Independent University **Bangladesh (IUB)**

IUB Academic Repository

Research Articles 2023

2023-05

SMOTE Oversampling and Near Miss Undersampling Based Diabetes Diagnosis from Imbalanced Dataset with XAI Visualization

Nayan, Nasim Mahmud

Independent University, Bangladesh

https://ar.iub.edu.bd/handle/11348/593

Downloaded from IUB Academic Repository

SMOTE Oversampling and Near Miss Undersampling Based Diabetes Diagnosis from Imbalanced Dataset with XAI Visualization

Nasim Mahmud Nayan*, Ashraful Islam†¶, Muhammad Usama Islam ‡, Eshtiak Ahmed†§,
Mohammad Mobarak Hossain*, Md Zahangir Alam¶
*Department of Computer Science and Engineering, University of Information Technology and Sciences, Bangladesh
†Center for Computational and Data Sciences, Independent University, Bangladesh
‡School of Computing and Informatics, University of Louisiana at Lafayette, USA
§Department of Computer Science and Engineering, Daffodil International University, Bangladesh
¶Department of Computer Science and Engineering, Independent University, Bangladesh
Corresponding Author: Ashraful Islam (ashraful@iub.edu.bd)

***Abstract*—This study investigated the predictive ability of ten different machine learning (ML) models for diabetes using a dataset that was not evenly distributed. Additionally, the study evaluated the effectiveness of two oversampling and undersampling methods, namely the Synthetic Minority Oversampling Technique (SMOTE) and the Near-Miss algorithm. Explainable Artificial Intelligence (XAI) techniques were employed to enhance the interpretability of the model's predictions. The results indicate that the extreme gradient boosting (XGB) model combined with SMOTE oversampling technique exhibited the highest accuracy and an F1-score of 99% and 1.00 respectively. Furthermore, the utilization of XAI methods increased the dependability of the model's decision-making process, rendering it more appropriate for clinical use. These results imply that integrating XAI with ML and oversampling techniques can enhance the early detection and management of diabetes, leading to better diagnosis and intervention.**

***Index Terms*—Machine learning, SMOTE, Near Miss, Oversampling, Undersampling, Diabetes, XAI, SHAP**

I. Introduction

The fourth industrial revolution (fourth IR) or Industry 4.0 has seen enormous growth in machine learning (ML), a subset of artificial intelligence (AI), which is generally considered as the most well-liked new technology. ML has made significant strides in the healthcare industry, especially in diabetes control and treatment, where ML has been used for predicting the onset of diabetes based on a person's genetic makeup, lifestyle, and other factors [1], [2].

There have been numerous studies conducted on diabetes prediction using different ML-based algorithms and methodologies. For instance, Devi et al. [3] investigated different mining strategies to predict diabetes, utilizing Random Forest, Decision Tree, Naïve Bayes, and J48 algorithms. Similarly, Yuvaraj et al. proposed a diabetes prediction application utilizing three different ML algorithms: Decision Tree, Random Forest, and Naïve Bayes [4], while Kandhasamy et al. compared ML classifiers, such as J48 Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVMs), to categorize patients with diabetes mellitus [5].

This research aimed to assess the efficacy of ten different ML models (i.e., Logistic Regression, Decision Tree, Linear Kernel SVM, Radial Basis Function (RBF) Kernel SVM, KNN, Gaussian Naïve Bayes (GNB), Extreme Gradient Boosting (XGB), Multilayer Perceptron (MLP), AdaBoost, and Random Forest) in predicting diabetes using an imbalanced dataset. Moreover, we combined oversampling and undersampling techniques like Synthetic Minority Oversampling Technique (SMOTE) and Near Miss respectively with the ten ML techniques to evaluate and compare the efficacy. The findings revealed that the model's performance significantly improved when SMOTE and Near Miss were incorporated, achieving accuracy and F1-score of 99% and 1.00 respectively., particularly for XGB with SMOTE model.

Additionally, the study aimed to investigate the use of Explainable Artificial Intelligence (XAI) methods for improving the interpretability of the employed models' predictions. The use of XAI methods facilitated a more comprehensive understanding of the model's decision-making process, enhancing its reliability and usefulness for clinical applications. These results suggest that the combination of XAI with ML and oversampling techniques has the potential to enhance the early detection and management of diabetes, resulting in improved patient outcomes. Unlike other works that solely prioritize achieving higher levels of accuracy, our study goes beyond this metric to also evaluate precision score, recall score, and the different types of errors encountered.

II. Materials and Methods

A. Dataset Description

In this study, we utilized a publicly available dataset of diabetes patients, sourced from Mendeley [6]. This dataset

included 1,000 patient records containing various health metrics, e.g., blood sugar levels, age, gender, creatinine ratio (Cr), body mass index (BMI), urea, cholesterol (Chol), fasting lipid profile, and HBA1C. The dataset was classified into three groups: Diabetic, Non-Diabetic, and Pre-Diabetic, with 844 instances of Diabetic, 103 instances of Non-Diabetic, and 53 instances of Pre-Diabetic. Table I describes a total of 14 attributes of the dataset briefly.

B. Data Pre-processing

In the first step of our data pre-processing procedure, we addressed the issue of missing values and duplicate data by removing them, specifically NaNs. To optimize the efficiency of subsequent processing steps, we employed label encoding to convert the data into a suitable format. Further, we standardized the data to eliminate the influence of varying scales of the different components on step sizes and to avoid unnecessary overhead. Standardization ensures a mean of 0 and a standard deviation of 1, which is beneficial for statistical analysis. In addition, outliers are preserved, making standardization a more appropriate choice than normalization. The standardization equation utilized in this study is represented by Equation 1.

$$X_{new} = \frac{X - \mu}{\sigma} \tag{1}$$

Here, the original data, the mean, and the standard deviation of the dataset are represented by X, μ, and σ respectively. X_{new} represents the standardized data after transformation.

TABLE I
AVAILABLE ATTRIBUTES OF THE DATASET

Serial No.	Attribute	Attribute Type	Details
1	ID	Integer	Identification No.
2	No_Pation	Integer	The number of patient
3	Gender	Object	Gender of a patient(Male or Female)
4	AGE	Integer	Age(years)
5	Urea	Float	Greater urea levels are linked to a greater risk of developing diabetes mellitus
6	Cr	Integer	Creatinine ratio
7	HbA1c	Float	Average blood glucose levels for the last two to three months
8	Chol	Float	Cholesterol (Diabetes and high cholesterol often occur together)
9	TG	Float	Triglycerides
10	HDL	Float	Lipoprotein (high-density)
11	LDL	Float	Lipoprotein (low-density)
12	VLDL	Float	Very-low-density lipoprotein
13	BMI	Float	BMI (weight in kilogram/height in square meter)
14	CLASS	Object	Class (Diabetic, Non-Diabetic, or Predict-Diabetic)

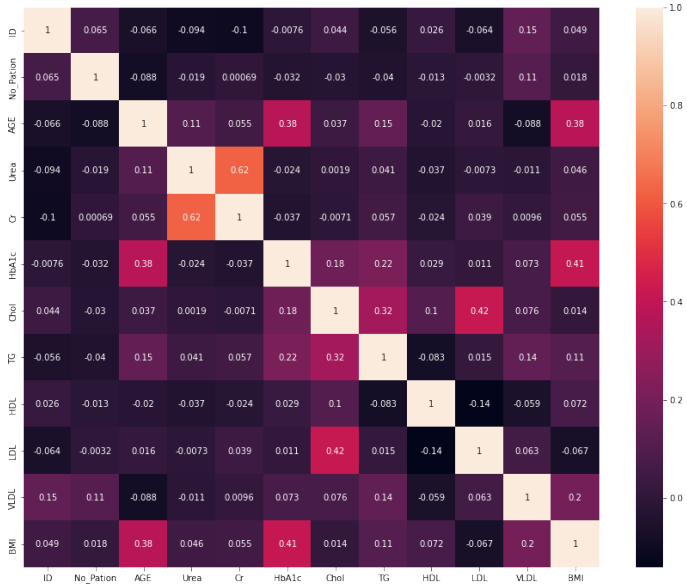


Fig. 1. Heat map for checking correlated columns/attributes in the dataset. Here, 'ID' represents the serial of instances, and 'No_Pation' stands for the respective number of patients in the dataset.

C. Exploratory Data Analysis

Our analysis involved a comprehensive exploratory data examination by implementing two widely recognized approaches: Heatmap and Correlations. The Heatmap approach is utilized to assess the degree of correlation between various factors, with the correlation coefficients being represented in the form of a heatmap. This method aids in identifying characteristics that are optimal for constructing machine learning models. The Heatmap approach transforms the correlation matrix into a color-coded representation. Fig. 1 provides meaningful insights into the dataset by generating a heatmap that illustrates the correlations between different variables.

D. Feature Selection

The feature selection process was conducted proficiently by utilizing a combination of several methods, including Featurewiz [7], the chi-square test [8], wrapper and filter methods [9], and expert domain knowledge.

In Featurewiz, a Python library (open-source), Searching for the Uncorrelated List of Variables (SULOV) is used to identify correlations between variables and Mutual Information Score (MIS) is utilized for quantifying information that can be obtained from one random variable given another variable. The least correlated variables and highest MIS scores are recursively passed to XGBoost to determine the optimal feature.

To assess the discrepancy between actual and predicted values for the independence of two occurrences, the chi-square test is frequently utilized. The formula utilized in our experiment is provided in equation 2, and it offers insights into our approach.

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where, c= degrees of freedom, O= observed values, E= Expected values;

Moreover, we utilized a combination of wrapper and filter methods in conjunction with expert domain knowledge to identify the features for the experimentation.

E. Class Imbalance Problem

Our dataset contains a class imbalance, thus to overcome this problem, we used two main strategies: SMOTE [10], which is an oversampling method, and the NearMiss algorithm [11], which involves an undersampling method.

F. Evaluation Metrics

1) Accuracy: In the evaluation of a model's performance, accuracy is often used as a metric to indicate the proportion of correctly predicted observations out of all observations. While accuracy is a useful indicator, it is important to note that it is only reliable when the false positive and false negative rates are similar across the dataset. In cases where there is a significant imbalance between the two rates, accuracy may not be the most informative metric to evaluate model performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

2) Precision: Precision is a performance metric that measures the proportion of correctly predicted positive observations to all predicted positive observations. It evaluates how frequently the model is accurate when making positive predictions. Precision becomes particularly useful when the costs of false positives are high. For example, when identifying the risk of diabetes, a model with poor precision would result in a large number of patients receiving a diagnosis of diabetes, including those who should not have been diagnosed, leading to unnecessary treatments and costs.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

3) Recall: Recall is a crucial performance metric that measures a model's ability to accurately identify all positive instances among the total number of actual positive instances. It quantifies the proportion of actual positive instances that the model correctly identifies as positive. Recall assumes significance when the cost of false negatives is substantial as it guarantees that all positive instances are identified correctly and not overlooked by the model. The formula for calculating recall is:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

4) F-1 Score: Taking into account erroneous positives and false negatives, the F-1 Score combines Precision and Recall. When dealing with an unbalanced class distribution, the F-1 Score is especially useful, but it may not be as simple to understand as accuracy. When the costs of false positives and false negatives are comparable, accuracy is higher; however, if the costs are significantly different, it is best to consider Precision and Recall combined.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{6}$$

5) Mean Absolute Error (MAE): By averaging the absolute variance across the entire data set, MAE represents the difference between the initial and anticipated values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}| \tag{7}$$

where N represents the number of observations, y_i represents the actual value, and $\hat{y}$ represents the predicted value.

6) Mean Squared Error (MSE): MSE indicates the difference between the original and predicted values that were calculated by squaring the mean difference throughout the data set.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{8}$$

where N represents the number of observations, y_i represents the actual value, and $\hat{y}$ represents the predicted value.

7) Mean Squared Logarithmic Error (MSLE): MSLE is a loss function in machine learning that evaluates the difference between actual and predicted values of a continuous target variable. It is similar to MSE, but instead of computing the squared difference between actual and predicted values, it calculates the squared difference between the logarithm of both values. Using logarithm is advantageous when the target variable has a wide range of values because it scales down the differences between large values and amplifies the differences between small values, resulting in greater sensitivity to small errors.

$$MSLE = \frac{1}{n} \sum_{i=1}^{n} \left(\log(y_i + 1) - \log(\hat{y}_i + 1)\right)^2 \tag{9}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

8) Root Mean Squared Error (RMSE): RMSE is the error rate calculated using the MSE's square root.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2} \tag{10}$$

III. Experiments and Findings

A. Experimental Setup

Our entire dataset was divided into two parts, one of which was utilized for model testing and the other for model training. We set aside 80% of the entire amount of data for the training dataset and used the remaining 20% for testing. To evaluate the effectiveness of our models, we employed a range of standards, including accuracy, precision, recall, and F-1 scores. We also used various error rates to properly visualize the results. In our results, we have included the following error metrics: MSE, MSLE, MAE, and RMSE, which are presented in Table VI. Additionally, we have also calculated the weighted precision, recall, and F-1 score for our experimental models and reported these values in Table VII.

B. Experimental Results

Table II provides the accuracy in the percentage of our experimented models. Table III, Table IV, and Table V provide the class-wise (Class 0: Non-Diabetic, Class 1: Pre-Diabetic, and Class 2: Diabetic) precision, recall, and F-1 score value of our experimented models respectively.

In our results, we have incorporated the following error values which are MSE, MSLE, MAE, and RMSE which can be visualized in Table VI.

Furthermore, we have calculated the weighted precision, recall, and F-1 score of our experiment and noted them down in table VII.

C. XAI Visualization

We evaluate the output of all algorithms and approaches and select the one (XGB+SMOTE) with the highest overall score based on accuracy, precision, and recall. This model is utilized in several XAI Visualizations. This explains models' precision, equity, openness, and results in decision-making supported by AI. Our XAI visualization is based on SHapley Additive exPlanations (SHAP) [12] and is illustrated in Fig. 2.

D. Comparison with Prior Research

Compared to two similar works conducted on the same dataset as this work and published earlier, our study contributed significantly and outperformed in terms of higher accuracy, number of employed ML models, and employing XAI visualization. In [13], Nuankaew et al. employed four ML models and obtained an accuracy of 98.95% for the average weighted objective distance method. In another similar work in [14], Rajput et al. employed five ML models and obtained an accuracy of 97.04% for stochastic gradient boosting. None of these two works presented XAI or any other visualization. Table VIII represents the comparison with the prior works.

TABLE II
TABLE DEPICTING THE ACCURACY (%) OF OUR EXPERIMENTED MODELS

Algorithm Name	Usual Model	With SMOTE	With NearMiss
Logistic Regression	92	90	88
Decision Tree	98	99	97
SVM(Linear)	93	92	88
SVM(RBF)	92	92	69
KNN	90	91	87
GNB	90	90	86
XGB	99	99	96
MLP	94	96	87
AdaBoost	95	91	85
Random Forest	94	94	95

TABLE III
TABLE DEPICTING THE CLASS-WISE (CLASS 0: NON-DIABETIC, CLASS 1: PRE-DIABETIC, AND CLASS 2: DIABETIC) PRECISION OF OUR EXPERIMENTED MODELS

Algorithm Name	Class 0	Class 1	Class 2
Logistic Regression	0.75	0.25	0.96
Decision Tree	0.95	0.90	0.99
SVM(Linear)	0.77	0.75	0.95
SVM(RBF)	0.75	0.67	0.94
KNN	0.70	0.33	0.95
GNB	0.60	0.55	0.97
XGB	0.95	1.00	1.00
MLP	0.82	0.67	0.97
AdaBoost	0.95	0.56	1.00
Random Forest	0.81	0.00	0.96
Logistic Regression + SMOTE	0.72	0.35	0.99
Decision Tree + SMOTE	0.95	1.00	0.99
SVM(Linear) + SMOTE	0.71	0.53	0.99
SVM(RBF) + SMOTE	0.71	0.53	0.99
KNN + SMOTE	0.70	0.41	0.99
GNB + SMOTE	0.61	0.64	0.96
XGB + SMOTE	0.95	1.00	1.00
MLP + SMOTE	0.86	0.78	0.98
AdaBoost + SMOTE	0.73	0.50	0.98
Random Forest + SMOTE	0.75	0.69	0.99
Logistic Regression + Near Miss	0.75	0.23	0.99
Decision Tree + Near Miss	0.95	0.71	1.00
SVM(Linear) + Near Miss	0.68	0.33	1.00
SVM(RBF) + Near Miss	0.78	0.12	1.00
KNN + Near Miss	0.65	0.21	0.99
GNB + Near Miss	0.48	0.47	0.99
XGB + Near Miss	0.95	0.59	1.00
MLP + Near Miss	0.73	0.30	0.99
AdaBoost + Near Miss	0.94	0.29	0.97
RandomForest + Near Miss	0.84	0.67	1.00

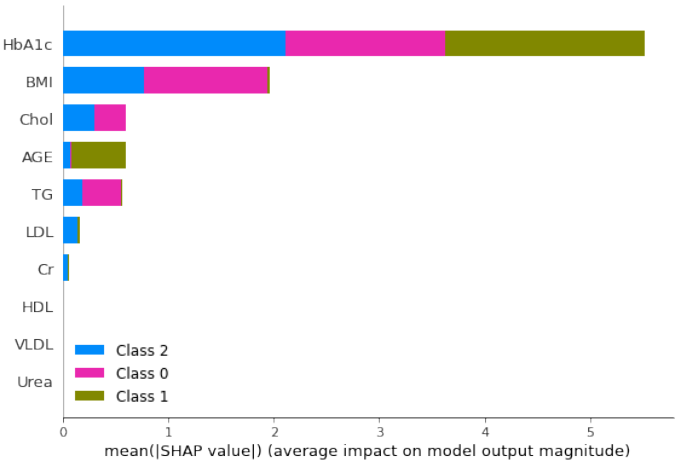


Fig. 2. The SHAP summary bar plot found during the XAI visualization. Here, Class 0, Class 1, and Class 2 represent Non-Diabetic, Pre-Diabetic, and Diabetic class labels respectively.

TABLE IV
TABLE DEPICTING THE CLASS-WISE (CLASS 0: NON-DIABETIC, CLASS 1: PRE-DIABETIC, AND CLASS 2: DIABETIC) RECALL OF OUR EXPERIMENTED MODELS

Algorithm Name	**Class 0**	**Class 1**	**Class 2**
Logistic Regression	0.86	0.10	0.98
Decision Tree	0.95	0.90	0.99
SVM(Linear)	0.81	0.30	0.98
SVM(RBF)	0.71	0.20	0.99
KNN	0.67	0.30	0.96
GNB	0.86	0.60	0.92
XGB	1.00	1.00	0.99
MLP	0.86	0.40	0.98
AdaBoost	1.00	1.00	0.95
Random Forest	1.00	0.00	0.99
Logistic Regression + SMOTE	0.86	0.60	0.92
Decision Tree + SMOTE	1.00	0.90	0.99
SVM(Linear) + SMOTE	0.95	0.80	0.92
SVM(RBF) + SMOTE	0.95	0.80	0.92
KNN + SMOTE	0.76	0.70	0.93
GNB + SMOTE	0.81	0.70	0.92
XGB + SMOTE	1.00	1.00	0.99
MLP + SMOTE	0.86	0.70	0.99
AdaBoost + SMOTE	0.90	0.90	0.91
Random Forest + SMOTE	1.00	0.90	0.93
Logistic Regression + Near Miss	0.86	0.50	0.91
Decision Tree + Near Miss	1.00	1.00	0.97
SVM(Linear) + Near Miss	0.90	0.80	0.88
SVM(RBF) + Near Miss	0.67	0.80	0.69
KNN + Near Miss	0.81	0.40	0.91
GNB + Near Miss	0.71	0.90	0.88
XGB + Near Miss	1.00	1.00	0.95
MLP + Near Miss	0.76	0.80	0.89
AdaBoost + Near Miss	0.76	1.00	0.85
RandomForest + Near Miss	1.00	1.00	0.95

TABLE V
TABLE DEPICTING CLASS-WISE (CLASS 0: NON-DIABETIC, CLASS 1: PRE-DIABETIC, AND CLASS 2: DIABETIC) F-1 SCORE OF OUR EXPERIMENTATION

Model	**Class 0**	**Class 1**	**Class 2**
Logistic Regression	0.80	0.14	0.97
Decision Tree	0.95	0.90	0.99
SVM(Linear)	0.79	0.43	0.97
SVM(RBF)	0.73	0.31	0.97
KNN	0.68	0.32	0.96
GNB	0.71	0.57	0.95
XGB	0.98	1.00	1.00
MLP	0.84	0.50	0.97
AdaBoost	0.98	0.71	0.97
Random Forest	0.89	0.00	0.97
Logistic Regression + SMOTE	0.78	0.44	0.95
Decision Tree + SMOTE	0.98	0.95	0.99
SVM(Linear) + SMOTE	0.82	0.64	0.96
SVM(RBF) + SMOTE	0.82	0.64	0.96
KNN + SMOTE	0.73	0.52	0.96
GNB + SMOTE	0.69	0.67	0.94
XGB + SMOTE	0.98	1.00	1.00
MLP + SMOTE	0.86	0.74	0.99
AdaBoost + SMOTE	0.81	0.64	0.94
Random Forest + SMOTE	0.86	0.78	0.96
Logistic Regression + Near Miss	0.80	0.31	0.95
Decision Tree + Near Miss	0.98	0.83	0.98
SVM(Linear) + Near Miss	0.78	0.47	0.93
SVM(RBF) + Near Miss	0.72	0.21	0.81
KNN + Near Miss	0.72	0.28	0.94
GNB + Near Miss	0.58	0.62	0.93
XGB + Near Miss	0.98	0.74	0.98
MLP + Near Miss	0.74	0.43	0.94
AdaBoost + Near Miss	0.84	0.45	0.91
RandomForest + Near Miss	0.91	0.80	0.97

IV. CONCLUSION AND FUTURE DIRECTIONS

In summary, this study has shown the promise of ML, oversampling, undersampling and XAI approaches for diabetes prediction. The outcomes of our experiment demonstrate that the models had higher accuracy, precision, and recall in their ability to predict diabetes, and XGD with SMOTE performed the best. Furthermore, the XGD with SMOTE model's decision-making process was better understood and more reliable for clinical use thanks to the application of XAI methodologies such as SHAP. These results imply that XAI and ML have the potential to enhance the early detection of diabetes and management, improving all stakeholder outcomes in the process. To further validate these findings, additional studies should concentrate on analyzing the model's performance over different datasets with larger and more diverse patient groups.

REFERENCES

[1] S. SK, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," *Journal of Medical Systems*, vol. 41, pp. 1–12, 2017.

[2] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, pp. 1–22, 2021.

[3] M. R. Devi and J. M. Shyla, "Analysis of various data mining techniques to predict diabetes mellitus," *International journal of applied engineering research*, vol. 11, no. 1, pp. 727–730, 2016.

[4] N. Yuvaraj and K. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster," *Cluster Computing*, vol. 22, no. Suppl 1, pp. 1–9, 2019.

[5] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science*, vol. 47, pp. 45–51, 2015.

[6] A. Rashid, "Diabetes dataset," Jul 2020, last accessed on 01 March 2023. [Online]. Available: https://data.mendeley.com/datasets/wj9rwkp9c2/1

[7] AutoViML, "Autoviml/featurewiz: Use advanced feature engineering strategies and select best features from your data set with a single line of code." 2020, last accessed on 01 March 2023. [Online]. Available: https://github.com/AutoViML/featurewiz

[8] R. J. Tallarida, R. B. Murray, R. J. Tallarida, and R. B. Murray, "Chi-square test," *Manual of pharmacologic calculations: with computer programs*, pp. 140–142, 1987.

[9] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006*. Springer, 2006, pp. 731–740.

[12] M. U. Islam, M. Mozaharul Mottalib, M. Hassan, Z. I. Alam, S. Zobaed, and M. Fazle Rabby, "The past, present, and prospective future of xai: A comprehensive review," *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*, pp. 1–29, 2022.

[13] P. Nuankaew, S. Chaising, and P. Temdee, "Average weighted objective distance-based method for type 2 diabetes prediction," *IEEE Access*, vol. 9, pp. 137 015–137 028, 2021.

[14] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis

TABLE VI
TABLE DEPICTING MSE, MSLE, MAE, AND RMSE OF OUR EXPERIMENTATION

Model	**MSE**	**MSLE**	**MAE**	**RMSE**
Logistic Regression	0.14	0.04	0.1	0.37
Decision Tree	0.05	0.01	0.03	0.22
SVM(Linear)	0.16	0.05	0.1	0.4
SVM(RBF)	0.19	0.06	0.11	0.43
KNN	0.21	0.06	0.14	0.45
GNB	0.29	0.08	0.17	0.53
XGB	0.02	0.01	0.01	0.14
MLP	0.12	0.04	0.08	0.35
AdaBoost	0.06	0.01	0.05	0.24
Random Forest	0.09	0.03	0.07	0.3
Logistic Regression + SMOTE	0.18	0.05	0.13	0.42
Decision Tree + SMOTE	0.03	0.01	0.02	0.16
SVM(Linear) + SMOTE	0.19	0.05	0.12	0.43
SVM(RBF) + SMOTE	0.19	0.05	0.12	0.43
KNN + SMOTE	0.19	0.07	0.13	0.43
GNB + SMOTE	0.32	0.09	0.18	0.56
XGB + SMOTE	0.02	0.01	0.01	0.14
MLP + SMOTE	0.09	0.03	0.06	0.29
AdaBoost + SMOTE	0.23	0.06	0.14	0.48
Random Forest + SMOTE	0.17	0.05	0.10	0.40
Logistic Regression + Near Miss	0.15	0.04	0.13	0.39
Decision Tree + Near Miss	0.04	0.01	0.03	0.2
SVM(Linear) + Near Miss	0.23	0.06	0.16	0.48
SVM(RBF) + Near Miss	0.34	0.07	0.32	0.58
KNN + Near Miss	0.21	0.06	0.17	0.45
GNB + Near Miss	0.40	0.12	0.23	0.63
XGB + Near Miss	0.06	0.01	0.05	0.23
MLP + Near Miss	0.21	0.06	0.16	0.45
AdaBoost + Near Miss	0.24	0.06	0.18	0.49
RandomForest + Near Miss	0.11	0.03	0.07	0.32

using medical attributes: A machine learning approach," *Journal of Xian University of Architecture and Technology*, 2022.

TABLE VII
TABLE DEPICTING THE WEIGHTED PRECISION, RECALL AND F-1 SCORE. HERE, P_w IS WEIGHTED PRECISION, R_W IS WEIGHTED RECALL, AND $F1S_w$ IS WEIGHTED F-1 SCORE.

Algorithm Name	P_w	R_W	$F1S_w$
Logistic Regression	0.90	0.92	0.91
Decision Tree	0.98	0.98	0.98
SVM(Linear)	0.92	0.93	0.92
SVM(RBF)	0.91	0.92	0.91
KNN	0.90	0.90	0.90
GNB	0.91	0.90	0.90
XGB	1.00	0.99	1.00
MLP	0.93	0.94	0.94
AdaBoost	0.97	0.95	0.96
Random Forest	0.90	0.94	0.92
Logistic Regression + SMOTE	0.93	0.90	0.91
Decision Tree + SMOTE	0.99	0.99	0.99
SVM(Linear) + SMOTE	0.94	0.92	0.93
SVM(RBF) + SMOTE	0.94	0.92	0.93
KNN + SMOTE	0.93	0.91	0.91
GNB + SMOTE	0.91	0.90	0.90
XGB + SMOTE	**1.00**	**0.99**	**1.00**
MLP + SMOTE	0.96	0.96	0.96
AdaBoost + SMOTE	0.93	0.91	0.91
Random Forest + SMOTE	0.95	0.94	0.94
Logistic Regression + Near Miss	0.93	0.88	0.90
Decision Tree + Near Miss	0.98	0.97	0.98
SVM(Linear) + Near Miss	0.93	0.88	0.89
SVM(RBF) + Near Miss	0.93	0.69	0.77
KNN + Near Miss	0.91	0.87	0.89
GNB + Near Miss	0.91	0.86	0.88
XGB + Near Miss	0.97	0.96	0.96
MLP + Near Miss	0.93	0.87	0.89
AdaBoost + Near Miss	0.93	0.85	0.88
RandomForest + Near Miss	0.97	0.95	0.96

TABLE VIII
COMPARING OUR RESULT WITH PRE-EXISTING WORKS ON THIS DATASET

Research Work	**Accuracy**	**How Many ML Algorithms Employed**	**Is XAI Available?**
Paper-1 ([13])	98.95%	4	No
Paper-2 ([14])	97.04%	5	No
Our Work	99%	10	Yes