Vol.64: e21210181, 2021
https://doi.org/10.1590/1678-4324-2021210181
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

Regression Imputation and Optimized Gaussian Naïve Bayes Algorithm for an Enhanced Diabetes Mellitus Prediction Model

Dhilsath Fathima Mohammed Mohideen[1]
https://orcid.org/0000-0002-4491-4352

Justin Samuel Savari Raj[2]
https://orcid.org/0000-0002-4322-3621

Raja Soosaimarian Peter Raj[3]
https://orcid.org/ 0000-0002-7216-2207

[1]Research Scholar, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu, India; [2]PSN Engineering College, Department of Computer Science and Engineering, Tirunelveli, Tamilnadu, India; [3]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Fabio Alessandro Guerra

Received: 2021.03.24; Accepted: 2021.04.14.

*Correspondence: dilsathveltech123@gmail.com (D.F.M.M.).

HIGHLIGHTS

- The aim of this study is to build a diabetes mellitus disease prediction model using machine learning algorithms.
- To build a diabetes mellitus model for diabetes early prediction, the optimized gaussian naive bayes algorithm is proposed and used as a classifier.

Abstract: Diabetes mellitus (DM) is a category of metabolic disorders caused by high blood sugar. The DM affects human metabolism, and this disease causes many complications like Heart disease, Neuropathy, Diabetic retinopathy, kidney problems, skin disorder and slow healing. It is therefore essential to predict the presence of DM using an automated diabetes diagnosis system, which can be implemented using machine learning algorithms. A variety of automated diabetes prediction systems have been proposed in previous studies. Even so, the low prediction accuracy of DM prediction systems is a major issue. This proposed work developed a diabetes mellitus prediction system to improve the diabetes mellitus prediction accuracy using Optimized Gaussian Naive Bayes algorithm. This proposed model using the Pima Indians diabetes dataset as an input to build the DM predictive model. The missing values of an input dataset are imputed using regression imputation method. The sequential backward feature elimination method is used in this proposed model for selecting the relevant risk factors of diabetes disease. The proposed machine learning classifier named Optimized Gaussian Naïve Bayes (OGNB) is applied to the selected risk factors to create an enhanced Diabetes diagnostic system which predicts Diabetes in an individual. The performance analysis of this prediction architecture shows that, over other traditional machine learning classifiers, the Optimized Gaussian Naïve Bayes achieves an 81.85% classifier accuracy. This proposed DM prediction system is effective as compared to other diabetes prediction systems found in the literature. According to our

experimental study, the OGNB based diabetes mellitus prediction system is more appropriate for DM disease prediction.

Keywords: Optimized Gaussian Naïve Bayes classifier; Regression imputation; Sequential backward feature elimination; Diabetes mellitus diagnosis.

INTRODUCTION

Diabetes mellitus (DM) is a prevalent chronic condition which causes a significant risk to human health in developing countries [1,2]. The abnormal secretion of insulin causes blood sugar levels to rise above the normal range, which is a sign of diabetes. DM is associated with the obesity rate, arteriosclerosis, hypertension, and a variety of disorders, is more prevalent in middle-aged and elderly people. DM is becoming more common in people's daily lives because of unhealthy living standards. Therefore, the reliable and successful diagnosis and testing of diabetes is a concern worth addressing. The earlier the diagnosis is provided, the better can monitor and control the consequences of the diabetes by increasing physical activity, adopting healthier lifestyle and improving the life quality. It is necessary to build a decision-making model to diagnose the diabetes mellitus for healthcare experts. Machine learning algorithms can help to a build a decision-making framework of diabetes mellitus which uses patient's physical test data to enable a conceptual decision about diabetes [3,4]. This proposed model is built on the basis of a machine learning concepts to identify the diabetes mellitus in a patient and can determine an individual as a diabetic patient or a non-diabetic patient. The Optimized Gaussian Naive Bayes Algorithm (OGNB) is a novel classification model which is used in this proposed work to predict diabetes mellitus in an individual by evaluating risk factors associated with diabetes.

Related Work

Many researchers in recent years have suggested many decision-support frameworks for diabetic prediction which are based on machine learning techniques. The standard machine learning classifier for developing a diabetes diagnosis model is logistic regression (LR) [5], decision tree (DT) [6], support vector machine (SVM) [7], naive bayes algorithm [8], K-nearest neighbor model [9], Random forest [10], etc. Different data imputation techniques are available to increase the efficiency of the proposed diabetic prediction model like simple imputation, mean imputation, conditional mean imputation [11]. Edla and coauthors [12] framed diabetes mellitus forecast system which uses the techniques called radial basis function neural network algorithm for developing a prediction model. This framework makes use of the PIMA Indian diabetes (PID) dataset. Radial basis function neural network (RBFNN) is a variant of neural network with three layers like input layer for processing input which is connected to hidden layer consist of gaussian activation function in each neuron of hidden layer. This hidden layer is connected to output layer which has two neurons. To boost the efficiency of the classifier, the Bat optimization technique is utilized to RBFNN. This model achieved high of accuracy of 73.91% accuracy than other ML techniques like Gini, Time Delay Network, Cascade Forward Network, Multi-Layer Feed Forward Neural Network, Learning Vector Quantization, Artificial Immune System, Probabilistic Neural Network. Sisodia and coauthors [6] designed a diabetic decision support system using three ML algorithms such as naïve bayes, support vector machine and decision tree. Classification performance metrics are used to assess the results of these three algorithms. The Naive Bayes classifier outperforms the SVM and decision tree algorithms at an accuracy rate of 76.30%. Zou and coauthors [13] developed a framework which diagnose the diabetes using random forest classifier, neural network, decision tree for training the model using input dataset. For validating the trained model k-fold cross validation and hold-out method are used. Minimum redundancy maximum relevance (mRMR) and principal component analysis are feature selection approaches that are used to minimize the dimensionality of a function. The random forest algorithm outperforms other diabetic prediction classifiers with an accuracy of 77.21 percent for the PID dataset.

Dwivedi and coauthors [14] compared six computational models for diabetic mellitus prediction, including SVM, classification tree, logistic regression, artificial neural network (ANN), naive bayes algorithm, and K-nearest neighbor technique. The outcome of these computational model is evaluated using classification measures like confusion matrix, recall, misclassification, precision, specificity, sensitivity. 78 and 77 percent classification accuracy is achieved by logistic regression and Artificial neural networks, respectively. Sivakumar and coauthors [15] finds answers to diabetic disease diagnosis problems by analyzing the meaningful patterns in the given input data to provide patients with early diagnosis utilizing various machine

learning algorithm like KStar, naïve bayes, oneR, ZeroR, random forest. They achieved significant accuracy of 76.3 percent and 75.7 percent using the Naive Bayes model and random forest technique, respectively, as compared to other classifiers. Kumari and coauthors [16] trained the Diabetes prediction model with a radius basis function as a kernel method in SVM and tested its output with 10-fold cross validation. According to the findings of this analysis, RBF kernel-based SVM has a training accuracy of 75.5 percent on PIMA diabetes dataset. From this related research, we concluded that machine learning models give more contribution in the prediction of diabetes disease.

Motivation and Justification of the Proposed Work

In the treatment of chronic diseases, specifically, diabetes mellitus, Machine learning techniques may be useful. Machine learning provide many conceptual techniques like data scaling, data imputation, data reduction, feature selection techniques and classification algorithm for building the diabetes diagnosis model. The focus of this research work is to build a diabetes mellitus disease prediction model that can be evaluated for accuracy and useful conclusions using the proposed Optimized Gaussian Naïve Bayes algorithm.

Contributions

The following three contributions are considered by this proposed work: (i) This model uses regression data imputation for imputing the missing value of an input sample, (ii) sequential backward feature elimination method is used for selecting the relevant features of an input dataset, and (iii) For building a diabetes mellitus model for early diagnosis, the optimized gaussian naive bayes algorithm is proposed and used as a classifier.

PROPOSED METHODOLOGY

The purpose of this proposed method is to build a machine learning framework that can accurately predict diabetic mellitus disease in its early stages, helping health care professionals to better monitor and treat diabetic mellitus disease in its earliest stages with better certainty.

The proposed diabetes mellitus diagnosis model includes a data preprocessing system, sequential backward feature elimination for selecting risk factors from the input dataset, and a novel optimized gaussian naive bayes algorithm (OGNB) for training and validating a classification model. The proposed OGNB classifier is trained and validated using the PIMA Indian diabetes dataset. Figure 1 depicts the process flow of the proposed system.

Dataset Description

The proposed model is trained and validated using the publicly available PIMA Indian Diabetes (PID) dataset [17]. This dataset provides information on 768 individual people (768 samples), 8 input features of Diabetes mellitus and one output class with binary values such as 1 (Diabetes mellitus - Present) and 0 (Diabetes mellitus - Absent). Diabetes mellitus-Positive means patients have a problem of Diabetes, and Diabetes mellitus -Negative means that patients do not have Diabetes disease. The data types of input features are numeric.

Data Preprocessing System

The techniques of data preprocessing assist in obtaining a clean and smooth dataset which enhances the consistency of collected patterns [18,19]. This proposed model uses a data preprocessing system to apply the data preprocessing techniques in the PID dataset. The components of preprocessing system are statistical summary analysis, regression imputation, Min-max normalization.

Statistical Summary Analysis

The statistical summary analysis is the method used for analyze the characteristics of an input dataset [20]. This provides a simple overview of each feature's missing data and a range of values. Using statistical summary analysis, we looked carefully at the features of the PID dataset and showed the result in Table.3, this table indicated the presence of missing values for the input features such as Blood pressure, Glucose, Insulin, BMI, Skin Thickness as the minimum value is zero. Based on clinical intuition, these features should not be zero. We assume, therefore, that certain input values for these features are missing. To identify the missing values of a PID dataset, statistical summary analysis is used.

Regression Imputation Technique

Missing values in a dataset will minimize the results of the training and prediction score of the proposed classifier and may provide skewed results, leading to false hypotheses [21]. Using the data imputation method, missing values of an input attribute can be filled with an estimated value. The proposed diabetes prediction model uses regression imputation strategy [22] for imputing the missing values of the input dataset. The value of other input features is used in imputation technique to estimate missing values in a feature using a regression model. A regression model assumes an input feature with missing values as a dependent attribute y and the remaining features as an independent attribute $x=(x_1,x_2,\ldots,x_n)$. Calculate the line of best fit using the regression equation as in Equation (1).

$$y_i=w_i x_i+w_0, \tag{1}$$

Where y_i is a dependent feature with missing value, x_i is an independent attribute, w_i is a gradient, w_0 is the bias. The method of least square is applied to find the values of y-intercept and gradient. Compute the gradient using the Equation (2).

$$w_i=\frac{N\sum(x_i y_i)-\sum x_i \sum y_i}{N\sum(x_i^2)-(\sum x_i)^2}, \tag{2}$$

Where N refers to the total number of samples of the input dataset. Find a bias value w_0 using Equation (3).

$$w_0=\frac{\sum y_i-w_i\sum x_i}{N}, \tag{3}$$

Substitute the values of w_i, w_0, x_i in the Equation.1 to impute the missing value of a dependent features.

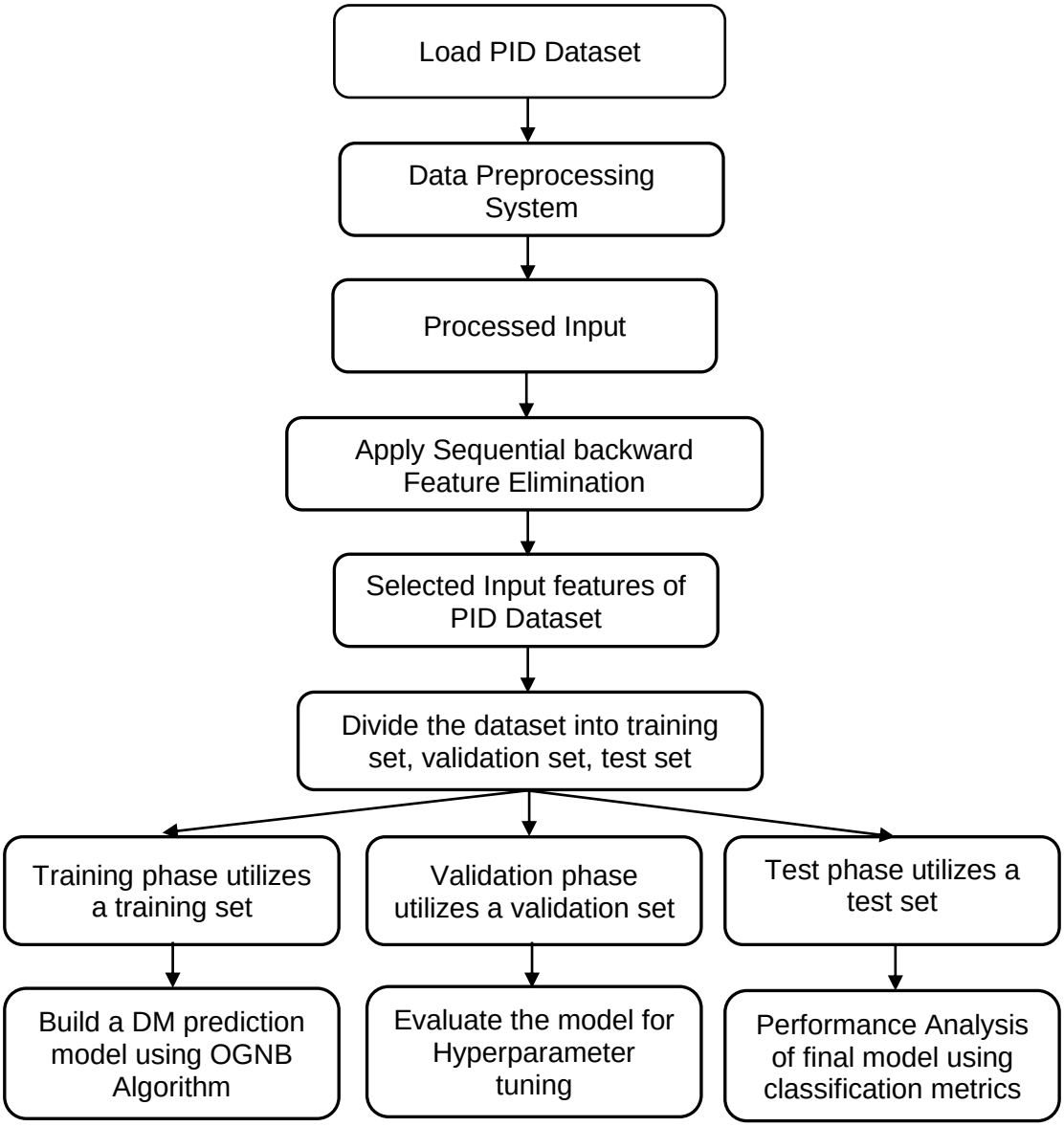

Figure 1. Outline of the proposed framework.

Min-Max Normalization

Data scaling is a method of data preprocessing that allows the features of an input dataset to be standardized. As a result of standardization, the input values fall within a specified range of values, like [0,1] or [-1,1]. It essentially allows the data to be organized within a given limit. Mostly, it helps accelerate the computation of a classifier. We analyzed the statistical summary analysis that indicates that there is a wide range of data values for several input features. Therefore, to standardize the dataset, data scaling is necessary. For data scaling, this proposed model incorporates Min-Max normalization, which preserves the relationships between the original data values of an input dataset [18]. In the Equation (4), the mathematical model of min-max normalization is given.

$$x_{new} = \frac{x_i - oldmin_x}{oldmax_x - oldmin_x} \left(New_{max_x} - New_{min_x}\right) + New_{min_x}, \quad (4)$$

Have X is an input feature with N data values such as x_1, x_2, … x_N. This approach transforms the input data (x_i) linearly and preserves the relationship between the original data values by changing the minimum value ($oldmin_x$) and maximum value ($oldmax_x$) of feature x_i to x_{new} in the value range of New_{max_x} and New_{min_x}.

Sequential backward feature elimination

The proposed model uses a sequential backward feature elimination (SBFE) [23,24] for selecting the significant features of PID dataset for building the classification model using the proposed OGNB classifier. SBFE is used to reduce classification modeling computation time and increase model efficiency by eliminating irrelevant features from the input data collection. In each iteration, SBFE begins with the complete input dataset (D) and eliminate irrelevant features using the objective function's selection mechanism. SBFE has an objective function called the Gini coefficient [18] which decides the best feature of D.

Steps of Sequential backward feature elimination

Step 1→ Start the feature selection process with full feature set D. D consist of N training samples which is denoted as $D=\left((x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\right)$ where $x_i \in X$ is a total input features and $y_i \in Y$ is a class label of the corresponding input features.
Step 2→ Construct a decision tree of D with Gini index measure for eliminating the irrelevant features of D. Gini index or gini impurity value is computed for all input features and eliminate the feature with high gini index (Gini impurity) value from the D during the first iteration. The mathematical derivation of Gini impurity of an input feature x is given in the following Equation (5).

$$\text{Gini index}(x)=\text{Gini}(D)-\text{Gini}_x(D), \tag{5}$$

Where Gini index(x) is the total gini impurity of an input feature x of D, Gini(D) is the gini index of input dataset D which is calculated as in equation (6) and the $\text{Gini}_x(D)$ is the gini index of an input feature x.

$$\text{Gini}(D)=1-\sum_{m=1}^{y} p_m^2, \tag{6}$$

Where p_m is the probability of class m in x to be determined as $p_m=\frac{|c_m,M|}{|M|}$ where |M| indicates number of data samples of a feature x and $|c_m,M|$ means the count of class label in m. If m on feature x is divided into 2 sets like m_1 and m_2 , $\text{Gini}_x(D)$ is specified in Equation (7).

$$\text{Gini}_x(D)=\frac{|m_1|}{|m|}\ \text{Gini}(m_1)+\frac{|m_2|}{|m|}\ \text{Gini}(m_2), \tag{7}$$

Step 3→ Eliminate features with a high Gini index(x) in the first iteration, based on the Gini impurity value.
Step 4→ Iterate the step from 2 to 3 until the number of features to be selected is reached.

Classification Modelling using optimized gaussian Naïve Bayes Algorithm

Classification modelling is the next step of feature selection phase in the proposed system, this phase takes the selected input features of PID dataset which are selected by the sequential backward feature elimination technique. This input dataset is divided into three sets like 80% of an input dataset is called training data, 10% of input data is named validation data, and remaining 10% data is called test data. The Diabetes mellitus predictive model is built using an Optimized Gaussian naïve bayes algorithm (OGNB). The optimized Gaussian naïve bayes algorithm is developed by utilizing Adaboost algorithm for boosting the incorrect prediction of Gaussian naïve bayes classifier and random search optimizer for increasing the performance of OGNB classifier by tuning the hyperparameters of an Adaboost and Gaussian naïve bayes algorithm to make an ensemble algorithm [25, 26, 27]. The OGNB classifier combines Gaussian naive bayes, Adaboost, and the random search method to create an efficient classifier that maximizes prediction score while minimizing overfitting issues. The overall diagrammatic representation of the OGNB classifier is represented in the Figure 2.

Training phase of optimized gaussian Naïve Bayes algorithm

The training process uses the training data to build a diabetes prediction model using the OGNB algorithm. Optimized Gaussian naive bayes (OGNB) based on the following norms to construct a machine learning model: the significance of a particular input feature is independent of the relevance of all other input features. The Adaboost algorithm is used in the OGNB classifier to combine the outputs of many Gaussian naïve bayes (GNB) classifiers to create a robust gaussian naive bayes model.

Validation phase of optimized gaussian Naïve Bayes algorithm

The proposed model uses a validation dataset D_{valid} to tune the OGNB classifier's hyper parameters to significantly improve classification modelling efficiency by reducing the misclassification using the random search optimizer which finds the optimal values of hyper parameters from the hyperparameter search grid. The use of D_{valid} to evaluate a trained model provides an unbiased prediction about what effectiveness would have been when the model is applied in real-world predictions. Hyperparameters are pivotal in classification modeling for regulating the efficiency of training algorithms and also have a significant effect on the development of machine learning models [28]. The hyper-parameter tuning of the proposed OGNB classifier helps to further increase the accuracy of the model.

In this step, the trained model's output is assessed using D_{valid}, after which the OGNB classifier hyperparameter is tuned using the random search optimizer, and the model is retrained to improve the training and validation scores. Random search is much more reliable for solving real-world optimization problems in a high-dimensional space. In this proposed diabetes prediction model, a random search approach has been used to optimize the OGNB classifier's accuracy, as shown in Equation (8).

$$OGNB_{validated}=\text{argmax}_X\, f(OGNB,Hyp,D_{train},D_{valid}), \tag{8}$$

Where $OGNB_{validated}$ yields a range of optimized hyperparameters that boost the performance of OGNB algorithm, Hyp specifies the hyper-parameter search space, D_{train} denotes the training dataset, D_{valid} is a validation dataset, $\text{argmax}_X\, f$ is random search optimization function on OGNB, Hyp, D_{train}, D_{valid} to improve the training and validation phase accuracy score.

Test phase of optimized gaussian Naïve Bayes algorithm

This phase tests the proposed diabetes prediction model's final output using different classification performance using 10% of the input dataset named D_{test}.

Highly detailed steps of the proposed optimized gaussian Naïve Bayes algorithm

Step 1: Gaussian naïve bayes algorithm is used as a base learner in the Adaboost algorithm to build a predictive model using the training dataset D_{train}. D_{train} consist of weighted N training samples $D=\left((x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\right)$ where $x_i \in X$, $y_i \in Y$ is an associated class labels of an input instance x_i. This method makes use of binary classification, whereas the response vector has two possible values of 0 and 1. As in equation (9), this can be denoted

$$y=\begin{cases}0,\text{if a diabetes mellitus absent}\\1,\text{if a diabetes mellitus present}\end{cases}, \tag{9}$$

Step 2: Bootstrap sampling is applied to the D_{train} to divide it into the d subset where $d<D_{train}$. For a number of iterations (s_i), the Adaboost algorithm iteratively calls a Gaussian naïve bayes model. Each input instance x_i of D_{train} is weighted in all iterations, denoting the likelihood of the samples being selected for an input classifier. As in the equation (10), the initial weight w_i is assigned to each training sample x_i.

$$w_i = {}^{1}/_{N} \text{ for all } x_i;\ i=1,2,\ldots.N, \tag{10}$$

Where N represents the overall training instances of D_{train}.

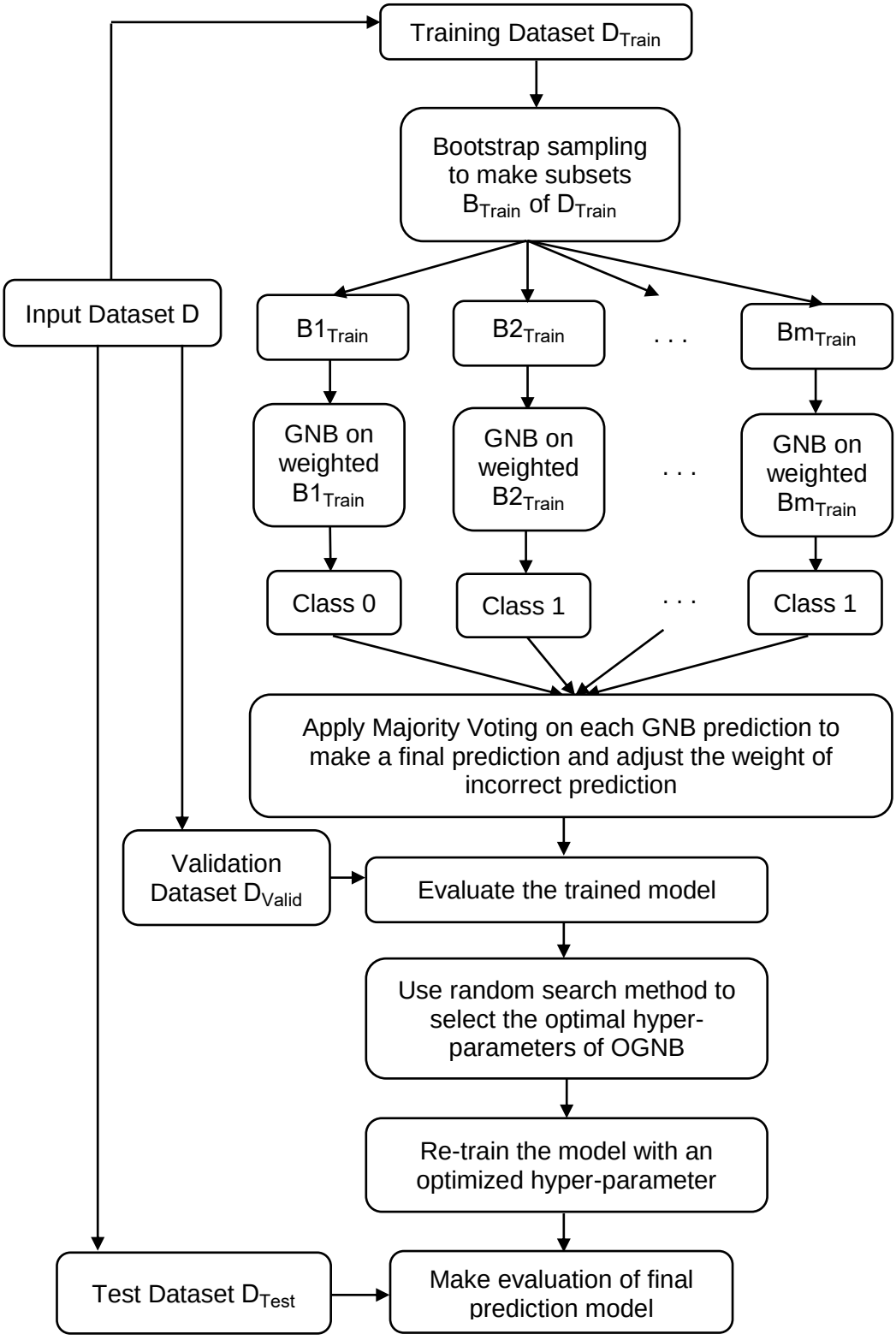

Figure 2. Graphical representation of an Optimized Gaussian naive bayes classifier.

Step 3: Apply the Gaussian Naïve Bayes (GNB) algorithm across all subset d. GNB is a type of Gaussian distribution function with a standard deviation σ and mean μ which is used to calculate the Gaussian probability density function (GaussianPDF) for measuring the probability of an input samples $P(x_i \mid y)$ utilizing equation (11) and equation (12).

$$\text{GaussianPDF}(x,\sigma,\mu)=\frac{1}{\sigma\sqrt{2\pi}}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (11)$$

So that class probability of given sample x_i is calculated using GaussianPDF function.

$$P(x_i \mid y)=\text{GaussianPDF}\left(x_i, \mu_{classi},\sigma_{classi}\right), \tag{12}$$

Where the Gaussian probability density function of a sample x is the GaussianPDF(x, σ, μ), π is called the mathematical constant, μ and σ represent the mean and standard deviation of an input features, the mathematical constant is exp. The standard deviation value of an input samples (σ_{y_i}) for each target class can be computed using the Equation (13).

$$\sigma_{y_i}=\sqrt{\frac{\sum(x_i-\mu)^2}{N}}, \tag{13}$$

Where N is the overall training instances of sample of D_{train} and x_i is the input samples of D_{train}. The mean value of an input samples(μ_{yi}) for each target class can be determined using the Equation (14).

$$\mu_{yi}=\frac{1}{N}\sum_{i=1}^{N} x_i, \tag{14}$$

Where N is the overall training instances of sample of D_{train} and x_i is the input samples of D_{train}.

Step 4: All training samples in the D_{train} would be given same weight w during the first iteration of the training process. An efficiency of the trained model is computed during the first iteration by using test samples D_{test}. The sample of the D_{test} is independently predicted by an individual GNB classifier.

Step 5: Construct a final hypothesis h_{final}:X→Y on the D_{test} by implementing the majority voting function. Use the equation (15) for computing the misclassification rate $error_{test}$ of h_{final}.

$$error_{test}=\frac{\text{sum}\left(w_i(h_{final}(x_i)\neq y_i)\right)}{\text{sum}(w_i)}, \tag{15}$$

During the second iteration of the training step, the weight of the misclassified samples is updated to motivate the incorrect samples in the training set and the process will continue on each iteration s_i of the training phase. Determine and adjust incorrect samples' weight as in Equation (16).

$$w_{i+1}(i)=\frac{w_i(i)^* \exp\left(-a_w y_i hyp_{gnb}(x_i)\right)}{z_w}, \tag{16}$$

Where a_w is the factor that used to prevent overfitting, which increase an algorithm generalization and normalization constant is z_w. The value of a_w is calculated using an Equation (17).

$$a_w=\frac{1}{2}\ln\left(\frac{1-e_{test}}{e_{test}}\right), \tag{17}$$

Step 6: Calculate the misclassification of the trained model $train_{error}$ using validation dataset D_{valid} for further tuning the OGNB classifier hyper-parameters which optimize the model's performance.

Step 7: Create a hyper-parameter search space after the validation phase and choose the optimal OGNB classifier hyper-parameter using the random search optimization method.

Step 8: Re-train the OGNB classifier with an optimized hyperparameter for predefined iterations to determine a validation error $valid_{error}$. Stop the re-training process of the proposed model when $valid_{error} > train_{error}$.

Step 9: Validate the final model h_{final} with test dataset D_{test} and interpret the outcomes using classification performance measures.

EXPERIMENTAL RESULTS

PID Dataset

The Diabetes mellitus prediction model uses PID dataset for training and validating OGNB classifier. Table 1 describes the risk factors (features) of Diabetes disease which are used in the PID dataset.

Table 1. Features of PIMA diabetes dataset.

S.No	Input features	Range of Values	Feature Description
1	Pregnant (Preg)	From 0 to 5	Number of Pregnancy
2	Glucose (Gluc)	From 0 to 199	Two-hour plasma glucose concentration
3	Blood Pressure (BP)	From 0 to 122	Diastolic BP
4	Skin Thick (ST)	From 0 to 99	Skin fold Thickness
5	Insulin (Insulin)	From 0 to 846	Two-hours serum insulin
6	BMI	From 0 to 67	Body mass index
7	Diabetes pedigree function (DPF)	From 0.078 to 2.42	Diabetes pedigree function
8	Age	From 21 to 81	An individual's age
9	Output feature	Either 0 or 1	1-Diabetes mellitus–Present 0-Diabetes mellitus–Absent

Table 2 provides the characteristics of the PID dataset. The PID dataset's input sample count, and also the number of positive and negative samples, are shown in this table.

Table 2. Characteristics of PIMA Indian Diabetes Dataset.

Dataset Name	Count of Input Features	Count of labels in output Class	Count of test samples	Count of positive samples	Count of negative samples	Count of missing values
PID Dataset	8	2	768	268	500	652

Performance measurement of the Diabetes mellitus Diagnosis system

This novel Diabetes mellitus diagnosis framework uses a number of classification performance metrics. Table 3 outlines the emphasis of the evaluation metrics. Almost all the evaluation criteria for the proposed work are based on a Confusion matrix that assesses the classifier performance via four components named True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP are correctly labeled positive samples, FP are falsely labeled negative samples, TN are the correctly labeled negative samples, and FN are falsely labeled positive samples. The components of the confusion matrix (cm) are given in Equation (18) below.

$$cm = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}, \tag{18}$$

Table 3. Evaluation metrics for the proposed diabetes prediction model

Metrics	Code	Formula	Evaluation Focus
Classifier Accuracy	Acc	$Acc=\frac{\text{Correctly predicted samples}}{\text{Total samples in the training set}}$	Score of Correct Predictions
Misclassification rate	MCR	$MCR=\frac{\text{Incorrect predictions}}{\text{Total samples in the training set}}$	Calculating the Misclassification Error
Sensitivity	Sen	$Sen=\frac{TP}{TP+FN}$	Estimation of how many samples is correctly identified as positive compare to how many samples are positive.
Specificity	Spe	$Spe=\frac{TN}{TN+FP}$	The percentage of samples correctly labeled as negative compared to the total negative samples.
Precision	Pre	$Pre=\frac{TP}{TP+FP}$	Positive samples given by the classifier out of all positive samples in the training set
F1-Score	F1	$F1=2*\frac{Precision*Recall}{Precision+Recall}$ $Recall=\frac{TP}{TP+FN}$	The weighted harmonic means of recall and precision
Matthew's correlation coefficient	MCC	$MCC=\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$	Combination of the Confusion matrix components
Receiver operator characteristic curve	ROC	$\text{True Positive Rate}=\frac{TP}{TP+FN}$ $\text{False Positive Rate}=\frac{FP}{FP+TN}$	ROC focuses on TPR and FPR

Experimental Result of Statistical Summary Analysis

Table 4 shows the statistical summary analysis for the input dataset. This statistical overview is used to understand that five features consist of missing values, such as blood pressure, glucose, insulin, skin thickness, BMI, out of eight input features.

Table 4. Statistical summary Analysis of an input dataset.

Statistical Property	Input Features of PID Dataset							
	Preg	Gluc	BP	ST	Insulin	BMI	DPF	Age
Sample count	768	768	768	768	768	768	768	768
Standard Deviation	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76
Mean	3.84	120.9	69.10	20.53	79.79	31.99	0.47	33.24
Maximum	17.00	199.0	122.0	99.00	846.0	67.10	2.42	81.00
Minimum	0.00	0.00	0.00	0.0	0.0	0.00	0.078	21.00

Experimental Result of Regression Imputation model

In the PID dataset, some of the input features have missing values. The missing value information of the PID dataset are shown in Table 5. To prevent inaccurate prediction and increase the accuracy of the proposed OGNB classifier, this proposed Diabetes mellitus prediction model uses a regression imputation method for imputing these missing values of PID dataset. Table 6 compares the output of the OGNB classifier with that of other conventional classifiers before using the regression imputation technique, and Table 7 illustrates the output comparison of the proposed OGNB classifier with other typical ML classifiers on PID input dataset after the regression imputation method has been used.

Table 5. Missing values in the PIMA Indian Diabetes dataset.

Predictor Attribute	Count of missing value
Gluc	5
BP	35
ST	227
Insulin	374
BMI	11
Total missing value count is	652

Table 6. Output comparison of the Optimized Gaussian Naïve Bayes classifier on PIMA Indian Diabetes dataset with other typical ML algorithms before using the regression imputation technique.

ML Models	Acc	MCR	Sen	Spe	Prec	F1	MCC	ROC
LR	76.76	23.24	67.23	80.12	60.23	63.45	46.78	73.45
NB	74.23	25.77	71.69	75.45	56.14	63.47	44.89	73.08
KNN	67.15	32.85	67.52	67.18	48.97	56.14	32.16	67.71
SVM	73.52	26.75	63.48	77.45	55.19	59.45	39.56	70.57
Decision Tree	67.32	32.68	64.25	68.13	47.02	54.13	30.06	66.26
Proposed OGNB	77.13	22.87	73.48	78.67	60.45	66.19	49.82	75.31

Table 7. Output comparison of the Optimized Gaussian Naïve Bayes classifier on PIMA Indian Diabetes dataset with other typical ML algorithms after applying the regression imputation technique.

ML Models	Acc	MCR	Sen	Spe	Prec	F1	MCC	ROC
LR	77.85	22.15	57.23	85.35	64.17	60.89	44.26	71.09
NB	75.56	24.44	69.45	78.45	59.46	63.19	45.67	73.46
KNN	73.42	26.58	61.10	78.14	56.18	59.18	39.46	70.34
SVM	74.02	25.98	60.43	81.45	58.49	59.07	41.27	70.93
Decision Tree	70.45	29.55	60.41	75.69	51.34	55.37	34.71	67.34
Proposed OGNB	78.74	21.26	67.87	83.09	64.89	65.17	49.37	75.48

Experimental Result of Sequential backward feature elimination

The sequential backward feature elimination technique (SBFE) being used to assess important risk factors in the input dataset. The performance of the SBFE is compared to other typical methods of feature selection, such as Analysis of variance (ANOVA) [29], Chi-square [30], Mutual Information [31], and Sequential Forward Feature Selection [24], with results shown in Table 8.

Table 8. The output of an Optimized Gaussian Naive Bayes classifier with various feature selection methods

Feature Selection Method	Acc	MCR
ANOVA	79.82	21.18
Chi-square	77.10	22.9
Mutual Information	78.56	21.44
SFFS	80.23	19.77
SBFE	81.85	18.15

Table 8 demonstrated that the sequential backward feature elimination technique performs better for the proposed OGNB classifier than other ML feature selection algorithms.

Experimental Result of OGNB classifier and performance analysis

A diabetes prediction model is developed using the proposed Optimized Gaussian Naive Bayes classifier and a training dataset. During the validation process, the trained model's output is further enhanced by using the random search method. The random search method tweaks the OGNB algorithm's hyperparameters to improve the classification model's predictability. The test dataset is used to prove the evaluation of a final tuned model when compared to other ML models like K-Nearest neighbor (KNN), Logistic regression (LR), Naïve Bayes (NB), Decision tree (DT), and Support vector machine (SVM). The hyperparameter range of the OGNB classifier is defined in Table 9.

Table 9. Optimized Gaussian Naïve Bayes classifier's hyperparameter configuration space.

Proposed Classifier	Hyperparameter	Description of Hyperparameter	Hyperparameter configuration space	Selected Hyper parameter
Optimized Gaussian Naïve Bayes classifier	Number of weak learners (GNB)	The number of GNB classifiers for train the Adaboost algorithm	[10, 50, 100, 500]	50
	Learning rate	Learning rate of the Adaboost algorithm	[0.0001, 0.001, 0.01, 0.1, 1.0]	0.1
	Randomness	Random state	[50,30,40]	50

Table 10 compares the efficiency of the suggested OGNB classifier to that of other ML classification algorithms after using the sequential backward feature elimination process.

Table 10. Output comparison of the Optimized Gaussian Naïve Bayes classifier on PIMA Indian Diabetes dataset with other ML algorithms after utilizing the sequential backward feature elimination method

ML Models	Acc	MCR	Sen	Spe	Prec	F1	MCC	ROC
LR	79.12	20.88	69.14	84.36	66.23	67.37	52.37	76.04
NB	77.04	22.96	66.47	82.17	62.47	64.17	47.17	74.19
KNN	74.27	25.73	64.19	78.49	56.09	60.99	41.76	71.46
SVM	78.67	21.33	56.78	88.66	67.47	61.40	47.75	72.07
Decision Tree	72.37	27.37	66.47	75.84	54.19	59.04	39.46	70.17
Proposed OGNB	81.85	18.15	81.17	89.47	81.46	72.47	59.47	78.49

The output of the proposed diabetes prediction model is compared to other diabetes prediction models' output in Table 11. This analysis (Table 11) is intended to demonstrate how the proposed classifier, the Optimized Gaussian Naive Bayes classifier, outperforms the existing studies in terms of prediction accuracy.

Table 11. The proposed diabetes prediction model's output is compared to the output of other diabetes prediction models.

Author(s)	Year of publication	Classifier	Highest Accuracy (in %)
Edla et al. [12]	2017	RBFNN	73.91
Deepti Sisodia et al. [6]	2018	Naïve Bayes	76.03
Quan Zou et al. [13]	2018	Random Forest	77.21
Dwivedi et al. [14]	2018	Logistic Regression	78
Sarwar et al. [32]	2018	SVM and KNN	77
Faruque et al. [33]	2019	C4.5 decision tree	73.5
Vigneswari et al. [34]	2019	Logistic model tree	79.31
Sivakumar et al. [15]	2020	Naïve Bayes	76.03
Pradhan et al. [35]	2020	ANN	85.09
Tigga et al. [36]	2020	Random Forest	75
Proposed Work	-	OGNB	81.85

DISCUSSIONS

The aim of this study has been to show that using the PID dataset, the proposed regression imputation + SBFE + OGNB model would accurately predict diabetes mellitus disease. The proposed model demonstrated the classification capabilities of the OGNB classifier with many machines learning classifiers,

including Logistic regression, K-nearest neighbor, naive Bayes, Support vector machine, and Decision tree classifiers, using the PIMA Indian Diabetes dataset, which contains 768 data samples.

To increase the efficiency of the proposed diabetes model, missing values from the PID dataset are imputed using regression imputation. The impact of regression imputation is shown in Tables 6 and 7. The proposed OGNB classifier has an output accuracy of 77.13% before regression imputation, which is increased to 78.74% after regression imputation. As a result, regression imputation reduces ML classifier misclassification and improves classifier accuracy.

It is evident from Table 8 that the sequential backward feature elimination technique is suitable for selecting an optimal feature of PID dataset to obtain good predictive performance over other feature selection techniques. The proposed classifier has an output accuracy of 78.74 percent before using SBFE, which is increased to 81.85 percent by using a SBFE method to obtain the input dataset's optimal features. As a result, SBFE boosts the accuracy of the proposed classifier by 3%. This analysis shows that the sequential backward feature elimination outperforms other feature selection methods in aspects of choosing the best features from the input dataset.

Table 10 shows that, in terms of different performance measures, the proposed OGNB classifier gives good efficiency in the training set and validation set than other ML classifiers for the PIMA Indian Diabetes dataset. The OGNB classifier's accuracy score on the PID dataset is 81.85%, is higher than other comparable ML models. Table 11 shows that the proposed prediction model (Optimized Gaussian Naïve Bayes classifier for Diabetes mellitus prediction) outperformed most of the existing literature in terms of increasing prediction accuracy (excluding Pradhan and coauthors [35] using ANN, which is a category of deep neural network). However, in this proposed model, the efficiency of the OGNB classifier is compared to that of other machine learning classifiers). As a consequence, the proposed prediction model can be applied to a variety of ML classification tasks.

These conclusions are important in identifying that positive and negative samples can be accurately classified by the proposed OGNB classifier. On the basis of a comparative analysis, it is confirmed that the present approach (Regression imputation + SBFE + OGNB) has effectively performed over the other classifiers as well as it is the best model for classifying new diabetes mellitus disease data samples. The proposed OGNB classifier achieves 81.85% classifier accuracy score on the proposed PID dataset attained by this suggested approach. The proposed OGNB classifier significantly outperforms most of the existing literature on the PIMA Indian diabetes dataset. This suggested model could be implemented on the automatic diabetic diagnosis system, although the accuracy level needs to be improved using regularization method by extending this proposed methodology.

CONCLUSION

The focus of this suggested work is to build a prediction system that uses the proposed OGNB classifier for early diagnosis of the diabetes in people. This study applied the regression imputation technique for the prediction of missing values of an input sample. In this analysis, a relevant risk factors of diabetes mellitus were identified using sequential backward feature elimination (SBFE) method. SBFE's output shows that identifying the most relevant features is beneficial because it reduces the number of irrelevant features while also increasing the classification results. In this prediction model, a novel algorithm called OGNB is implemented to incorporate effectiveness of the proposed model. The result of the proposed system indicates that the proposed OGNB with regression imputation + SBFE provides better outcomes on PID dataset than other conventional ML models.

REFERENCES

1. Misra A, Gopalan H, Jayawardena R, Hills AP, Soares M, Reza-Albarrán AA, et al. Diabetes in developing countries. J Diabetes. 2019;11(7):522–39.
2. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019;6(1).
3. Chaki J, Thillai Ganesh S, Cidham SK, Ananda Theertan S. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. J King Saud Univ - Comput Inf Sci. 2020;
4. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017; 15:104–16.
5. Tabaei BP, Herman WH. A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care. 2002;25(11):1999–2003.

6. Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Comput Sci. 2018; 132:1578–85.
7. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10(1):16.
8. Nai-arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci. 2015; 69:132–42.
9. Saxena K, Khan Z, Singh S. Diagnosis of diabetes mellitus using K nearest neighbor algorithm. Ijcstjournal. 2014 Jul;2(4):36-43.
10. VijiyaKumar K, Lavanya B, Nirmala I, Caroline SS. Random Forest Algorithm for the Prediction of Diabetes. In: Proceedings of 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN 2019); India: IEEE 2019 Mar 29; pp. 1-5.
11. Masconi KL, Matsha TE, Erasmus RT, Kengne AP. Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa. PloS one 2015 Sep 25;10(9): e0139210.
12. Damodar Reddy E, Ramalingaswamy C. Diabetes-finder: A bat optimized classification system for type-2 diabetes. Procedia Comput Sci. 2017; 115:235–42.
13. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018; 9:515.
14. Dwivedi AK. Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Comput Appl. 2018;30(12):3837–45.
15. Sivakumar S, Venkataraman S, Bwatiramba A. Classification algorithm in predicting the diabetes in early stages. J Comput Sci. 2020;16(10):1417–22.
16. Anuja Kumari V, Chitra R. Classification of diabetes disease using support vector machine. International J of Engineering Research and Applications. 2013 Mar;3(2):1797-801.
17. UCI Machine Learning. Pima Indians Diabetes Database. Available from: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Cited 2021 Jan 23]
18. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011
19. Siddiqui MK, Morales-Menendez R, Ahmad S. Application of receiver operating characteristics (ROC) on the prediction of obesity. Braz Arch Biol Technol. 2020;63.
20. Atrey K, Sharma Y, Bodhey NK, Singh BK. Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype. Braz Arch Biol Technol. 2019;62.
21. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402–6.
22. Shao J, Wang H. Sample correlation coefficients based on survey data under regression imputation. J Am Stat Assoc. 2002;97(458):544–52.
23. Guyon I. An introduction to variable and feature selection. Jmlr.org. 2003 March;1157-82.
24. Kumar V. Feature Selection: A literature Review. The smart comput rev. 2014;4(3)
25. Pérez A, Larrañaga P, Inza I. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. Int J Approx Reason. 2006;43(1):1–25.
26. Yoav Freund RES. Experiments with a new boosting algorithm. In: In Proceedings of the Thirteenth International Conference on Machine Learning. 1996.
27. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Jmlr.org. 2012 Feb 1;13(2).
28. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. J Electron Sci Technol. 2019;17(1):26–40.
29. Tabachnick, B.g. and Fidell, L.s. Experimental designs using ANOVA. Thomson/Brooks/Cole, Belmont. 2007.
30. Liu H, Setiono R. Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the International Conference on Tools with Artificial Intelligence. IEEE; 1995. p. 388–91.
31. El Akadi A, El Ouardighi A, Aboutajdine D. A Powerful Feature Selection approach based on Mutual Information. Ijcsns.org. 2008 Apr 30;8(4):116.
32. Sarwar MA, Kamal N, Hamid W, Shah MA. Prediction of diabetes using machine learning algorithms in healthcare. In: 2018 24th International Conference on Automation and Computing (ICAC). IEEE; 2018.
33. Faruque MF, Asaduzzaman, Sarker IH. Performance analysis of machine learning techniques to predict diabetes mellitus. In: 2019 International Conference on Electrical, Computer and Communication Engineering. IEEE; 2019.
34. Vigneswari D, Kumar NK, Ganesh Raj V, Gugan A, Vikash SR. Machine learning tree classifiers in predicting diabetes mellitus. In: 2019 5th International Conference on Advanced Computing & Communication Systems. IEEE; 2019. p. 84–7.

35. Pradhan N, Rani G, Dhaka VS, Poonia RC. Diabetes prediction using artificial neural network. In: Deep Learning Techniques for Biomedical and Health Informatics. 2020. p. 327–39.
36. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci. 2020; 167:706–16.