ISSN 2723-6471

Early Detection of Female Type-2 Diabetes using Machine Learning and Oversampling Techniques

Lana Al-Dabbasa[1], Ahmad Adel Abu-Sharehaa[2,*]

[1,2] *Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan*

(Received: May 20, 2024; Revised: June 19, 2024; Accepted: July 24, 2024; Available online: August 10, 2024)

Abstract

Early diabetes prediction is crucial as it can save numerous lives and prevent diabetes-related complications. The experiments conducted on diabetes prediction are keen on the limited samples of diabetes and non-diabetes cases provided in the available dataset. Various techniques have been implemented, focusing on the classification technique to improve the accuracy of prediction results. As a significant technique, oversampling has been implemented using SMOTE, which improved the results yet posed limitations due to its naïve technique. In this paper, a framework for diabetes prediction is developed, integrating an advanced oversampling technique using SVMSMOTE with various machine-learning algorithms to achieve the best performance. The proposed framework aims to overcome the problem of inaccurate data and limited samples using preprocessing and oversampling techniques. Besides, these techniques are integrated with other data mining and machine learning algorithms to improve the performance of diabetes prediction. The framework consists of four main stages: data exploration, data preprocessing, data oversampling, and classification. The experiments were conducted on the Pima Indian diabetes dataset, which comprises 768 samples and 9 columns. The results showed that the proposed framework achieved an accuracy of 91%, which improved the accuracy compared to using classification without oversampling, which achieved an accuracy of 90%. In comparison, the best results addressed in the literature were an accuracy of 85.5%. As such, the proposed framework improves the results by approximately 6.4% compared to the existing frameworks. Besides, the proposed framework achieved the best f-measure using the XGBoost classifier and SVMSMOTE, equal to 0.879. The best recall was achieved using RF and SVMSMOTE, which was 0.931. Finally, the best precision was achieved using FR without oversampling, with a value of 0.918.

Keywords: Diabetes, Classification, Oversampling

1. Introduction

Diabetes is a non-communicable disease that severely affects and significantly influences the functionality of the body's systems [1], [2]. Diabetes causes higher blood glucose levels than healthy bodies [3]. For example, the American Diabetes Association states that a typical fasting blood glucose level is below 100 mg/dL. In contrast, a fasting blood glucose level of 126 mg/dL or higher indicates diabetes. Similarly, a typical level two hours after eating is below 140 mg/dL, and a level of 200 mg/dL or higher suggests diabetes. Glucose, a type of sugar, is essential for optimal metabolism in the body and serves as a source of energy for all cells in the body. Increased blood glucose levels due to the lack of insulin hormone leads to severe damage to various body parts, including the eyes, heart, kidneys, and many more [4]. The number of diabetes cases worldwide is increasing rapidly, calling for an urgent need for effective techniques to manage and mitigate the risk of this disease [5].

Generally, diabetes is classified into two categories: Type-1 and Type-2. Type-1 diabetes is common among individuals below 40 years old. Clinical symptoms for Type-1 include increased thirst and frequent urination. Type-1 is a critical condition that cannot be managed by oral medication, thus requiring insulin therapy and is therefore called insulin-dependent or juvenile diabetes. On the other hand, Type-2 diabetes is common in middle-aged individuals. Those with Type-2 diabetes produce insulin, but it does not work efficiently. Type-2 diabetes is often associated with arteriosclerosis, obesity, dyslipidemia, and hypertension [6]. Type-2 diabetes mellitus is more common among men than women [7]. However, Kautzky-Willer, et al. [8] reported that women with Type-2 diabetes have a greater relative risk of cardiovascular disease (CVD) and mortality compared to men.

***Corresponding author: Ahmad Adel Abu-Sharehaa (a.abushareha@ammanu.edu.jo)**
DOI: https://doi.org/10.47738/jads.v5i3.298

ISSN 2723-6471

The demands for Female Type-2 diabetes detection are multi-fold: 1) the risk of CVD, as mentioned earlier. 2) the widespread of the disease. The International Diabetes Federation (IDF) revealed that there are currently 537 million adults aged 20-79 living with diabetes, projected to increase to 643 million by 2030 and 783 million by 2045 [7]. 3) Early diagnosis greatly influences controlling the disease's consequences. Evidence indicates that diabetes may manifest for 4-12 years before an official diagnosis [9]. During this period, the immunity of diabetic individuals gradually diminishes, making them susceptible to various diseases, with cardiovascular conditions being a primary cause of mortality [6]. Early diabetes prediction is crucial as it can save numerous lives and prevent diabetes-related complications. Diabetes detection helps in preventing heart diseases, blindness, vascular complications, stroke, kidney failure, and limb amputations [10]. Besides, early prediction plays a pivotal role in mitigating the overall impact of the disease, thereby enhancing the quality of life for patients [11]. Like many other diseases, diabetes is characterized by multiple tests and examinations. The results of these tests are reported and saved, forming a valuable historical dataset that can be used to improve detection processes. Such extensive data serves as a valuable resource for analyzing and predicting diabetes at an early stage.

Accordingly, while various datasets have been collected for diabetes research, the Pima Indian Diabetes dataset specifically addresses female Type-2 diabetes. The data were collected from residents of Mexico and Arizona, USA, a Native American community identified for their elevated prevalence of diabetes [12]. Consequently, studying this group is crucial and reflective of global health trends. The Pima Indian Diabetes dataset, focusing on females aged 21 years and older, is widely recognized as a benchmark dataset in diabetes research. The valuable data generated from diabetes-related tests and examinations presents an opportunity to utilize advanced data science and artificial intelligence techniques. Machine learning (ML) algorithms can be trained on historical data to identify patterns and correlations that are not immediately apparent to human observers. For instance, predictive models can be developed to analyze blood glucose levels, HbA1c readings, insulin sensitivity, and other relevant biomarkers related to diabetes, thereby enhancing early detection of diabetes. Moreover, using such data enables the creation of personalized medicine approaches, where individual patient data can be used to tailor treatment plans specific to their unique physiological responses and health history. ML can help identify risk factors and comorbidities associated with diabetes, providing a comprehensive understanding of the disease's progression [13].

While various techniques have been implemented, the classification technique has been focused on improving the accuracy of prediction results. Oversampling, a significant technique used to solve the Imbalance problem in the dataset, has been implemented using the Synthetic Minority Over-sampling Technique (SMOTE) [14]. This paper proposes a framework for diabetes prediction and provides insights into the main factors that cause this health condition. The framework uses various classification algorithms and an advanced oversampling using Support Vector Machine (SVM)-based SMOTE (SVM-SMOTE). The structure of this paper is as follows: Section 2 focuses on the literature review. Section 3 discusses the proposed framework and its components: exploration, preprocessing, oversampling and classification. Section 4 elucidates the methodology employed for dataset analysis and evaluations. Lastly, Section 5 presents the concluding remarks and discussions on future research directions.

2. Literature Review

The existing diabetes prediction approaches use different machine-learning algorithms to improve the results. Deep Learning, particularly Convolutional Neural Network (CNN), has shown promise in achieving high accuracy using the early-stage diabetes risk prediction dataset, based on the research conducted by Ergun and Ilhan [11]. Meanwhile, the Pima Indians Diabetes Dataset commonly resulted in lower accuracy than the early-stage diabetes risk prediction dataset. Karegowda, et al. [15] used a back propagation network (BPN) optimized using a genetic algorithm (GA) for classifying the diabetes samples without any oversampling technique. The results showed that the accuracy of the proposed approach was 84.07%. Wei, et al. [16] used a Deep Neural Network (DNN), and the results showed that the accuracy of the DNN was 77.86%. The results of Karegowda, et al. [15] and Wei, et al. [16] cannot be compared, as the former used 40% of the data for testing and computing the results, while the latter used cross-validation of 10-fold.

Reza, et al. [17] used SMOTE with SVM for diabetes prediction. The effect of SMOTE was demonstrated as the results of the experimented approach were improved compared to Wei, et al. [16], as both used cross-validation to obtain the results. Accordingly, the results showed that the accuracy of the SVM classifier with SMOTE oversampling was 85.5%.

ISSN 2723-6471

Accordingly, the results of the approach proposed by Perdana, et al. [18] using KNN without oversampling was 83.12%. The results were obtained by splitting the dataset into 90% and 10% for training and testing. As such, the reviewed recent literature indicated the following: 1) Using SMOTE improved the results of diabetes prediction. 2) Various machine learning were utilized, and SVM has shown potential in predicting diabetes. 3) The evaluation procedure among the existing techniques is not unified, as cross-validation and percentage split were utilized. A summary of the reviewed literature is given in table 1.

Table 1. Description of The Relevant Studies Conducted in The Context of Diabetes Prediction

Ref.	Classifier	Oversampling	Accuracy	Dataset Evaluation	Dataset
[10]	CNN	None	99.04 %	10-fold cross-validation	Early-stage
[11]	BPN	None	84.07%	Train and Test split using a 60-40 ratio	PIMA
[12]	DNN	None	77.86%	10-fold cross-validation	PIMA
[13]	SVM	SMOTE	85.5%	10-fold cross-validation	PIMA
[14]	KNN	**None**	**83.12%**	Train and Test split using a 90-10 ratio	PIMA

3. The Proposed Framework

The proposed framework for diabetes prediction using the Pima dataset consists of four components, as illustrated in figure 1. The handled problem is formulated as a binary classification problem (diabetic or not). The components of the framework are described accordingly.

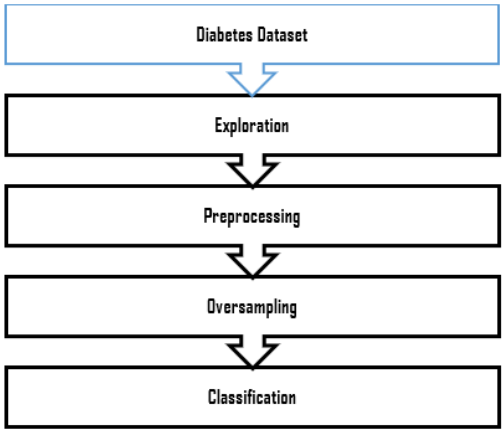

Figure 1. The Components of the Proposed Framework

3.1. Dataset Exploration

The PIMA dataset [19] was collected by the National Institute of Diabetes, Digestive, and Kidney Diseases (NIDDK). The data can be downloaded from Kaggle and UCI data repositories. PIMA comprises samples representing females of Pima Indian heritage who are of a minimum age of 21. The dataset consists of 768 instances and 9 columns. Table 2 describes the columns of the PIMA Indian Diabetes dataset in terms of the name, data type, and mean.

Table 2. Description of the PIMA Indian Diabetes Dataset

Variable Name	Data Type	Mean
Pregnancies	Integer	3.84
Glucose	Integer	120.89
Blood Pressure	Integer	69.10
Skin Thickness	Integer	20.53
Insulin	Integer	79.79

BMI	Float	31.99
DPF	Float	0.47
Age	Integer	33.24
Outcome	Integer	0.34

The PIMA dataset exploration showed that, on average, participants have 3.84 pregnancies and a glucose level of 120.89 mg/dL, indicating a risk for diabetes. The average blood pressure is 69.10 mm Hg, which falls within the normal range, while the mean BMI is 31.99, suggesting that most participants are obese, a known risk factor for diabetes. Skin thickness and insulin levels at an average of 20.53 mm and 79.79 mu U/ml, respectively, with considerable variability likely indicating diverse profiles. The diabetes pedigree function (DPF) mean is 0.47, reflecting a significant familial risk. The mean age of participants is 33.24 years, and 34% of them have diabetes, underscoring a high prevalence of the disease. These observations highlight the influence of multiple factors, such as obesity, high glucose levels, and family history, on diabetes prevalence in this population.

3.2. Data Preprocessing

The exploration step revealed the diversity of the column values in the PIMA dataset. Accordingly, the dataset underwent scaling using a min-max scaler to augment its overall quality and integrity. The min-max scaler is calculated as given in (1). The results of this scaler will be a value in the range [0-1], which unifies the variation of the columns in the dataset.

$$x' = (x - x_{min})/(x_{max} - x_{min}) \quad (1)$$

Min-max scaling is particularly useful to ensure that all features contribute equally, as the dataset does not follow a Gaussian distribution (see figure 2). In contrast, standard scaling, which standardizes data to have a mean of 0 and a standard deviation of 1, is more suitable for datasets with a Gaussian distribution. Normalization, which adjusts values to a standard range, can be less effective as there are outliers presented in the dataset.

The outliers in the dataset are also highlighted using the box plot, as illustrated in figure 2. The insulin column seems to have some outliers that need to be processed. Besides, all columns in the PIMA dataset exhibited zero values, which forms errors in capturing or entering the data. These values were replaced with the median of the column to solve these issues. Replacing zero values with the median might seem simplistic, but it is a well-considered choice for the small dataset. Given the limited size of the dataset, removing outliers could lead to significant data loss, which is not desirable. The median is a robust statistic less sensitive to extreme values than the mean, making it an effective method for handling outliers without distorting the overall data distribution. Transforming outliers acknowledges their benefits, but the decision to use the median ensures that the integrity of the dataset is maintained while effectively managing outliers.

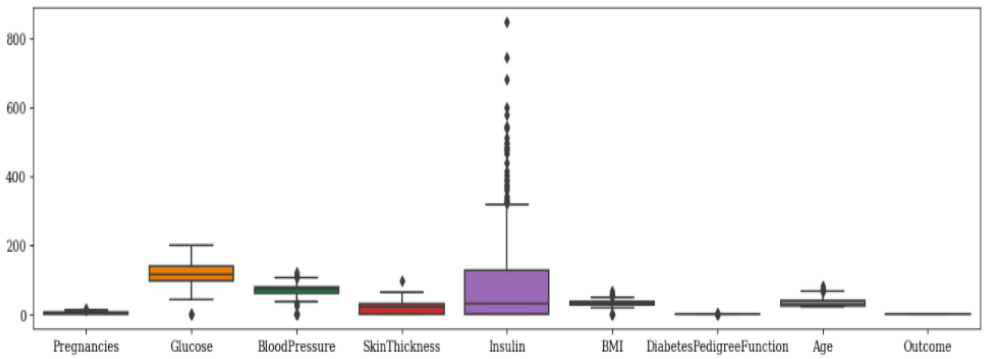

Figure 2. Box-Plot of the PIMA Dataset

3.3. Data Oversampling

The PIMA dataset comprises 500 non-diabetic samples and 268 samples with diabetes, as illustrated in figure 3. The prevalent class imbalance introduces a potential bias in machine learning models favouring the majority class during

 ISSN 2723-6471

training. Accordingly, the Synthetic Minority Over-sampling Technique (SMOTE) and its extended SVM-SMOTE were implemented to oversample the minority class. These methods ease the class imbalance problem by generating synthetic instances for the minority class, thus fostering a more equitable training data distribution.

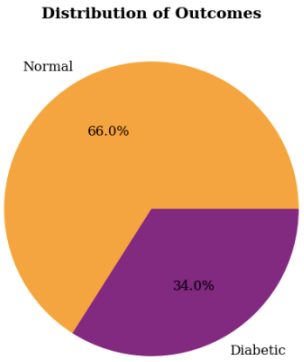

Figure 3. Samples Distribution of the PIMA Dataset

SVM-SMOTE extends the original SMOTE technique based on the support vectors generated by the SVM. The original SMOTE algorithm creates synthetic samples of the minority class by interpolating between existing minority class instances. SVM-SMOTE, however, combines SMOTE with Support Vector Machines (SVM) to improve the selection of samples for synthetic generation. First, the SVM classifier is trained to determine the support vectors that define the decision boundary between classes. These support vectors are then used to guide the generation of synthetic samples. The rationale is that support vectors are more informative for the classification task. Thus, generating synthetic samples near these vectors should improve the classifier's performance on imbalanced datasets. As such, SVM-SMOTE results in improving overall model performance. The steps to implement SVM-SMOTE are described as follows: (1) Training the SVM: Train an SVM classifier on the training set to identify the support vectors and the decision boundary of the minority class. (2) Selecting Support Vectors: Identify the support vectors from the trained SVM that belong to the minority class. These vectors are crucial as they lie close to the decision boundary and represent the minority class well. (3) Generating Synthetic Samples: Apply SMOTE to the support vectors of the minority class. SMOTE works by selecting pairs of minority class samples, finding the line segment between them, and generating synthetic samples along this line. The synthetic samples are interpolated points between the existing minority class samples. (4) Incorporating Synthetic Samples: Add the generated synthetic samples to the original dataset. This step increases the representation of the minority class in the dataset, reducing the imbalance. (5) Re-training the Classifier: Train a new classifier on the augmented training set, which now includes both the original and synthetic samples. The increased representation of the minority class helps the classifier to predict minority class instances.

3.4. Machine Learning Algorithms

While various classifiers exist, three classifiers were selected for the proposed framework. The criteria for selection were 1) Robustness in performance and against overfitting, which can be achieved using an RF classifier. 2) Superior performance and ability to handle imbalanced datasets, which can be achieved using EXtreme Gradient Boosting (XGBoost) classifier. 3) An SVM classifier can achieve effectiveness in high-dimensional spaces and high performance with limited samples. (1) Random Forest (RF): The RF is an ensemble method consolidating the predictions from multiple decision trees into a singular outcome. RF excels in addressing both classification and regression challenges. The popularity of RF stems from its robust performance and the ability to handle complex datasets, making it a preferred choice in various domains. (2) Support Vector Machine (SVM): SVM is a machine-learning algorithm for classification and regression tasks. The primary goal of SVM is to find the optimal hyperplane that best separates different classes in the data space, ensuring a clear distinction between them. SVM is particularly effective in high-dimensional spaces and excels when the margin between classes needs to be maximized. (3) EXtreme Gradient Boosting (XGBoost): The XGBoost is a high-performance machine-learning algorithm widely recognized for its predictive accuracy. Belonging to the ensemble learning family, XGBoost constructs a series of decision trees to refine predictions, highlighting efficiency through parallel processing and adept handling of missing data.

ISSN 2723-6471

4. Experimental Results

In the experiments, two goals were carried out: 1) exploring the significant factors that contribute to the development of diabetes as equivalent to feature selection in many developed frameworks. 2) Evaluate and compare the developed framework and the utilized machine learning models.

4.1. Factors Significances

As illustrated in table 3, the mean values for both non-diabetic and diabetic patients were calculated, revealing a substantial difference between these groups. The Fasting Plasma Glucose (FPG) was significantly elevated in individuals who developed diabetes compared to non-diabetic's individuals. The average blood glucose was 141.2 for diabetic individuals and 109.9 for those without diabetes, resulting in a significant difference of 31. According to Sagesaka, et al. [20], this often happens at least ten years before the diagnosis of diabetes. This suggests that there are crucial indicators affecting the human body before a diabetes diagnosis, and without proper care, individuals may end up being diagnosed with diabetes. Besides, blood pressure is also higher in people with diabetes, and their insulin levels are markedly elevated, with a mean of 100.3 compared to 68.7 in those without diabetes, highlighting a difference of 31.6. Additionally, older age groups may have a higher likelihood of developing diabetes. A comparison of diabetes and non-diabetics is illustrated in figure 4.

Table 3. The Mean Difference Between Diabetic and Healthy Person

Outcome	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Age
Diabetic	4.86	141.25	70.82	22.16	100.33	35.14	37.06
Healthy	3.29	109.98	68.18	19.66	68.792	30.30	31.19

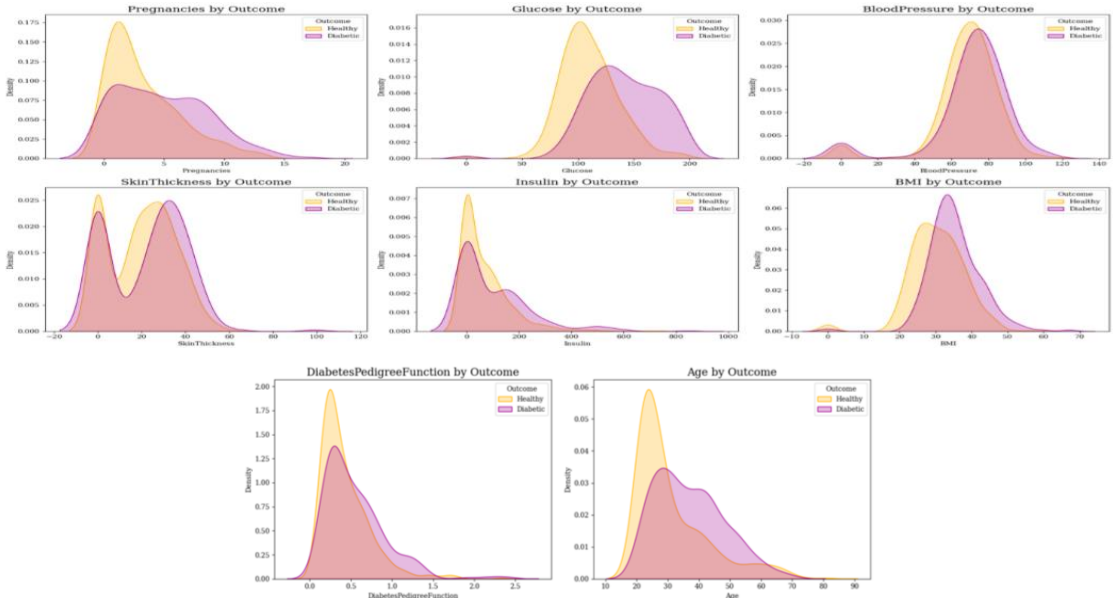

Figure 4. Density Comparison between Diabetic and Non-Diabetic

4.2. Classification Results

The experiments were conducted by splitting the data into 80% and 20% for training and testing, respectively. As shown in figure 5, the results showed that the XGBoost combined with SVM-SMOTE for oversampling stands out with an impressive accuracy of 91%. Additionally, RF achieves an accuracy of 90.2%, showing its effectiveness. Notably, SVM with the Polynomial kernel also demonstrates low performance with an accuracy of 83.11%. Table 4 illustrates the performance of these classifiers in terms of accuracy, precision, recall and F-measure. The results also showed that the proposed framework outperformed the existing approaches, as summarized in table 1.

 ISSN 2723-6471

As such, the proposed framework achieved an accuracy of 91% using the XGBoost classifier and SVMSMOTE. These results improved the accuracy compared to using classification without oversampling, which achieved an accuracy of 90%, while the best results addressed in the literature were 85.5%. The proposed framework showed the best performance. As such, the proposed framework improves the results by approximately 6.4% compared to the existing frameworks. Besides, the proposed framework achieved the best f-measure using the XGBoost classifier and SVMSMOTE, equal to 0.879. The best recall was achieved using RF and SVMSMOTE, which was 0.931. Finally, the best precision was achieved using FR without oversampling, with a value of 0.918.

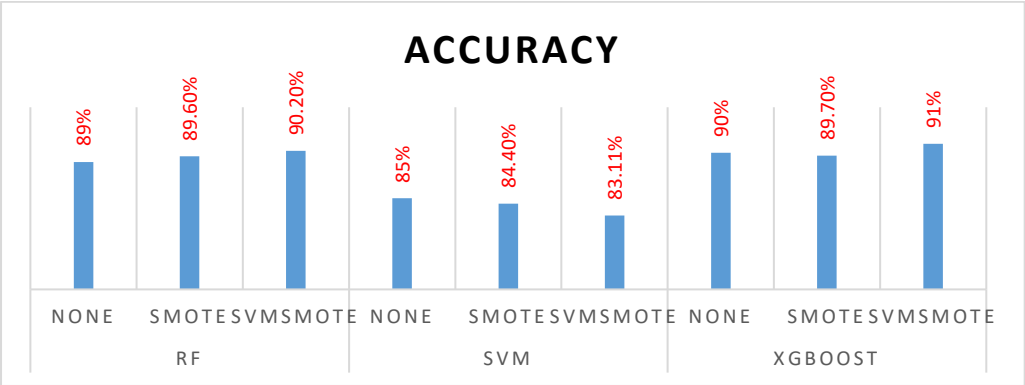

Figure 5. Accuracy of Diabetes Prediction

Table 4. Results of Diabetes Prediction

Model	Data Balancing	Accuracy	Precision	Recall	F-Measure
RF	-	89.00%	0.918	0.7750	0.841
	SMOTE	89.60%	0.860	0.8400	0.852
	SVMSMOTE	90.20%	0.831	0.9310	0.878
SVM with Polynomial Kernel	-	85.00%	0.807	0.7930	0.800
	SMOTE	84.40%	0.774	0.8280	0.800
	SVMSMOTE	83.11%	0.750	0.8276	0.787
XGBoost	-	90.00%	0.893	0.8630	0.877
	SMOTE	89.70%	0.847	0.8620	0.855
	SVMSMOTE	91.00%	0.879	0.8790	0.879

5. Conclusion

In conclusion, this study has demonstrated the significant impact of machine learning techniques, particularly Random Forest, SVM, and XGBoost, in predicting diabetes outcomes when combined with oversampling methods. Notably, XGBoost, in combination with SVM-SMOTE, achieved the highest accuracy of 91%. Besides, the proposed framework achieved the best f-measure using the XGBoost classifier and SVMSMOTE, equal to 0.879. The best recall was achieved using RF and SVMSMOTE, which was 0.931. Finally, the best precision was achieved using FR without oversampling, with a value of 0.918. Expanding the dataset size holds promise for uncovering deeper insights and refining predictive capabilities. This research underscores the potential of machine learning in revolutionizing diabetes management and highlights avenues for future exploration in this critical healthcare domain. Future work will focus on using other oversampling techniques to improve the results while experimenting with more classification and preprocessing algorithms.

6. Declaration

6.1. Author Contributions

Conceptualization: L.A.D. and A.A.A.S.; Methodology: A.A.A.S.; Software: L.A.D.; Validation: L.A.D. and A.A.A.S.; Formal Analysis: L.A.D. and A.A.A.S.; Investigation: L.A.D.; Resources: A.A.A.S.; Data Curation: A.A.A.S.; Writing Original Draft Preparation: L.A.D. and A.A.A.S.; Writing Review and Editing: A.A.A.S. and L.A.D.; Visualization: L.A.D.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. 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] R. Zolfaghari, "Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm," *The International Journal of Computational Engineering and Management*, vol. 15, no. 4, pp. 2230-7893, 2012.

[2] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm," *presented at the 21st international conference of computer and information technology (ICCIT), Dhaka, Bangladesh*, vol. 1, no. 1, pp. 1-5, 21-23 December, 2018.

[3] Y. Guo, G. Bai, and Y. Hu, "Using bayes network for prediction of type-2 diabetes," *presented at the 2012 International conference for internet technology and secured transactions, London, United Kingdom*, vol. 1, no. 1, pp. 471-472, 10-12 December, 2012.

[4] J. Han, J. C. Rodriguez, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," *presented at the Second international conference on future generation communication and networking, Hainan, China*, vol. 1, no. 1, pp. 96-99, 13-15 December, 2008.

[5] J. Davis, A.H. Fischl, J. Beck, L. Browning, A. Carter, J. E. Condon, M. Dennison, T. Francis, P. J. Hughes, S. Jaime, and K. H. K. Lau, "2022 National standards for diabetes self-management education and support," *The science of diabetes self-management and care*, vol. 48, no. 1, pp. 44-59, 2022.

[6] M. M. Hassan, Z. J. Peya, S. Mollick, M. A. M. Billah, M. M. H. Shakil, and A. U. Dulla, "Diabetes prediction in healthcare at early stage using machine learning approach," *presented at the 12th International conference on computing communication and networking technologies (ICCCNT), Khargpur, India*, vol. 1, no. 1, pp. 1-5, 6-8 July, 2021.

[7] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, and M. E. Pavkov, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes research and clinical practice*, vol. 183, no. 1, pp. 109-119, 2022.

[8] A. Kautzky-Willer, J. Harreiter, and G. Pacini, "Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus," *Endocrine reviews*, vol. 37, no. 3, pp. 278-316, 2016.

[9] M. I. Harris, R. Klein, T. A. Welborn, and M. W. Knuiman, "Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis," *Diabetes care*, vol. 15, no. 7, pp. 815-819, 1992.

 ISSN 2723-6471

[10] R. Ambady and S. Chamukuttan, "Early diagnosis and prevention of diabetes in developing countries," *Reviews in Endocrine and Metabolic Disorders*, vol. 9, no. 1, pp. 193-201, 2008.

[11] Ö. N. Ergün, "Early stage diabetes prediction using machine learning methods," *Avrupa Bilim ve Teknoloji Dergisi*, vol. 2021, no. 29, pp. 52-57, 2021.

[12] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes and Metabolic Disorders*, vol. 19, no. 1, pp. 391-403, 2020.

[13] L. Xie, "Pima Indian Diabetes Database and Machine Learning Models for Diabetes Prediction," *Highlights in Science, Engineering and Technology*, vol. 88, no. 1, pp. 97-103, 2024.

[14] 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, no. 1, pp. 321-357, 2002.

[15] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes," *International Journal on Soft Computing*, vol. 2, no. 2, pp. 15-23, 2011.

[16] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," *presented at the IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore*, vol. 88, no. 1, pp. 291-295, 5-8 February, 2018.

[17] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Computer Methods and Programs in Biomedicine Update*, vol. 4, no. 1, pp. 100-118, 2023.

[18] A. Perdana, A. Hermawan, and D. Avianto, "Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 1, pp. 70-75, 2023.

[19] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," *in Proceedings of the annual symposium on computer application in medical care, Washington, DC, USA*, vol. 1, no. 1, pp. 261-265, 6-9 November, 1988.

[20] H. Sagesaka, Y. Sato, Y. Someya, Y. Tamura, M. Shimodaira, T. Miyakoshi, K. Hirabayashi, H. Koike, K. Yamashita, H. Watada, and T. Aizawa, "Type 2 diabetes: when does it start?," *Journal of the Endocrine Society*, vol. 2, no. 5, pp. 476-484, 2018.