Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:45

https://doi.org/10.28919/cmbn/8492

ISSN: 2052-2541

PRINCIPAL COMPONENT ANALYSIS IMPLEMENTATION ON MACHINE LEARNING IN DIABETES CLASSIFICATION

MICHAEL TANTOWEN[1,*], KRISNA PUTRA[1], MAHMUD ISNAN[2], BENS PARDAMEAN[1,2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia

[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract. Diabetes Mellitus, a global health burden linked to increased cancer risks, can be identified through variables like BMI, age, blood sugar, and HbA1c. This study explored diverse machine learning techniques for diabetes prediction, emphasizing dimensionality reduction and feature selection's role in enhancing model accuracy. Our motive is to compare the performance of multiple machine learning algorithms measures between original data and original data on which the handling sampling method or principal component analysis (PCA) was applied. The study utilizes Kaggle's "Diabetes Prediction Dataset" with 100,000 entries, employing eight features and one target variable related to diabetes. In the experiment, the dataset was divided into three distinct datasets: 1) whole dataset, 2) dataset containing males only, and 3) dataset containing females only. Those datasets were trained with multiple machine learning models: K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machines (SVM), XGBoost (XGB), and Random Forest (RF). The findings revealed that XGB outperformed other models with f1-score

[*]Corresponding author

E-mail address: michael.tantowen@binus.ac.id

Received February 13, 2024

of 80.87 for an imbalanced dataset. Moreover, in diabetes classification based on gender, the random forest model was better for males with 80.34 as the f1-score while XGB was good for females 81.9 as the f1-score.

Keywords: diabetes; imbalanced dataset; data pre-processing; PCA; machine learning.

2020 AMS Subject Classification: 92C50.

1. INTRODUCTION

Diabetes Mellitus poses a significant health burden affecting millions of individuals globally [1]. Moreover, it is correlated with an increased risk of various cancers [1–3]. The identification of diabetes can be facilitated through the consideration of multiple variables, including Body Mass Index (BMI), age, blood sugar level, and HbA1c level [4]. Leveraging machine learning models to analyze these variables allows for the extraction of patterns, enabling the development of predictive insights through inference such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), XGBoost (XGB), and Random Forest (RF) [5–10]. This approach holds promise in enhancing our understanding of diabetes and contributes to the potential for more effective diagnostic and predictive tools in the realm of healthcare.

In recent studies, researchers have utilized the application of machine learning algorithms and various datasets such as PIMA Indian to classify diabetes [11–13]. This dataset was created by National Institute of Diabetes and Digestive and Kidney Diseases, published in 1988 [14]. On another occasion, Zhang [4] employed datasets from Kaggle's Electronic Health Records (EHRs). This dataset comprises a compilation of medical and demographic information from patients, coupled with their respective diabetes statuses. They reported accuracy rates ranging from 38.98% to 51.22%, employing algorithms such as CatBoost, RF, LightGBM, XGB, and Deep Neural Network (DNN). In the study, they found the challenge in increasing the model performance due to the imbalanced dataset, prompting the implementation of pre-processing procedures to rectify this imbalance [8–10]. The undertaking of such pre-processing measures is crucial for ensuring the robustness and reliability of subsequent analyses and model training within the research. Interestingly, we found this dataset has not been explored yet to its full potential.

In the pursuit of enhancing the accuracy and reliability of diabetes prediction models, advanced techniques such as Principal Component Analysis (PCA) and resampling methods were explored. PCA has been utilized in medical data to train machine learning models. Gárate-Escamila et al. [15] proposed the use of a chi-square (CHI) with PCA to improve the prediction of machine learning models when classifying heart disease. They mentioned that complete features were feasible when the system resources needed to be considered. They applied dimensionality reduction techniques to improve the raw data result by reducing 74 features given to three groups of features and achieved better performance. The study found the best dimensionality reduction method for the prediction of heart disease in terms of performance, for this reason, CHI-PCA was the most consistent and preferable method.

On the other hand, Reddy *et al.* [16] investigated the effect of two pioneer dimensionality reduction techniques, PCA and Linear Discriminant Analysis (LDA). The reduced dataset was experimented with DT, NB, RF, and SVM, and concluded that the model had better performance when combined with PCA compared to LDA. Similarly, Bhattacharya *et al.* [17] also successfully used PCA to choose the most significant features, eliminating irrelevant ones, which have a negative effect on the accuracy of the prediction. So, this study aimed to evaluate the implementation of dimensionality reduction using PCA on model performance across diverse algorithms. The handling sampling methods for the imbalance dataset were also evaluated in classifying diabetes.

2. DATASET AND METHODOLOGY

In this study, the raw dataset used went through the data pre-processing first. After that, the pre-processed dataset was trained using machine learning models to predict diabetes. We also trained the models with the dataset which was done by handling sampling method or reducing the dimension using PCA. Finally, we evaluated and compared the performance of the diabetes prediction models. The overall flow of the research used can be seen in Figure 1. Each stage is explained in more detail in the next subchapter.

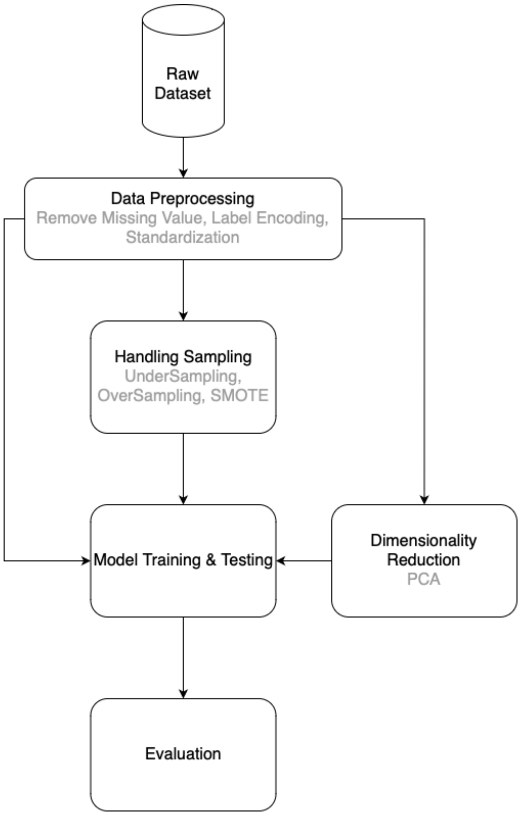


Figure 1. Research workflow in this study

2.1. Dataset

The study employed Kaggle's "Diabetes Prediction Dataset" to facilitate the training and assessment of the model under examination. The dataset, consisting of 100,000 entries, encompasses 8 distinct features and 1 target variable. These features divide patients into two categories: those afflicted with diabetes and those without. The dataset's information emanates from EHRs derived from a healthcare service provider. The data collection methodology encompasses the extraction of medical and demographic information from patients diagnosed with or identified as being at risk of diabetes [18]. Notably, the dataset encompasses two categorical features, namely gender and smoking history, with the remaining features being numerical, as depicted in Table 1.

TABLE 1. List of Features in the Dataset and Their Data Types

Feature Name	Data Types	Description
gender	object	Gender of the patient (Male, Female, and Other)
age	float64	Age of the patient
hypertension	int64	Information on whether the patient has hypertension (0 for no and 1 for yes)
heart_disease	int64	Information on whether the patient has heart disease (0 for no and 1 for yes)
smoking_history	object	Information about the patient's smoking status (divided into 6, namely: 'No Info', 'Never', 'Former', 'Current', 'Not Current', and 'Ever')
bmi	float64	BMI (Body Mass Index) is a measure of a patient's body fat based on weight and height.
HbA1c_level	float64	HbA1c, or hemoglobin A1c, is a measure of a patient's average blood glucose level over a certain period of time.
blood_glucose_level	int64	Blood glucose levels refer to the amount of glucose in a patient's bloodstream.
Diabetes	int64	Information on whether the patient suffers from diabetes (0 for no and 1 for yes).

2.2. Data Pre-processing

The data quality assurance process involves carefully checking for missing values, and ensuring that there are no null or empty entries in the data set. The encoding process is carried out on features that have the 'object' data type. Missing values and outliers handling were also carried out on the dataset. Following the standardization of data entry to 'int64' or 'float64', an important step in optimizing model performance, especially for distance-dependent algorithms such as KNN, has been implemented [19]. This standardization ensures fair feature contribution during model training [20,21]. Correlation analysis is also carried out to check the correlation between each feature and the target variable.

2.3. Handling sampling

Acknowledging the potential deleterious impact of an imbalanced dataset, this study has mitigated its negative effects by implementing three distinct methods: Under-sampling, Over-sampling, and Synthetic Minority Over-sampling Technique (SMOTE) on the utilized dataset [18,22]. These techniques aim to address the imbalance and enhance the model’s ability to generalize across both diabetic and non-diabetic classes.

2.4. Dimensionality Reduction

PCA was applied to the dataset. PCA is a dimensionality reduction technique that aims to capture the most critical features of the dataset while minimizing information loss. By implementing PCA, the study sought to evaluate whether a reduced set of principal components can effectively capture the underlying patterns in the data, potentially improving the model's efficiency [23]. The subsequent analysis involved training and evaluating the algorithm on the dataset processed through PCA, allowing for a comparative assessment of model performance before and after dimensionality reduction.

2.5. Model Experiment

This research delved into various machine learning algorithms for diabetes classification. Among these, the KNN algorithm emerged as a commonly employed method within the expansive spectrum of machine learning algorithms [24]. Noteworthy for their effectiveness in representing classifiers in data classification contexts, DT classifiers were also explored [25]. The RF methodology, utilizing a bagging technique, was employed to generate decision tree ensembles from a randomly selected subset of the training set. The amalgamation of individual decisions through a majority voting mechanism determined the ultimate prediction [26]. Acknowledged for its rapid and efficient handling of high-dimensional data, the SVM was highlighted, utilizing various kernel techniques for effective modelling of non-linear feature combinations and enhancing classifier performance [27,28]. Additionally, the research recognized the XGB algorithm as an efficient implementation of the Gradient Boosting Decision Tree (GBDT) technique, seamlessly integrating software and hardware optimization methodologies to yield superior outcomes with fewer computing resources compared to alternative methods [29].

2.6. Evaluation

Model performance was assessed using the F1-Score, a pertinent metric in medical classifications like diabetes, emphasizing equal importance on predicting both positive and negative cases. The F1-Score, derived from precision and recall values (equations 1, 2, and 3), measures the model's accuracy in predicting true positive patients and its effectiveness in identifying the majority of such instances. The best model's characterization involves presenting its AUC (Area Under Curve) and a confusion matrix, offering insights into its proficiency in predicting positive and negative patients [30,31].

$$F1\,Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3. RESULT AND DISCUSSION

The experiment begins by applying a data pre-processing stage to the dataset. Start by checking for missing values and outliers first. In the dataset used, no data was found with 'null' values or missing values so there was no need for handling missing values. An irregularity was identified in the gender feature, where 18 patients were categorized as 'other' prompting the systematic removal of rows with this classification. Subsequently, label encoding was applied to 'Object' data types such as gender and smoking history, encoding males as 1 and females as 0 for the gender feature, and employing predefined rules for encoding smoking history categories.

Following the standardization of data entries to either 'int64' or 'float64,' a critical step in optimizing model performance, especially for distance-dependent algorithms like KNN, was implemented [19]. This standardization ensures equitable feature contributions during model training [20,21].

A correlation analysis between each feature and the target variable was conducted, visualized through a heatmap in Figure 2. The results provide a comprehensive overview of the relationships between individual features and the target variable, contributing to a nuanced understanding of the dataset's characteristics [22-23].

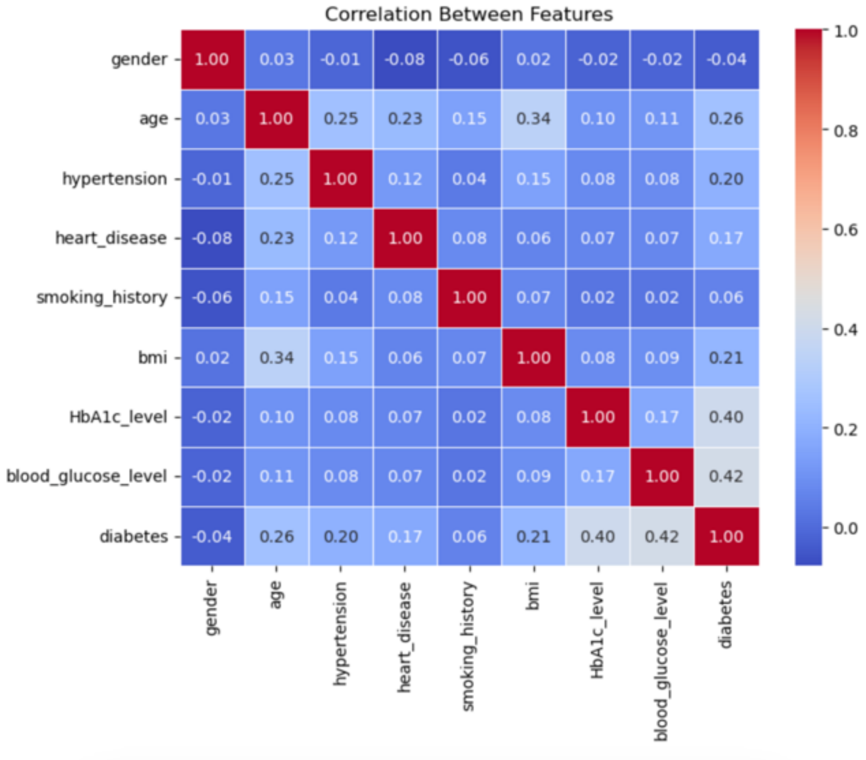


Figure 2. Heatmap of the correlation between each feature in the dataset

The correlation analysis in Figure 2 indicates a weak correlation between gender and the target outcome. Despite this, lifestyle factors like physical activity, alcohol, smoking, and diet can influence diabetes development [32-33]. Consequently, the current study stratified the dataset by gender, resulting in 58,552 female and 41,430 male instances. Both sets were evaluated using the same model to identify effective diabetes classification strategies for each gender. However, all datasets showed a notable imbalance, with non-diabetic cases exceeding 90%, potentially introducing bias in model performance [34]. After that, all models were trained with the dataset.

Table 2 presents the f1-score for each model, trained and tested on a dataset inclusive of both male and female genders. The models were subjected to be trained using four distinct dataset combinations: an imbalanced dataset, and a dataset subjected to under-sampling, oversampling, and SMOTE methods. The findings revealed that, among the models, KNN, SVM, XGB, and RF exhibit noteworthy reliability in diabetes classification when training on an imbalanced dataset. In contrast, a significant reduction in f1-scores was seen in all four models when training and testing on datasets manipulated for balance via three sampling methods namely undersampling, oversampling, and SMOTE. This decline in f1-score signifies an information loss in specific classes due to dataset manipulation, resulting in suboptimal model performance during testing. Based on these results, for these four models, the imbalance dataset is divided by gender for the next experiments. This decision is motivated by the critical nature of diabetes classification, where misclassifying a patient as negative when they are, in fact, positive can pose significant risks. Mitigating this risk is achievable by employing a combination of models and datasets that prioritize higher recall values.

TABLE 2. F1-Score From Both Gender Dataset with Handling Sampling Methods

	KNN	**DT**	**SVM**	**XGB**	**RF**
Imbalanced	**65.91**	73.05	**72.11**	**80.87**	**79.80**
Undersampling	56.95	56.06	58.22	62.90	62.11
Oversampling	63.82	**73.89**	58.20	66.58	78.30
SMOTE	62.48	70.43	58.37	80.42	76.54

Unlike the other four models, the DT exhibited a performance improvement when training on a dataset oversampled from its diabetes-positive patient class. Although the increment is very small, falling below 1%, it shows the potential of oversampling to mitigate bias in the majority class (in this case, diabetes-negative patients), thereby enhancing the model's overall accuracy. To prove this, a detailed investigation into the DT model was conducted, specifically focusing on the presentation of precision and recall values. When undergoing training and testing with an imbalanced dataset, the DT model exhibited precision and recall values of 71.57 and 74.59, respectively. In contrast, under the scenario of training and testing with an oversampled dataset, the DT model yielded precision and recall values of 73.64 and 74.14. A decline in the recall value implies the model's inability to capture a substantial number of true positive cases, while an increase in precision indicates a reduction in false positive predictions.

From the data presented in Table 2, it is evident that judging by the f1-score, the XGB model stands out as the most dependable classifier for diabetes when trained on a dataset that encompasses both genders. The XGB model achieved a remarkable f1-score of 80.87.

The XGB model exhibited outstanding performance with an AUC of 0.98 for both diabetes and non-diabetes classes in Figure 3. This exceptional AUC score indicates the model's robust ability to discriminate between positive and negative instances, showcasing its effectiveness in capturing relevant patterns despite class imbalance. The high AUC values suggest that the model achieves near-perfect true positive rates and low false positive rates for both diabetes and non-diabetes predictions.

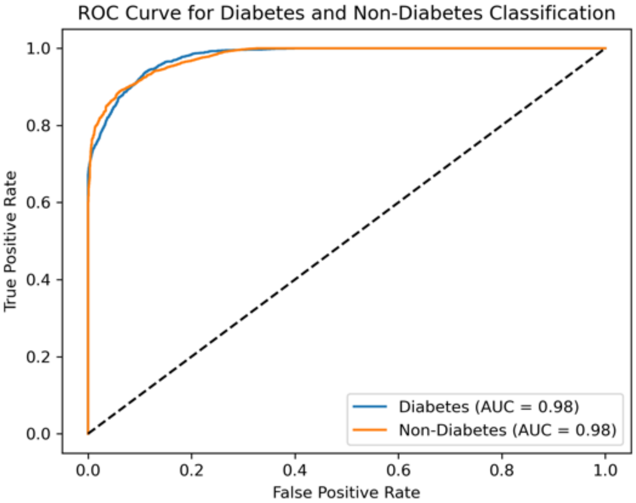


Figure 3. ROC curve of XGB with both gender imbalanced dataset

In Figure 4, the confusion matrix for diabetes classification using the XGB model shows promising performance, correctly identifying 1,239 cases of diabetes (True Positives) and 18,172 cases of non-diabetes (True Negatives). However, there were 50 instances of false positives, where non-diabetic cases were misclassified as diabetic, and 536 instances of false negatives, indicating cases of diabetes that were overlooked. While the model exhibits overall strong predictive capabilities, attention should be given to addressing false positives and false negatives to further enhance its precision and recall.

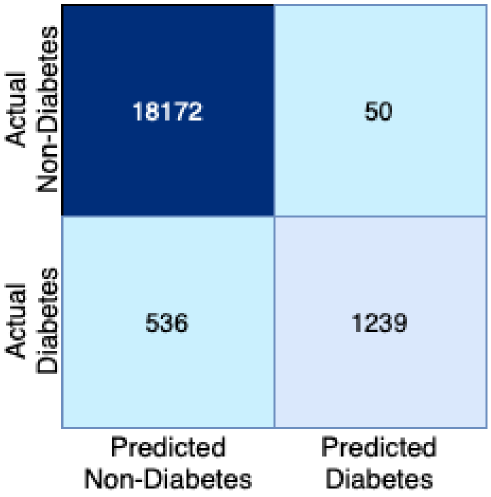


Figure 4. Confusion Matrix of XGB with both gender imbalanced dataset

The next step involved training the model using datasets exclusively comprising male and female genders. The f1-score for each model is provided in Table 3. For the dataset consisting solely of male gender, a marginal decrease is observed in the f1-score for KNN, DT, and XGB models when compared to training with a dataset encompassing both genders. Conversely, SVM and RF models exhibited enhanced performance, evidenced by increased f1 scores, particularly when classifying diabetes data among male patients. On the other hand, the dataset featuring female patient data manifested a decline in the f1-score for KNN, DT, and SVM models. In contrast, both XGB and RF models demonstrate an increase in f1-score values. These findings suggested that the RF model excels in diabetes classification when considering gender separately, as opposed to a combined gender approach. Meanwhile, KNN and DT consistently exhibit superior performance when trained on datasets inclusive of all genders, as opposed to gender-specific datasets. Furthermore, the XGB model demonstrates superior performance in classifying female patients exclusively compared to both gender and male-only datasets, in contrast to RF, which excels in classifying male patients.

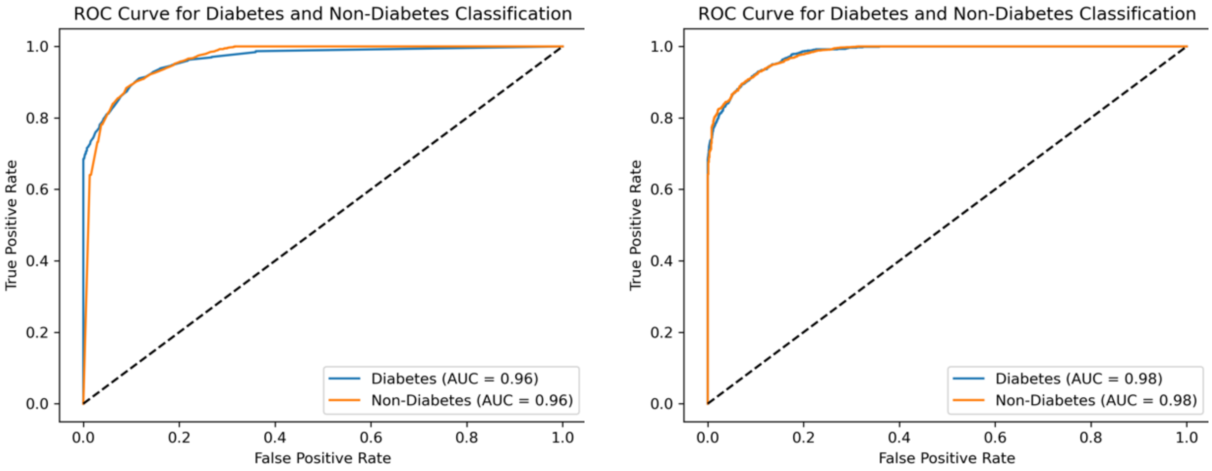
TABLE 3. F1-Score From Divided Male and Female Using imbalance Dataset

	KNN	**DT**	**SVM**	**XGB**	**RF**
Male	65.01	72.57	73.27	80.15	**80.34**
Female	62.91	72.74	70.95	**81.90**	80.08

This observation may suggest a stronger correlation between the features in the utilized dataset for male patients, contributing to the models' enhanced ability to classify male patients accurately. Despite the larger volume of data from female patients, the models exhibit a better proficiency in classifying male patients.

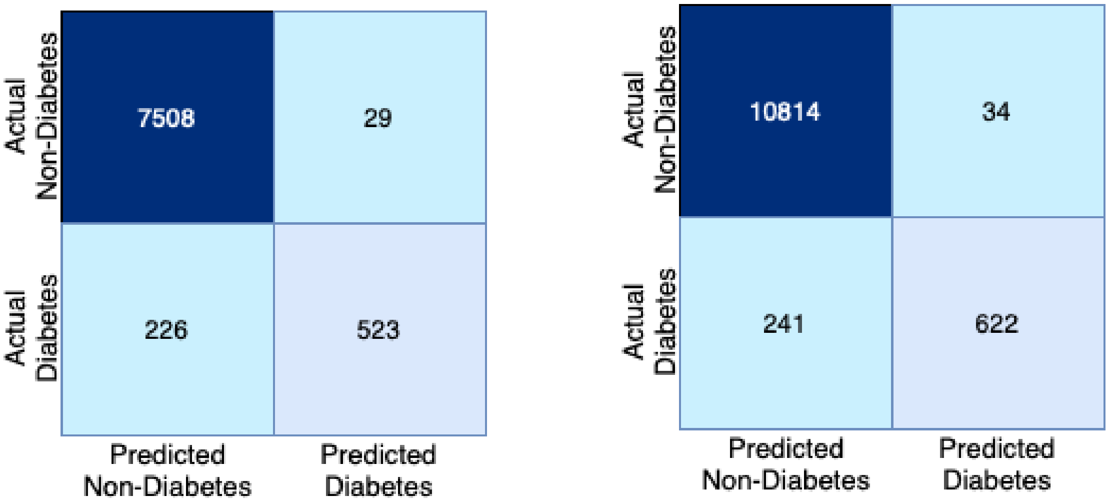
In Figure 5 (a), the RF model trained on a dataset of male patients demonstrates excellent performance, achieving an AUC of 0.96 for both diabetes and non-diabetes classes. This high AUC indicates the model's robust ability to distinguish between positive and negative instances, showcasing its effectiveness despite class imbalance. The model exhibits strong true positive rates and low false positive rates, emphasizing its precision. The confusion matrix in Figure 6 (a) for male patient further highlights the model's strength, correctly identifying 523 cases of diabetes (True Positives) and 7508 cases of non-diabetes (True Negatives). However, attention is needed to address 29 false positives and 226 false negatives, aiming to enhance precision and recall.

In Figure 5 (b), the XGB model trained on female patient data shows outstanding performance with an AUC of 0.98 for both diabetes and non-diabetes classes. This underscores the model's ability to capture relevant patterns despite class imbalance, achieving near-perfect true positive rates and low false positive rates. The confusion matrix in Figure 6 (b) for female patients reveals strong performance, correctly identifying 622 cases of diabetes (True Positives) and 10814 cases of non-diabetes (True Negatives). However, addressing 34 false positives and 241 false negatives is crucial to further enhance precision and recall. Overall, both models exhibit strong predictive capabilities with room for improvement in addressing misclassifications.



(a) (b)

Figure 5. ROC curve of RF for (a) male and (b) ROC curve of XGB for female



	Predicted Non-Diabetes	Predicted Diabetes
Actual Non-Diabetes	7508	29
Actual Diabetes	226	523

	Predicted Non-Diabetes	Predicted Diabetes
Actual Non-Diabetes	10814	34
Actual Diabetes	241	622

(a) (b)

Figure 6. Confusion Matrix of RF for (a) male and (b) female dataset

Next, dimensionality reduction was carried out on the three datasets using PCA and then trained and tested against the same model. Table 4 shows the f1-score of each model that was trained and tested using a dataset that has gone through the PCA process. Table 4 illustrates that the KNN model exhibits enhanced performance when trained on a dataset that has undergone dimensionality reduction using the PCA algorithm for all datasets, as evidenced by an increase in the f1-score. This improvement could be attributed to the ability of PCA to capture and retain the most significant features of the dataset while reducing its dimensionality. By focusing on the principal components that contribute the most to the variance in the data, PCA may facilitate a more robust and efficient representation of the underlying patterns, potentially benefiting the KNN model [23].

The performance of the SVM model seemed to have enhanced following dimensionality reduction using PCA on datasets exclusive to male and female genders. Conversely, the same model exhibited a decline in performance when trained and tested on a combined gender dataset that had undergone PCA. This underscores the dependency of PCA's effectiveness on the specific characteristics of the provided data. Even datasets with identical features can exert distinct influences on a given model. In this context, the division of the dataset into gender-specific subsets may lead to clearer patterns, enabling PCA to identify new linear combinations and address the multicollinearity issue within the dataset.

TABLE 4. F1-Score From Each Model Trained and Evaluated With Dataset + PCA

	KNN	**DT**	**SVM**	**XGB**	**RF**
Both Gender + PCA	73.33	68.45	71.24	71.73	**76.43**
Male + PCA	73.43	69.00	73.34	**77.31**	75.76
Female + PCA	73.61	69.09	71.73	**77.26**	77.16

In the context of the DT, XGB, and RF models, after their training and evaluation using datasets subjected to PCA-induced dimensionality reduction, there was a consistent decrement in performance observed across the three provided datasets. This trend served as an indicative signal that these models exhibited heightened sensitivity toward linear transformations. This heightened sensitivity, when PCA yielded novel linear transformations from the employed dataset, results in a diminished capacity of the models to discern prevalent patterns within the dataset, consequently causing a decline in performance.

These findings indicate that among the models examined, the KNN model demonstrated the most significant positive influence when employing PCA for dimensionality reduction on the utilized dataset. Conversely, the SVM model exhibited a favorable impact when applying to single-gender datasets but experiences a decline in performance when confronting with datasets encompassing both genders. The remaining models demonstrated an overall reduction in performance when utilizing the PCA-transformed dataset. Consequently, it is evident that the application of PCA in the context of diabetes classification, using the provided dataset, does not uniformly yield positive effects on the model.

4. CONCLUSIONS

Conducted experiments revealed that employing sampling techniques such as undersampling, oversampling, and SMOTE on the dataset aimed at addressing imbalances led to a decline in the f1-score performance of KNN, SVM, XGB, and RF models. Interestingly, the DT model exhibited a slight increase in f1-score when trained and tested with an oversampled dataset. Notably, the XGB model achieved the highest f1-score of 80.87 in diabetes classification, particularly when dealing with a both-gender dataset. However, performance varied when models were trained on gender-specific datasets; XGB excelled in classifying female patients with 81.9 as f1-score, while RF outperformed in classifying male patients with 80.34 as f1-score. Application of the PCA algorithm for dimensionality reduction yielded mixed results, with only the KNN model consistently improving performance. The study emphasized the intricate nature of model optimization, highlighting the importance of tailored approaches based on dataset characteristics and algorithm selection in diabetes prediction research. In particular, this study acknowledged the limitations of an imbalanced data set and suggests future research to try to overcome this problem by using other methods to handle imbalanced data such as assigning weights to each class and virtualization in computing [35-39].

ACKNOWLEDGEMENT

This work is supported by Bina Nusantara University.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] D. Tomic, J.E. Shaw, D.J. Magliano, The burden and risks of emerging complications of diabetes mellitus, Nat. Rev. Endocrinol. 18 (2022), 525-539. https://doi.org/10.1038/s41574-022-00690-7.

[2] N. Dimou, A.E. Kim, O. Flanagan, et al. Probing the diabetes and colorectal cancer relationship using gene – environment interaction analyses, Br. J. Cancer. 129 (2023), 511-520. https://doi.org/10.1038/s41416-023-02312-z.

[3] J. Baurley, A. Perbangsa, A. Subagyo, et al. A web application and database for agriculture genetic diversity and association studies, Int. J. Bio-Sci. Bio-Technol. 5 (2013), 33-42. https://doi.org/10.14257/ijbsbt.2013.5.6.04.

[4] S. Zhang, Advancing diabetes prediction: a nuanced six-class classification system and risk factor interactions investigation, in: P. Kar, J. Li, Y. Qiu (Eds.), Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023), Atlantis Press International BV, Dordrecht, 2023: pp. 677-686. https://doi.org/10.2991/978-94-6463-300-9_71.

[5] H. Kaur, V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, Appl. Comput. Inform. 18 (2020), 90-100. https://doi.org/10.1016/j.aci.2018.12.004.

[6] I. Ibrahim, A. Abdulazeez, The role of machine learning algorithms for diagnosing diseases, J. Appl. Sci. Technol. Trends. 2 (2021), 10-19. https://doi.org/10.38094/jastt20179.

[7] G. Battineni, G.G. Sagaro, N. Chinatalapudi, et al. Applications of machine learning predictive models in the chronic disease diagnosis, J. Pers. Med. 10 (2020), 21. https://doi.org/10.3390/jpm10020021.

[8] R.E. Caraka, M. Noh, R.C. Chen, et al. Connecting climate and communicable disease to penta helix using hierarchical likelihood structural equation modelling, Symmetry 13 (2021), 657. https://doi.org/10.3390/sym13040657.

[9] I.D. Kurniawan, R.C.H. Soesilohadi, C. Rahmadi, The difference on arthropod communities' structure within show caves and wild caves in Gunungsewu Karst area, Indonesia, Ecol. Environ. Conserv. 24 (2018), 72-81.

[10] R.E. Caraka, S. Shohaimi, I.D. Kurniawan, et al. Ecological show cave and wild cave: negative binomial Gllvm's arthropod community modelling, Procedia Computer Sci. 135 (2018), 377-384. https://doi.org/10.1016/j.procs.2018.08.188.

[11] N. Abdulhadi, A. Al-Mousa, Diabetes detection using machine learning classification methods, in: 2021 International Conference on Information Technology (ICIT), IEEE, Amman, Jordan, 2021: pp. 350-354. https://doi.org/10.1109/ICIT52682.2021.9491788.

[12] U.M. Butt, S. Letchmunan, M. Ali, et al. Machine learning based diabetes classification and prediction for healthcare applications, J. Healthc. Eng. 2021 (2021), 9930985. https://doi.org/10.1155/2021/9930985.

[13] H. Kaur, V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, Appl. Comput. Inform. 18 (2022), 90-100. https://doi.org/10.1016/j.aci.2018.12.004.

[14] K. Lakhwani, S. Bhargava, K.K. Hiran, et al. Prediction of the onset of diabetes using artificial neural network and pima Indians diabetes dataset, in: 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), IEEE, Jaipur, India, 2020: pp. 1-6. https://doi.org/10.1109/ICRAIE51050.2020.9358308.

[15] A.K. Gárate-Escamila, A.H. El Hassani, E. Andrès, Classification models for heart disease prediction using feature selection and PCA, Inform. Med. Unlocked. 19 (2020), 100330. https://doi.org/10.1016/j.imu.2020.100330.

[16] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, et al. Analysis of dimensionality reduction techniques on big data, IEEE Access. 8 (2020), 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942.

[17] [1] S. Bhattacharya, S.R.K. S, P.K.R. Maddikunta, R. Kaluri, S. Singh, T.R. Gadekallu, M. Alazab, U. Tariq, A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU, Electronics 9 (2020) 219. https://doi.org/10.3390/electronics9020219.

[18] M. Mustafa, Diabetes prediction dataset, 2023. https://www.kaggle.com/Datasets/Iammustafatz/Diabetes-Prediction-Dataset.

[19] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Appl. Soft Comput. 97 (2020), 105524. https://doi.org/10.1016/j.asoc.2019.105524.

[20] I.N. Setiawan, R. Kurniawan, B. Yuniarto, et al. Parameter optimization of support vector regression using Harris Hawks optimization, Procedia Computer Sci. 179 (2021), 17-24. https://doi.org/10.1016/j.procs.2020.12.003.

[21] N. Dominic, Daniel, T.W. Cenggoro, et al. Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification, Commun. Math. Biol. Neurosci. 2021 (2021), 39. https://doi.org/10.28919/cmbn/5565.

[22] D. Elreedy, A.F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance, Inform. Sci. 505 (2019), 32-64. https://doi.org/10.1016/j.ins.2019.07.070.

[23] K.M. Alalayah, E.M. Senan, H.F. Atlam, et al. Effective early detection of epileptic seizures through eeg signals using classification algorithms based on t-distributed stochastic neighbor embedding and K-means, Diagnostics. 13 (2023), 1957. https://doi.org/10.3390/diagnostics13111957.

[24] S. Uddin, I. Haque, H. Lu, et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction, Sci. Rep. 12 (2022), 6256. https://doi.org/10.1038/s41598-022-10358-x.

[25] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends. 2 (2021), 20-28. https://doi.org/10.38094/jastt20165.

[26] A. Hessane, A. El Youssefi, Y. Farhaoui, et al. A machine learning based framework for a stage-wise classification of date palm white scale disease, Big Data Min. Anal. 6 (2023), 263-272. https://doi.org/10.26599/BDMA.2022.9020022.

[27] R.E. Caraka, N.T. Nugroho, S.K. Tai, et al. Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using Boruta and Bayesian MCMC, Commun. Math. Biol. Neurosci. 2020 (2020), 22. https://doi.org/10.28919/cmbn/4584.

[28] R.E. Caraka, M. Tahmid, R.M. Putra, et al. Analysis of plant pattern using water balance and cimogram based on oldeman climate type, IOP Conf. Ser.: Earth Environ. Sci. 195 (2018), 012001. https://doi.org/10.1088/1755-1315/195/1/012001.

[29] J. Ma, Z. Yu, Y. Qu, et al. Application of the XGBoost machine learning method in PM2.5 prediction: A case study of Shanghai, Aerosol. Air Qual. Res. 20 (2020), 128-138. https://doi.org/10.4209/aaqr.2019.08.0408.

[30] M. Isnan, G.N. Elwirehardja, B. Pardamean, Sentiment analysis for TikTok review using VADER sentiment and SVM model, Procedia Computer Sci. 227 (2023), 168-175. https://doi.org/10.1016/j.procs.2023.10.514.

[31] H.H. Muljo, B. Pardamean, G.N. Elwirehardja, et al. Handling severe data imbalance in chest X-Ray image classification with transfer learning using SwAV self-supervised pre-training, Commun. Math. Biol. Neurosci. 2023 (2023), 13. https://doi.org/10.28919/cmbn/7526.

[32] M. Muilwijk, R. Bolijn, H. Galenkamp, et al. The association between gender-related characteristics and type 2 diabetes risk in a multi-ethnic population: The HELIUS study, Nutr. Metab. Cardiovasc. Dis. 32 (2022), 142-150. https://doi.org/10.1016/j.numecd.2021.09.015.

[33] B. Tramunt, S. Smati, N. Grandgeorge, et al. Sex differences in metabolic regulation and diabetes susceptibility, Diabetologia. 63 (2020), 453-461. https://doi.org/10.1007/s00125-019-05040-3.

[34] V.W. de Vargas, J.A.S. Aranda, R.D.S. Costa, et al. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study, Knowl. Inform. Syst. 65 (2023), 31-57. https://doi.org/10.1007/s10115-022-01772-8.

[35] B. Pardamean, A. Budiarto, B. Mahesworo, et al. Supervised learning for imbalance sleep stage classification problem, Commun. Math. Biol. Neurosci. 2023 (2023), 131. https://doi.org/10.28919/cmbn/8297.

[36] T.W. Cenggoro, B. Mahesworo, A. Budiarto, et al. Features importance in classification models for colorectal cancer cases phenotype in Indonesia, Procedia Computer Sci. 157 (2019), 313-320. https://doi.org/10.1016/j.procs.2019.08.172.

[37] K. Muchtar, F. Rahman, T.W. Cenggoro, et al. An improved version of texture-based foreground segmentation: block-based adaptive segmenter, Procedia Computer Sci. 135 (2018), 579-586. https://doi.org/10.1016/j.procs.2018.08.228.

[38] M.F. Kacamarga, B. Pardamean, H. Wijaya, Lightweight virtualization in cloud computing for research, in: R. Intan, C.-H. Chi, H.N. Palit, L.W. Santoso (Eds.), Intelligence in the Era of Big Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015: pp. 439-445. https://doi.org/10.1007/978-3-662-46742-8_40.

[39] J.W. Baurley, A. Budiarto, M.F. Kacamarga, et al. A web portal for rice crop improvements, Int. J. Web Portals. 10 (2018), 15-31. https://doi.org/10.4018/IJWP.2018070102.