Healthcare Analytics 2 (2022) 100092

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators

Shahid Mohammad Ganie, Majid Bashir Malik *

Department of Computer Sciences BGSB University, Rajouri, J&K (UT), India

ARTICLE INFO

Keywords:
Ensemble Machine Learning
Diagnostic analytics
Exploratory data analysis
Type II diabetes mellitus
Feature engineering
K-fold cross-validation

ABSTRACT

Machine Learning (ML) is a branch of artificial intelligence that allows computers to learn without being explicitly programmed. ML has been widely used in healthcare to predict various chronic diseases. Prediction of diabetes at earlier stages is crucial for better clinical pathways to reduce the complications and delay the occurrence of diabetes. In this study, a new ensemble learning-based framework is proposed for the early predicting of Type-II diabetes mellitus using lifestyle indicators. Different ensemble learning techniques like Bagging, Boosting, and Voting are employed. Exploratory data analysis is used to improve the quality assessment of the dataset. The synthetic minority oversampling technique is used for class balancing, and the K-fold cross-validation technique is employed to validate the results. A feature engineering process is applied to calculate the contribution of lifestyle parameters. Among all the classification techniques, the bagged decision tree achieved the highest accuracy rate (99.41%), precision (99.13%), recall (95.83%), specificity (99.11%), F1-score (99.15%), misclassification rate (MCR) (0.86%), and receiver operating characteristic (ROC) curve (99.07%), respectively. The proposed framework can be used in the healthcare industry for the early prediction of diabetes. Also, it can be used for other datasets which share a commonality of data with diabetes.

1. Introduction

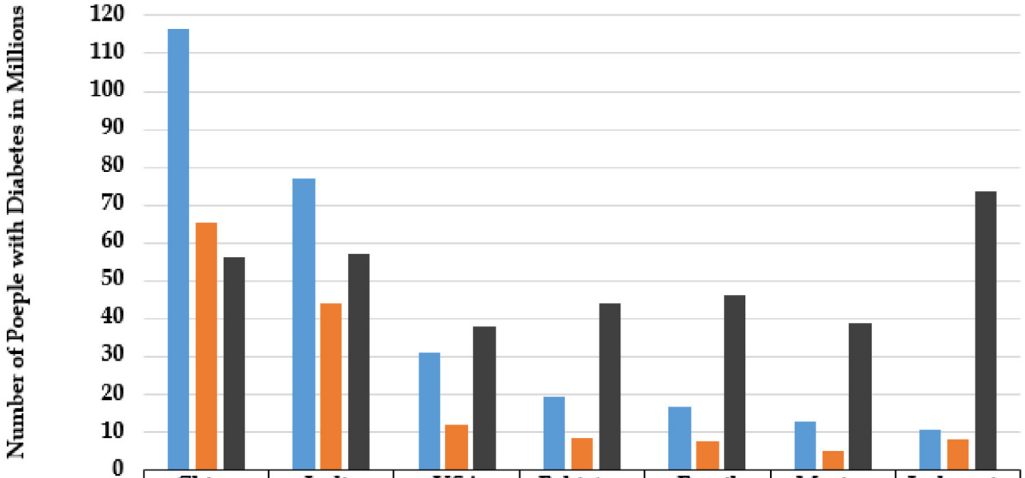
Diabetes mellitus is a collection of metabolic diseases where a person suffers from an extreme level of blood glucose in the body, which is the production of insulin is either insufficient or does not respond properly in a human body [1]. Almost there are 40 types of diabetes and people over the globe are not aware of the complications of diabetes disease because of the under-resourced healthcare system [2]. Some common types are [3]: Type 1 Diabetes Mellitus is insulin and medical treatment dependent and can be developed at any stage especially in children [4]. Type 2 Diabetes mellitus is insulin-independent but lifestyle dependent and is most common type of diabetes which affects all age groups [5]. Gestational Diabetes is caused due to hyperglycemia conditions, it is found in pregnant ladies and may even affect born babies [6]. Pre-diabetes is caused because of increased production of insulin due to genetic disorders [7]. Diabetes can directly affect different organs of the human body like kidney, brain, liver, etc. The common early symptoms and signs found in diabetic and potentially diabetic patients are excessive thirst, excessive fatigue, skin discolouring, frequent urination, etc. [3].

Diabetes is a long-lasting fatal disease that has affected millions of people all over the globe [8]. A substantial increase in the number of diabetes affected people in the last decade has made the disease a global threat. Type 2 Diabetes Mellitus (TIIDM) accounts for 90% of the affected population among all types of diabetes [2]. Some of the major statistical reports from various health organizations about diabetes mellitus that signifies the risk of developing life-threatening, severe, and serious complications [9,10]. According to a report generated by the International Diabetes Federation (IDF) Atlas 2019, 463 million adults having an age range between 20–79 years (9.3% of the world's population) are currently living with diabetes over the globe. There are also projections that it will affect 578 million by 2030 and 700 million by 2045 [11]. It has been estimated in 2019 that 4.2 million deaths happened worldwide because of diabetes mellitus [2]. By IDF Atlas 2019, it has been expected that in 2045 India will top the list with [134.3–165.2] million diabetic people [2]. Fig. 1 presents the top 7 countries or regions suffering from diabetes and the undiagnosed rate of people in millions [2].

In recent studies, Machine Learning (ML) and Ensemble Learning (EL) techniques play a major role in prediction of different chronic diseases like diabetes and have achieved good results in terms of various statistical measurements [12–15]. The earlier prediction of diabetes mellitus is essential and obtaining a higher accuracy rate by using machine learning techniques is decisive [16,17]. The main objective of this research study is to design a computational model based on machine learning and ensemble learning techniques using lifestyle data for the prediction of TIIDM disease. In this work, Bagged Decision

* Corresponding author.
E-mail addresses: Shahidmohammad@bgsbu.ac.in (S.M. Ganie), majidbashirmalik@bgsbu.ac.in (M.B. Malik).

https://doi.org/10.1016/j.health.2022.100092
Received 25 January 2022; Received in revised form 9 August 2022; Accepted 9 August 2022



	China	India	USA	Pakistan	Brazil	Mexico	Indonesia
Diabetes	116.4	77	31	19.4	16.8	12.8	10.7
Undiagnosed	65.2	43.9	11.8	8.5	7.7	4.9	7.9
Undiagnosed (%)	56	57	38.1	43.8	46	38.6	73.7

Fig. 1. Top 7 countries suffering from diabetes.

Tree (BDT), Random Forest (RF), and Extra Tree (ET) algorithms have been explored using the Bagging method. Adaboost (AB) and Stochastic Gradient Boosting (SGB) have been used through Boosting method. Then finally Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT) have been used via Voting method.

The dataset used for the study has been collected from different geographical regions based on the expert advice of specialists like diabetologists, endocrinologists, etc. Different statistical/ML tasks for Exploratory Data Analysis (EDA) have been performed for improving the data quality assessment. Lastly, advanced ML techniques were employed to develop the framework, and performance evaluation of these classifiers has been done through various statistical measurements. The remaining section is structured as follows. Related works on TIIDM prediction using ML techniques are discussed in Section 2. Section 3 describes the dataset, Exploratory Data Analysis, proposed methodology, and ML/EL algorithms. In Section 4, the results of experimental study have been presented and discussed by using statistical/ML metric analysis for prediction of disease. Finally, this research work is concluded along with future directions in Section 5.

2. Related works

This section demonstrates some previous research works for prediction, and detection of TIIDM using machine learning/ensemble learning techniques. The algorithms, datasets, and methodology used by researchers have been discussed. Recent research studies in science have proven with their experimental procedure that lifestyle, demography, psychosocial, and genetic risk factors play an important role in controlling and managing diabetes (especially Type-2 Diabetes Mellitus) at earlier stages. For example, **Liying et al.** [18] developed a framework for TIIDM using ML/EL algorithms like LR, Classification and Regression Technique (CART), Artificial Neural Network (ANN), SVM, RF, and Gradient Boosting Machine (GBM) based on lifestyle data for the Chinese population. The data was collected from the Henan rural cohort containing a total of 36652 instances and 10 variables. Among all the classifiers, the GBM performed best with the highest accuracy. In another approach, **Neha and Shruti** [19] proposed a model for prediction of TIIDM using machine learning techniques. The dataset (952 records and 18 features) used has been collected related to health, lifestyle, and family background. Different machine learning algorithms like LR, K- Nearest Neighbour (KNN), SVM, DT, and RF were employed. Among all the classifiers, RF achieved the highest accuracy rate of 94.10%. **Leon et al.** [20] implemented ML-based framework for TIIDM disease using an ensemble learning method to monitor the glucose level corresponding to various independent features. Firstly, dataset comprised of 27050 instances and 111 features was collected by examination of preventive healthcare populations in 10 Slovenian health institutes. Preprocessing and feature engineering was employed and only 59 variables were selected for developing the framework. Among all the classifiers LightGBM achieved better results in terms of accuracy, precision, recall, Area Under Curve (AUC), Area Under The Precision–Recall Curve (AUPRC), and Root Mean Square Error (RMSE). **Ayush and Divya** [6] proposed ML model for diabetes based on persons daily lifestyle activities. Data was collected through questioners and parameters were decided on the interaction of practitioners/doctors. Various ML classifiers have been used for model building and among all the classifiers CART obtained the highest accuracy rate of 75%. **Kamrul et al.** [21] proposed a robust framework for TIIDM by employing machine learning classifiers like k-nearest neighbour, decision tree, AdaBoost, naïve bayes, XGBoost, and multi-layer perceptron. Outlier detection, filling of missing values, standardization of data, selection of features, and validation of results have been done using Exploratory Data Analysis (EDA). Among all the algorithms, ensembling classifier AdaBoost and XGBoost outperformed with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC as 0.789, 0.934, 0.092, 66.234, and 0.950 respectively. **Rawat and Suryakant** [22] proposed work employed five machine learning techniques viz., AdaBoost, logic boost, robust boost, naïve bayes, and bagging aimed at prediction and analysis of diabetes mellitus patients. The dataset used is PIMA Indian diabetes sourced from the UCI machine learning repository. The results so computed were found to be very accurate with a classification accuracy of 81.77% followed by 79.69% by bagging before AdaBoost approaches respectively.

Based on the existing literature, it can be found easily that lifestyle/biological data can be explored for early prediction of Type-II Diabetes Mellitus. To reduce hospital re-admissions, visiting clinical labs, costs for medical check-ups, and facilitate the medical practitioners to make better decisions in real-time about diabetes. The system can also help the patients and probable patients because the prediction can be made at earlier stages to delay the conception of disease. As per the studies conducted there are 232 million [2] people in world who even do not know if they are suffering from diabetes just because of unawareness

Table 1
Dataset description used for research.

S. no	Parameter	Description	Values	Type
1	Age	Age of the subject in years.	Between 5 to 83	numeric
2	Sex	Gender of the subject (Male/Female).	Male 1, Female 0	categorical
3	FamilyHistory	Whether any family member of the subject is/was suffering from diabetes.	No 0, Yes 1	categorical
4	Smoking	Whether the subject is a smoker or not.	No 0, Yes 1	categorical
5	Drinking	Whether the subject is liquor or non-liquor.	No 0, Yes 1	categorical
6	Thirst	The number of times the subjects drink water in a day/night.	Min 1, Max 15	numeric
7	Urination	How many times the subject passes urine in a day/night.	Min 2, Max 15	numeric
5	Height	Height of the subject in centimeter (cm).	Between 61 to 766	numeric
9	Weight	Weight of the subject in kilogram (Kg).	Min 15, Max 96	numeric
10	Fatigue	If the subject feels fatigued or not.	No 0, Yes 1	categorical
11	Outcome	If the subject is diabetic or not.	0 Non-diabetic 1 Diabetic	categorical

and under-resourced healthcare system. If such technological aid is provided to common people, it shall be of great use.

3. Data collection and methodology

3.1. Dataset description

The dataset used for this study has been collected through offline as well as online modes based on the recommendations of domain experts. The lifestyle parameters were decided in consultation with the experts of diabetes specialists called Endocrinologists, Diabetologists, etc. The authors were involved actively to collect data for carrying out this research work over two years from 2019 to 2021. Survey forms have been distributed to different hospitals of Jammu and Kashmir-UT to collect data from different departments like inpatient, outpatient, and emergency based on the different demographic regions. Also, google forms were designed to get data about subjects working in various organizations so that data cover a good mixture of classes. It contains people from different regions (rural and urban areas), adults from different age groups, balanced male–female ratio, etc. The dataset comprised of 1939 records and 11 biological/lifestyle parameters, where first 10 parameters are predicate/independent variables and last 1(Outcome) is target/dependent variable. The statistical description of dataset is shown in Table 1.

3.2. Proposed framework

Machine learning techniques are performing well whenever used in the healthcare system for diagnosis of different diseases [19,23–25]. The proposed framework has been developed through machine learning classifiers based on ensemble method for classification of TIIDM disease. The working flow of methodology used for this research work is shown in Fig. 2. It depicts the step-by-step procedural operations to build the realistic framework using machine learning techniques based on the ensemble method.

The step-by-step procedure for developing the intelligent framework for TIIDM using the ML/EL methods based on the lifestyle indicators is given in Algorithm 1. The algorithm represents the overall work from input data up to the generation of desired results.

3.3. Data pre-processing

Data pre-processing is a key activity to acquire better results before building the machine learning models. The collected dataset was pre-processed using techniques such as resampling and discretization via various statistical libraries through an Integrated Development Environment Spyder with Python (3.9.1) as a programming tool [26,27]. The required libraries for data quality assessment have been imported where missing values were filled by averaging particular attribute values like Age, Sex, Height, Weight, Thirst, Fatigue, etc. by data imputation method. The outlier detection has been done using boxplot, where the interquartile range method was applied to replace the outlier with feasible sampling values. Data transformation has been performed to improve the efficiency of data before building the machine learning models. Also, duplication, inconsistency, and corrupted data have been neutralized from the dataset by using different data exploratory analytical techniques [28].

3.4. SMOTE for data balancing

Synthetic Minority Oversampling Technique (SMOTE) is powerful and is widely used for high dimensional imbalanced data to make class balancing for solving different real-life problems [29]. The challengeable task while working with imbalanced datasets is that building models will turn into poor performance in terms of various statistical measurements. In this study, SMOTE method was used for class balancing before developing the ML/EL models to maximize the prediction capabilities of framework. This technique helps to oversample and augment the minority class present in dataset. It selects instances of minority class randomly and then finds their k nearest minority class neighbour, then chooses one of the neighbours to form a line segment in the feature space as generated through a convex combination of two synthetic instances say A and B. Fig. 3 depicts the Outcome (class variable) count before and after applying SMOTE technique [30].

3.5. Dataset distribution

FacetGrid method (Seaborn package) has been used to plot the distribution of predicate parameters like *Age, Sex, Family History, Smoking, Drinking, Thirst, Urination, Height, Weight, and Fatigue* towards the target variable *Outcome*. In this method, Kernel Density Estimate (KDE) plot function has been used for visualization of distribution of observation in the dataset. It represents the data samples using a continuous probability curve in one or more dimensions. The horizontal or x-axis represents the range of data samples in dataset and vertical or y-axis represents the probability density function (frequencies) of a random variable. Total shaded are of curve under two points x1 and x2 is probability of value and on every data point x_i, we place a kernel function K. The kernel density estimate is calculated by using equation1 as:

$$\boldsymbol{P}(\boldsymbol{x}) = \frac{1}{\boldsymbol{N}\boldsymbol{h}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{K} \frac{(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{i}})}{\boldsymbol{h}} \tag{1}$$

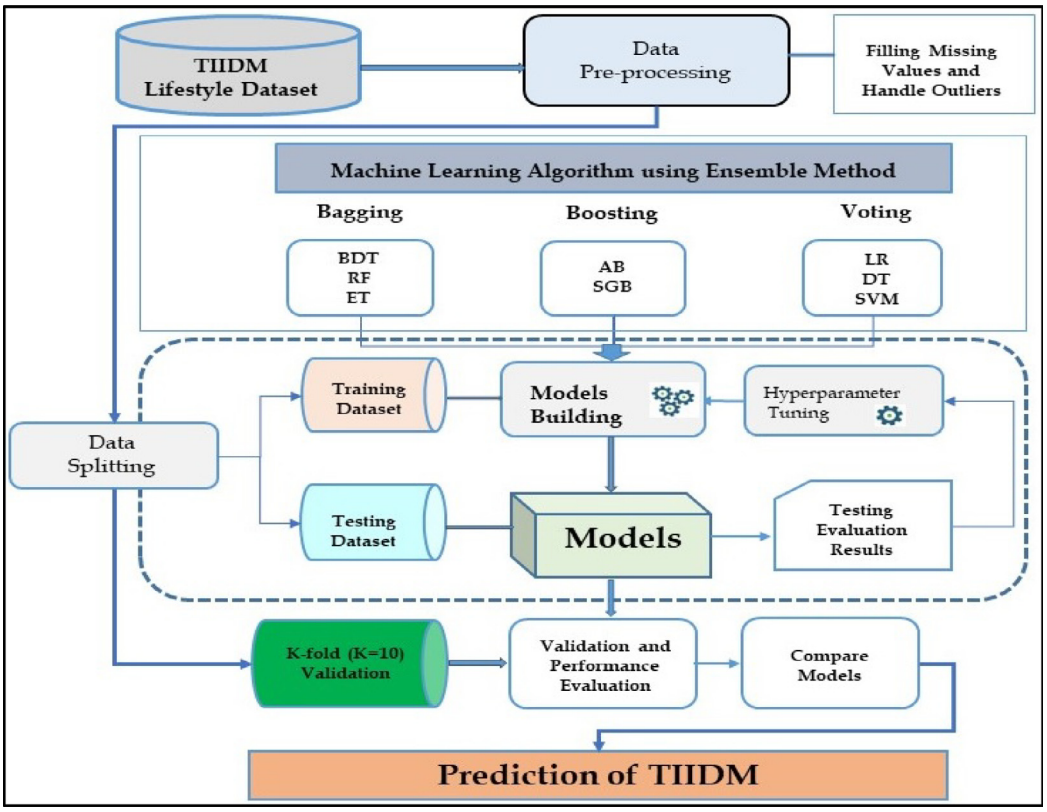


Fig. 2. Proposed framework for prediction of TIIDM.

*Where, **'P'** is density at a given point x*
***'K'** is kernel non-negative function*
***'N'** number of steps*
***'h'** represents the smoothing parameter*
'x' is the maximum random value
'x_i' is variable that denotes vary rate of data samples

The distribution of all lifestyle parameters is shown in Figs. 4–13, where blue and orange colour denotes non-diabetes and diabetes class respectively. For example, in '*Age*' parameter 20–60 years of candidate patients are having highest probability with 20–30 years for non-diabetic and 30–60 years for diabetic patients.

3.6. Dataset splitting and K-fold cross-validation

The K-fold cross-validation technique is widely used by researchers and practitioners for model building process to remove the biasness in the dataset [31]. The pictorial representation of data splitting (10-fold cross-validation) used in this work is shown in Fig. 14. The K-fold cross-validation method has been used with a k value of 10. The entire dataset was randomly subdivided into 10 equal-sized partitions. Of the 10 partitions, a single partition was retained to validate (testing set) the model, and the remaining 10 - 1 partitions are used to train model as training data. The overall process has been repeated 10 times, with each of the 10 partitions used exactly once as the validation data. The results of all the iterations have been aggregated by summation function. The problem of overfitting and underfitting have been reduced in the dataset to meet the performance of both training and testing datasets. The benefit of this technique was that it removes the biasness of data for developing the ML models to achieve realistic results. All data samples are used for both training and testing, and each bin of testing data has been used for validation of the results exactly once.

4. ML models and ensembling

Machine learning techniques whenever implemented have provided better results in almost every sphere of the world especially healthcare analytics [32–35]. Machine learning models are being combined strategically to improve the analytical capabilities of different developed frameworks in order to solve real-life problems. In Bagging various models typically of same learning nature from a different subsample of training set have been developed. On the other hand, in Boosting method different models of same type in which each learns to fix the prediction errors of a prior model in chain have been developed. Finally, in Voting method several models of different type have been developed and their prediction has been combined by calculating the mean to optimize the various statistical/ML metrics for better prediction of TIIDM disease.

4.1. Bagging method:

Bagging is an ensemble technique that explores the bootstrap aggregation method and creates multiple training sets for model-building purposes [36]. The training sets are constructed from the original dataset in a random repeated fashion. After the creation of different training sets, various models are applied to resampling process with ensemble structure. Finally, all the results of learners are aggregated to make the final prediction. It helps to reduce the variances from the models at training time to tackle the problem of overfitting. The three main steps used in bagging method are bootstrapping, parallel training, and aggregating. Some of the algorithms of bagging method are Bagged Decision Tree, Random Forest, Extra Tree, etc. The algorithm and equation for bagging method are given as:

$$B_{bag} = \sum_{i=1}^{n} B_i(X) \tag{2}$$

Algorithm 1: Workflow of the proposed framework.

Input: *Lifestyle Dataset*
Output: *T2DM lifestyle disease prediction using EL models*
BEGIN
STEP 1: *Import the dataset*
STEP 2: *Preprocess the dataset*
STEP 2.1: *Data integration*
STEP 2.2: *Data transformation*
STEP 2.3: *Data cleaning*
STEP 3: Xtrain, ***Ytrain****--70% of dataset*
STEP 4: Xtest, ***Ytest****--30% of dataset*
STEP 5: *Ensemble Learning Methods and their Algorithms*
mn=[BDT(), RF(), ET(), AB(), SGB(), LR(), DT(), SVM()]
for(i=0; i<8; i++) do
Model= mn[i];
Model.fit();
model.predict();
print(Accuracy(i), confusion_matrix, classification_report, roc_curve);
End
STEP 6: *Deployment of framework*
STOP

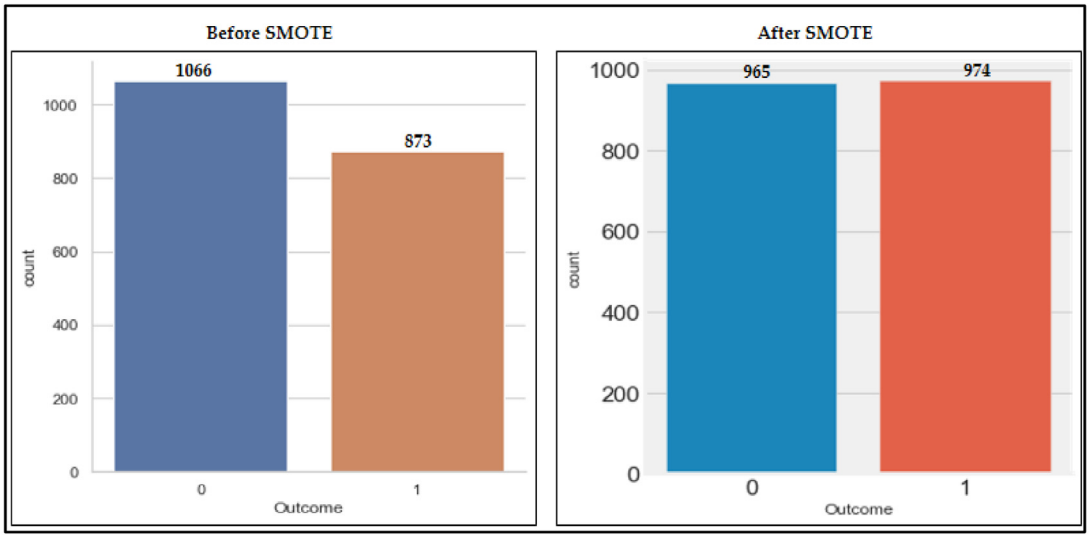


Fig. 3. Class balancing of dataset using SMOTE.

Where, 'B_{bag}' represents bagged prediction (Final Model)
'B_i' is denotes weak/individual learning algorithm
'X' is training example over data sample
'n' is number of bootstrap samples

4.2. Boosting method

Boosting is a process of converting weak learners into strong learners [37]. In this method, all traditional/weak classifiers are combined to form a strong model to improve the predictive capabilities of final

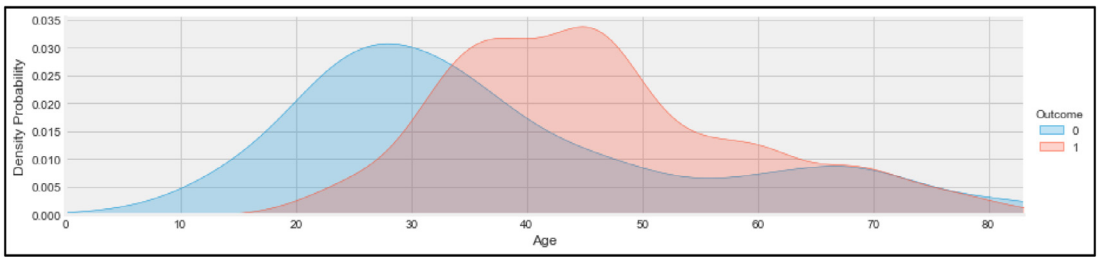


Fig. 4. Age with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

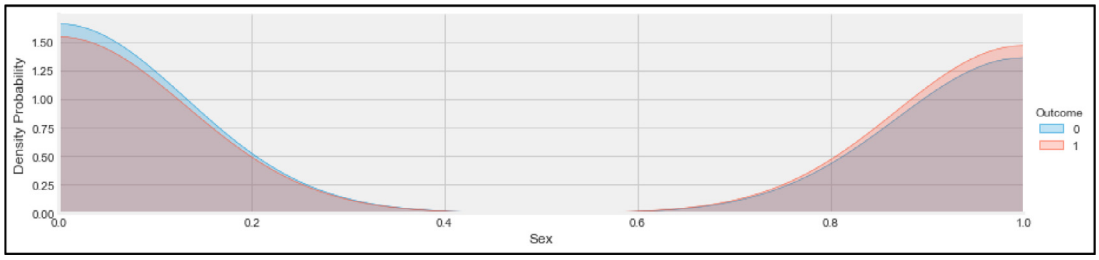


Fig. 5. Sex with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

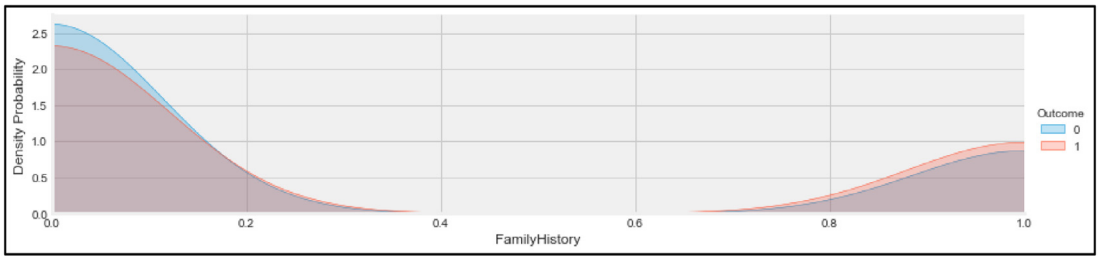


Fig. 6. Family History with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

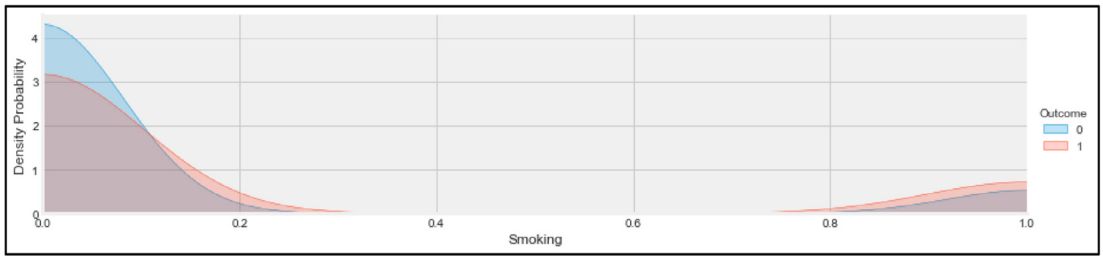


Fig. 7. Smoking with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

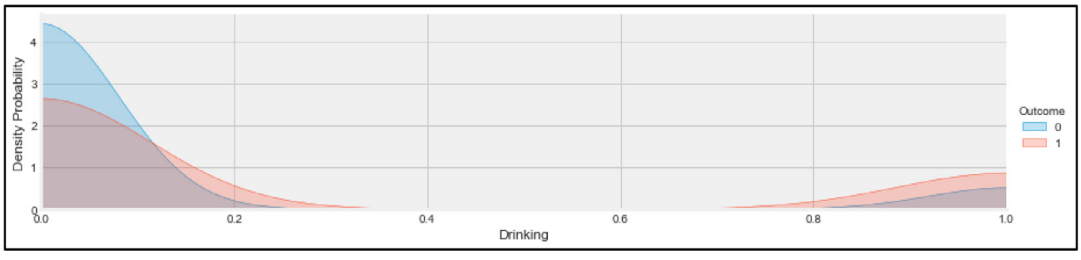


Fig. 8. Drinking with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

model [38]. It uses the aggregation results of various weak classifiers $\{c_1, c_2, c_3, c_4, c_5, \ldots\ldots\ldots, c_k\}$ to improve the forecasting capability of model C*(Ensemble Model). The base learning classifier is applied many times in order to generate a new prediction rule. The process will be repeated iteratively, after n number of iterations the boosting method will combine the results from weak learners and convert them

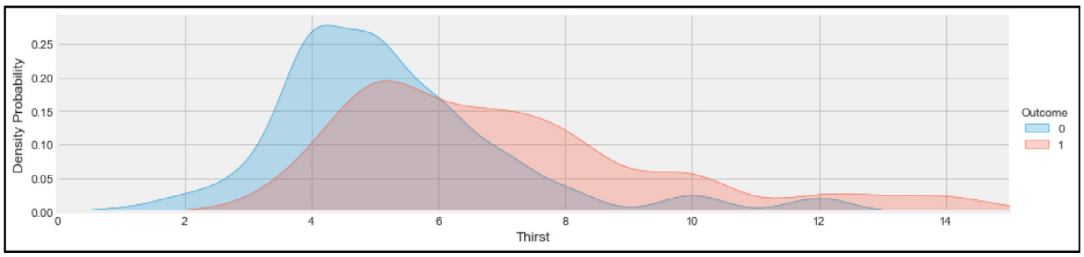


Fig. 9. Thirst with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

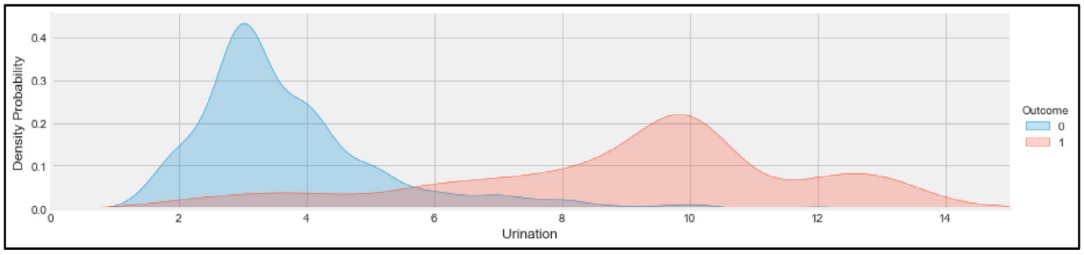


Fig. 10. Urination with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

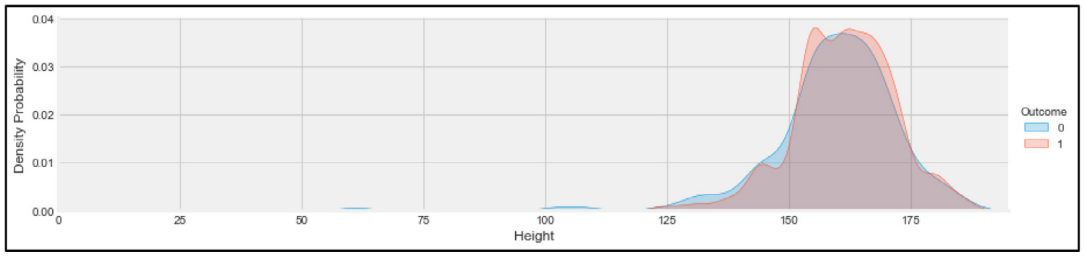


Fig. 11. Height with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

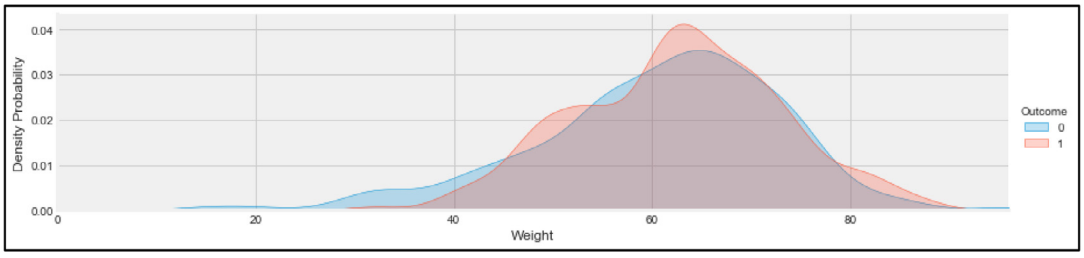


Fig. 12. Weight with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

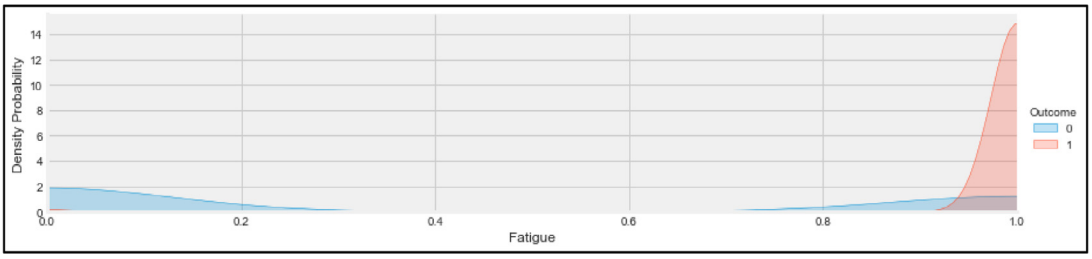


Fig. 13. Fatigue with respect to Outcome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

into a single strong prediction model. There are different types of boosting algorithms like AdaBoost, Gradient Boosting, XGBoost, Random Forest, etc. The general algorithm for boosting method is given as in Algorithm 3.

4.3. Voting method

In this ensemble learning method, the predictions of base learners are combined to make new features for training sets to enhance the

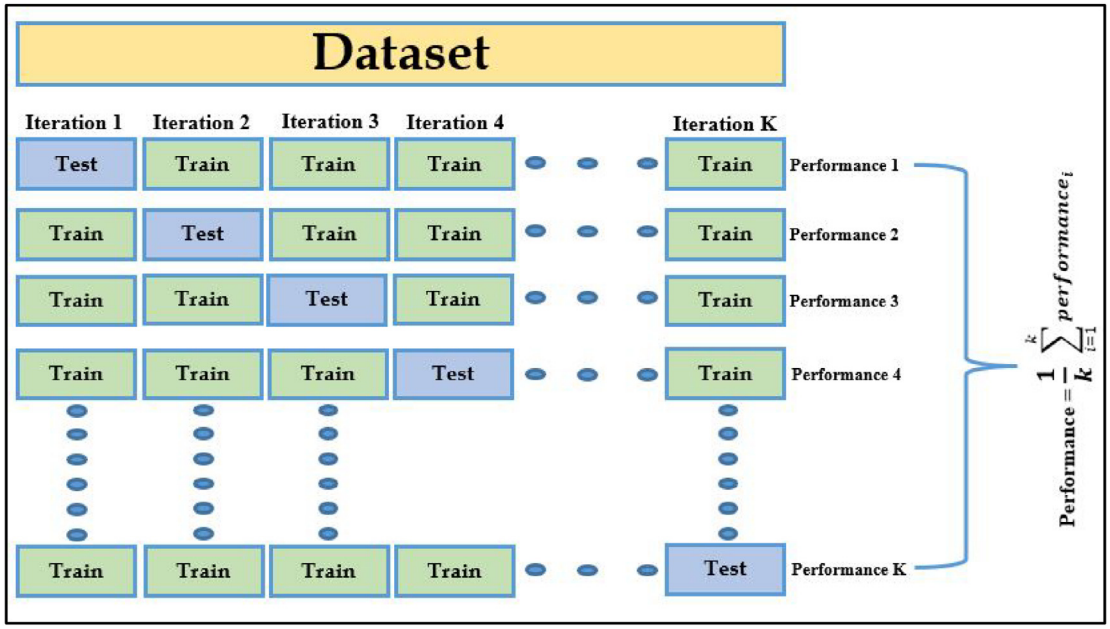


Fig. 14. K-fold Cross-Validation technique.

Algorithm 2: *Bagging Method for Prediction.*

1. *Let* $\boldsymbol{n}$ *be the number of bootstrap samples*
2. *For* $\boldsymbol{i=1}$ *to* $\boldsymbol{n}$ ***do***
   Formulate the bootstrap sample ‘$\boldsymbol{S_i}$’ *of size* $\boldsymbol{K}$
   Train the base learner $\boldsymbol{L_i}$ *with bootstrap sample*
   End
3. $\boldsymbol{L^*(x) = agrmax_{y\in Y} \sum_{i=1}^{n} 1(y = L_i(x))}$ *//*$\boldsymbol{L^*}$ *is Ensemble model and* $\boldsymbol{agrmax_{y\in Y}}$ *is aggregation of all majority voting values of weak learner* ‘$\boldsymbol{L_i}$’ *on training sample* ‘$\boldsymbol{x}$’*//*
4. $\{\boldsymbol{L_i}$ *= 1 if it is true and 0 otherwise }*

Algorithm 3: *Boosting Method for Prediction.*

1. *Fit estimators* $\boldsymbol{E^1}$
2. *For* $\boldsymbol{i}$ *in* $\boldsymbol{[1, C]}$ *weak estimators // i is number of iterations //*
3. $Loss^i = \sum_{j=1}^{n} (Y_j - E^i(X_j))^2$ *// loss in* i^{th} *iteration //*
4. *Compute neg gradient:* $-\frac{\partial L^i}{\partial X_j} = -\frac{2}{n} * (Y_j - E^i(X_j))\, Yi$
5. *Fit a weak estimator:* H^i ***on*** $(X, \frac{\partial L}{\partial X})$ *// ρ changes the step size //*
6. *Forecast:* $E^m(X) = E^i(X) + \rho * H^i(X) = E^1 + \rho * \sum_{i=1}^{m} H^i(X)$

desired results [39]. This method is used to combine the traditional as well as advanced classifiers to develop the meta-features for final prediction. The output of base classifiers is aggregated based on majority vote and weighted techniques. This ensemble method is particularly used to combine heterogeneous base classifiers and make an evaluation for final results. The general algorithm for voting method is given as in Algorithm 4.

5. Results and discussion

In this section, outcomes achieved through the experimental setup by using ML/EL techniques for prediction of TIIDM based on lifestyle predictors have been discussed and presented. The PC (Work Station) used for experimental process was HP Z60. The technical specification of hardware is processor Intel XEON with speed 2.4 GHz (12

Algorithm 4: *Voting Method for Prediction*

1. *Let T be the training dataset,* $\boldsymbol{T = \{x_i, y_i\}_{i=1}^{n}}$
2. *Train the base learners* $\boldsymbol{L = \{ l_1 + l_2 + l_3 + l_4 + \ldots\ldots\ldots\ldots + l_k \}}$

   For $\boldsymbol{l=1}$ *to* $\boldsymbol{k}$ ***do***

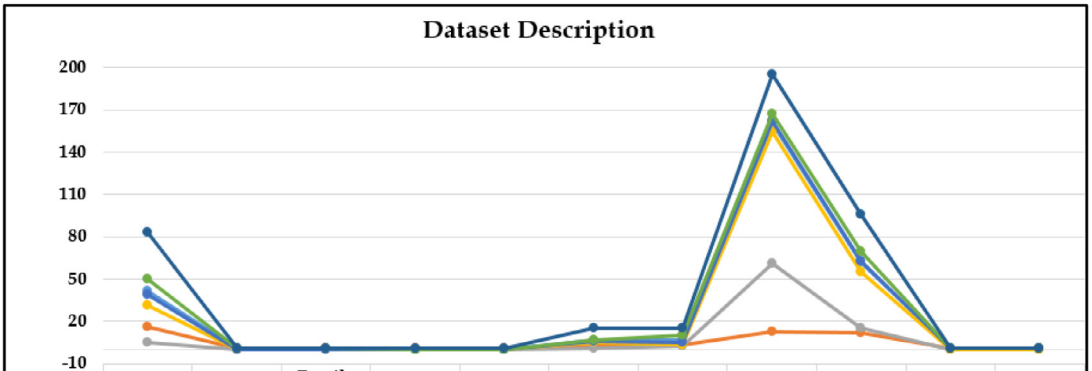
   Train $\boldsymbol{e_l}$ *using* $\boldsymbol{T}$ *// Training of case learners //*

   End
3. *Design the new dataset for predictions*

   For $\boldsymbol{i=1}$ *to* $\boldsymbol{n}$ ***do***

   $\boldsymbol{T_e = \{x_i', y_i\}}$, ***where*** $\boldsymbol{x_i' = \{ e_l(x_i), \ldots\ldots\ldots\ldots, e_k(x_i) \}}$

   End
4. *Train a meta-classifier*
5. *Train* $\boldsymbol{E}$ *using* $\boldsymbol{T_l}$
6. *Return* $\boldsymbol{E}$ *// E is an ensemble Classifier //*



	Age	Sex	Family History	Smoking	Drinking	Thirst	Urination	Height	Weight	Fatigue	Outcome
mean	41.77	0.47	0.27	0.15	0.18	6.18	6.4	160.07	61.57	0.69	0.5
std	15.84	0.5	0.45	0.35	0.38	2.43	3.46	12.43	11.48	0.46	0.5
min	5	0	0	0	0	1	2	61	15	0	0
25%	31	0	0	0	0	5	3	154	55	0	0
50%	39	0	0	0	0	6	5	162	62.5	1	1
75%	50	1	1	0	0	7	10	167	69.25	1	1
max	83	1	1	1	1	15	15	195	96	1	1

Fig. 15. Description of parameters used in dataset.

CPUs) along with GPU NVIDIA Quadro K2200. The system RAM and display RAM is 4 GB each. The storage capacity of system is 1TB and operating system installed is Windows 10 pro-64-bit. The basic descriptive statistics of lifestyle parameters with their measures like mean, std, min, max, etc. are shown in Fig. 15. For example, in Age parameter the values for mean, std, min, 25%, 50%, 75% and max are as 41.77, 15.84, 5, 31, 39, 50 and 83 respectively. The same calculations have been made for other parameters as well to summarize the main characteristics of a dataset.

5.1. Correlation coefficient analysis

Correlation Coefficient Analysis (CCA) method is used to find a relationship between pairs of variables present in dataset [40]. The main idea behind the CCA is to find the relevance of the features in the dataset. The feature set is considered well for building the machine learning models if there is a good relationship between independent and dependent variables. Fig. 16 presents the correlation matrix between the set of features, where range of the numerically calculated values has been presented by a finite number ranging from +1 and -1. CCA matrix is presented by a finite number between +1 and −1 along the x-axis and y-axis and their values are interpreted as so:

- **+1** *represents complete positive correlation*
- **+0.8** *represents strong positive correlation*

+0.6 *represents moderate positive correlation*

- **+0** *represents no correlation*
- **-0.6** *represents moderate negative correlation*
- **-0.8** *represents strong negative correlation*
- **-1** *represents complete negative correlation*

The features are evaluated by the equation below:

$$CCA = \frac{k\,avg\,(corr_{fc})}{\sqrt{k + k\,(k-1)\,avg\,(corr_{ff})}} \tag{3}$$

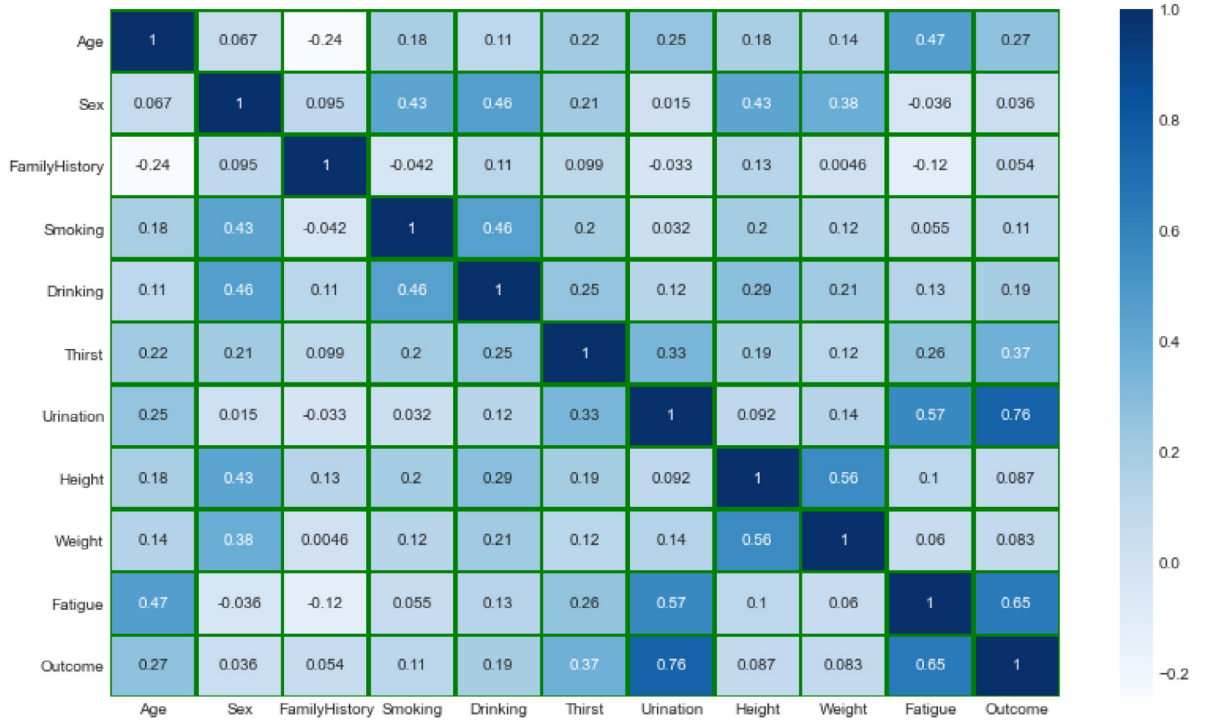


Fig. 16. Correlation coefficient matrix.

Table 2
Confusion matrix.

		Predicted values	
		No (0)	Yes (1)
Actual values	No (0)	True Negative (TN)	False Positive (FP)
	Yes (1)	False Negative (FN)	True Positive (TP)

Where CCA is importance between dependent and independent biological features and is used to define the rank transformation for exploring the set of parameters/features. The (avg(corr $_{fc}$)) represents average of correlation between predicate/independent and target/dependent variables and (corr$_{ff}$) defines average of correlation among parameters. Also, K denotes the number of features present in dataset. The parameters *Urination, Fatigue, Thirst, Age, Drinking, Smoking, Weight, Height, FamilyHistory, and Sex* are having positive relationship with *Outcome* class for the prediction of TIIDM disease. For example, relationship among independent and dependent variables like Urination w.r.t. Outcome is 0.76, Fatigue w.r.t. Outcome is 0.65, Thirst w.r.t. Outcome is 0.37, etc. Also, relationship among independent variables like Fatigue w.r.t. Urination is 0.57, Age w.r.t. Fatigue is 0.47, Drinking w.r.t. Age is 0.46, Height w.r.t. Sex is 0.43, etc.

5.2. Confusion matrices of algorithms

The confusion matrix presented in Table 2 is used to check the performance evaluation of ML/EL models for identification of mislabelled/errors in predicting the TIIDM disease. It matches the actual values with predicted ones based on four elements like *True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)*.

The confusion matrices (Training and Testing) of ensemble learning classifiers to measure the performance evaluation towards prediction of T2DM disease are shown in Figs. 17–28. The confusion matrices have been used to evaluate the various ML/EL classifiers through statistical/ML measurements like accuracy, precision, recall, specificity, f1-score, false positive rate, false negative rate, negative predicted values, kappa, miss classification rate, ROC curve, etc.

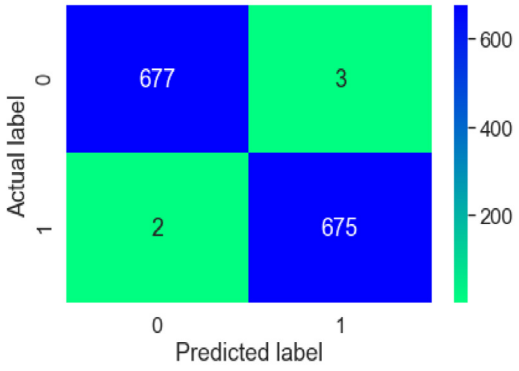


Fig. 17. BDT Training Confusion Matrix.

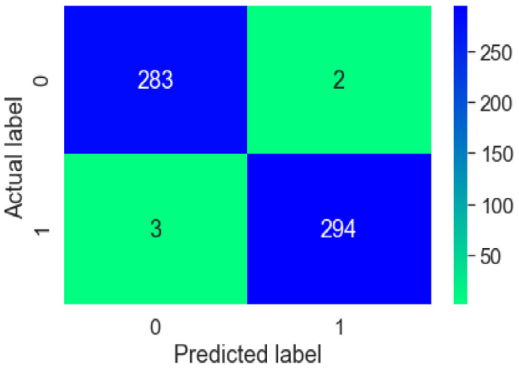


Fig. 18. BDT Testing Confusion Matrix.

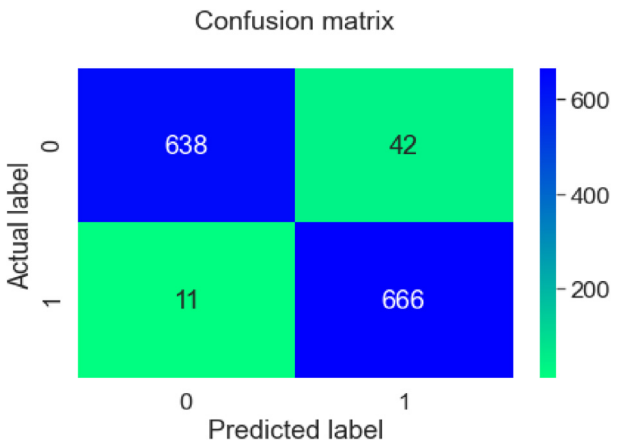


Fig. 19. RF Training Confusion Matrix.

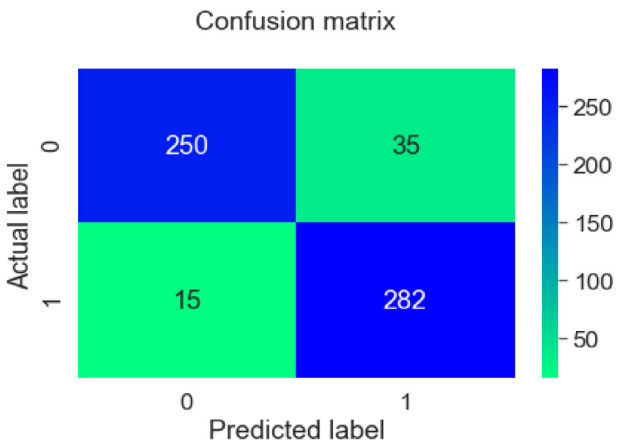


Fig. 22. ET Testing Confusion Matrix.

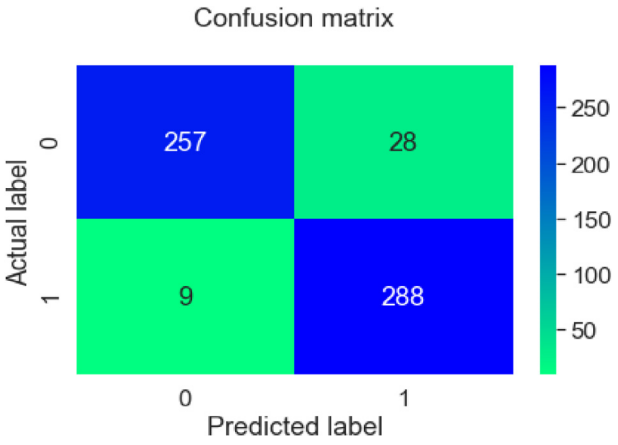


Fig. 20. RF Testing Confusion Matrix.

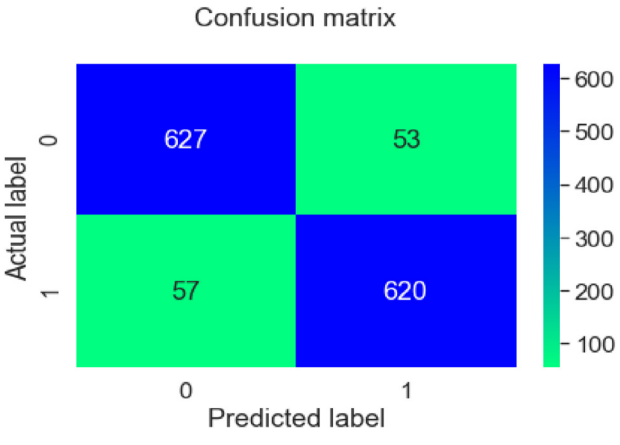


Fig. 23. AB Training Confusion Matrix.

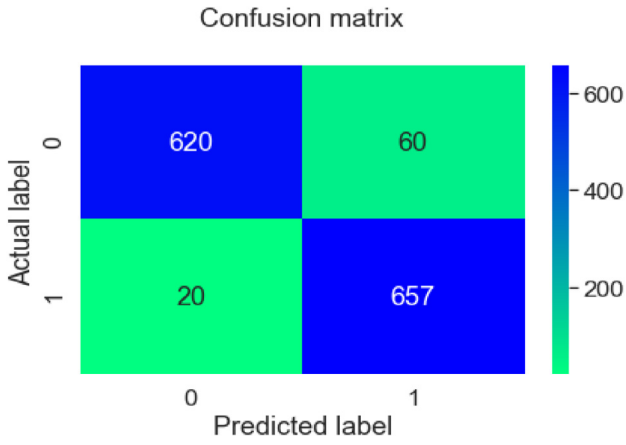


Fig. 21. ET Training Confusion Matrix.

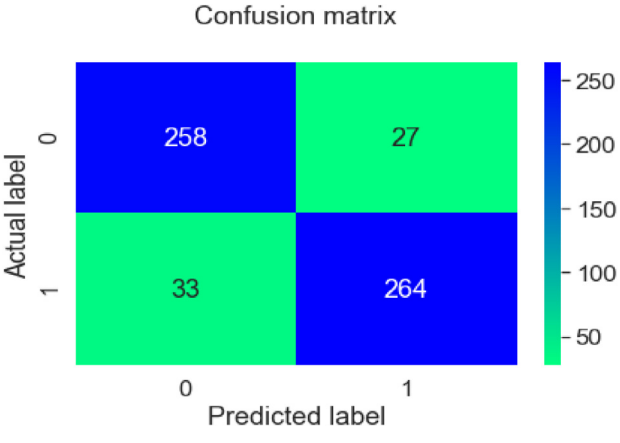


Fig. 24. AB Testing Confusion Matrix.

5.3. Performance evaluation of EL algorithms

The aggregation/summation of results achieved using 10-fold cross-validation method for different ML/EL models has been presented in Table 3. It describes results of different evaluation metrics like Training Accuracy, Testing Accuracy, Miss Classification Rate, Kappa, and Running Time (RT). Among all the models, BDT algorithm achieved the highest testing accuracy rate of 99.14% followed by SGB, RF, ET, AB, and Voting classifiers (LR, DT, SVM), as 98.45%, 93.63%, 91.41%, 89.69%, and 89.51% respectively. However, in terms of MCR (number of misclassifications on test dataset), BDT achieved lowest rate of 0.86% and Voting achieved the highest rate of 10.49%. In kappa statistical analysis, BDT achieved highest performance rate of 98.17% and ET achieved the lowest rate of 84.60%. In addition, AB takes a minimum running time of 0.0330 s and the Voting method takes a maximum running time of 0.0990 s while executing algorithms.

Table 4 is the representation of other statistical/ML measurements of the test dataset namely precision, recall, specificity, false positive

Table 3
Performance measure of EL models.

Algorithms	Training accuracy	Testing accuracy	MCR (%)	Kappa (%)	RT (s)
Bagged Decision Trees	**99.63%**	**99.14%**	**0.86**	**98.17**	**0.0740**
Random Forest	96.09%	93.64%	6.36	87.68	0.0640
Extra Trees	94.10%	91.41%	8.59	84.60	0.0600
AdaBoost	91.89%	89.69%	10.31	86.70	0.0330
Stochastic Gradient Boosting	99.56%	98.45%	1.55	91.08	0.0340
Voting (LR, DT, SVM)	91.23%	89.51%	10.49	89.68	0.0990

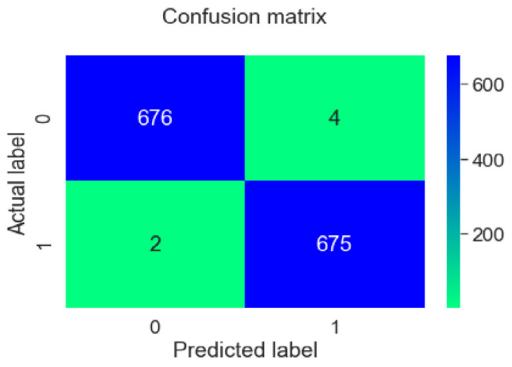


Fig. 25. SGB Training Confusion Matrix.

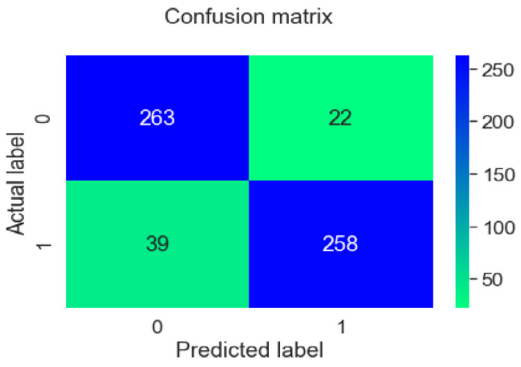


Fig. 28. Voting Classifiers Testing Confusion Matrix.

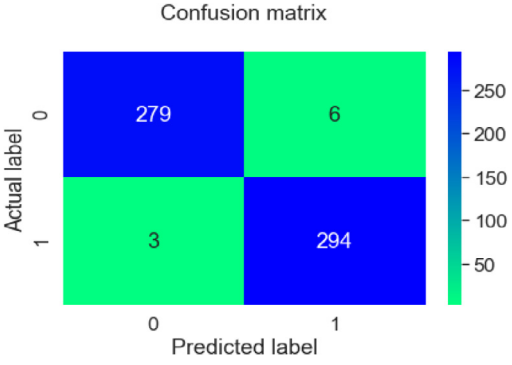


Fig. 26. SGB Testing Confusion Matrix.

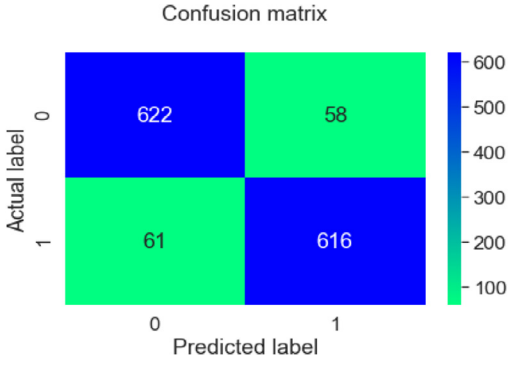


Fig. 27. Voting Classifiers Training Confusion Matrix.

rate, false negative rate, F1-score, and negative predicted values. However, in terms of precision, recall, specificity, FPR, FNR, F1-score, and NPV, BDT achieved desirable performance rate of 98.98%, 99.32%, 98.95%, 1.04%, 0.67%, 99.15%, and 99.29% respectively. In addition, Voting method achieved the lowest performance in terms of precision as 86.86%. In terms of recall, ET algorithm achieved the lowest rate of 88.95%. BDT acquired a maximum NPV rate of 99.29%, on the other hand, ET and RF acquired a minimum NPV rate of 87.71% and 90.17%.

5.4. Feature engineering

Feature engineering plays an important role in ML/EL model building process. Inconsequential or inappropriate features can adversely affect model execution [41]. Proper feature selection enhances accuracy and decreases the training time. Some of the feature selection techniques in machine learning paradigms are embedded, filter, wrapper, embedded, and hybrid methods [42]. In this work, Information Gain and Correlation methods were employed for feature selection. It has been found that almost all the selected features except *'Sex'* have a good amount of contribution towards the prediction of TIIDM showcasing for Bagged Decision Tree (BDT) classifier in Fig. 29. The ranking/importance from highest to lowest of all features are *Urination, Weight, Thirst, Age, Fatigue, Family History, Smoking, Drinking, and height* towards *Outcome*. Although *Sex* parameter does not contribute towards outcome class but it has a good amount of relationship with independent variables and is a key lifestyle parameter.

6. Comparative analysis with existing work

The performance of our proposed framework has been compared with several relevant kinds of literature in terms of techniques used, dataset, and analysis as shown in Table 5. Most of the lifestyle indicators are common for all the studies carried out for comparison with the proposed work. It has been found that our considered framework yielded good results in terms of different evaluation metrics particularly accuracy for the prediction of TIIDM. The techniques like data imputation for handling missing values, detection and replacement of outliers using boxplot method, and standardization and normalization of data using transformation method have been used to achieve better results than other related works. Also, SMOTE method, hyperparameter tuning, and K-fold cross-validation technique were employed while developing the proposed framework to achieve more valid results than other related studies.

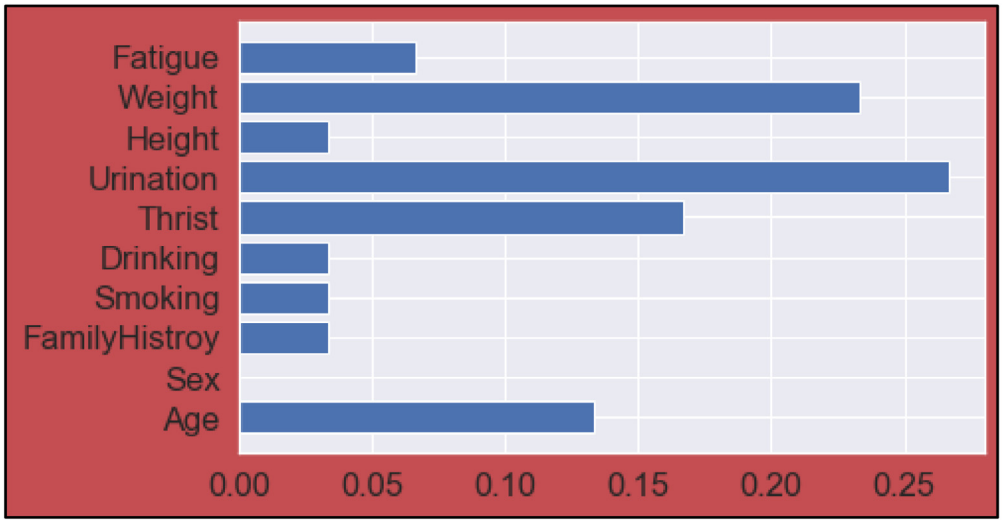


Fig. 29. Feature Importance towards prediction of TIIDM.

Table 4
Classification performance measurements of test dataset..

Algorithms	Results (%)						
	precision	recall	specificity	FPR	FNR	F1-Score	NPV
Bagged Decision Trees	**98.98**	**99.32**	**98.95**	**1.04**	**0.67**	**99.15**	**99.29**
Random Forest	96.96	91.13	96.61	3.38	8.86	93.96	90.17
Extra Trees	94.94	88.95	94.33	5.66	11.04	91.85	87.71
AdaBoost	88.88	90.72	88.65	11.34	9.27	89.79	90.52
Stochastic Gradient Boosting	98.98	98.00	98.93	1.06	2.00	98.49	97.89
Voting (LR, DT, SVM)	86.86	92.14	87.08	12.91	7.85	89.42	92.22

Table 5
Comparison with existing systems.

Authors	Technique used	Dataset	Analysis
[6]	CART (Classification and Regression Trees)	Collected dataset through questionnaire	75% for CART
[43]	SVM, RF and LR	Demographic web-based questionnaire	80.17% for SVM
[44]	LR, GBC, LDA, ABC, ETC, NB, Bagging, RF, DT, SVC, Perceptron and KNN	Collected dataset from hospital	96% for LR
[19]	LR, KNN, SVM, NB, DT, RF	Offline and online questionnaire	94.10% for RF
[45]	LR, LDA, KNN, DT, NB, SVM, RFC and ANN	Noakhali Medical College Bangladesh	94.07% for ANN
[46]	LR, SVM, KNN, RF, NB, GB	Murtala Mohammed Specialist Hospital, Kano	88.76% for RF
Our Proposed Study	**BDT, RF, ET, AB, SGB, LR, SVM, and DT**	**Lifestyle dataset from geographical regions**	**99.14% for BDT**

7. Conclusion and future scope

Diabetes a chronic/fatal disease has affected millions of people all over the globe at an alarming rate. In this research work, early prediction of TIIDM based on lifestyle/biological features has been accomplished using ML/EL techniques. A detailed analysis of patients' lifestyle data has been done for the development of framework. The EDA phase plays an important role in better prediction by improving the quality assessment of the dataset, where filling of missing values, detection and replacing of outliers, and SMOTE for class balance was a core concern. CCA was employed for choosing the optimum set of lifestyle features. Finally, eight different machine learning algorithms based on ensemble methods were applied using 10-fold cross-validation for the prediction of disease. The results achieved in this proposed work are realistic and robust by metric analysis and to our knowledge, this is the newly introduced framework produces much better prediction results compared to existing research work.

In the future, this proposed framework can be used to identify the probability of disease in patients and probably patients at earlier stages. Based on the inclination of the disease different categories of patients can be advised to make appropriate changes to their lifestyle (Diet Plans and Physical Exercise Charts). Additionally, mobile applications can be developed to help healthcare providers to detect and predict TII2M disease. It will be also useful for end users in reducing the complications of diabetes at earlier stages and avoiding hospital readmissions, unnecessary medical check-ups, and regularity of visiting clinical labs, etc.

Declaration of competing interest

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

References

[1] N. Sneha, T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, J. Big Data 6 (1) (2019) http://dx.doi.org/10.1186/s40537-019-0175-6.

[2] Diabetes Federation International, IDF, DF Diabetes Atlas 2019, ninth ed., 2019.

[3] P. Kaur, M. Sharma, Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: A review, Int. J. Pharm. Sci. Res. 9 (7) (2018) 2700–2719, http://dx.doi.org/10.13040/IJPSR.0975-8232.9(7).2700-19.

[4] R. Sengamuthu, R. Abirami, D. Karthik, Various data mining techniques analysis to predict, Int. Res. J. Eng. Technol. 5 (5) (2018) 676–679, [Online]. Available: https://www.irjet.net/archives/V5/i5/IRJET-V5I5134.pdf.

[5] D. J. D. M.D., Diabetes mellitus diabetes mellitus, Ferri's Clin. Advis. 2020 512 (January) (2020) 432–441, http://dx.doi.org/10.1016/B978-0-323-67254-2.00255-2.
[6] A. An, D. Shakti, Prediction of diabetes based on personal lifestyle indicators, in: Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol., no. September, NGCT 2015, 2016, pp. 673–676, http://dx.doi.org/10.1109/NGCT.2015.7375206.
[7] S. Vyas, R. Ranjan, N. Singh, A. Mathur, Review of predictive analysis techniques for analysis diabetes risk, in: Proc. - 2019 Amity Int. Conf. Artif. Intell., AICAI 2019, 2019, pp. 627–631, http://dx.doi.org/10.1109/AICAI.2019.8701236.
[8] D.M. Chan, Director-General, WHO, Global Report on Diabetes World Health Organization, 2018, p. 88.
[9] S. Edition, IDF Diabetes Atlas, seventh ed., 2015.
[10] International Diabetes Federation and Nam Han Cho (chair), et al., Eighth edition 2017, 2017.
[11] IDF, IDF Diabetes Atlas 2019, ninth ed., 2019.
[12] B. Davazdahemami, H.M. Zolbanin, D. Delen, An explanatory analytics framework for early detection of chronic risk factors in pandemics, Healthc. Anal. 2 (January) (2022) 100020, http://dx.doi.org/10.1016/j.health.2022.100020.
[13] M. Samieinasab, S.A. Torabzadeh, A. Behnam, A. Aghsami, F. Jolai, Meta-health stack: A new approach for breast cancer prediction, Healthc. Anal. 2 (October 2021) (2022) 100010, http://dx.doi.org/10.1016/j.health.2021.100010.
[14] S.M. Ganie, M.B. Malik, T. Arif, Machine learning techniques for diagnosis of type 2 diabetes using lifestyle data, in: International Conference on Innovative Computing and Communications, in: Advances in Intelligent Systems and Computing, vol. 1394, Springer, Singapore, 2022, pp. 487–497.
[15] N. Nissa, S. Jamwal, S. Mohammad, Early detection of cardiovascular disease using machine learning techniques an experimental study, Int. J. Recent Technol. Eng. 9 (3) (2020) 635–641.
[16] F. Anwar, Qurat-Ul-Ain, M.Y. Ejaz, A. Mosavi, A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey, Inf. Med. Unlocked 21 (April) (2020) 100482, http://dx.doi.org/10.1016/j.imu.2020.100482.
[17] S.M. Ganie, M.B. Malik, T. Arif, Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches, J. Diabetes Metab. Disord. 2022 (2022) 339–352.
[18] L. Zhang, Y. Wang, M. Niu, C. Wang, Z. Wang, Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan rural cohort study, Sci. Rep. 10 (1) (2020) 1–10, http://dx.doi.org/10.1038/s41598-020-61123-x.
[19] N.P. Tigga, S. Garg, Prediction of type 2 diabetes using machine learning classification methods, Procedia Comput. Sci. 167 (2019) (2020) 706–716, http://dx.doi.org/10.1016/j.procs.2020.03.336.
[20] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, G. Stiglic, Early detection of type 2 diabetes mellitus using machine learning-based prediction models, Sci. Rep. 10 (1) (2020) 1–13, http://dx.doi.org/10.1038/s41598-020-68771-z.
[21] K. Hasan, A. Alam, D. Das, E.H. Senior, Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, VOl. X, 2020, pp. 1–19, http://dx.doi.org/10.1109/ACCESS.2020.2989857.
[22] V. Rawat, Suryakant, A classification system for diabetic patients with machine learning techniques, Int. J. Math. Eng. Manag. Sci. 4 (3) (2019) 729–744, http://dx.doi.org/10.33889/IJMEMS.2019.4.3-057.
[23] S.M. Ganie, M.B. Malik, T. Arif, Early prediction of diabetes mellitus using various artificial intelligence techniques: A technological review, Int. J. Bus. Intell. Syst. Eng. 1 (4) (2021) 1–22.
[24] S. Jamwal, S.M. Najmu Nissa, Heart disease prediction using machine learning, in: Lect. Notes Networks Syst., vol. 203 LNNS, (no. 67) 2021, pp. 653–665.
[25] V. Chang, V.R. Bhavani, A.Q. Xu, M. Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, Healthc. Anal. 2 (2021) (2022) 100016, http://dx.doi.org/10.1016/j.health.2022.100016.
[26] Anaconda Inc, Anaconda distribution, Anaconda, 2019, [Online]. Available: https://www.anaconda.com/distribution/.
[27] S. Raschka, J. Patterson, C. Nolet, Machine learning in python: Main developments and technology trends in data science, Mach. Learn., Artif. Intell., Inf. 11 (4) (2020) http://dx.doi.org/10.3390/info11040193.
[28] A. Jazayeri, O.S. Liang, C.C. Yang, Imputation of missing data in electronic health records based on patients' similarities, J. Healthc. Inform. Res. 4 (3) (2020) 295–307, http://dx.doi.org/10.1007/s41666-020-00073-5.
[29] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018) http://dx.doi.org/10.3390/app8081325.
[30] Y. Li, H. Li, H. Yao, Analysis and study of diabetes follow-up data using a data-mining-based approach in New Urban Area of Urumqi, Xinjiang, China, 2016–2017, Comput. Math. Methods Med. 2018 (2018) http://dx.doi.org/10.1155/2018/7207151.
[31] M.K. Hasan, M.A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531, http://dx.doi.org/10.1109/ACCESS.2020.2989857.
[32] S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, Int. J. Cogn. Comput. Eng. 2 (November 2020) (2021) 40–46, http://dx.doi.org/10.1016/j.ijcce.2021.01.001.
[33] S.M. Ganie, M.B. Malik, Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus, Int. J. Med. Eng. Inform., in press.
[34] M.B. Malik, S.M. Ganie, T. Arif, Machine learning techniques in healthcare informatics: Showcasing prediction of type 2 diabetes Mellitus disease using lifestyle data machine learning in healthcare, Predict. Model. Biomed. Data Min. Anal. (2022).
[35] M. Izadikhah, A fuzzy stochastic slacks-based data envelopment analysis model with application to healthcare efficiency, Healthc. Anal. 2 (February) (2022) 100038, http://dx.doi.org/10.1016/j.health.2022.100038.
[36] M.N. Algedawy, Detecting diabetes mellitus using machine learning ensemble, 670 | Int. J. Comput. Syst. ISSN 03 (12) (2017) 670–677, [Online]. Available: http://www.ijcsonline.com/.
[37] P. Doupe, J. Faghmous, S. Basu, Machine learning for health services researchers, Value Heal. 22 (7) (2019) 808–815, http://dx.doi.org/10.1016/j.jval.2019.02.012.
[38] A. Sarwar, M. Ali, J. Manhas, V. Sharma, Diagnosis of diabetes type-II using hybrid machine learning based ensemble model, Int. J. Inf. Technol. (December) (2018) http://dx.doi.org/10.1007/s41870-018-0270-5.
[39] S.K. Dehkordi, H. Sajedi, Prediction of disease based on prescription using data mining methods, Health Technol. (Berl) 9 (1) (2019) 37–44, http://dx.doi.org/10.1007/s12553-018-0246-2.
[40] A. Hussain, S. Naaz, Prediction of Diabetes Mellitus: Comparative Study of Various Machine Learning Models, Vol. 1166, Springer Singapore, 2021.
[41] D. Dutta, D. Paul, P. Ghosh, Analysing feature importances for diabetes prediction using machine learning, in: 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2018, 2019, pp. 924–928, http://dx.doi.org/10.1109/IEMCON.2018.8614871.
[42] M. Maniruzzaman, et al., Accurate diabetes risk stratification using machine learning: Role of missing value and outliers, J. Med. Syst. 42 (5) (2018) http://dx.doi.org/10.1007/s10916-018-0940-7.
[43] R. Patil, K. Shah, Assessment of risk of type 2 Diabetes Mellitus with stress as a risk factor using classification algorithms, Int. J. Recent Technol. Eng. 8 (4) (2019) 11273–11277, http://dx.doi.org/10.35940/ijrte.d9509.118419.
[44] A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, Procedia Comput. Sci. 165 (2019) 292–299, http://dx.doi.org/10.1016/j.procs.2020.01.047.
[45] M. Kowsher, M.Y. Turaba, T. Sajed, M.M. Mahabubur Rahman, Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers, in: 2019 22nd Int. Conf. Comput. Inf. Technol., no. December, ICCIT 2019, 2019, pp. 18–20, http://dx.doi.org/10.1109/ICCIT48885.2019.9038574.
[46] L.J. Muhammad, E.A. Algehyne, S.S. Usman, Predictive supervised machine learning models for Diabetes Mellitus, SN Comput. Sci. 1 (5) (2020) 1–10, http://dx.doi.org/10.1007/s42979-020-00250-8.