Asia-Pacific Journal of Information Technology and Multimedia 13(02) 2024: 253 – 265
https://doi.org/10.17576/apjitm-2024-1302-07

A Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction

Analisis Perbandingan Algoritma Pembelajaran Mesin untuk Ramalan Diabetes

*Waseem Abdulmahdi Alansari, Masnizah Mohd**

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

**Corresponding author: masnizah.mohd@ukm.edu.my*

Received 19 June 2024
Accepted 4 October 2024, Available online 13 October 2024

ABSTRACT

Diabetes mellitus is a chronic metabolic disorder with significant global health implications. The accurate prediction and detection of diabetes using artificial intelligence are crucial for preventing complications and improving patient outcomes. This study focuses on comparing the performance of three machine learning algorithms, namely Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF), in predicting diabetes using two datasets: Pima Indians Diabetes Dataset (PIDD) and the Diabetes 2019 Dataset (DD2019), and the need to identify the most accurate and effective algorithm for diabetes prediction. Nine features which are Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, Pregnancy, BMI, Insulin level, and Outcome been used for the prediction of diabetes. The methodology involves data collection, pre-processing, and training the algorithms using k-fold cross-validation. The results indicate that pre-processing steps and dataset characteristics significantly impact algorithm performance. We discovered that the model with RF consistently achieves the highest accuracy. As per the findings, the RF algorithm attained the maximum accuracy of 77% in the context of PIDD. During the DD2019 experiment, the RF and SVM algorithms demonstrated the highest levels of accuracy, achieving 96.65% and 93.93%, respectively. The study contributes insights into the importance of pre-processing and feature selection in improving algorithm performance. The findings have implications for developing accurate predictive models and improving diabetes detection.

Keywords: Machine Learning, Diabetes Prediction, Naïve Bayes, Support Vector Machines

ABSTRAK

Diabetes mellitus adalah gangguan metabolik kronik dengan implikasi kesihatan global yang ketara. Ramalan tepat dan pengesanan diabetes menggunakan kecerdasan buatan adalah penting untuk mencegah komplikasi dan meningkatkan kesihatan pesakit. Kajian ini memberi tumpuan perbandingan prestasi tiga algoritma pembelajaran mesin, iaitu Naive Bayes (NB), Mesin Vektor Sokongan (SVM), dan Hutan Rawak (RF), dalam meramalkan diabetes menggunakan dua set data: Pima Indians Diabetes Dataset (PIDD) dan Set Data Diabetes 2019 (DD2019). Perbandingan ini bertujuan mengenal pasti algoritma yang tepat dan berkesan untuk ramalan diabetes. Sembilan ciri iaitu Umur, Tekanan Darah, Ketebalan Kulit, Glukosa, Keturunan, Kehamilan, BMI, Tahap Insulin, dan Hasil telah digunakan untuk ramalan diabetes. Metodologi ini melibatkan pengumpulan data, pra-pemprosesan dan latihan algoritma menggunakan pengesahan silang *k-fold*. Keputusan menunjukkan bahawa langkah pra-pemprosesan dan ciri set data memberi kesan ketara kepada prestasi algoritma. Kajian mendapati bahawa model dengan RF secara konsisten mencapai ketepatan tertinggi. Malah, algoritma RF mencapai ketepatan maksimum 77% dalam konteks PIDD. Semasa percubaan DD2019, algoritma RF dan SVM menunjukkan tahap ketepatan tertinggi, masing-masing mencapai 96.65% dan 93.93%. Kajian ini menyumbang kepada kepentingan proses pra-pemprosesan dan pemilihan ciri dalam meningkatkan prestasi algoritma. Penemuan ini mempunyai implikasi untuk membangunkan model ramalan yang tepat dan meningkatkan pengesanan diabetes.

Kata kunci: Pembelajaran Mesin, Ramalan Diabetes, Naive Bayes, Mesin Vektor Sokongan

INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels due to defects in insulin secretion or action. This condition poses a significant risk, as individuals with diabetes have a higher mortality rate compared to those without the condition. The global prevalence of diabetes is steadily increasing, with projections indicating a concerning escalation in the coming years. By 2025, it is estimated that there will be 570.9 million people affected by diabetes, with 1.59 million diabetes-related deaths and 79.3 million Disability-Adjusted Life Years (DALYs) lost. Urgent public health interventions are needed to address the growing burden of diabetes and its associated health risks (Lin et al. 2022). In the field of data mining and machine learning, powerful tools are available for managing large datasets and extracting knowledge. Various machine learning classifiers, such as J48, SVM, KNN, Decision Tree, Random Forest, and Naïve Bayes, have shown promise in predictive analysis and have been used successfully in the medical field for disease diagnosis (Tripathi & Kumar 2022, Sisodia & Sisodia 2023). Leveraging the intelligence of computers, these algorithms enable more accurate prediction and diagnosis of diseases, including diabetes. Particularly, the Pima Indians Diabetes Dataset and the Diabetes 2019 Dataset have been utilized for developing and evaluating machine learning algorithms for early detection and prediction of diabetes. The Pima Indians Diabetes Dataset is a well-known dataset collected from Pima Indian women residing in the United States. It contains medical records with

attributes such as glucose levels, blood pressure, body mass index, and diabetes status. This dataset is widely accessible and has been frequently used in machine learning research (Bhoi 2021, Bhulakshmi & Gandhi 2020). On the other hand, the Diabetes 2019 Dataset was specifically collected for research purposes from individuals based on their lifestyle and family background. The dataset aims to investigate the risk of diabetes among individuals (Tigga & Garg 2020). Investigation into various machine learning classifiers across two datasets can tailor predictions and recommendations to individual patients based on their unique data. This personalized approach can improve the effectiveness of treatment plans and patient management strategies, thus leading to earlier and more accurate detection of diabetes. This has triggered our research to investigate in detail. Naive Bayes, SVM, and Random Forest, have shown promising results in predicting diabetes, exhibiting high accuracy, sensitivity, and specificity in the scope of Pima Indians Diabetes Dataset and the Diabetes Dataset 2019 (Tigga & Garg 2020, Pokala & Kumar 2022, Jain 2022, Islam et al. 2020, Gupta et al. 2021, Sneha & Gangil 2019, Ghosh et al. 2021). Therefore, this study aims to investigate various machine learning algorithms and identify the best models for predicting diabetes using the Pima Indians Diabetes Dataset and the Diabetes Dataset 2019.

This paper presents a comprehensive study on the application of machine learning algorithms for diabetes prediction. Section II provides an overview of related work, highlighting the utility and performance comparison of Naive Bayes, Support Vector Machine, and Random Forest algorithms in diabetes prediction. In Section III, we detail the methodology used in building our prediction model, which addresses the challenges identified in the literature, including dataset quality, feature selection, class imbalance, and overfitting. Section IV presents the experimental results and analysis. Section V summarize and concludes our paper, and outlining potential avenues for future research and improvements in diabetes prediction models.

RELATED WORKS

The successful application of Machine Learning in the field of diabetes prediction has resulted in a substantial body of research literature. This section provides an overview of several research studies conducted in this field. In the study by Costea et al. (2021), three ML methods (Naive Bayes, Random Forest, and Support Vector Machine) were compared for diabetes prediction. Two datasets, the Pima Indians Diabetes Dataset and the Diabetes Dataset 2019, were utilized for evaluation. Preprocessing involved converting non-numeric data into numerical encoding and replacing missing values. Both Random Forest and Support Vector Machine achieved accuracy levels surpassing 80%, with Random Forest exceeding 90% on the Diabetes Dataset 2019. Pokala and Kumar (2022) focused on the performance of Support Vector Machine and Random Forest for diabetes prediction. Each algorithm underwent separate training with a confusion matrix obtained for evaluation. The Matlab-based Random Forest outperformed Support Vector Machine with an accuracy of 79.02% compared to 77.67%. The performance of Random Forest improved with increased data, indicating its potential for accurate prediction of diabetes. Jain (2022) highlighted the effectiveness of ML algorithms in estimating the likelihood of diabetes based on physical symptoms. The Random Forest model exhibited the highest prediction accuracy of 88.14% among SVM and Naive

Bayes models. Important features such as blood pressure, glucose level, insulin, BMI, and pregnancy were selected by medical professionals, further validating the results. Islam et al. (2020) focused on predicting the development of type 2 diabetes. The ensemble Naive Bayes achieved the highest accuracy of 95.94% among evaluated models. Random Forest and Support Vector Machine exhibited acceptable accuracy levels but relatively low sensitivity in predicting diabetes. Results were computed using a 10-fold cross-validation approach. Tigga and Garg [6] utilized their Diabetes 2019 Dataset and the Pima dataset to predict diabetes using various classifiers. Random Forest achieved an accuracy rate of 94.10% for the Diabetes 2019 Dataset and 75% for the Pima dataset. Models were evaluated based on various measures and 10-fold cross-validation.

Ismail and Materwala (2021) proposed an intelligent diabetes mellitus prediction framework and evaluated decision tree-based random forest and support vector machine models. After feature selection, random forest achieved the highest accuracy among the models evaluated. Selected features included age, blood pressure, cholesterol, gender, and obesity. Gupta et al. (2021) focused on diabetes classification using the PIMA Indian Diabetes dataset. SVM outperformed naive Bayes in terms of accuracy (81.17%). Preprocessing steps involved replacing missing values and feature scaling. Sneha and Gangil (2019) proposed a method for early detection of diabetes based on optimal feature selection. They identified 11 relevant attributes through correlation analysis. Decision tree and random forest had the highest specificity values, while naive Bayes achieved the highest accuracy of 82.30%. Ghosh et al. (2021) classified diseases using machine learning algorithms on the Pima Indians diabetes dataset. Random Forest achieved the best results with an accuracy of 99.35% and the highest sensitivity, specificity, and negative predictive value. SVM had the lowest performance, while AdaBoost and Gradient Boosting performed better but were outperformed by Random Forest.

The literature review reveals that machine learning algorithms, including Naive Bayes, SVM, and Random Forest, have shown promising results in predicting diabetes, exhibiting high accuracy, sensitivity, and specificity. However, consensus on the performance comparison of these algorithms is lacking due to several factors, including dataset quality, feature selection, class imbalance, and overfitting, which have been identified as potential limitations and challenges. To address these issues and advance current understanding, we propose a comparative analysis of Naive Bayes, SVM, and Random Forest algorithms for predicting diabetes. Drawing on insights from existing literature and taking into account considerations like dataset scale, feature selection methods, and data pre-processing techniques has contributed in understanding the factors involved. Thus, our study aims to identify the best Machine Learning model for predicting diabetes using the Pima Indians Diabetes Dataset and the Diabetes Dataset 2019.

METHODOLOGY

It is crucial to collect a comprehensive dataset and select suitable algorithms. To develop an effective model for a specific domain, the model's accuracy is evaluated by applying statistical metrics to correctly and incorrectly classified instances. Pre-processing stages, such as data cleaning and normalization, are implemented to ensure data quality. Feature selection

techniques are employed to identify highly correlated features that contribute to improved accuracy. There are seven steps involved in the methodology, to facilitate the selection and optimization of classifiers tailored to the medical condition, ultimately resulting in the best possible accuracy for the model.

Step 1 Dataset Selection: Two datasets were utilized in this study, namely the Pima Indian Diabetes Dataset (PIDD) with a size of 768 instances and 9 features, and the Diabetes 2019 dataset with a larger size of 952 instances and 18 features. The Diabetes 2019 dataset provides a broader range of features compared to PIDD, allowing for a more comprehensive analysis of the predictors associated with diabetes as shown in Table 1.

TABLE 1. Description of dataset features

	Diabetes 2019 Dataset		Pima Indian Diabetes Dataset (PIDD)	
No	Feature (18)	Description	Feature (9)	Description
1	Age	Age in years	Age	Age in years
2	Gender	Male or Female	-	-
3	Family_Diabetes	Family history with diabetes (Yes or No)	Diabetes Pedigree Function	Diabetes pedigree function, which provides a genetic measure of diabetes.
4	highBP	Diagnosed with high blood pressure (Yes or No)	-	-
5	PhysicallyActive	Walk/run or can be physically active	Skin Thickness	Shows the triceps skin thickness in mm
6	BMI	Body Mass Index	BMI	Body Mass Index
7	Smoking	Whether the person smokes or not (Yes or No)	-	-
8	Alcohol	Alcohol consumer (Yes or No)	-	-
9	Sleep	Hours of sleep	-	-
10	SoundSleep	Hours of sound sleep	-	-
11	RegularMedicine	Regular intake of medicine (Yes or No)	-	-
12	JunkFood	Junk food consumer (Yes or No)	-	-
13	Stress	How much stress taken	-	-
14	BPLevel	High/normal/low	Blood Pressure	Shows the diastolic blood pressure in mm Hg.
15	Pregnancies	No. of pregnancies	Pregnancies	No. of pregnancies
16	Pdiabetes	Gestation diabetes (Yes or No)	Insulin	Shows 2-Hour serum insulin (mu U/ml)
17	UrinationFreq	Frequency of Urination (Not much or Quite much)	Glucose (mg/dl)	Shows the plasma glucose concentration level (2 hours)
18	Diabetic	Yes or No	Outcome	Shows either 0 or 1. 0 means non-diabetic and 1 means diabetes.

Step 2 Preprocessing: The preprocessing stage involved several steps to ensure data quality and prepare the data for modeling. Missing values were addressed using imputation techniques,

which involved estimating missing values based mean of the feature information. Imputation was chosen over data deletion because the approach aimed to avoid removing any samples and potentially losing important information that could affect the learning process.

Step 3 Data Normalization: To handle the varying units and scales of features, the z-score normalization method was employed. This method calculates the feature's z-score by subtracting the mean and dividing by the standard deviation. Normalization using z-scores helps mitigate the effect of outliers and ensures that all features contribute proportionally to the model's training process.

Step 4 Class Imbalance Handling: The datasets exhibited class imbalance, where the number of positive and negative cases differed significantly. To address this, oversampling technique was employed. This approach ensured that both positive and negative cases were adequately represented during model training, preventing the loss of important information that could impact the learning process.

Step 5 Feature Selection: The feature selection process aimed to identify the most informative variables for the prediction models. The Exhaustive Feature Selector method was chosen for its ability to systematically evaluate all possible feature combinations. By exhaustively considering different subsets of features and selecting the best one based on accuracy, this method ensured that the models were built with the most relevant features for each classifier. This reduced complexity, improved reliability, stability, and classification accuracy, and contributed to the overall performance of the approach.

Step 6 Model Construction: Three classification algorithms, SVM, NB, and RF, were employed to build the prediction models. These algorithms were selected based on their proven track record of providing accurate results in various classification scenarios. Additionally, they could handle both numerical and categorical data, making them suitable for the diverse feature types present in the diabetes datasets.

Step 7 Evaluation Technique: The evaluation technique employed for assessing the performance of the prediction models was k-fold cross-validation. This technique was chosen to address the potential issue of overfitting caused by oversampling and normalization during the preprocessing stage. By dividing the data into multiple subsets and iteratively training and testing the models, k-fold cross-validation provided a robust estimation of their performance. In this study, a value of K equal to 20 was used, ensuring comprehensive coverage of the entire dataset during the evaluation process. To evaluate the efficacy of different classifiers, various statistical metrics such as accuracy, sensitivity (recall), specificity, and the confusion matrix were computed. These metrics relied on the classification labels, such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which represented the agreement or disagreement between the predicted outcomes and the actual values present in the dataset. Figure 1 shows the steps sequence of the suggested approach using the Pima Indians dataset and using the Diabetes2019 dataset in the prediction process.

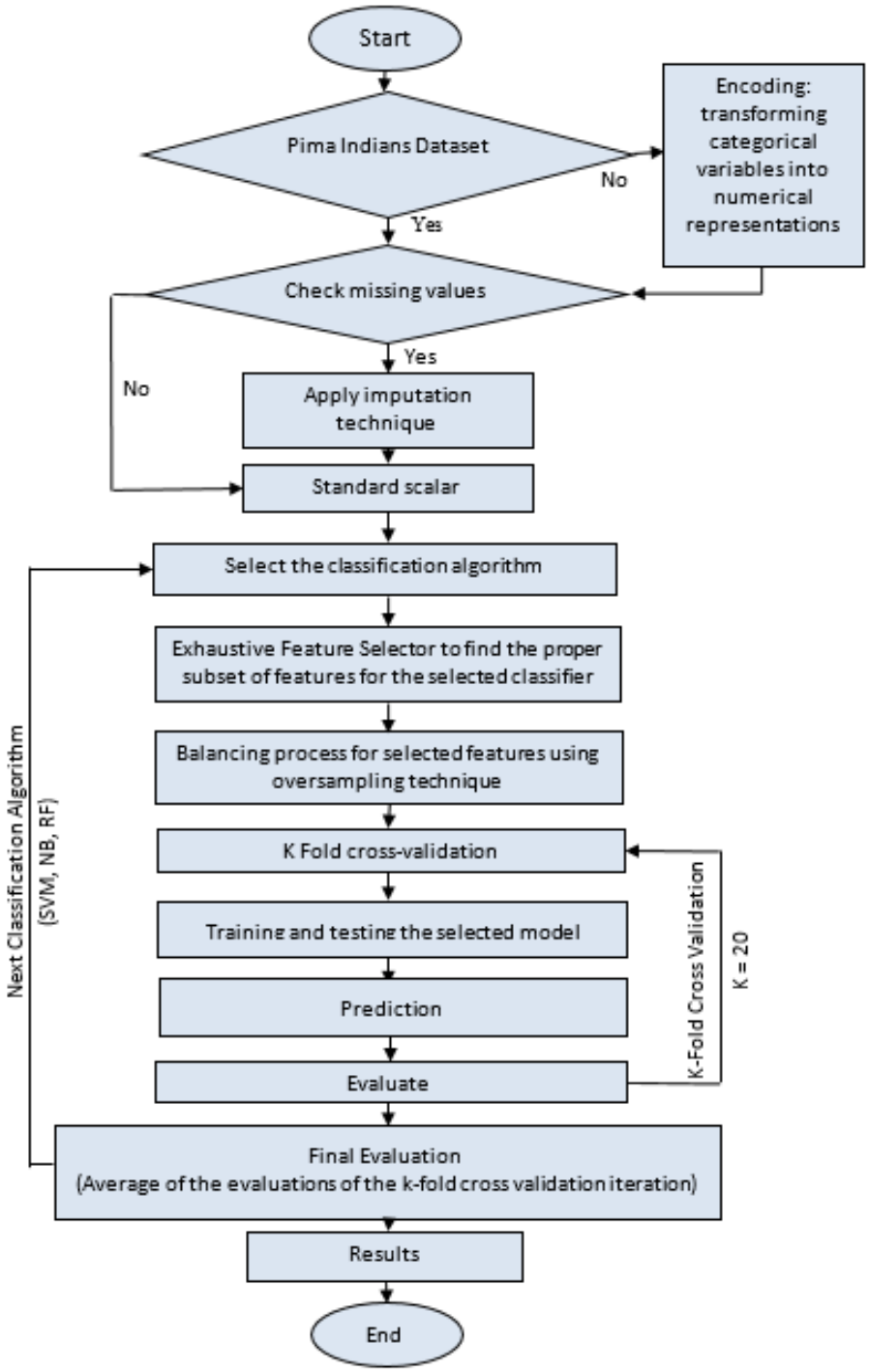

FIGURE 1. Flowchart of the methodology

RESULTS AND DISCUSSION

This section provides a comparative and analytical assessment of the outcomes obtained from three machine learning-based diabetes prediction systems: Support Vector Machine, Naive Bayes, and Random Forest. Accuracy Metric. The accuracy metric provides an overall assessment of algorithm performance, and in this study, the SVM, Naive Bayes, and Random

Forest algorithms were evaluated on the Pima Indians and Diabetes 2019 datasets. The results as shown in Figure 2 revealed that SVM achieved a moderate accuracy of 70.89% on the Pima Indians dataset but showed excellent performance with an accuracy of 93.39% on the Diabetes 2019 dataset. Naive Bayes achieved an accuracy of 75.54% on the Pima Indians dataset and 85.81% on the Diabetes 2019 dataset. Random Forest outperformed the other algorithms, achieving an accuracy of 77% on the Pima Indians dataset and an outstanding accuracy of 96.65% on the Diabetes 2019 dataset. These results indicate that Random Forest effectively captures complex relationships, making it a promising algorithm for diabetes prediction. The exhaustive feature selection process, coupled with pre-processing techniques like imputation and normalization, contributes to the improved accuracy of the algorithms. Additionally, the larger number of features in the Diabetes 2019 dataset provides more information for the algorithms to learn from, enhancing their accuracy. Overall, the combination of feature selection, pre-processing, algorithm complexity, and the data characteristics contributes to the higher accuracy achieved by the models.

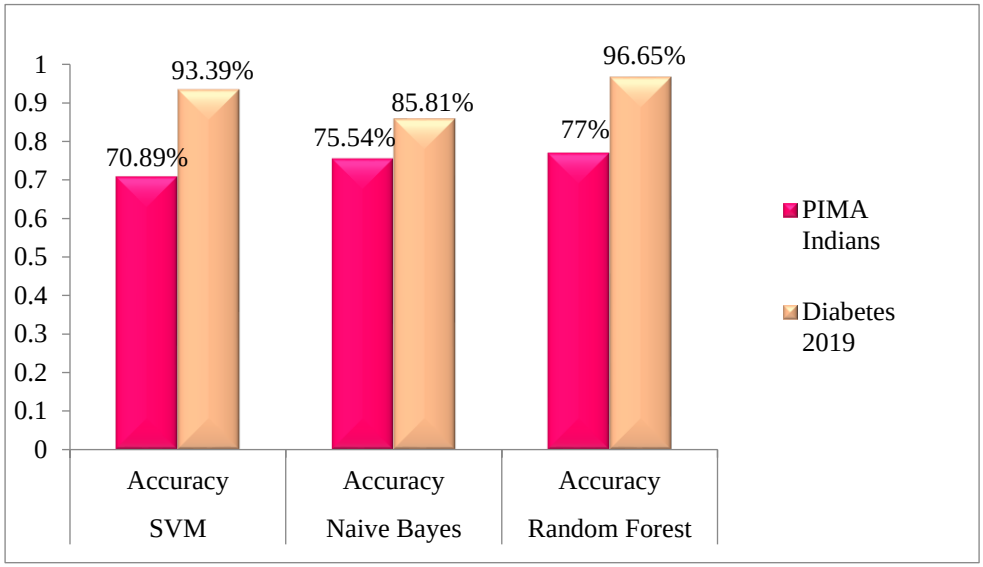

FIGURE 2. Comparative Accuracy of Classifiers on the Pima Indians Diabetes and Diabetes 2019 Datasets.

Sensitivity Metrics

The sensitivity metric, also known as the true positive rate or recall, measures the ability of algorithms to correctly identify positive cases as shown in Figure 3.

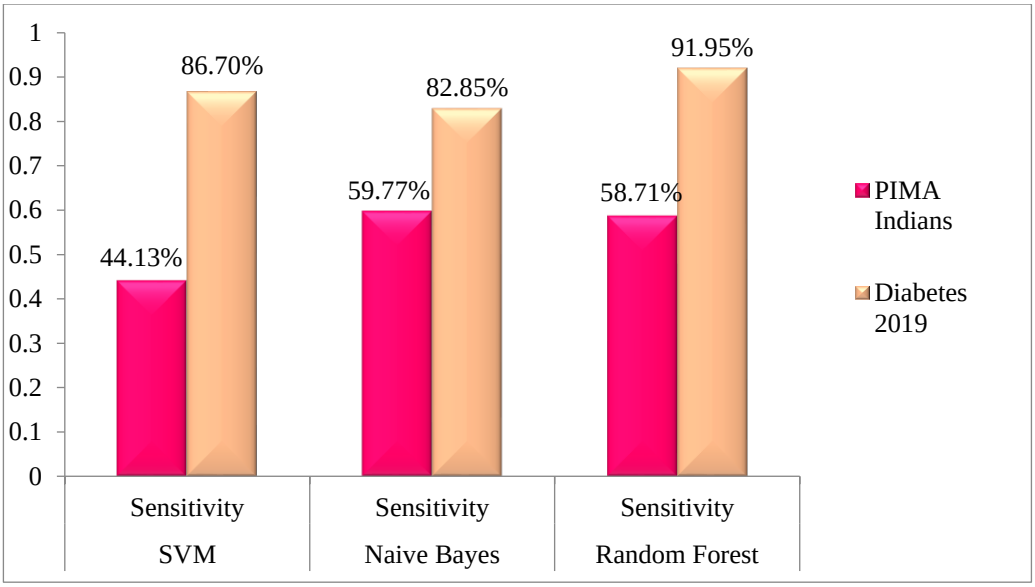

FIGURE 3. Comparative Sensitivity of Classifiers on the Pima Indians Diabetes and Diabetes 2019 Datasets.

On the Pima Indians dataset, SVM achieved a sensitivity of 44.13%, indicating its ability to accurately identify individuals with diabetes. Naive Bayes demonstrated a higher sensitivity of 59.77%, while Random Forest achieved a sensitivity of 58.71%. Moving to the Diabetes 2019 dataset, SVM exhibited a significantly higher sensitivity of 86.7%, followed by Naive Bayes with 82.85%. Random Forest achieved the highest sensitivity of 91.95%. The differences in sensitivity performance can be attributed to the algorithms' characteristics and their suitability for the datasets. Naive Bayes, despite assuming feature independence, demonstrates good sensitivity due to the alignment between the selected features and its probabilistic nature. Random Forest, with its ensemble approach and complexity, excels in capturing intricate relationships among features, resulting in high sensitivity. SVM benefits from separability in higher-dimensional space, allowing it to achieve good sensitivity. The exhaustive feature selection process and preprocessing stages play a crucial role in enhancing sensitivity by providing informative features and appropriate data handling. The findings highlight the significance of feature selection and preprocessing in improving sensitivity for diabetes prediction.

Specificity Metrics

Figure 4 shown the specificity metric measures the algorithms' ability to correctly identify negative cases. SVM achieved a specificity of 85.48% on the Pima Indians dataset. On the Diabetes 2019 dataset, SVM exhibited a higher specificity of 91.79%, further highlighting its ability to accurately identify negative cases. Naive Bayes demonstrated a specificity of 83.46% on the Pima Indians dataset. On the Diabetes 2019 dataset, Naive Bayes exhibited a relatively lower specificity of 74.73%, suggesting a higher false positive rate in distinguishing individuals without diabetes in this dataset. Random Forest achieved a specificity of 86.13% on the Pima Indians dataset and an impressive specificity of 94.13% on the Diabetes 2019 dataset.

The specificity results highlight the effective performance of SVM, Naive Bayes, and Random Forest in correctly classifying negative cases in both datasets. Random Forest consistently achieved the highest specificity values, indicating its ability to avoid false positives and accurately identify individuals without diabetes. This can be attributed to its modeling technique, including the use of an ensemble approach and the exhaustive feature selection process. By considering a larger number of features and capturing complex relationships, Random Forest benefits from the informative features selected through the exhaustive feature selection, leading to strong specificity in both datasets.

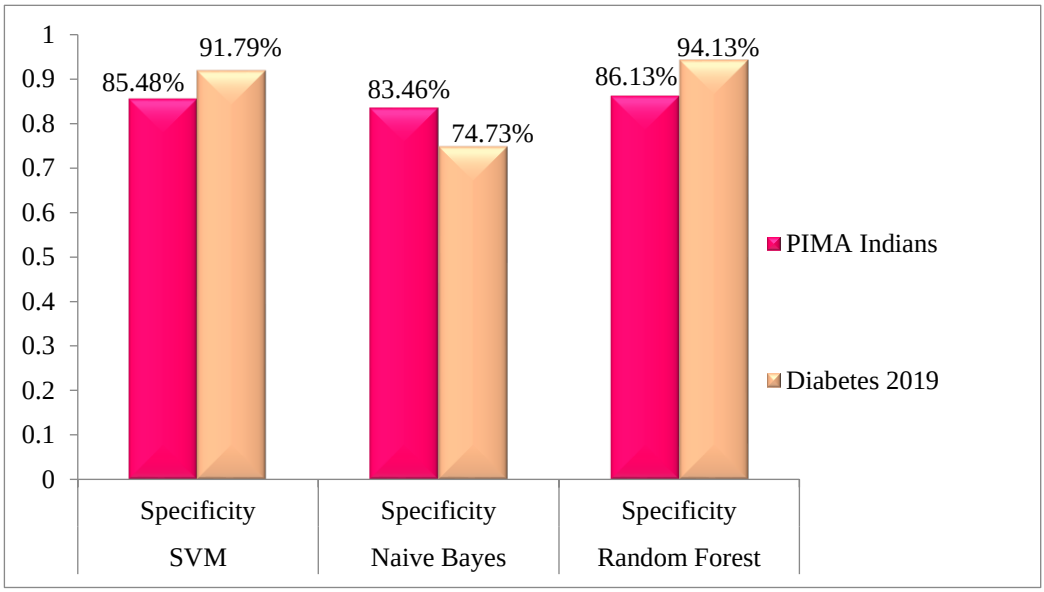

FIGURE 4. Comparative Specificity Assessment of Classifiers on the Pima Indians Diabetes and Diabetes 2019 Datasets.

Confusion Matrix

The confusion matrix results as illustrated in Table 2 showed that SVM, Naive Bayes, and Random Forest exhibited varying performance on the Pima Indians and Diabetes 2019 datasets. SVM demonstrated better performance on the Diabetes 2019 dataset, with higher numbers of true positives and true negatives, indicating its ability to accurately predict both diabetic and non-diabetic instances. Naive Bayes also performed well on the Diabetes 2019 dataset, achieving higher numbers of true positives and true negatives. Random Forest exhibited exceptional performance on the Diabetes 2019 dataset, with the highest numbers of true positives and true negatives, highlighting its strong predictive capabilities. The low numbers of false positives and false negatives for all algorithms indicate a good balance between sensitivity and specificity. The exhaustive feature selection process enhanced the performance of Random Forest by selecting informative features that align well with its modeling technique. Overall, all three algorithms benefited from the selected features, enabling them to capture relevant patterns and make accurate predictions.

TABLE 2. Support Vector Machine, Naïve Bayes, and Random Forest classifiers confusion matrixes values

	Support Vector Machine		Naïve Bayes		Random Forest	
	Pima Indians	Diabetes 2019	Pima Indians	Diabetes 2019	Pima Indians	Diabetes 2019
TP	6	12	10	11	12	13
TN	18	32	17	32	19	33
FP	3	2	4	2	2	1
FN	11	1	7	2	5	0

The results mentioned in Table 2 are taken as an average of different feature groups that have been selected by exhaustive feature selection for the individual classification models against the performance metrics. The groups have been determined as shown in Table 3.

TABLE 3. Comparison of dataset features

PIMA Indians	Diabetes 2019
• Selected features for SVM (SVM Pima): Glucose', 'BloodPressure', 'BMI', DiabetesPedigreeFunction'. • Selected features for NB (NB Pima): 'Glucose', 'BMI', DiabetesPedigreeFunction'. • Selected features for RF (RF Pima): 'Pregnancies', 'Glucose', 'BloodPressure', 'Insulin', 'BMI', 'Age'.	• Selected features for SVM (SVM D2019): 'Age', 'Gender', 'Family_Diabetes', 'PhysicallyActive', 'BMI', 'Smoking', 'Alcohol', 'Sleep', 'SoundSleep', 'RegularMedicine', 'JunkFood', 'Stress', 'BPLevel', 'Pregancies'. • Selected features for NB (NB D2019): 'Age', 'highBP', 'PhysicallyActive', 'Smoking', 'Sleep', 'SoundSleep', 'RegularMedicine', 'JunkFood', 'UriationFreq'. • Selected features for RF (RF D2019): 'Gender', 'Family_Diabetes', 'highBP', 'PhysicallyActive', 'BMI', 'Sleep', 'SoundSleep', 'JunkFood', 'Stress'.

Table 4 shows the summary of the performance scores (Accuracy, Sensitivity, Specificity) obtained by applying the algorithms (RF, SVM, NB) using the above selected features groups as shown in Table 3 from the two datasets Pima Indians and Diabetes 2019 separately.

TABLE 4. Summary of scores of three machine learning classification models on Pima Indians and Diabetes 2019 datasets using the three metrics**.**

Model Name	Performance Metric	Group of Features		PIMA Indians	Diabetes 2019
		PIMA Indians	Diabetes 2019		
SVM	Accuracy	SVM Pima	SVM D2019	70.89%	93.39%
	Sensitivity	SVM Pima	SVM D2019	44.13%	86.70%
	Specificity	SVM Pima	SVM D2019	85.48%	91.79%
Naïve Bayes	Accuracy	NB Pima	NB D2019	75.54%	85.81%
	Sensitivity	NB Pima	NB D2019	59.77%	82.85%
	Specificity	NB Pima	NB D2019	83.46%	74.73%
Random Forest	Accuracy	RF Pima	RF D2019	77.00%	96.65%
	Sensitivity	RF Pima	RF D2019	58.71%	91.95%
	Specificity	RF Pima	RF D2019	86.13%	94.13%

CONCLUSION

In conclusion, this study employed SVM, Naive Bayes, and Random Forest (RF) algorithms to develop machine learning models for predicting diabetes. The models underwent comprehensive pre-processing, feature selection, and evaluation using performance metrics and k-fold cross-validation. The contribution of this works are the comparative analysis of various machine learning classifiers across two datasets, Pima Indians Diabetes Dataset (PIDD) and the Diabetes 2019 Dataset (DD2019). It provides insights into the importance of pre-processing and feature selection in improving algorithm performance. Thus, the findings have implications for developing accurate predictive models and improving diabetes detection. The results highlighted the influence of pre-processing steps, dataset size, feature selection, and algorithm characteristics on the models' performance. Random Forest demonstrated the highest accuracy and sensitivity, making it the most accurate algorithm for diabetes prediction in this study. This is because RF is an ensemble learning method that combines the predictions of multiple decision trees, which helps reduce overfitting and increase the overall robustness of the model. In addition, RF can assess the importance of different features during training, which helps in selecting the most relevant features and reducing noise from irrelevant data. Recommendations for future research include expanding the datasets by considering complex diabetes dataset, and involving healthcare professionals in model development. Overall, this study contributes to accurate diabetes prediction and has implications for diagnosis and management.

ACKNOWLEDGEMENT

This study was supported by the Fundamental Research Grant Scheme (FRGS/1/2021/ICT02/UKM/02/1) from the Ministry of Higher Education, Malaysia.

REFERENCES

Bhoi, S.K., 2021. Prediction of diabetes in females of Pima Indian heritage: a complete supervised learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), pp.3074-3084.

Bhulakshmi, D. & Gandhi, G., 2020. The prediction of diabetes in Pima Indian women mellitus based on XGBoost ensemble modeling using data science. *Technical Report*, EasyChair.

Costea, N.E., Moisi, E.V. & Popescu, D.E., 2021. Comparison of machine learning algorithms for prediction of diabetes. In *2021 16th International Conference on Engineering of Modern Electric Systems (EMES)*. IEEE.

Ghosh, P., et al., 2021. A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192, pp.467-477.

Gupta, S., Verma, H.K. & Bhardwaj, D., 2021. Classification of diabetes using Naive Bayes and support vector machine as a technique. In *Operations Management and Systems Engineering: Select Proceedings of CPIE 2021*. Springer.

Islam, M.S., et al., 2020. Advanced techniques for predicting the future progression of type 2 diabetes. *IEEE Access*, 8, pp.120537-120547.

Ismail, L. & Materwala, H., 2021. IDMPF: intelligent diabetes mellitus prediction framework using machine learning. *Applied Computing and Informatics*, (ahead-of-print).

Jain, V., 2022. Diabetes prediction using support vector machine, naive Bayes, and random forest machine learning models. In *2022 6th International Conference on Electronics, Communication, and Aerospace Technology*. IEEE.

Lin, X., et al., 2022. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Scientific Reports*, 10(1), pp.1-11.

Pokala, V.S.K. & Kumar, N.S., 2022. Analysis and comparison for prediction of diabetic among pregnant women using innovative support vector machine algorithm over random forest algorithm with improved accuracy. *Cardiometry*, (25), pp.956-962.

Sisodia, D. & Sisodia, D.S., 2023. Prediction of diabetes using classification algorithms. *Procedia Computer Science*, 132, pp.1578-1585.

Sneha, N. & Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal feature selection. *Journal of Big Data*, 6(1), pp.1-19.

Tigga, N.P. & Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167, pp.706-716.

Tripathi, G. & Kumar, R., 2022. Early prediction of diabetes mellitus using machine learning. In *2020 8th International Conference on Reliability, Infocom Technologies, and Optimization (Trends and Future Directions) (ICRITO)*. IEEE.