Original Article

JHSMR Journal of Health Science and Medical Research

Follow-up Period Classification of Type 2 Diabetes Patients using Data Mining Techniques

Ilham Chapakiya, B.Sc.[1], Attasuntorn Traisuwan, Ph.D.[2], Supichaya Chumpong, M.D.[3], Kittisak Chumpong, Ph.D.[1,4,5]

[1]Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.

[2]Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.

[3]Department of Medicine, Pak Phanang Hospital, Pak Phanang, Nakhon Si Thammarat 80140, Thailand.

[4]Mathematics and Statistics with Applications Research Center, Prince of Songkla University, Songkhla 90110, Thailand.

[5]Financial Mathematics, Data Science and Computational Innovations Research Unit (FDC), Department of Mathematics, Faculty of Science, Kasetsart University, Chatuchak, Bangkok 10900, Thailand.

Received 28 February 2024 • Revised 2 April 2024 • Accepted 2 April 2024 • Published online 5 September 2024

Abstract:

Objective: This study investigates the use of high-performance data mining techniques to predict the follow-up period of diabetes patients.

Material and Methods: The diabetes dataset was obtained from Pak Phanang hospital in Nakhon Si Thammarat, Thailand. The hospital acquired the data between January 1 and December 31, 2022. The hospital-based retrospective study was based on 2,042 records, featuring 14 independent factors; including age, gender, systolic blood pressure, diastolic blood pressure, body mass index, pulse, weight, height, waist, smoking, drinking, parental history of diabetes, and fasting blood sugar and creatinine levels. To predict the follow-up period of diabetes patients, six well-known classification models were employed: Random Forest (RF), Extra Trees Classifier (ETC), Adaptive Boosting (Adaboost), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN). Class imbalances were addressed by using the Synthetic Minority Oversampling Technique (SMOTE), and feature importance was handled using the RF model.

Results: The experimental results demonstrated that, by applying SMOTE together with Random forest feature selection, the Support vector machine outperformed the other models; exhibiting the highest performances with a weighted precision of 0.9296.

Contact: Kittisak Chumpong, Ph.D.
Division of Computational Science, Faculty of Science, Prince of Songkla University,
Hat Yai, Songkhla 90110, Thailand.
E-mail: kittisak.ch@psu.ac.th

J Health Sci Med Res 2025;43(2):e20241083
doi: 10.31584/jhsmr.20241083
www.jhsmr.org

Conclusion: The results indicated that incorporating both SMOTE and feature selection resulted in significantly improved accuracy in predicting the follow-up period of diabetes patients for most models. Therefore, doctors and related healthcare providers could employ our proposed web-based tool to effectively schedule follow-up care for diabetes patients.

Keywords: classification, diabetes, feature selection, follow-up period, SMOTE

Introduction

Diabetes is a chronic disease caused by the insufficient production of the hormone insulin, leading to high blood sugar levels. Diabetes has an impact on the overall health of the body and can lead to complications, such as cardiovascular diseases, stroke, high blood pressure and chronic kidney disease. According to the International Diabetes Federation (IDF), in 2021, there were 537 million people with diabetes, and it is estimated that by the years 2030 and 2045, the number of people with diabetes will have increased to 643 and 783 million, respectively. Currently, up to 6.7 million diabetes-related deaths occur annually, or one every 5 seconds. Additionally, diabetes and diabetes-related healthcare account for up to 11.0% of global healthcare spending. Therefore, diabetes has a significant impact on the global economy and wider society[1,2].

According to a public health statistics report from the ministry of public health in Thailand, diabetes is one of the six leading causes of death in the country, with the incidence of diabetes continuously rising. There are 300,000 new cases annually, and 3.3 million are currently registered diabetes patients. In 2020, there were a total of 16,338 deaths from diabetes, which equals to a death rate of 25.1 per 100,000 population. The average annual public healthcare expenditure for diabetes treatment is 47.596 billion Baht. When the costs of diabetes-related cardiovascular disease, stroke and high blood pressure are included, the overall healthcare costs amount to approximately 302.367 billion Baht per year[3,4].

In general, diabetes patients are categorized into three groups for treatment. Patients in each group are given a target range for fasting blood sugar levels. For adults with good functional status and without complications, the target fasting blood sugar level ranges from 80 to 130 milligrams per deciliter. For patients with complications, reduced functionality, life-limiting comorbid illnesses, or substantial cognitive or functional impairments, the target fasting blood sugar level ranges from 90 to 150 milligrams per deciliter. For diabetes patients receiving palliative care and end-of-life care, the target fasting blood sugar level ranges from 100 to 180 milligrams per deciliter, so as to avoid hypoglycemia and symptomatic hyperglycemia, while reducing the burdens of glycemic management[5].

Continued follow-up care is necessary to sustain the effects of a good treatment plan as well as to monitor target fasting blood sugar levels. This can reduce mortality rates and diminish and avert complications caused by diabetes, resulting in an improved quality of life[6]. The follow-up period is therefore critical for diabetes patients and has been extensively researched. For diabetes patients treated with oral hypoglycemic agents, monthly follow-ups have been shown to indicate better levels of Diabetes Mellitus Quality of Life (DMQoL) and fasting blood sugar than three-month follow-ups[7]. Zhao et al. found that diabetes patients who underwent more than two follow-up periods a year showed better results in fasting blood sugar, hemoglobin A1C, waist circumference, blood pressure, cholesterol and low-density lipoprotein than patients who underwent less frequent monitoring, especially in younger patients or those with high hemoglobin A1C levels[8]. According to the Thailand diabetes practice guidelines 2023, the follow-up for diabetes treatment is determined by fasting blood sugar. The initial follow-up period is every 1–4 weeks. Once fasting blood sugar is controlled within the target range, the follow-up period is every 2–6 months or every 3 months on average[9].

Today a huge amount of valuable raw medical data is available to healthcare providers. Conversely, so much data is available that providers may have difficulty extracting the most appropriate information from a database. Handling big datasets usually requires the use of data mining techniques, which explore unobserved patterns in data. The use of these techniques improves the performance of predictive models; thereby, aiding medical decision-making. Classification is a fundamental, widely used data mining technique, which is an essential decision-making tool for building diabetes prediction models. In the literature, many studies have been conducted to accurately forecast the diagnostic result for patients. Perveen et al. utilized J48 DT, Bagging, and Adaboost on the dataset obtained from the Canadian Primary Care Sentinel Surveillance Network (CPSSN). Their findings indicated that the Adaboost ensemble methodology yielded the most effective results[10]. Mujumdar and Vaidehi applied various classification techniques; including Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), Extra Tree Classifier (ETC), Adaptive Boosting (AdaBoost), Multi-Layer Perceptron (MLP), Linear Discriminant Analysis (LDA), Logistic Regression (LR), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (NB), Bagging and Gradient Boost Classifier, to the Pima Indian Diabetes Datasets (PIDDs). They concluded that LR achieved the highest accuracy, while the use of a pipeline resulted in AdaBoost classifier being identified as the best model[11]. In 2020, Kazerouni et al. utilized four classification models; namely: KNN, SVM, LR and Artificial Neural Network (ANN), to predict type-2 diabetes mellitus using 6-lncRNA. They employed privately collected datasets from other authors. The results showed that SVM and LR achieved the highest area under the curve (AUC), indicating superior performance. On the other hand, KNN and ANN exhibited high mean AUC and low standard deviation of AUC. Additionally, KNN demonstrated the highest mean sensitivity, while SVM showed the highest specificity[12]. Pranto et al. have developed a model using machine learning approaches; such as DT, KNN, RF, and NB on the PIDDs, and a dataset obtained from Kurmitola general hospital in Dhaka, Bangladesh. They have concluded that both the RF and NB classifiers performed effectively on both datasets[13].

When dealing with imbalanced datasets, classifiers often exhibit a bias towards the majority class, resulting in high accuracy for that class, but poor accuracy for the minority classes[14]. Researchers have presented the Synthetic Minority Oversampling Technique (SMOTE) technique for addressing imbalanced data, which has demonstrated superior performance in existing literature. Sooklal and Hosein have applied various LR models, including LR with SMOTE, benefit-based LR using a cost-based model and benefit-based LR using a life-expectancy model to the PIDDs. They have reported that the LR model, with a simple modification achieved the maximum accuracy[15]. Nnamoko and Korkontzelos utilized the SMOTE technique to address class imbalance. They applied various classifiers; including NB, SVM, RIPPER, and C4.5 DT, on the PIDDs, German credit dataset, and biodegradation dataset. Their findings indicated that using SMOTE in conjunction with the C4.5 DT classifier resulted in superior performance compared to the other classifiers[16]. Hairani, Saputro, and Fadli employed the SMOTE technique to address the issue of imbalanced classes in on the PIDDs. The SVM has the highest accuracy and sensitivity; whereas, the naïve Bayes technique yields the highest specificity[17]. Erlin et al. addressed the issue of unbalanced data by employing SMOTE to augment the minority class in the PIDDs with synthetic data samples. The model was assessed and demonstrated a satisfactory level of performance[18].

In addition, feature selection is essential in diabetes datasets, to determine the most crucial characteristics for classification or prediction. High-dimensional datasets frequently include unnecessary or redundant features, which can lead to overfitting and reduced accuracy. Feature selection enhances classification accuracy, by reducing dimensionality and eliminating non-essential or redundant features; hence, simplifying the model and enabling better real-time performance. Researchers in the literature have explored the use of feature selection techniques in diabetes datasets to tackle this issue. These techniques include: genetic algorithm[19], RF importance[20], Fast correlation-based filter[21] and Binary wheal optimization algorithm[22], combined with different classification models for improved performance.

However, the use of classification techniques to predict the most appropriate follow-up period for diabetes patients with RF feature selection and SMOTE still remains challenging, due to the lack of research.

In this work, we utilized a diabetes dataset from Pak Phanang Hospital in Nakhon Si Thammarat, Thailand, to develop a classification model that can predict the follow-up period of diabetes patients. The SMOTE and feature selection by RF were applied to the acquired dataset to improve classification performances. The impact of both processes was analyzed. Six well-known classification techniques were then applied to the preprocessed data and their performances were compared.

Material and Methods

Ethical considerations

The study was approved by the PSU Human Research Ethics Committee, Prince of Songkla University (PSU-HREC-2023-044-1-3), which met the criteria for an Exempt Research Determination.

Dataset description

The diabetes dataset was obtained from Pak Phanang Hospital in Nakhon Si Thammarat, Thailand, between January 1 and December 31, 2022. The dataset consists of 2,042 records of patients over 35 years old that were diagnosed with diabetes based on ICD-10 codes E110–E119. In the data preprocessing stage, we removed any records that were missing or duplicated. In addition, we transformed qualitative data; such as gender, smoking status, alcohol consumption, into quantitative data. The specifications of 14 independent, potential risk factors and one multiclass follow-up period outcome variable are presented in the following table.

SMOTE (imbalanced data to balanced data)

The SMOTE is used in machine learning to address class imbalance in datasets when the minority class is underrepresented compared to the majority class or classes. SMOTE generates new samples in a class by randomly selecting one (or more, depending on the over-sampling rate) from the KNN of the existing samples in the minority class. The performance of predictions is improved, because class imbalance can lead to predictive models that under-perform due to a lack of examples[23]. In the context of medicine, SMOTE can be applied to tasks; such as disease prediction, medical image analysis, clinical decision support systems and drug discovery[24,25].

Feature selection

Feature selection based on the RF classifier is a dimensionality reduction technique used to improve the quality and efficiency of machine learning in various fields; including healthcare and finance. It identifies and ranks the importance of factors in a dataset. It is employed to determine which factors have the most significant impact on the predictions of the model, and aids in selecting a subset of the most relevant factors for building predictive models[25,26].

Diabetes prediction model

In computer science and machine learning, classification is a significant process that predicts or categorizes objects based on multiple factors. Classification is accomplished using supervised learning computer models that divide the data into a training set and a testing set. Individual classifiers rely on the training data, and the effectiveness of these classifiers is evaluated using the test data. We employed the following six different classification models, using the Scikit.learn 1.4.1 package[27].

Table 1 Dataset specifications

No.	Factors	Values	Type
1	Age	35–102 years	Numeric
2	Gender	The value is 0 when the patient is male. The value is 1 when the patient is female.	Binary
3	SBP	78–185 mm/hg	Numeric
4	DBP	45–109 mm/hg	Numeric
5	BMI	12.17–57.81 kg/m^2	Numeric
6	Pulse	50–126 bpm	Numeric
7	Weight	28.2–148.0 kg	Numeric
8	Height	133–190 cm	Numeric
9	Waist	60–144 cm	Numeric
10	Smoking	The value is 0 when the patient does not smoke. The value is 1 when the patient either smokes or has quit smoking for less than a month. The value is 2 when the patient has quit smoking for greater than a month. The value is 3 when there is no smoking information available about the patient.	Multiclass
11	Drinking	The value is 0 when the patient does not drink alcohol. The value is 1 when the patient drinks alcohol. The value is 2 when the patient has quit drinking alcohol. The value is 3 when there is no drinking alcohol information available about the patient.	Multiclass
12	PD	The value is 0 when the parents do not have diabetes. The value is 1 when one of the parents has diabetes. The value is 2 when the parents have diabetes.	Multiclass
13	FBS	49–572 mg/dl	Numeric
14	CR	0.30–9.74 mg/dl	Numeric
15	Follow-up period	The class is 0 when the follow-up period ranges between 1 and 4 weeks. The class is 1 when the follow-up period ranges between 5 and 8 weeks. The class is 2 when the follow-up period ranges between 9 and 12 weeks. The class is 3 when the follow-up period is greater than 12 weeks.	Multiclass

SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, PD=parent with diabetes, FBS=fasting blood sugar, CR=creatinine

Random forest

RF classification is based on aggregating votes from multiple decision trees, each generated by independently sampling data with consistent patterns. Despite potential differences in the factors considered, this process enhances diversity and independence among the trees. The decision tree with the most votes determines the classification outcome[28]. In this research, the number of trees in the forest was set to 15.

Extra Trees classifier

ETC algorithm operates quite similarly to the RF, except for how it constructs its trees. Each decision tree in ETC is built from the original training data. Random samples of the top *k* best factors are selected for decision making and the Gini index is used to identify the most important feature for splitting the data in trees. These randomly chosen feature samples contribute to creating several decision trees that are independent of each other[29]. In this research, the parameters of ETC were set to default in the Scikit.learn package.

Adaptive boosting

AdaBoost is frequently used in conjunction with other algorithms to improve their performance. Its primary function is boosting, which transforms weak learners into strong ones. The performance of each tree within the AdaBoost classifier is contingent on the error rate of the last built tree[30]. In this research, the base estimator was the decision tree classifier initialized with a maximum depth of 9.

Support Vector Machine

The objective of the SVM algorithm is to find hyperplanes in a multidimensional space that can effectively classify data points. These hyperplanes are generated iteratively by SVM, with the aim of minimizing errors[31]. In this research, the kernel type was set to linear.

K-nearest neighbor

The KNN algorithm categorizes data points based on the classification of their *k* closest neighbors, which can affect sensitivity to the choice of the parameter *k* and potentially reduce accuracy[32]. In this research, the parameters of KNN were set to default in the Scikit.learn package.

Artificial neural network

The ANN model is a network of interconnected nodes analogous to neurons in the biological neural network. Each neural network has three critical components: node characteristics, network topology, and learning rules. Node characteristics determine how signals are processed by nodes; including the number of inputs and outputs associated with each node, the weights associated with each input and output, and the activation function. Network topology defines how nodes are organized and connected. Learning rules specify how the weights are initialized and adjusted[33]. In this research, we set up a model with 2 hidden layers: the first with 200 neurons and the second with 100 neurons, with the maximum number of iterations being set to 1,000.

Assessment of performance metrics

The confusion matrix for multiclass classification is an *N* x *N* matrix; wherein, *N* represents the number of classes in the multiclass classification problem. Each row in the matrix represents the actual classification, and each column represents the predicted classification. Therefore, correctly classified elements are located on the main diagonal from top left to bottom right. It was used to determine model performances and gain insights into the classification results. Let TP_{ii} represent the number of observations correctly predicted as class *i* and FP_{ij} represent the number of observations from actual class *i* that were incorrectly predicted as class *j*, where $i \neq j$ and *i*, *j =1,2,...,N*.

Accuracy

This metric measures the overall correctness of predictions by calculating the ratio of correctly predicted instances to the total number of instances. It was calculated as follows:

$$Accuracy = \frac{\sum_{i=1}^{N} TP_{ii}}{\sum_{i=1}^{N} TP_{ii} + \sum_{i \neq j} FP_{ij}}.$$

Weighted precision

Weighted recall is a performance indicator that assesses the effectiveness of a classification model, accounting for the unequal distribution of classes by considering the significance or weight of each class. The calculation involves summing the recall values for each class, with each value weighted according to the proportion of that class in the dataset. It was calculated as follows:

$$Weighted\ precision = \sum_{i=1}^{N} w_i \left(\frac{TP_{ii}}{TP_{ii} + \sum_{i \neq j} FP_{ji}} \right),$$

Wherein: w_i is the weight or proportion of class in the dataset.

Weighted recall

Weighted recall is a performance metric for classification models that accounts for class imbalance, by considering the importance or weight of each class. The calculation involves summing the recall values for each class, with each value being weighted by the proportion of that class in the dataset. It was calculated as follows:

$$Weighted\ recall = \sum_{i=1}^{N} w_i \left(\frac{TP_{ii}}{TP_{ii} + \sum_{i \neq j} FP_{ij}} \right).$$

Wherein: w_i is the weight or proportion of class in the dataset.

Weighted F1-score

The weighted F1-score is the harmonic mean of weighed precision and weighted recall. It provides a balanced measure of precision and recall for each class, treating all classes equally. It was calculated as follows:

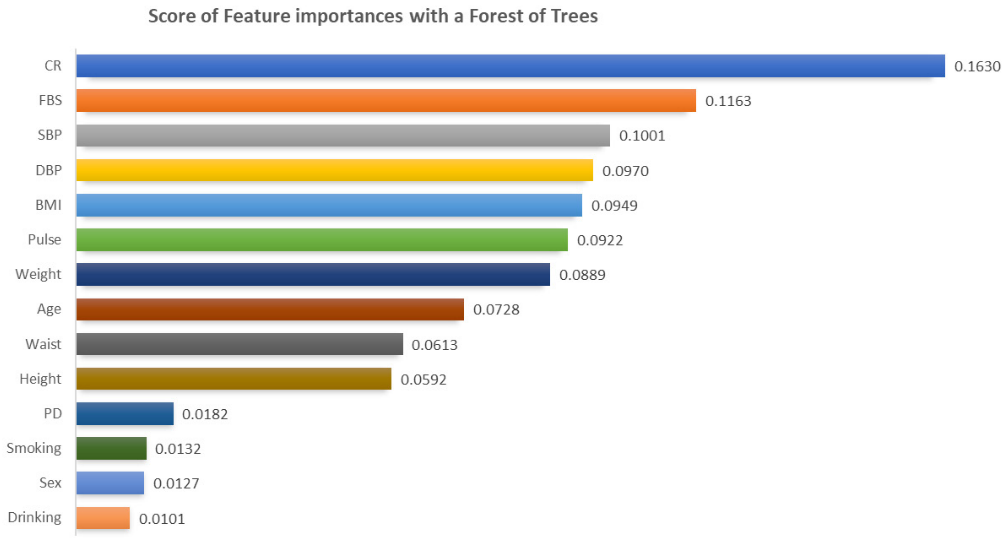
$$Weighted\ F1 - Score = \frac{2(Weighted\ precision)(Weighted\ recall)}{Weighted\ precision + Weighted\ recall}.$$

Results

The experiments were performed on a quad-core Intel Core i5 2.40 GHz processor with 8 GB of main memory, using various Python libraries within the Jupyter environment under Microsoft Windows 11 64-bit.

After performing data preprocessing, the dataset comprised of 2,042 records. The class imbalance was a multiclass classification problem. Out of 2,042 records, 37 records were categorized in class 0 (1.8%), 49 in class 1 (2.4%), 33 in class 2 (1.6%) and 1,923 in class 3 (94.2%). The normalized dataset with min-max scaling was then input to the feature selection process by the RF algorithm.

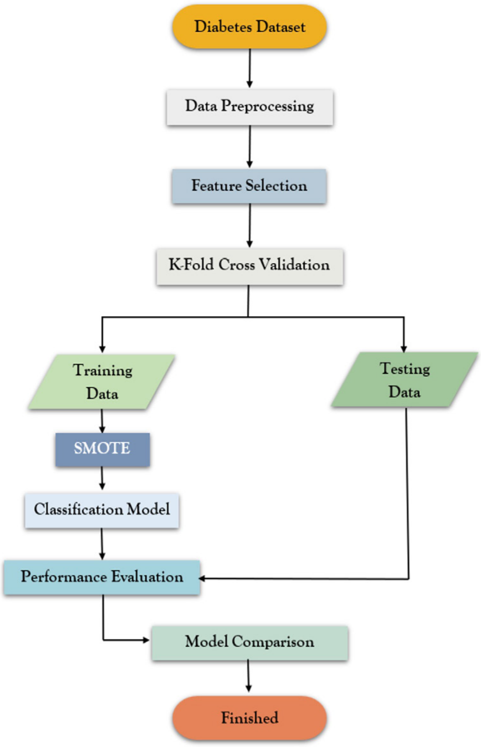
Figure 1 displays the factor ranking for the dataset generated by the RF classifier. We examined the importance of these factors to identify those that significantly contributed to the input of the classification model. Ultimately, we determined that ten factors maximized the model precision. These factors were: creatinine level, fasting blood sugar level, systolic blood pressure, diastolic blood pressure, body mass index, pulse, weight, age, waist and height.



CR=creatinine, FBS=fasting blood sugar, SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, PD=parent with diabetes

Figure 1 Feature importance score using the Random Forest classifier

Next, the study uses *k*-folds cross validation, a method of retesting random input attributes, to test an algorithm model. The data is divided into 10 subsets, with 90% for training and 10% for testing. The process is repeated 10 times until all data records are part of the testing data[34]. All classifiers were fitted to the SMOTE training data. The method is iterated 10 times and the data, after applying SMOTE, are as follows: 1,729, 1,732, 1,730, 1,734, 1,728, 1,729, 1,728, 1,734, 1,737 and 1,726, respectively. The performance of the machine learning classifiers was then assessed using the four performance evaluation metrics: accuracy, weighted precision, weighted recall and weighted F1-score. A flowchart of the proposed methodology is presented in Figure 2. The following subsection presents a comparison of the model performances.



SMOTE=Synthetic Minority Oversampling Technique

Figure 2 Flowchart of the proposed methodology

Diabetes model performances

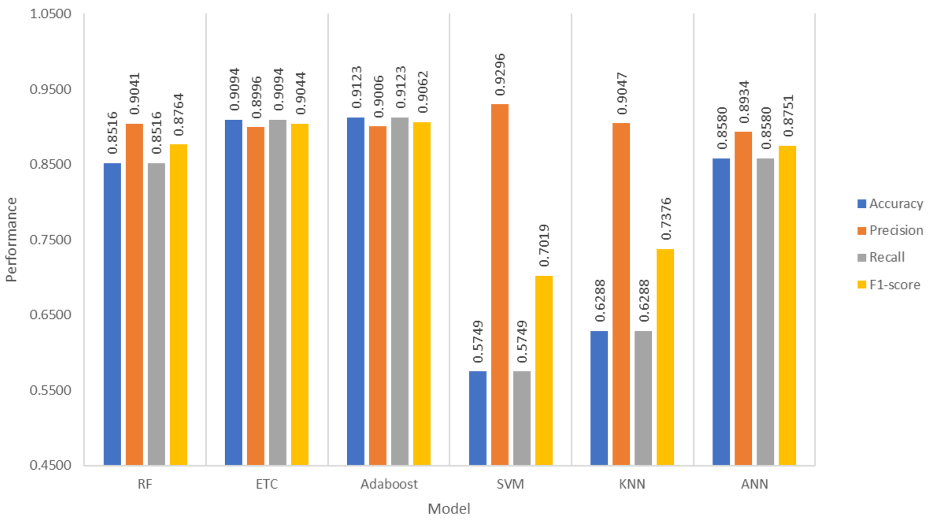
Figure 3 displays the performances of the six machine learning classifiers: RF, ETC, AdaBoost, SVM, KNN and ANN. The experimental results calculated, from the classifier confusion matrix in Table 2, showed that most of the classification models could accurately predict the follow-up period from the balanced dataset after feature selection. We focused on using the precision metric, which emphasizes accurate prediction of the follow-up period classification, corresponding to the confusion matrix. The classifiers exhibited varying performances, with RF, AdaBoost, SVM and KNN classifiers achieving good precision. The SVM classifier exhibited the highest performances, with a weighted precision of 0.9226. The ETC classifier performed less well, while the ANN classifier produced the lowest performance for predicting diabetes follow-up periods, with a weighted precision of 0.8934.

Discussion

Impact analysis of random forest feature selection and SMOTE

We conducted a comparative analysis of feature selection from the diabetes dataset without applying SMOTE to the dataset: precision was the performance metric. The RF, ETC, Adaboost and KNN models exhibited improved precision when feature selection by RF was applied (Figure 4). Specifically, the Adaboost model showed the most significant improvement, experiencing an increase in precision from 0.8868 to 0.8937. The precision of the SVM model was identical, 0.8868; whereas, ANN was the only model that exhibited a decrease in performance after the implementation of feature selection. Overall, adding feature selection to the six machine learning models failed to improve accuracy in only 16.67% of the diabetes dataset.

We also studied the impact of SMOTE on the diabetes dataset without employing RF feature selection: all models demonstrated enhanced performances. The SVM model showed the most significant improvement; increasing precision from 0.8868 to 0.9276 (Figure 5). This study indicated that all models exhibited improved precision when SMOTE was applied to the dataset.



RF=Random Forest, ETC=Extra Trees classifier, Adaboost=Adaptive boosting, SVM=Support Vector Machine, KNN=K nearest neighbor, ANN=Artificial neural network

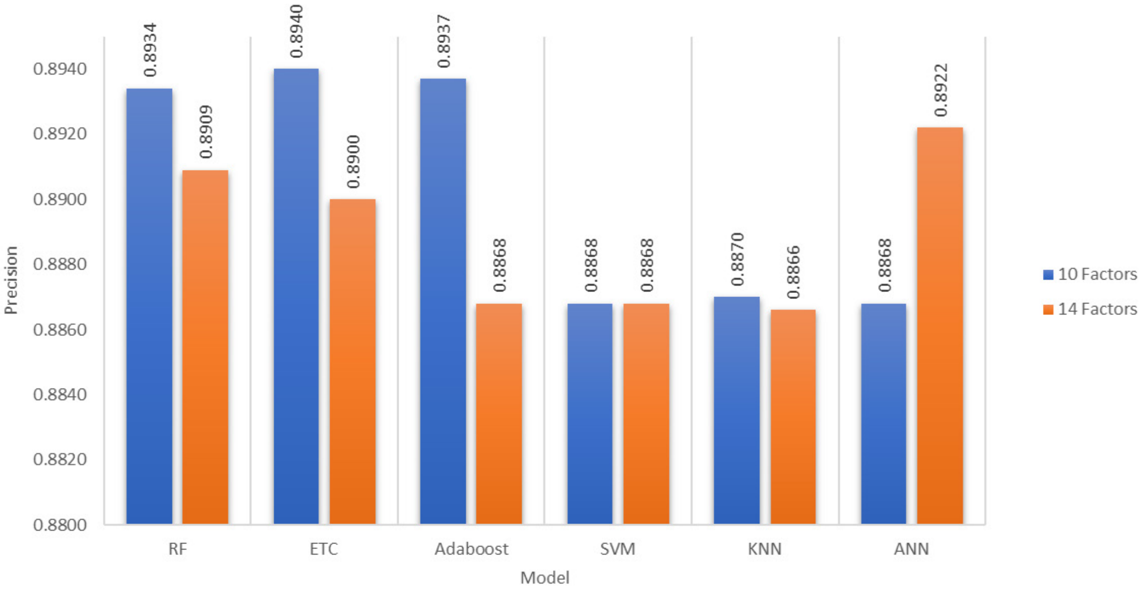
Figure 3 Performance comparison of all models using ten significant factors selected by RF, after data balancing with SMOTE

Table 2 Classifier confusion matrix

Confusion matrix			**Predicted classification**			
			Class 0	**Class 1**	**Class 2**	**Class 3**
Actual classification	Random forest	Class 0	2	6	3	26
		Class 1	5	10	5	29
		Class 2	3	6	1	23
		Class 3	69	79	49	1,726
	Extra trees classifier	Class 0	1	2	1	33
		Class 1	1	6	3	39
		Class 2	2	3	2	26
		Class 3	26	33	16	1,848
	Adaptive boosting	Class 0	2	3	0	32
		Class 1	5	4	2	38
		Class 2	1	1	3	28
		Class 3	23	33	13	1,854
	Support Vector Machine	Class 0	12	10	10	5
		Class 1	17	15	11	6
		Class 2	13	8	5	7
		Class 3	218	221	342	1,142

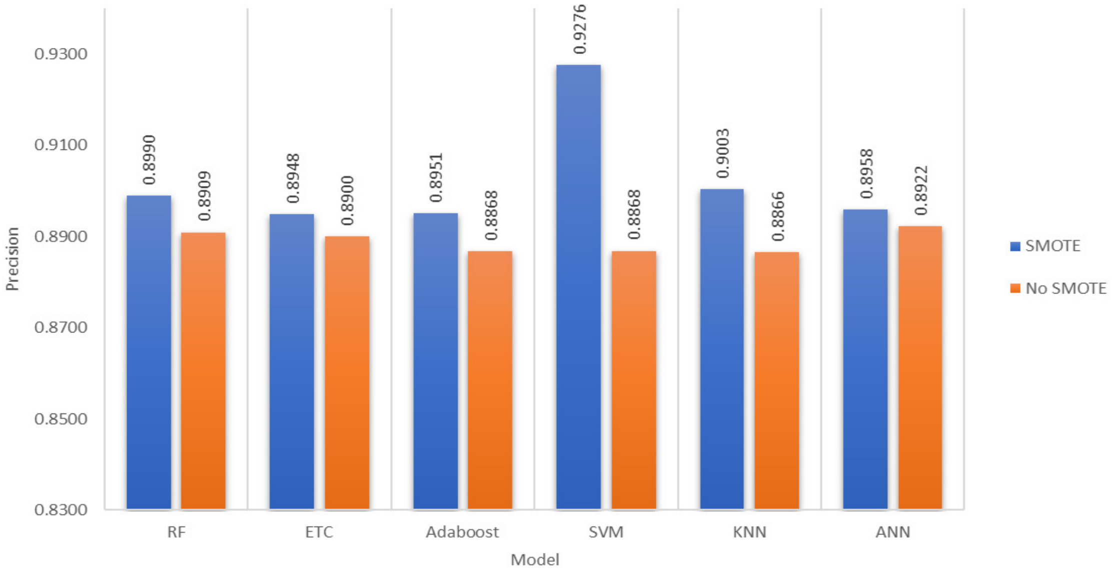
Table 2 (countinued)

Confusion matrix			Predicted classification			
			Class 0	**Class 1**	**Class 2**	**Class 3**
Actual Classification	K-nearest neighbor	Class 0	6	9	5	17
		Class 1	11	9	8	21
		Class 2	6	7	4	16
		Class 3	210	255	193	1,265
	Artificial neural network	Class 0	1	4	0	32
		Class 1	2	5	3	39
		Class 2	2	1	2	28
		Class 3	54	84	41	1,744



RF=Random Forest, ETC=Extra Trees classifier, Adaboost=Adaptive boosting, SVM=Support Vector Machine, KNN=K nearest neighbor, ANN=Artificial neural network

Figure 4 Precision comparison of all models with and without random Forest Feature selection

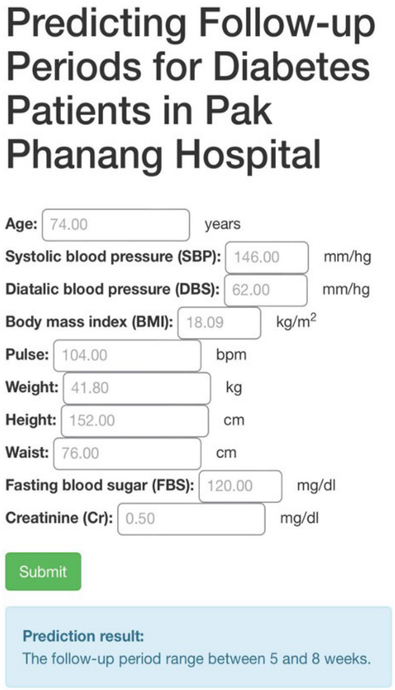


RF=Random Forest, ETC=Extra Trees classifier, Adaboost=Adaptive boosting, SVM=Support Vector Machine, KNN=K nearest neighbor, ANN=Artificial neural network

Figure 5 Precision comparison of all models with and without SMOTE, from the whole 14 features

Management implications

Web-based diagnostics are used by researchers, doctors, and related healthcare providers to facilitate decision-making in a range of contexts[35-40]. Therefore, we designed and implemented a web-based tool for predicting the follow-up period of diabetes patients, aiming to assist medical teams in effective follow-up scheduling. The web-based follow-up prediction system was implemented using Python V3.9.13 and Flask V2.3.2. The SVM prediction model, incorporating both RF feature selection and SMOTE data balancing, was implemented on the server side using Scikit.learn V1.3.2. The user can input feature data through a web browser on either a computer or a mobile device and then click the submit button. The input features are sent to a web server, where our model predicts the follow-up period and presents the result in the output interface: as shown in the Figure 6.


Predicting Follow-up Periods for Diabetes Patients in Pak Phanang Hospital

Age: 74.00 years
Systolic blood pressure (SBP): 146.00 mm/hg
Diatalic blood pressure (DBS): 62.00 mm/hg
Body mass index (BMI): 18.09 kg/m^2
Pulse: 104.00 bpm
Weight: 41.80 kg
Height: 152.00 cm
Waist: 76.00 cm
Fasting blood sugar (FBS): 120.00 mg/dl
Creatinine (Cr): 0.50 mg/dl

Prediction result:
The follow-up period range between 5 and 8 weeks.

Figure 6 Web-based input form and prediction output interface for predicting follow-up periods in diabetes patients

Conclusion

This study presented a novel tool for predicting the follow-up period of diabetes patients to help sustain the effects of a good care plan. An imbalanced hospital training dataset was preprocessed using the SMOTE. Important features in the dataset were identified by feature selection using the RF algorithm. The preprocessed dataset was then analyzed by various machine learning models that included: the RF, ETC, AdaBoost, SVM, KNN and ANN classifiers. Model performances were then compared. The experimental results demonstrated that the SVM outperformed the other models; achieving a score of 0.9296 for weighted precision. Furthermore, the predicted results were less accurate when the dataset was not balanced by the SMOTE and RF feature selection was not applied. In addition, we integrated our SVM prediction model into a web-based follow-up period prediction application, providing valuable support to the medical team in making informed decisions.

This research holds the potential to make significant contributions to health monitoring systems, developing a valuable tool for both service users and medical teams. The advancement of categorization techniques for predicting follow-up monitoring of diabetes patients can significantly improve healthcare management. However, as this study focused on a medical records dataset from a specific hospital, the results may not be readily generalized. Future studies should consider the sampling technique used in the sample group before inferences can be made about an entire population. In addition, the follow-up period in the literature varies depending on glycemic targets, individual circumstances, treatment plans, comorbidity and population groups; such as adults, older adults, children, and pregnant individuals. The findings derived from our study are appropriate to the wider population. In future research, we intend to focus on studying specific population groups with comorbidity in order to compare the findings with existing literature.

Acknowledgement

We gratefully acknowledge the Medical records department at Pak Phanang Hospital for the provision of their diabetes dataset. We would also like to express our gratitude to Thomas Duncan Coyne for improving the English language. All errors are the responsibility of the authors. Specifically, we thank Oab Chumpong for the wonderful encouragement that enriches the lives of the third and fourth authors, making this research and family more complete.

Conflict of interest

There are no potential conflicts of interest to declare.

References

1. International Diabetes Federation. IDF Diabetes Atlas. 8th ed. [homepage on the Internet]. Brussels: International Diabetes Federation; 2017 [cited 2024 Sep 4]. Available from: https://diabetesatlas.org/eighth-edition/
2. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. [homepage on the Internet]. Brussels: International Diabetes Federation; 2021 [cited 2023 Jul 24]. Available from: https://diabetesatlas.org/tenth-edition/
3. Strategy and Planning Division. Public Health Statistics A.D. 2021 [homepage on the Internet]. Nonthaburi: Strategy and Planning Division; 2021 [cited 2023 Jul 24]. Available from: https://dmsic.moph.go.th/index/detail/9127
4. Chumpong S, Chumpong K. A retrospective analysis of the relationship of non-communicable diseases. Princess Naradhiwas Univ J 2023;15:193–210.
5. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 13 older adults: standards of care in diabetes—2023. Diabetes Care 2022;46:S216–29.
6. Gregg EW, Geiss LS, Saaddine J, Fagot-Campagna A, Beckles G, Parker C, et al. Use of diabetes preventive care and complications risk in two African-American communities. Am J Prev Med 2001;21:197–202.
7. Hu M, Zhou Z, Zeng F, Sun Z. Effects of frequency of follow-up on quality of life of type 2 diabetes patients on oral hypoglycemics. Diabetes Technol Ther 2012;14:777–82.
8. Zhao Q, Li H, Ni Q, Dai Y, Zheng Q, Wang Y, et al. Follow-up frequency and clinical outcomes in patients with type 2 diabetes: a prospective analysis based on multicenter real-world data. J Diabetes 2022;14:306–14.

9. Diabetes Association of Thailand. Medical Practice Guidelines for Diabetes 2023 [homepage on the Internet]. Bangkok: Diabetes Association of Thailand; 2023 [cited 2023 Aug 12]. Available from: https://www.dmthai.org/new/index.php/activities-and-news/news-pr/naewthang-wech-ptibati-sahrab-rokh-bea-hwan-2566

10. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 2016;82:115–21.

11. Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Comput Sci 2019;165:292–9.

12. Kazerouni F, Bayani A, Asadi F, Saeidi L, Parvizi N, Mansoori Z. Type 2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches. BMC bioinformatics 2020;21:1-13.

13. Pranto B, Mehnaz SM, Mahid EB, Sadman IM, Rahman A, Momen S. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information 2020;11:374.

14. Vijayan V, Ravikumar A. Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. IJCA 2014;95:17.

15. Sooklal S, Hosein P. A benefit optimization approach to the evaluation of classification algorithms. In artificial intelligence and applied mathematics in engineering problems: proceedings of the international conference on artificial intelligence and applied mathematics in engineering. Antalya: Springer International Publishing 2019;2020:35-46.

16. Nnamoko N, Korkontzelos I. Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med 2020;104:101815.

17. Hairani H, Saputro KE, Fadli S. K-means-SMOTE for handling class imbalance in the classification of diabetes with C4. 5, SVM, and naive Bayes. J Teknol dan Sist Komput 2020;8:89-93.

18. Erlin, Marlim YN, Junadhi, Suryati L, Agustina N. Early detection of diabetes using machine learning with logistic regression algorithm. JNTETI 2022;11,88-96.

19. Pradhan M, Bamnote GR. Design of classifier for detection of diabetes mellitus using genetic programming. In proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications. FICTA 2015;1:763-70.

20. Saxena R, Sharma SK, Gupta M, Sampada GC. A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Comput Intell Neurosci 2022;1:3820360.

21. Ilango BS, Ramaraj N. A hybrid prediction model with F-score feature selection for type II diabetes databases. In proceedings of the 1st Amrita ACM-W celebration on women in computing in India; 2010 Sep 16-17; Coimbatore, India. New York: Association for Computing Machinery; 2010.p.1-4.

22. Astuti LW, Saluza I, Yulianti E, Dhamayanti D. Feature selection menggunakan binary wheal optimizaton algorithm (BWOA) pada klasifikasi penyakit diabetes. J Ilm Inform Glob 2022;13:7-12.

23. Gu Q, Wang XM, Wu Z, Ning B, Xin CS. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. J Digit Inf Manag 2016;14:92–103.

24. Chen YF, Lin CS, Wang KA, Rahman LOA, Lee DJ, Chung WS, et al. Design of a clinical decision support system for fracture prediction using imbalanced dataset. J Healthc Eng 2018;2018:13.

25. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 2021;9:39707–16.

26. Kumar R, Arora R, Bansal V, Sahayasheela VJ, Buckchash H, Imran J, et al. Accurate prediction of COVID-19 using chest X-Ray images through deep feature learning model with SMOTE and machine learning classifiers. medRxiv 2020; doi: https://doi.org/10.1101/2020.04.13.20063461.

27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. JLMR 2011;12:2825-30.

28. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 2009;10:213.

29. Breiman L. Random forests. Mach Learn 2001;45:5–32.

30. Sharaff A, Gupta H. Extra-tree classifier with metaheuristics approach for email classification. In: Bhatia SK, Tiwari S, Mishra KK, Trivedi MC, editors. Advances in intelligent systems and computing. Bangkok: Springer link; 2019;p.189-97.

31. Freund Y, Schapire R, Abe N. A short introduction to boosting. J-Jpn Soc Artif Intell 1999;14:771-80.

32. Suthaharan S. Support vector machine. In: Machine learning models and algorithms for big data classification. MA: Springer

US; 2016;p.207–35.

33. Gou J, Ma H, Ou W, Zeng S, Rao Y, Yang H. A generalized mean distance-based k-nearest neighbor classifier. Expert Syst Appl 2019;115:356–72.
34. Majid AM, Utomo WH. Application of discretization and adaboost method to improve accuracy of classification algorithms in predicting diabetes mellitus. ICIC express letters. Part B, Applications: an international journal of research and surveys 2021;12:1177-84.
35. Alfian G, Syafrudin M, Ijaz MF, Syaekhoni MA, Fitriyani NL, Rhee J. A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. Sensors 2018;18:2183.
36. Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, et al. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. Computers 2022;11:136.
37. Fitriyani NL, Syafrudin M, Alfian G, Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access 2020;8:133034–50.
38. Krebs J, Negatsch V, Berg C, Aigner A, Opitz-Welke A, Seidel P, et al. Applicability of two violence risk assessment tools in a psychiatric prison hospital population. Behav Sci Law 2020;38:471-81.
39. Syafrudin M, Alfian G, Fitriyani NL, Anshari M, Hadibarata T, Fatwanto A, et al. A self-care prediction model for children with disability based on genetic algorithm and extreme gradient boosting. Mathematics 2020;8:1590.
40. Yu CS, Lin YJ, Lin CH, Lin SY, Wu JL, Chang SS. Development of an online health care assessment for preventive medicine: a machine learning approach. J Med Internet Res 2020;22:e18585.