Journal of Modern Applied Statistical Methods
https://jmasm.com

Article

Optimizing Diabetes Classification: BOA-Enhanced ML with EDA and SMOTE

R. Harihara Krishnan [1] and Ananthi Sheshasaayee [2,*]

[1] PDF Scholar, IIIT, Delhi 110020, India
[2] QUAID-E-MILLATH Government College for Women (Autonomous), Chennai 600002, India
* Correspondence: ananthiseshu@gmail.com

How To Cite: Harihara Krishnan, R.; Sheshasaayee, A. Optimizing Diabetes Classification: BOA-Enhanced ML with EDA and SMOTE. *Journal of Modern Applied Statistical Methods* **2025**, *24*(1), 11. https://doi.org/10.56801/Jmasm.V24.i1.11

Abstract: Diabetes Mellitus, a chronic metabolic disorder stemming from fluctuations in blood glucose and insulin levels, exerts profound impacts on every organ, significantly compromising overall health. While a permanent cure remains elusive, proactive management can control the disease's extent. Early detection is pivotal in averting its onset. This research employs Exploratory Data Analysis (EDA), coupled with SMOTE analysis, to unveil patterns, correlation, characteristics, and data structures. For diabetes classification, Support Vector Machine (SVM), Extreme Gradient Boosting (XG Boost). Random Forest (RF), Logistic Regression (LR) and Decision Tree (DT) optimized by Bees Optimization, were employed. Metrics like the F1 Score, ROC curve, accuracy, precision, and recall are used to carefully evaluate the model's performance. In order to determine the parameters that support classification, this model was tested using the PIMA Indian dataset and real-time datasets. For the real- time dataset with BOA, the SVM model scored an astounding 98.86% accuracy, but for the PIMA dataset, it only managed a 96% accuracy. As a result, this study proves that, in comparison to cutting-edge techniques, combining EDA with SMOTE and ML with BOA produces better outcome.

Keywords: investment efficiency; bibliometric literature review; corporate governance; thematic evolution; ESG

1. Introduction

In modern times, Diabetes Mellitus, commonly referred to as diabetes, stands as an increasingly prevalent and vital health concern. Recognized by the abbreviated term "ADS diabetes", it represents a condition requiring continual attention and treatment due to its widespread impact on human health. Globally, diabetes affects about 422 million people, according to World Health Organization data. Diabetes can manifest itself in a variety of ways. Type 1 is brought on by an attack by the immune system on insulin-producing cells, while Type 2 is brought on by a combination of insulin resistance, lifestyle choices, and genetic predisposition. Although a long-term solution is still unattainable, early identification is essential to protecting people from the severity of this illness.

Using a real-time dataset compiled from nearly a thousand participants, machine learning algorithms are widely used in disease prediction. Selecting an appropriate machine learning algorithm is a challenging task for researchers. The data were thoroughly preprocessed to remove null values, clean, balance, and identify correlated features. This included using Exploratory Data Analysis (EDA) with upsampling analysis. For efficient disease prediction, the improved dataset was subsequently fed into the ML techniques [1] XG Boost, Random Forest, Logistic Regression, Decision Tree and Support Vector Machine. As a global health priority, ongoing research endeavors seek to deepen our comprehension of diabetes, fostering the development of innovative treatments and improving the general standard of living for people impacted by this widespread and impactful condition. This

document is formatted as follows: A thorough summary of the literature is given in Section 2; Section 3 explored the research methodology; Section 4 explains the experimental results and discussions; and Section 5 summarizes the main conclusions and findings.

2. Literature Review

In the research conducted by Bala et al. (2020), they used Random Forest feature selection in conjunction with a Deep Neural Network to predict Diabetic Mellitus. The assessment was conducted using the PID dataset, yielding a remarkable accuracy of 98.16% [2].

RF, GBM, and LGBM were suggested by Shamreen et al. (2022) for the classification of diabetes mellitus. The experiment was conducted using PID and carefully selected data sets, resulting in a high accuracy identification of LGBM [3].

Weiyi et al. (2022) used exploratory data analysis (EDA) to analyze data in-depth in their most recent study. The authors used XGBoost, LGBM, Hybrid Random Forests, and Random Forests among other classification models. Interestingly, auto parameter tuning techniques were used in their experiment. The Hybrid Random Forests model exhibited superior performance compared to other models, showcasing a remarkable accuracy of 86.4% in the experimental results [4].

Rahman et al. 2023, examined risk factors linked to diabetes, such as polyuria, delayed healing, and polydipsia, Rahman looked into a number of machine learning algorithms, including Random Forest (RF), Multi-Layer Perceptron (MLP), and Support Vector Machines (SVM), LightGBM (LGBM), XGBoost (XGB), and Decision Trees (DT). During training, Random Forest remarkably achieved an astounding accuracy rate of 99.36% [5].

In their investigation of the PID dataset, Khanam et al. (2021) applied various classification algorithms, including Neural Network (NN), AdaBoost (AB), Support vector machines (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), Naive Bayes (NB), and Logistic Regression (LR). Notably, the Neural Network exhibited a high accuracy of 88.6%, surpassing both SVM and LR, which obtained accuracies between 77 and 78% in comparison to the other techniques [6].

Maniruzzaman et al. used odds ratios and p-values in their 2020 study to determine the diabetic risk factors using Logistic Regression (LR). Random Forest (RF), Decision Trees (DT), AdaBoost (AB), Naive Bayes (NB), and Random Forest (RF) were used as partition protocols to predict diabetes. Notably, with an AUC of 0.95 and an accuracy of 94.25%, RF showed exceptional performance. In terms of diabetes prediction, the combination of LR and RF produced a remarkably high accuracy [7]. Wee et al. investigated invasive and non-invasive diabetes mellitus datasets in their 2023 study. They performed extensive preprocessing, which included feature selection, class balancing, data imputation, and normalization of the data. The authors examined and analyzed nearly fifty Machine Learning (ML) and Deep Learning (DL) algorithms in detail for classification purposes [8].

Sivaranjani et al. (2021) combined backward and forward feature selection methods in their study by utilizing Random Forest (RF) and Support Vector Machines (SVM) algorithms. The accuracy rate of the Random Forest model was 83% [9].

Findings

It is clear from a survey of the literature that most classification efforts have not included optimization techniques in their work. often resulting in accuracy levels plateauing around 90%. In response to this, the current research employs Bee Optimization to enhance the classification accuracy, aiming to push beyond the common ceiling observed in prior studies.

3. Proposed Work

Identifying the best machine learning algorithms for diabetes mellitus is the main goal of this study. In order to do this, the study balances the dataset using exploratory data analysis (EDA) and upsampling analysis (Figure 1). This is followed by an evaluation of the dataset using the Decision Tree (DT), XG Boost, Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) models [10–12].

- Dataset

In this research classification study involved the utilization of both the PIMA Indian dataset [13], and a real-time dataset gathered through a Google Form distributed among the general population. Nearly 2000 data entries were collected. For experimental purposes, a few unrecorded columns were removed, and the retained features included Glucose, Blood Pressure, Smoking Habit, Alcoholic Habit, HbA1c level, Insulin, BMI, Age, and Outcome.

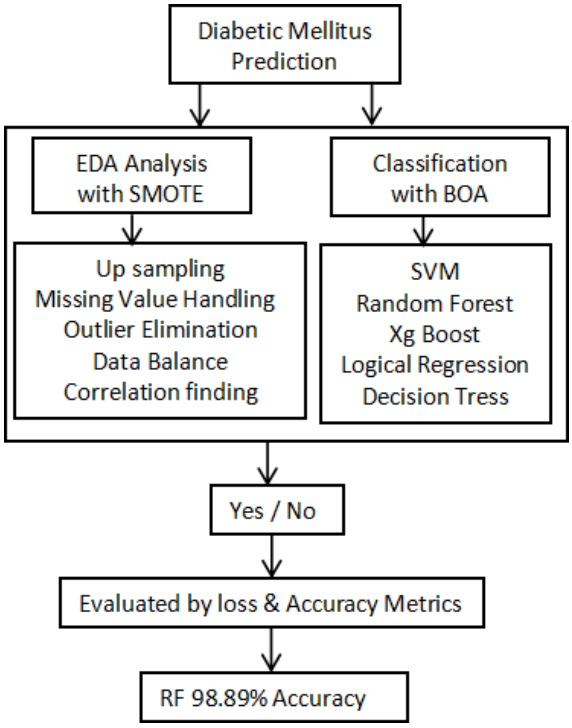


Figure 1. Proposed experiment flow.

3.1. Data Pre-Processing and Visualization

A crucial step in improving the caliber of data collected is pre-processing [14]. Cleaning, integration, transformation, and reduction are some of the processes used to address problems like missing values, outliers, redundant features, and formatting complexity. Up sampling the data is done using SMOTE analysis as well. The development of a model must be planned carefully at this stage in order to be effective. To illustrate intricate patterns and trends in the data, data visualization makes use of graphical formats.

3.2. Exploratory Data Analysis (EDA)

A thorough preprocessing approach for comprehending the nature of the data and resolving class imbalance combines Exploratory Data Analysis (EDA) with the Synthetic Minority Over-sampling Technique (SMOTE) [15].

By graphically examining statistical correlations and relationships between variables, EDA makes basic data analysis easier. Through visual aids, EDA provides insights into the dataset's underlying structure.

SMOTE, an oversampling technique, on the other hand, is essential for maintaining a balance between majority and minority classes. SMOTE addresses issues of class imbalance by creating artificial instances of the minority class, ensuring a more equitable distribution. By producing correlation matrices for both the original and over sampled datasets, this method facilitates a comprehensive comprehension of data patterns in various settings.

A critical stage in improving machine learning models is feature engineering, which converts unprocessed data into a format that improves the models' functionality. Categorical variables, which stand in for qualitative data like colors or types, are the subject of this process specifically. For these variables to be effectively incorporated into machine learning models, they require special handling.

By using data splitting, a dataset is split into training and test sets. The training set can be used to train a model., but it cannot perform well on new, unseen data, a situation known as over fitting. To avoid this, the testing set evaluates the model's performance. Twenty percent goes toward testing and eighty percent is used for training.

For datasets with features that span different ranges, scaling becomes essential. Because of its bigger scope, it allows no single feature to unduly impact the learning process. Unit variance scaling and mean removal are used to standardize, as well as min-max scaling, are two common methods for standardizing features., which normalizes values to a specified range (e.g., 0 to 1). The Min-Max Scaler is utilized in this project in order to scale (Figure 2).

	Glucose	BloodPressure	smoking habbit	Alcoholic habbit	HbA1c level	Insulin	BMI	Age	Outcome
0	148	72	1	1	1	0	33.6	50	1
1	97	66	0	0	0	0	26.6	31	0
2	183	64	0	0	0	0	23.3	32	1
3	100	66	0	0	0	94	28.1	21	0
4	137	40	1	1	1	168	43.1	33	1

Figure 2. Features in dataset.

3.3. Bee's Optimization Algorithm

Optimization helps to enhance the model performance [16,17], accuracy and abstraction of unseen data. Models must be optimized in order for them to perform well and adjust to the unique characteristics of different datasets. By mimicking honeybee gathering behavior, the Bee Optimization Algorithm (BOA), an optimization technique inspired by nature, efficiently explores and exploits solution spaces. The process consists of the four stages listed below to determine the best course of action.

(1) Initialize population of scout bees and employed bees

Assess each solution's fitness within the population.
While stopping criterion not met:

(2) Employed bee phase:

For each employed bee:
Choose a solution at random (a neighbor) from the population.
Change the existing solution to create a new one.
Determine whether the new solution is fit.
Update the current solution if the new solution is more fit.

(3) Onlooker bee phase:

For each onlooker bee:
Make a decision on a solution based on the likelihood that it will work.
Change the chosen solution to produce a new one.
Determine whether the new solution is fit.
Update the chosen solution if the new solution proves to be more fit.

(4) Scout bee phase:

Find solutions that, after a certain number of iterations, have not improved.
Put new, randomly generated solutions in place of these ones.
If needed, update the global best solution.
Return the best solution found.

3.4. Model Building

As mentioned before, building a model is combining different classification models and finding the model that has the best predictive power. This research experiment the Classification Methods as follows:

- Support Vector Machine (SVM)
- Random Forest (RF)
- XG Boost (XGB)
- Logistic Regression (LR)
- Decision Tree (DT)

3.4.1. Support Vector Machine (SVM)

Support vector machines, are a potent class of machine learning algorithms that find extensive use in tasks involving regression and classification [8]. Especially noteworthy for its efficacy in situations where a hyperplane can divide data points from different classes, SVM was first introduced by Vapnik and Cortes in the 1990s. Finding the best hyperplane to maximize the margin—that is, the distance between the hyperplane and the nearest data points from each class—is the main goal of support vector machines. The classification decision

function for Support Vector Machines (SVM) is determined based on the sign of the linear equation of the hyperplane. Given a data point a, the classification is performed using the decision function:

$$\text{Decision}(x) = \text{sign}\,(w \cdot a + b) \tag{1}$$

Here, the terms are as follows:

a—feature vector of input.

w—weight vector.

b—bias term.

The sign of w·a + b determines the class assignment. If w·a + b is positive, If the value is negative, the data point belongs to the other class; otherwise, it belongs to one class.

In mathematical terms:

$$\begin{gathered}\text{If } w \cdot a + b > 0, \text{ then} \\ \text{Decision}(x) = 1 \text{ (Positive class)} \\ \text{If } w \cdot a + b < 0, \text{ then} \\ \text{Decision}(x) = -1 \text{ (Negative class).}\end{gathered} \tag{2}$$

With the use of this decision function, SVM is able to categorize previously undiscovered data points using the support vectors and learned hyperplane. During Support Vector Machine (SVM) training, the goal is to ascertain optimal values for w and b, aiming to maximize the margin between classes while adhering to the constraints imposed by the training data.

3.4.2. Random Forest (RF)

Building multiple decision trees during training results in an output class that is the average of the classes (classification) of the individual trees. This is how the Random Forest classification ensemble learning technique operates. Traversing the tree from the root to a leaf node predicts a single decision tree, which is based on a specific feature, based on the input features.

The prediction of y_i of a single decision tree for an input feature vector xi is determined by the path it takes through the tree.

$$\text{RF} = \text{Decision Tree}\,(x_i) \tag{3}$$

Combining multiple decision trees make a final prediction

$$\text{Random Forest}\,(x_i) = \text{Decision Tree}\,(x_i) + \text{Decision Tree}\,(x_i) + \ldots\ldots \text{Decision Tree}\,(x_i) \tag{4}$$

3.4.3. Extreme Gradient Boosting (XG Boost)

XGBoost is an efficient and expandable gradient boosting implementation. The XGBoost classification formula for binary classification predicts the likelihood of the positive class ($p(y_{i=1})$)

$$p(y_{i=1}) = \frac{1}{1 + e^{-F(x_i)}}\pi r^2 \tag{5}$$

$P(y_{i=1})$ represents the estimated likelihood that instance I is a member of the positive class.

$F(x_i)$ is the XG Boost model's raw prediction score, for instance i. The individual tree scores in the ensemble are added together to get this score, which is weighted.

The raw prediction score $F(x_i)$ can be computed in the manner described below

$$F(x_i) = \sum_{k=1}^{K} \boldsymbol{f_k}\,(\boldsymbol{x_i}) \tag{6}$$

The total number of trees is K.

$\boldsymbol{f_k}(\boldsymbol{x_i})$ is the k-th tree's prediction.

The tree structure is used to compute the individual tree predictions, and for a given tree, the prediction for an instance (i) is based on which leaf node it belongs to. The final raw prediction score is derived from the combination of the leaf node predictions.

3.4.3. Decision Tree (DT)

By moving through the tree from the root node to a leaf node and using the majority class of the training instances in that leaf node as a guide, one can determine the predicted class label for a specific instance. The feature x_{ij} of the instance is compared to the threshold s at node t to determine the decision for an instance x_i.

$$\text{Decision at node t:if } x_{ij} <= s, \tag{7}$$

go to the left child; otherwise, go to the right child. Until a leaf node is reached, this process is repeated. A set of training instances are contained in the leaf node, and the majority class in that leaf node is usually used to predict the class label.

3.4.4. Logistic Regression (LR)

It simulates the likelihood that an instance will belong to a particular class. Sometimes the sigmoid function is used in place of the logistic function. It predicts the class 1 as follows

$$p\left(y = \frac{1}{x}\right) = \frac{1}{1 - e^{-\beta_n x_n + \beta_n x_n + \cdots \beta_n x_n}} \tag{8}$$

where $\mathrm{p}\left(\mathrm{y} = \frac{1}{\mathrm{x}}\right)$:s the positive class's probability in light of input feature x. e is the natural logarithm's base, $-\beta_n x_n + \beta_n x_n + \cdots \beta_n x_n$ are the coefficients (parameters) learned during training. $x_n + x_n + \cdots x_n$ are the input features.

3.4.5 Evaluation Metrics

F1 Score, Accuracy, Precision, and Recall were used to validate the model's performance. Precision highlights the accuracy of positive predictions, recall concentrates on finding all pertinent instances, and F1 Score finds a balance between recall and precision. Accuracy provides a comprehensive assessment of the model's correctness.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{9}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{10}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Positive} * 100 \tag{11}$$

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{12}$$

4. Result & Discussion

In this study, the research showcases the performance of crucial feature analysis.

The statistical analysis of the PIMA Indian dataset-1 and realtime dataset-2 is shown in the following figures (Figure 3).

Handling missing values (Figure 4) by filling 0's in the NaN palce.

Heatmap and correlation map used to identify the relationship among variable (Figure 5). From below observation this find the features Glucose, BMI, Insulin, Age is strongly correlated with outcome in dataset-1.

This study finds a strong correlation between diabetes and the features of glucose and insulin in Dataset-2, as supported by the observations listed below. Furthermore, smoking and alcohol consumption have strong relationships with diabetes and are important variables in the dataset.

	count	mean	std	min	25%	50%	75%	max
Glucose	147.0	127.795918	32.070970	71.0	103.0	119.0	147.0	197.0
BloodPressure	147.0	70.591837	20.348927	0.0	66.0	72.0	82.0	110.0
smoking habbit	147.0	0.224490	0.418672	0.0	0.0	0.0	0.0	1.0
Alcoholic habbit	147.0	0.360544	0.481800	0.0	0.0	0.0	1.0	1.0
Insulin	147.0	83.204082	153.211911	0.0	0.0	0.0	115.0	846.0
BMI	147.0	31.895918	7.950703	0.0	27.4	31.6	37.6	45.8
Age	147.0	37.551020	11.392249	21.0	29.0	33.0	48.0	60.0
Outcome	147.0	0.510204	0.501605	0.0	0.0	1.0	1.0	1.0

(a) Realtime dataset

	count	mean	std	min	25%	50%	75%	max
Glucose	1029.0	127.795918	31.977241	71.0	103.0	119.0	147.0	197.0
BloodPressure	1029.0	70.591837	20.289456	0.0	66.0	72.0	82.0	110.0
smoking habbit	1029.0	0.224490	0.417449	0.0	0.0	0.0	0.0	1.0
Alcoholic habbit	1029.0	0.360544	0.480392	0.0	0.0	0.0	1.0	1.0
Insulin	1029.0	83.204082	152.764140	0.0	0.0	0.0	115.0	846.0
BMI	1029.0	31.895918	7.927467	0.0	27.4	31.6	37.6	45.8
Age	1029.0	37.551020	11.358955	21.0	29.0	33.0	48.0	60.0
Outcome	1029.0	0.510204	0.500139	0.0	0.0	1.0	1.0	1.0

(b) PIMA dataset

Figure 3. Statistical Features in Dataset (**a**,**b**).

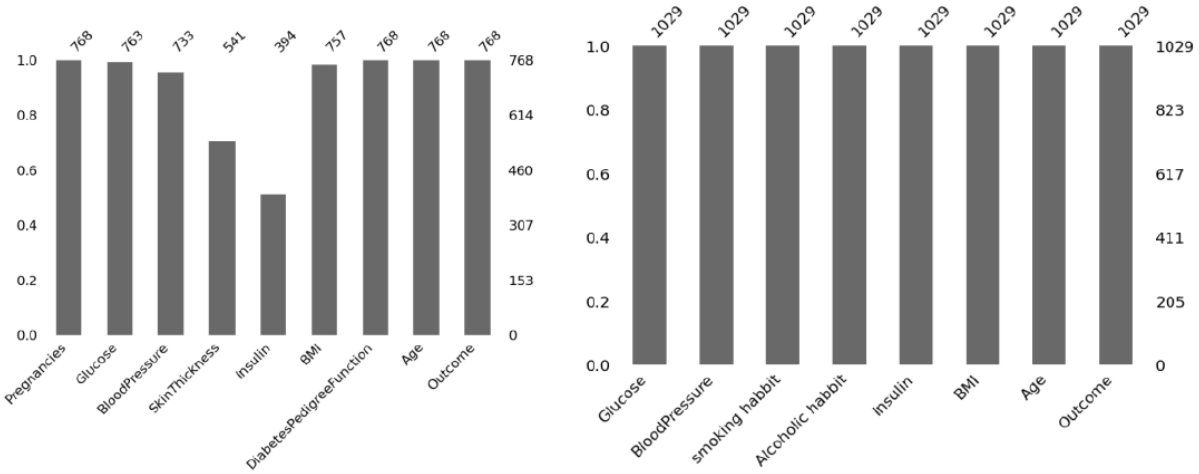


Figure 4. Handling missing values.

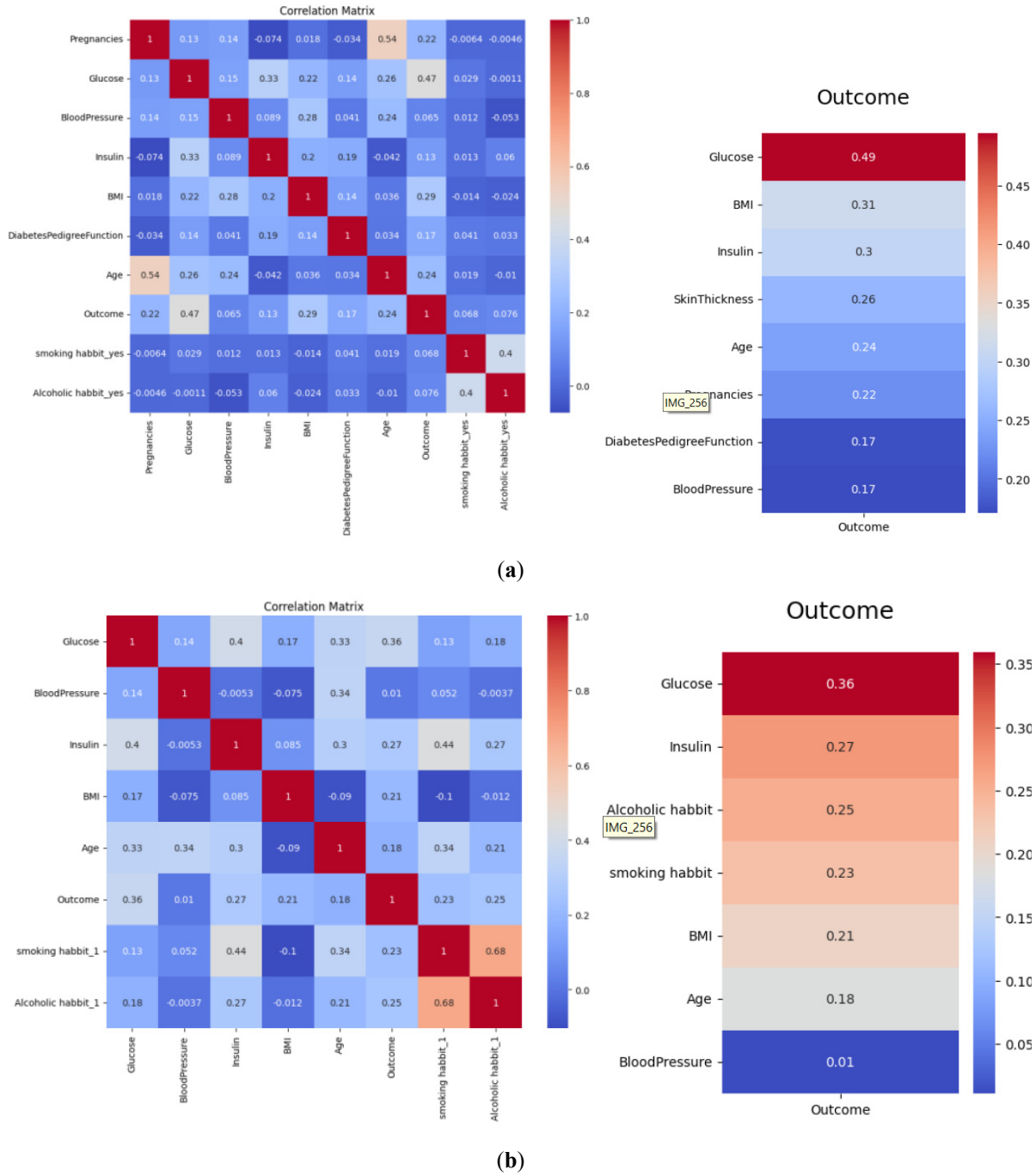


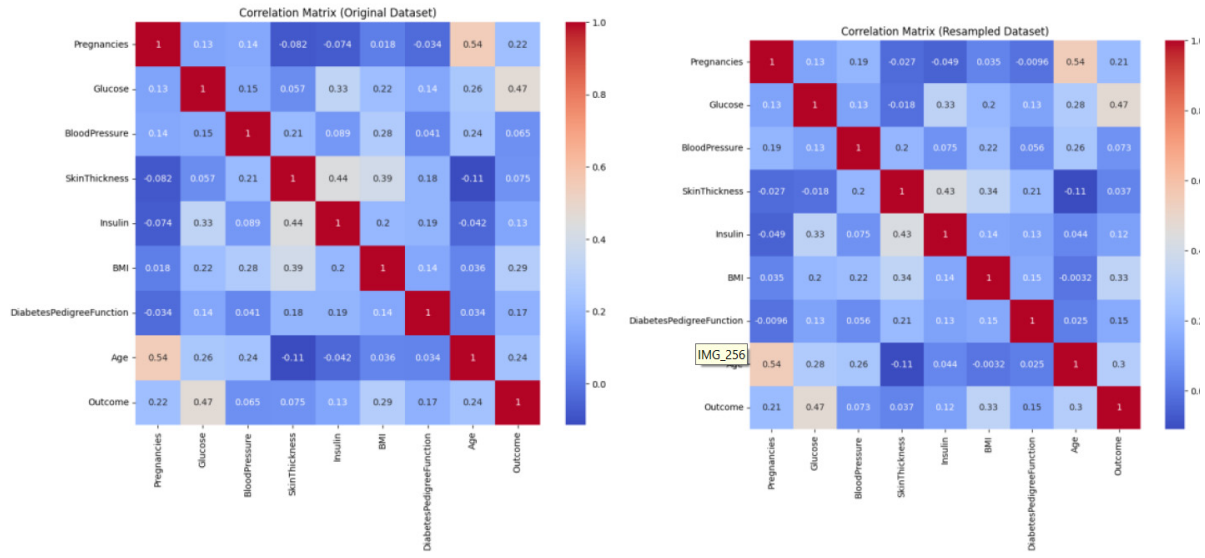
Figure 5. (**a**) Dataset-1 Correlation & Correlation among all variable with respect to Outcome. (**b**) Dataset-2 Correlation & Correlation among all variable with respect to Outcome.

SMOTE Analysis produce actual and upsampled correlation for experimented datasets (Figure 6).

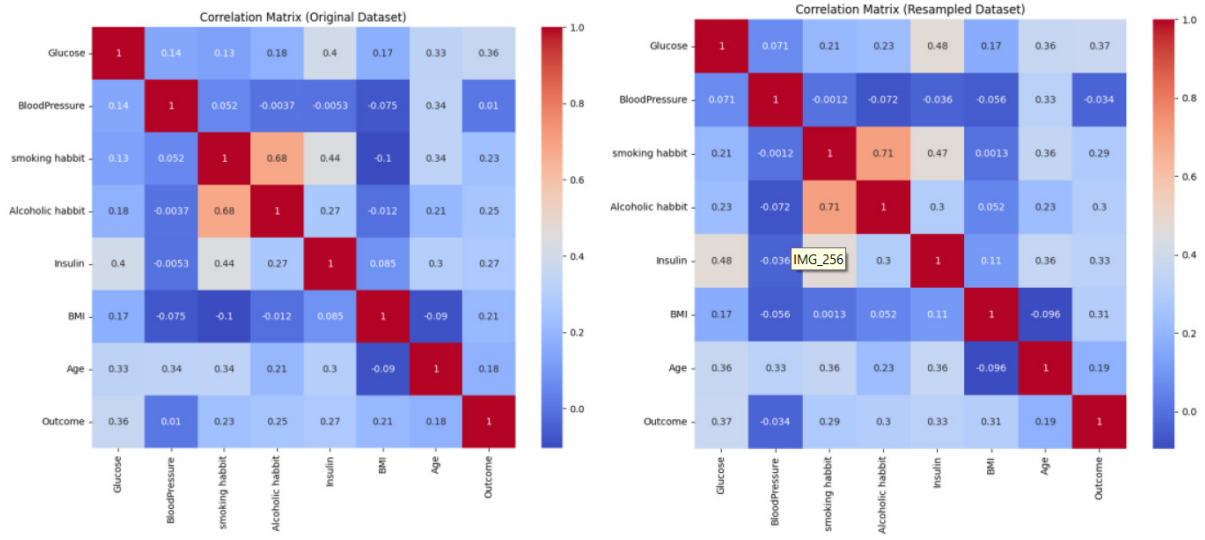
The depicted figure illustrates the ROC scores of Dataset-1 (Table 1).and Dataset-2 (Table 2), showcasing the performance of various classifiers. The results indicate that the real-time dataset achieved superior accuracy and ROC scores in comparison to Dataset-1. Among all classifiers, Random Forest (RF) outperformed others, achieving the highest accuracy.

The accuracy performance comparison of the tested models is shown in the presented figure, which shows that Dataset-2 produces better results than Dataset-1 (Figure 7).

A thorough overview of the combined performance of Datasets 1 and 2 is given in the illustrated in Figure 8.



(**a**)



(**b**)

Figure 6. (**a**) SMOTE Analysis for dataset-1. (**b**) SMOTE Analysis for dataset-2.

Table 1. ROC Score and Accuracy for Dataset-1.

Classifiers	ROC Score	F1 Score	Precision	Recall	Accuracy (%)
SVM	0.94	0.90	0.85	0.86	96
LR	0.92	0.83	0.81	0.83	95.27
RF	0.93	0.87	0.78	0.85	97.76
XG Boost	0.87	0.76	0.70	0.74	95.41
DT	0.85	0.65	0.55	0.62	91.88

Table 2. ROC Score and Accuracy for Dataset-2.

Classifiers	ROC Score	F1 Score	Precision	Recall	Accuracy (%)
SVM	0.96	0.94	0.89	0.87	97.82
LR	0.94	0.91	0.93	0.91	96
RF	0.97	0.93	0.91	0.87	98.89
XG Boost	0.95	0.87	0.84	0.82	97.01
DT	0.91	0.81	0.83	0.80	94.23

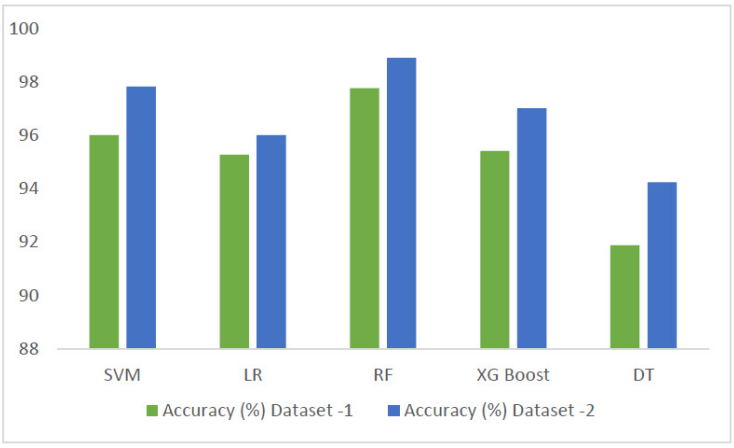
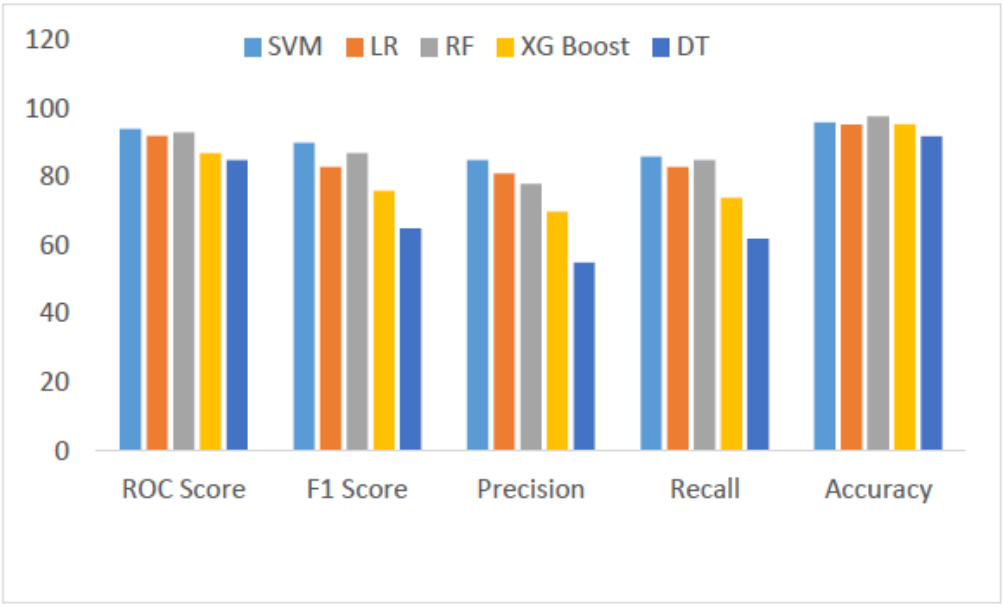
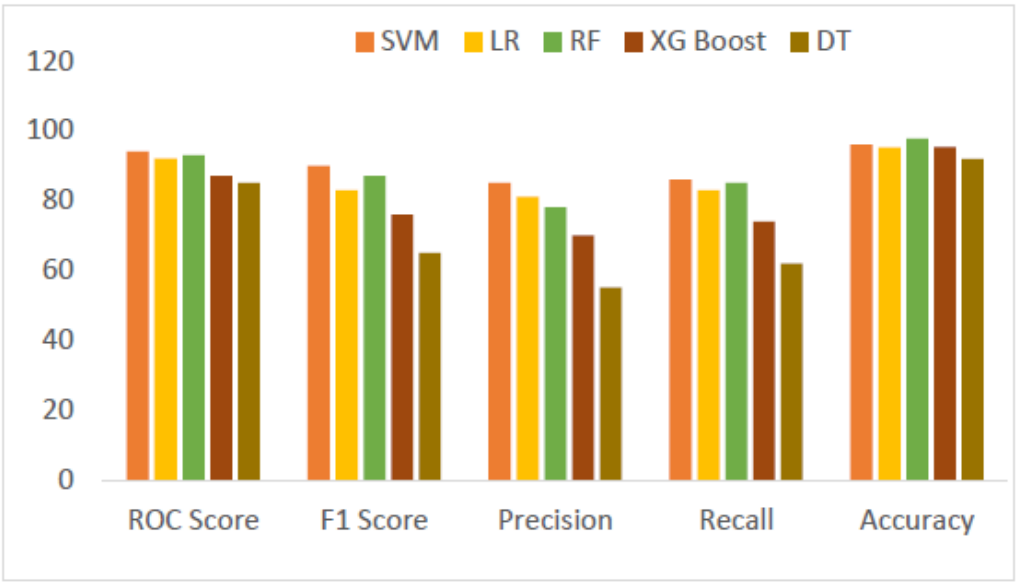


Figure 7. Accuracy performance comparison.



(**a**)

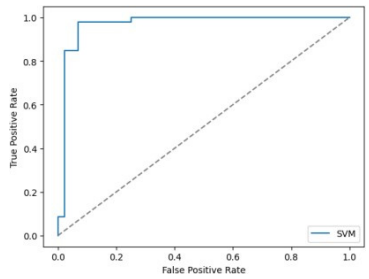
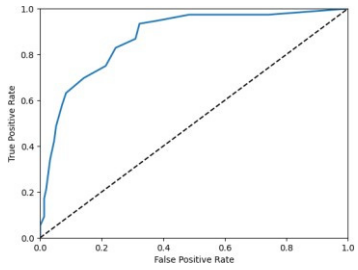
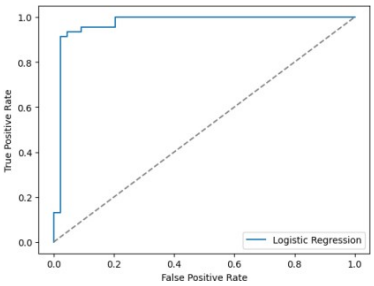
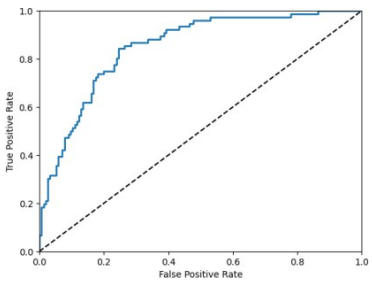
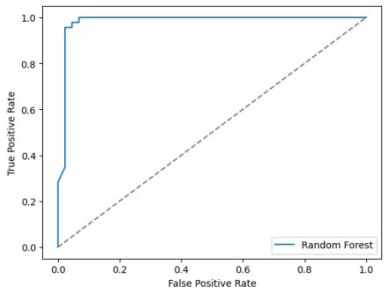
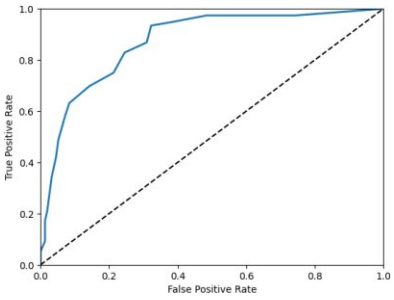
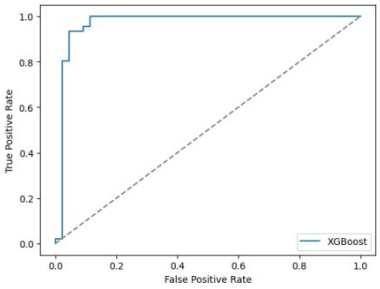
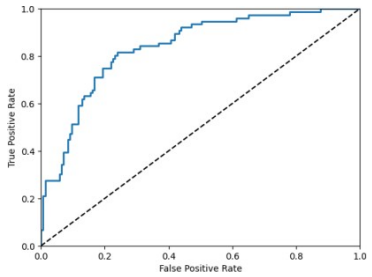
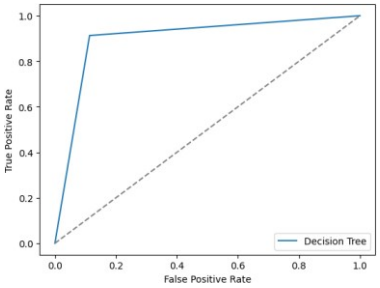
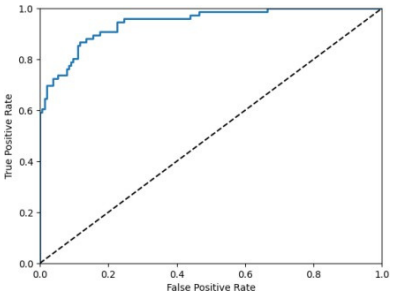


(**b**)

Figure 8. (**a**) Dataset-1 Model performance. (**b**) Dataset-2 Model performance.

The Table 3 presented below depicts a comparison of the ROC curves between Dataset-1 and Dataset-2.

Table 3. ROC curve comparison.

Classifiers	ROC for Dataset-2	ROC for Dataset-1
SVM		
LR		
RF		
XG Boost		
DT		

Findings

(1) The model's overall performance is significantly improved by implementing Bee's Optimization.
(2) Variations in the correlated features are seen between Datasets 1 and 2. The association between alcohol consumption and smoking has a significant effect on the analysis of diabetes, according to Dataset 2.
(3) To sum up, Dataset-2 achieves better ROC scores and accuracy than Dataset-1.

5. Conclusions

Diabetes mellitus consumes maximum people health due to changing food habbit and even children's easily get affected by it. So that to obtain supportive solution so many going to analyse the diabetic mellitus [18–20]. This research also try to predict the key factors which is reason for diabetes and provide better result for classification by machine learning algorithms. This research identifies several potential hypotheses for classification, and it is evident that combining Bees Optimization with ML algorithms presents an avenue for enhancing the classification process. This paper explore correlation features in dataset, by EDA and SMOTE analysis, hence based on highest correlation of outcome helps to select the important features. In Dataset-1, SVM achieved 96%, LR reached 95.27%, RF excelled with 97.76%, XG Boost attained 95.41%, and DT showed 91.88% accuracy. For Dataset-2, SVM led with 97.82%, LR held 96%, RF outperformed at 98.89%, XG Boost scored 97.01%, and DT achieved 94.23%. The findings highlight the SVM and RF models' superior performance on a variety of datasets. The real-time dataset, in particular, shows notably improved accuracy for RF in the diabetes mellitus classification.

Future Enhancement

Investigate deep learning algorithms to gain more insight into the different aspects related to diabetes. Determining the best course of action for the early diagnosis and classification of diabetes.

Author Contributions

R.H.K. and A.S.: conceptualization, methodology, software, data curation, writing—original draft preparation, visualization, investigation, supervision, validation. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

In this research classification study involved the utilization of both the PIMA Indian dataset [13], and a real-time dataset gathered through a Google Form distributed among the general population. Nearly 2000 data entries were collected.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Wee, B.F.; Sivakumar, S.; Lim, K.H.; et al. Diabetes detection based on machine learning and deep learning approaches. *Multimed. Tools Appl.* **2023**, *83*, 24153–24185. https://doi.org/10.1007/s11042-023-16407-5.
2. Bala Manoj Kumar, P.; Srinivasa Perumal, R.; Nadesh, R.K.; et al. Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. *Int. J. Cogn. Comput. Eng.* **2020**, *1*, 55–61. https://doi.org/10.1016/j.ijcce.2020.10.002.
3. Ahamed, B.S.; Arya, M.S.; Nancy, A.O.V. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation. *Adv. Hum.-Comput. Interact.* **2022**, *2022*, 9220560. https://doi.org/10.1155/2022/9220560.
4. Zhan, W. A Comparative Study on Machine Learning Based Type 2 Diabetes Mellitus Prediction. In Proceedings of the 2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022), Guangzhou, China, 28–30 October 2022. https://doi.org/10.2991/978-94-6463-108-1_95.
5. Rahman, M.A.; Abdulrazak, L.F.; Ali, M.M.; et al. Machine Learning-Based Approach for Predicting Diabetes Employing Socio—Demographic Characteristics. *Algorithms* **2023**, *16*, 503. https://doi.org/10.3390/a16110503.

6. Khanam, J.J.; Foo, S.Y. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* 2021, 7, 432–439. https://doi.org/10.1016/j.icte.2021.02.004.
7. Tigga, N.P.; Garg, S. Predicting type 2 Diabetes using Logistic Regression. In *Lecture Notes of Electrical Engineering*; Springer: Singapore, 2020.
8. Haffner, C.P.; Vapnik, V.N. Support vector machines for histogram-based image classification. *IEEE Trans. Neural Netw.* **1999**, *10*, 1055–1064. https://doi.org/10.1109/72.788646.
9. Maniruzzaman, M.; Rahman, M.J.; Ahammed, B.; et al. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf. Sci. Syst.* **2020**, *8*, 7. https://doi.org/10.1007/s13755-019-0095-z.
10. Banerjee, S. Machine Learning (ML) in Diet Planning for Type-1 Diabetes—An Overview. *J. Healthc. Treat. Dev. (JHTD)* **2022**, *2*, 1–5. https://doi.org/10.55529/jhtd25.1.5.
11. Zou, Q.; Qu, K.; Luo, Y.; et al. Predicting Diabetes Mellitus with Machine Learning Techniques. *Front Genet.* **2018**, *9*, 515. https://doi.org/10.3389/fgene.2018.00515.
12. Islam, M.R.; Banik, S.; Rahman, K.N.; et al. A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning. *Comput. Methods Programs Biomed. Update* **2023**, *4*, 100113. https://doi.org/10.1016/j.cmpbup.2023.100113.
13. Pima Indians Diabetes Database. Available online: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
14. Sivaranjani, S.; Ananya, S.; Aravinth, J.; et al. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 19–20 March 2021; Volume 1, pp. 141–146. https://doi.org/10.1109/ICACCS51430.2021.9441935.
15. Unwin, A. Exploratory Data Analysis. In *International Encyclopedia of Education,* 3rd ed.; Peterson, P., Baker, E., McGaw, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2010; pp. 156–161, ISBN 9780080448947. https://doi.org/10.1016/B978-0-08-044894-7.01327-0.
16. Sneha, N.; Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* **2019**, *6*, 13. https://doi.org/10.1186/s40537-019-0175-6.
17. Thamilarasi, V. Artificial Intelligence-Driven Smart Scenic Management: Automated Decision Making and Optimization. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 2731.
18. Kopitar, L.; Kocbek, P.; Cilar, L.; et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **2020**, *10*, 11981. https://doi.org/10.1038/s41598-020-68771-z.
19. Thamilarasi, V., Roselin. R. Automatic Classification and Accuracy by Deep Learning Using CNN Methods in Lung Chest, X.-Ray Image. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *1055*, 012099.
20. Chou, C.Y.; Hsu, D.Y.; Chou, C.H. Predicting the Onset of Diabetes with Machine Learning Methods. *J. Pers. Med.* **2023**, *13*, 406. https://doi.org/10.3390/jpm13030406.