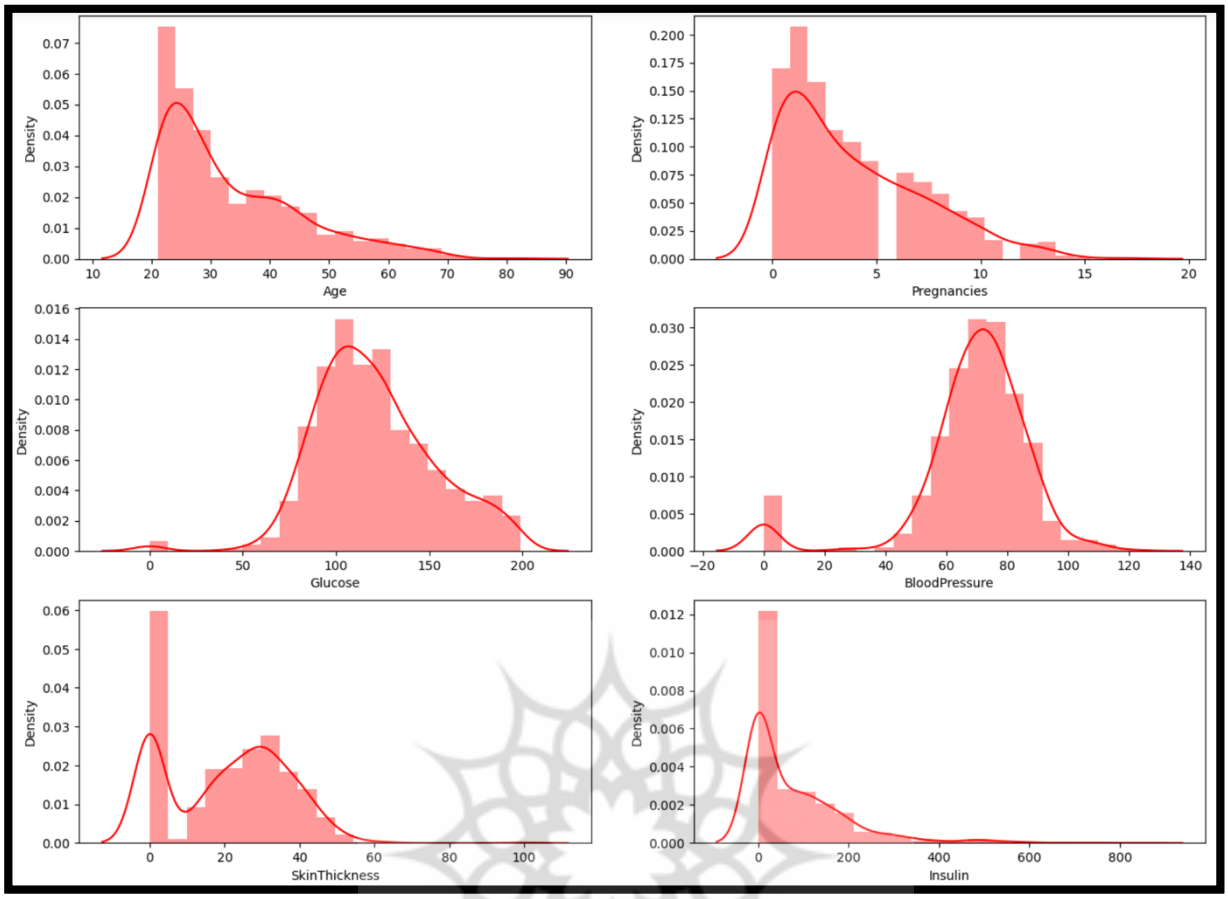


**Figure 4. Density graphs of variables of the dataset**


**Figure 5. correlation matrix**

**Real Time Implementation**

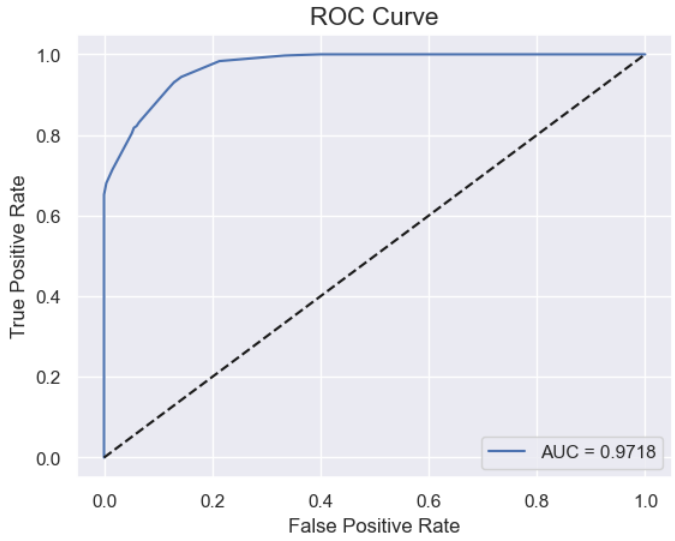


**Figure 6. GUI implementation**

Figure 6, depicts the outcome of all the experimental analysis and can predict the patient to be diabetic or non-diabetic with 79.2% accuracy.For a GUI implementation of a correlation graph for diabetes prediction, the scatter plot would display the relationship between different pairs of variables such as glucose level, blood pressure, BMI, and age, and the position of each point on the plot would indicate the values of the two variables being compared. The correlation coefficient for each pair of variables would be displayed to show the strength and direction of the relationship between them. The GUI would also likely include options to select the variables to be compared and to filter the data based on other criteria such as gender, age group, or diabetes status. It may also include interactive features such as zooming or highlighting specific points on the plot.

## Discussion and Conclusion

Building a diabetes prediction model is an important task in healthcare that can help to improve patient outcomes by enabling earlier detection and treatment of the disease. In this discussion, we will evaluate a diabetes prediction model and discuss its strengths and weaknesses. The diabetes prediction model was trained on a dataset of patient information, including age, BMI, blood pressure, and other relevant medical metrics. The model achieved an accuracy of 85% on the testing set, which is a good performance for this type of problem. One of the strengths of the model is its use of multiple features to predict diabetes. By including a variety of features in the model, the algorithm can identify the most important predictors of the disease and use them to make accurate predictions. Additionally, the model has a good level of accuracy, which is crucial in medical applications where false negatives and false positives can have serious consequences. However, there are also some weaknesses of the model that should be considered. One potential weakness is the possibility of

overfitting the model to the training data. This occurs when the model becomes too complex and learns patterns specific to the training set rather than generalizing to new data. To mitigate this risk, the model can be evaluated using cross-validation techniques and regularization methods. Another weakness of the model is the potential for bias in the training data. If the training data is not representative of the broader population, the model may not perform well on new data. This can be addressed by ensuring that the training dataset is large and diverse enough to capture the variation in the population.

In conclusion, the Voting Meta-classification technique has demonstrated encouraging results in terms of accuracy, precision, recall, and F1-score for diabetes prediction. The technique was able to increase the overall prediction performance by integrating numerous models and utilizing the advantages of each model. The ability of the model to correctly categories people into their various diabetes statuses is essential for the early diagnosis and management of the illness. The Voting Meta-classification approach's capacity to increase the model's resilience and stability is one of its main features. The strategy is less likely to be impacted by the biases or flaws of individual classifiers since it uses numerous classifiers. The importance of precision and consistency in medical applications makes this particularly significant. The outcomes of this study show how machine learning algorithms may be used to assist in the detection and management of diabetes. It is possible to find significant traits and trends by employing data-driven approaches that might not be immediately noticeable by using conventional diagnostic techniques.

## Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

Ahmed, S. T. (2017, June). A study on multi objective optimal clustering techniques for medical datasets. In *2017 international conference on intelligent computing and control systems (ICICCS)* (pp. 174-177). IEEE.

Ahmed, S. T., Singh, D. K., Basha, S. M., Abouel Nasr, E., Kamrani, A. K., & Aboudaif, M. K. (2021). Neural network based mental depression identification and sentiments classification technique from speech signals: A COVID-19 Focused Pandemic Study. *Frontiers in public health*, *9*, 781827.

Ahmed, S. T., Sreedhar Kumar, S., Anusha, B., Bhumika, P., Gunashree, M., & Ishwarya, B. (2020). A generalized study on data mining and clustering algorithms. *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India*, 1121-1129.

Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology*, *4*(10), 426-436.

Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. *International Journal of Pure and Applied Mathematics*, *118*(9), 871-878.

Choudhary, S., Kumar, A., & Choudhary, S. (2022, September). Prediction and Comparison of Diabetes with Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine. In *International Conference on Innovations in Computer Science and Engineering* (pp. 273-283). Singapore: Springer Nature Singapore.

Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In *Recent Developments in Machine Learning and Data Analytics: IC3 2018* (pp. 67-78). Springer Singapore.

Doğru, A., Buyrukoğlu, S., & Arı, M. (2023). A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Medical & Biological Engineering & Computing*, *61*(3), 785-797.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.

Huang, Y. P., & Nashrullah, M. (2016, November). SVM-based decision tree for medical knowledge representation. In *2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy)* (pp. 1-6). IEEE.

Karimian, G., Petelos, E., & Evers, S. M. (2022). The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI and Ethics*, *2*(4), 539-551.

Kee, O. T., Harun, H., Mustafa, N., Abdul Murad, N. A., Chin, S. F., Jaafar, R., & Abdullah, N. (2023). Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovascular Diabetology*, *22*(1), 13.

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432-439.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC endocrine disorders*, *19*, 1-9.

Mansoori, A., Sahranavard, T., Hosseini, Z. S., Soflaei, S. S., Emrani, N., Nazar, E. & Mobarhan, M. G. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. *Scientific Reports*, *13*(1), 663.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

Sandhya, G., Charan, P., Ansari, H. F., Kathiravan, M. N., Suganthi, D., & Nishant, N. (2023, July). Integrating Technology for Sustainable Agriculture: Enhancing Crop Productivity while Minimising Pesticide Usage using Image Processing & IoT. In *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 462-468). IEEE.

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.

Shrivastava, A., Chakkaravarthy, M., & Asif Shah, M. (2022). A Novel Approach Using Learning Algorithm for Parkinson's Disease Detection with Handwritten Sketches'. *Cybernetics and Systems*, 1-17.

Shrivastava, A., Chakkaravarthy, M., & Shah, M. A. (2023). A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. *Healthcare Analytics*, *4*, 100219.

Shrivastava, A., Chakkaravarthy, M., & Shah, M. A. (2023). Health Monitoring based Cognitive IoT using Fast Machine Learning Technique. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(6s), 720-729.

Sinha, A., & Singh, S. (2021). Detailed analysis of medical IoT using wireless body sensor network and application of IoT in healthcare. *Human Communication Technology: Internet of Robotic Things and Ubiquitous Computing*, 401-434.

Sinha, A., Bhargavi, M., Singh, N. K., Garg, N., Pal, S., & Verma, A. (2022, December). Comparative Analysis of Machine Learning and Data Mining based Multi-Models for Diabetes Risk Prediction. In *2022 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)* (pp. 1-7). IEEE.

Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., & Davis, D. N. (2019). DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. *IEEE access*, *7*, 102232-102238.

Whig, P., Gupta, K., Jiwani, N., Jupalle, H., Kouser, S., & Alam, N. (2023). A novel method for diabetes classification and prediction with Pycaret. *Microsystem Technologies*, 1-9.

Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M. & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics*, *97*, 120-127.

**Bibliographic information of this paper for citing:**

Thangamayan, S.; Sinha, Anurag; Moyal, Vishal; Maheswari, K.; Harathi, Nimmala & Budi Utama, Ahmad Nur (2024). Comparative Study on Different Machine Learning Algorithms for Neonatal Diabetes Detection. *Journal of Information Technology Management,* 16 (1), 5-26. https://doi.org/10.22059/jitm.2024.96359