Mathematical Sciences and Informatics Journal
Vol. 3, No. 2, Nov. 2022, pp. 25-34
http://www.mijuitmjournal.com DOI: 10.24191/mij.v3i2.20118

Diabetec Disease Classifier Based on Three Machine Learning Models

'Ayuni Zamri
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
2019291632@student.uitm.edu.my

Mohd Faaizie Darmawan
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
faaizie@uitm.edu.my

Shahirah Mohamed Hatim
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
shahirah88@uitm.edu.my

Ahmad Firdaus Zainal Abidin
Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Pahang, Malaysia
firdausza@ump.edu.my

Mohd Zamri Osman
Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Pahang, Malaysia
zamriosman@ump.edu.my

Article Info

Article history:

Received July 28, 2022
Revised Sept. 27, 2022
Accepted Oct 20, 2022

Keywords

Diabetes
Support Vector Machine
K-Nearest Neighbors
Random Forest
Classification

ABSTRACT

Diabetes is generally acknowledged as an increasing epidemic that affects nearly every country, age group, and economy on the earth. Without doubt, this worrisome statistic requires immediate response. The healthcare business produces vast volumes of complicated data on regular basis from a variety of sources, including electronic patient records, medical reports, hospital gadgets and billing systems. Traditional approaches cannot handle and interpret the massive volumes of data created by healthcare transactions because they are too complicated and numerous. Machine learning has been used to many sectors of medical health due to the rapid growth of the technology. The aim of this research is to aid the medical professionals to diagnose patients whether the patients is diabetic or not diabetic, by applying machine learning algorithms, and evaluate the results to find the best algorithm to predict diabetic diseases. Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Random Forest (RF) are implemented in this research. Performance measures which is accuracy score is utilized to determine the performance for each model. Based on the results for each model, the model with the highest accuracy score obtained is SVC Linear with the score of 78.62%. The proposed models are valuable to be used for medical practice or in assisting medical professionals in making treatment decisions.

Corresponding Author:
Name: Mohd Faaizie Darmawan
Affiliation: Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
email: faaizie@uitm.edu.my

PENERBIT PRESS UNIVERSITI TEKNOLOGI MARA

1. Introduction

Diabetes is one of the main sources of morbidity and is expected to rise considerably throughout the following many years. Diabetes is becoming more widespread in people's daily lives as their living standard rise. It is claimed that half of diabetic patients are oblivious of their condition, making them more vulnerable to diabetic complications [1]. This chronic illness is a major cause of mortality in people all over the world. In 2012, it was the fifth common cause of death for women and the eighth most common cause for both sexes. In the year 2013, global diabetes data indicated that 382 million people worldwide were affected by the disease. Chronic illness has a monetary burden attached to them. Governments and individuals spend a significant amount of money to control the chronic disease [2].

Diabetes diagnosis is regarded as difficult problem for data analysis. Also, when there are too many data that needs to be interpret in a time, human error may surface. The earlier diagnosis is detected, the easier it will be to control. In order to anticipate the disease at an early stage, machine learning algorithms are utilized [3]. Machine learning may assist patients in making a preliminary diagnosis of diabetes based on their daily physical examination data, and it can also be used by medical professionals as a reference [4].

There are another three sections in this paper which are Section Materials and Method, Methodology, Results and Discussions, and Conclusions. Section Materials and Method explain the data collection and preprocessing of the dataset. Section Methodology explain on the development of the three proposed models, Section Results and Discussion explain the results produced by each model and Section Conclusions conclude the paper generally.

2. Literature Review

Several methods, including the standard machine learning model, have been applied to predict diabetes such as SVM [4]–[7], KNN [8]–[11], and RF [3], [12]–[14] where the good results have been produced for each model. SVM model consist of support vector regressions (SVR) and support vector classifier (SVC). They are among most accurate and robust models in data mining algorithm. The general idea of SVM is to determine a hyperlane which separates the d-dimensional perfectly into two classes.

For the KNN model, the model is based on the nearest neighbors distance of each sample. It is a simple, easy to implement supervised machine learning model. The KNN algorithm believes that related things are located nearby. In other words, related things are located close to one another, thus the nearest neighbors of the target sample will be assumed as its group. While the RF model is also a supervised machine learning model, which can be used for both regression classification tasks. RF is another classification technique based on a tree-based algorithm that involves building several trees (decision trees), then combining their output to improve the generalization ability of the model. Thus, this research applying these three machine learning models is classify the diabetes and to select the best model among them. The development of all these four models is by using Phyton where the Scikit-Learn contains many machine learning libraries are utilized for the development process.

3. Methodology

3.1 Data Collection and Pre-processing of the Data

The diabetes dataset is originally from National Institute of Diabetes and Digestive and Kidney Diseases and obtained from Kaggle Datasets website. All of the patients are Pima Indian females at least 21 years old age. This dataset contains a total of 768 samples, composed of 8 attributes and 1 output outcome of the patients labelled (0:Not Diabetic, 1:Diabetic). The objective is to estimate whether the patient has diabetes or not. Table 1 summarize the datasets attributes description.

Table 1. The Attribute of the Diabetes Dataset

Attribute	Attribute Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum Insulin (mu U/ml)
BMI	Body Mass Index (weight in kg/(height in m)^2)
DiabetesPF	Diabetes Pedigree Function
Age	Age (years)
Outcome	Class variable (0 or 1)

To start the development of each model, the attributes in the dataset need to be divided into data input and data target. All the process will be done in Phyton. The data input is the attributes that will be use as input in developing the model while the data target is the attribute that will be classified from the developing model. For the data target, the attribute outcome will be chosen while the rest of the attributes will be chosen as data input. The data input and data target are then will be divided into training and testing dataset with ratio of 80:20. The dataset is divided using the *test_train_split* function that are available in Scikit-Learn library. The advantage of this function is it will randomly split the data into training and testing dataset, and it can minimize bias in evaluation and validation process.

In dividing the dataset into data input and data target, firstly, import the pandas library to read the *diabetes.csv* files that are in the same folder as the Phyton coding. Using the coding as shown in Figure 1, the *diabetes.cvs* file is then stored in variable "*data*" declared in the second line of coding.

```
import pandas as pd
data = pd.read_csv('diabetes.csv')
data
```

Figure 1. Import the dataset into the coding

Next, the data input is set by dropping the attribute 'Outcome' from the variable data and the data target is set by choosing only attribute Outcome. The coding is shown in Figure 2.

```
data_input = data.drop(columns = ['Outcome'])
data_input
data_target = data['Outcome']
```

Figure 2. Coding for selecting data input and data target

The data input and data target are then divided into training and testing dataset by ratio 80:20, using function *train_test_split* from library *sklearn.model_selection*, as shown in Figure 3. There are four new variables produced by this coding which are *input_train*, *input_test*, *target_train*

and *target_test*. The *input_train* and *target_train* will be used by the proposed models to train the model while the *input_test* and *target_test* is use to produced the results and calculate accuracy.

```
from sklearn.model_selection import train_test_split
input_train, input_test, target_train, target_test =
train_test_split(data_input, data_target, test_size=0.2)
```

Figure 3. Coding in dividing the data input and data target into training and testing dataset

3.2 Development of SVM Model

SVM model is one of the supervised learning algorithms and mostly used for classification problem. SVM is the best model against overfitting and provides sample data accuracy quickly [15]. There are four kernels to be used for the SVM kernels which are *linear*, *rbf*, *poly* and *sigmoid*. Each of the kernels produce accuracy score where the kernel that produce the highest accuracy score will be chosen to be compared with the other model. Figure 4 shows the flowchart of the SVM model while Figure 5 shows the example of coding in training and testing the model the SVM model for kernel linear.

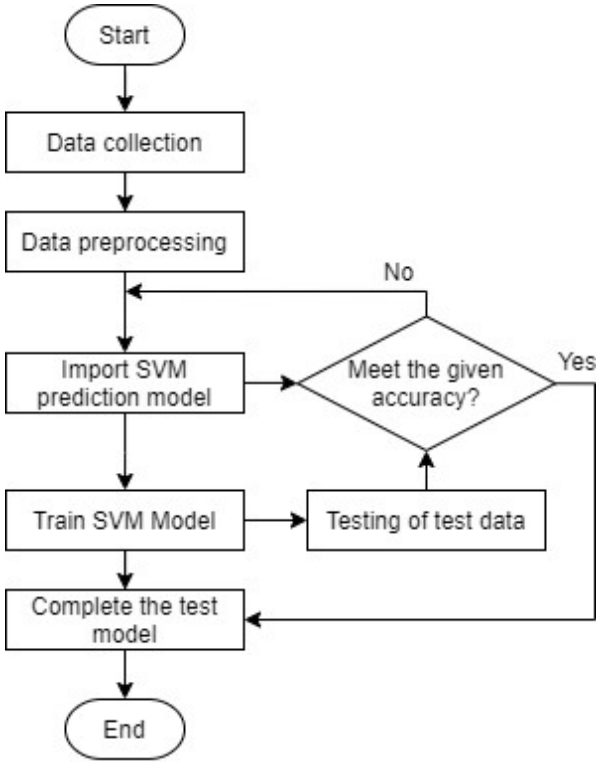

Figure 4. Flowchart of SVM model

```
from sklearn.svm import SVC
svcLinearTrain = SVC(kernel = 'linear')

svcLinearTrain.fit (input_train, target_train)
predictionLinear = svcLinearTrain.predict(input_test)
```

Figure 5. Coding for the development of SVM model

From the Figure 5, the first line is to import the library SVM model and the first kernel which is linear kernel is chosen for the development of SVM model. The model is then trained using data *input_train* and *target train*, and the trained model is then tested using the data *input_test* as shown in the last line in the Figure 5. The predicted outcome for each sample in *input_test* is then stored in variable '*predictionLinear*' to be used for calculation of accuracy score for linear kernel.

3.3 Development of KNN Model

K-Nearest Neighbour is one of the most well-researched supervised machine learning classifiers. KNN classifies each data instance into its closest neighbourhood using distance metrics (e.g, Euclidean, Manhattan, and Minkowski functions) [15]. KNN is a continuous data classifier. It is also called a case-based reasoning method, and it is utilized in a variety of applications including sample popularity and statistical estimate. The KNN algorithm is a non-parametric approach that may be applied to a classification and regression problems. For this research, the number of neighbour used are 10, 20 and 30, where the accuracy score of each number of neighbors will be compared and the highest accuracy will be chosen to be compared with other models. Figure 6 shows the flowchart of the development of KNN model while Figure 7 shows the coding using Phyton. The coding is similar with the coding for SVM model where the Scikit-Learn library is used for the development of the model.

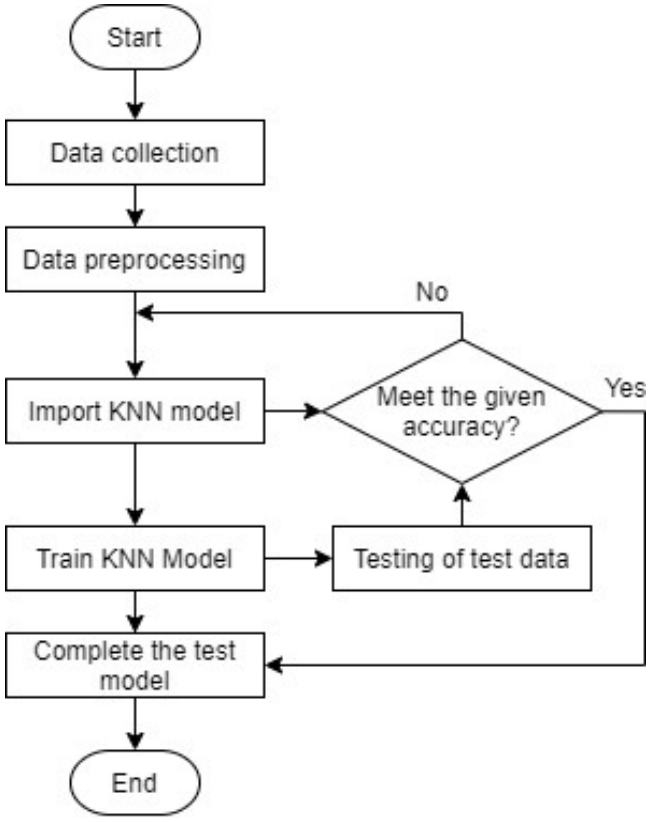

Figure 6. Flowchart of KNN model

```
from sklearn.neighbors import KNeighborsClassifier
knn10 = KNeighborsClassifier (n_neighbors = 10)
knn10.fit (input_train, target_train)
precitionKNN10 = knn10.predict (input_test)
```

Figure 7. Coding for the development of KNN model

3.4 Development of RF Model

The RF model may be thought of as a composite of numerous Decision Tree occurrences. The Random Forest algorithm is mostly used for classification issues and inherits certain from the Decision Tree approach. As name implies, Random Forest exploits the collective intelligence of several Decision Trees, and its output is the class or category chosen by a majority of the trees [16]. Based on the foregoing, Random Forest may be used to categorize. Figure 8 shows the flowchart in developing the RF model while Figure 9 shows the coding utilized the Scikit-Learn library for the development. The coding is similar with the coding for SVM and KNN model.

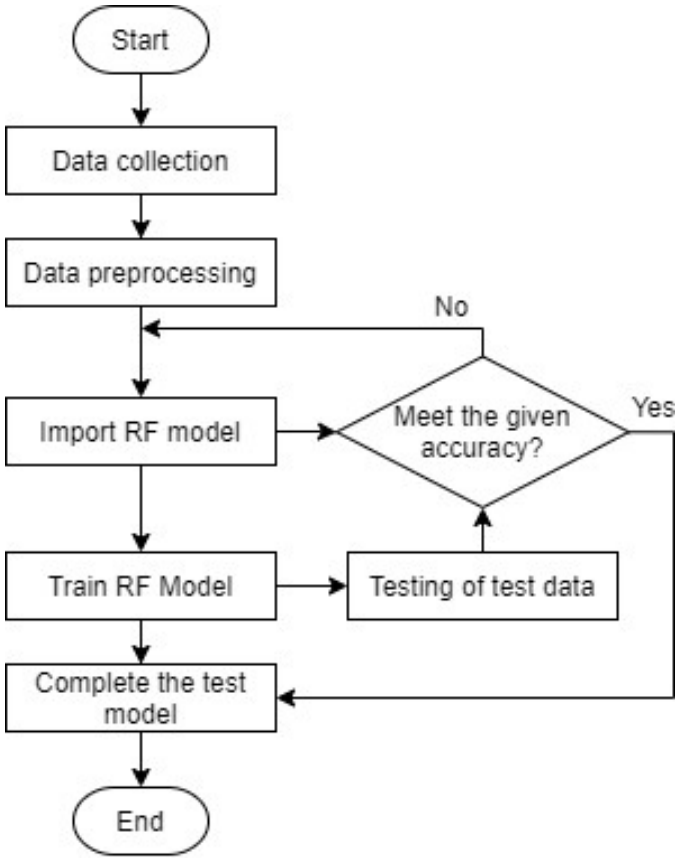

Figure 8. Flowchart of RF model

```
from sklearn.ensemble import RandomForestClassifier
RF100 = RandomForestClassifier (n_estimators = 100)
RF100.fit (input_train, target_train)
precitionRF100 = RF100.predict (input_test)
```

Figure 9. Coding for the development of RF model

4. Results and Discussion

The results of applying the three classification models; SVM, KNN, and RF on the diabetes dataset are shown in Table 2, Table 3 and Table 4, respectively. The results are based on the

accuracy score where the highest accuracy score will be chosen as the best model in will be summarized in Table 5 for comparison purpose.

Table 2. The accuracy score of SVM model for each kernel

Kernels	**Accuracy (%)**
Linear	78.62
RBF	75.86
Poly	76.55
Sigmoid	48.28

Table 3. The accuracy score of KNN model for each number of neighbors

Number of nearest neighbour	**Accuracy (%)**
10	70.34
20	73.72
30	73.79

Table 4. The accuracy score of RF model for each number of trees

Number of Trees	**Accuracy (%)**
100	73.10
200	70.89
300	71.35

From the Table 2, SVC Linear kernel records the highest accuracy of 78.62% among other kernels. Thus, Linear kernel is chosen and summarize in Table 5 to be compare with the KNN, and RF models.

Overall results for the KNN model in Table 3 shows that the KNN with 30 number of neighbours produced the highest accuracy score of 73.79% compare to the other number of neighbors. Thus, KNN with number of neighbours 30 is chosen and tabulated in Table 5 to be compared with SVM, and RF models.

For the RF model, the Table 4 shows that the RF with number of trees 100 produce the highest accuracy score compare to the other number of trees. Thus, the RF model with number of trees 100 is chosen to be summarize in Table 5 for the comparison purpose with the SVM and KNN model.

Table 5. The accuracy score produced by the best of each model

Number of Trees	**Accuracy (%)**
SVC Linear	78.62
KNN (n_neigbors = 30)	73.79
RF (number of trees = 100)	73.10

Table 5 displays the combination of the best-chosen results of each model in one table. Based on the table, the results show that the SVC Linear produced the highest accuracy of 78.62% compared to the other model followed by KNN with accuracy of 73.79% and RF with accuracy of 73.10%.

5. Conclusion

Diabetes is one of the chronic diseases need to be considered for the fast diagnosis before it's too late. The earlier diagnosis can prevent the disease spread widely and easy to be controlled. In order to anticipate the disease at an early stage, machine learning model which are SVM, KNN and RF model are applied. From the results produced by each model, the SVM model using kernel linear shows the highest accuracy compared to the KNN and RF model. Thus, it can be concluded that the SVM linear model is the best model in classifying the outcome of the diabetes disease followed by the KNN model and RF model. The model is then will be enhanced for a better result in the future and for now, it is suggested for this model to be used for medical practice in early diagnosis of the diabetes disease.

Acknowledgements

The authors gratefully acknowledge the Universiti Teknologi MARA (UiTM), Perak branch.

Conflict of Interest

The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] K. Papatheodorou, M. Banach, E. Bekiari, M. Rizzo, and M. Edmonds, "Complications of Diabetes 2017," *J. Diabetes Res.*, pp. 1–4, 2018.

[2] T. M. Alam *et al.*, "A model for early prediction of diabetes," *Informatics Med. Unlocked*, vol. 16, pp. 1–6, 2019.

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Genet.*, vol. 9, no. 515, 2018.

[4] N. I. A. Kader, U. K. Yusof, and S. Naim, "A Study of Diabetic Retinopathy Classification Using Support Vector Machine," *Int. J. Eng. Technol.*, vol. 7, pp. 521–527, 2018.

[5] A. Viloria, Y. Herazo-Beltran, D. Cabrera, and O. B. Pineda, "Diabetes Diagnostic Prediction Using Vector Support Machines," *Procedia Comput. Sci.*, vol. 170, pp. 376–381, 2020.

[6] S. Sistla, "Predicting Diabetes using SVM Implemented by Machine Learning," *Int. J. Soft Comput. Eng.*, vol. 12, no. 2, pp. 16–18, 2022.

[7] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 16, pp. 1–7, 2010.

[8] R. Saxenaa, S. K. Sharmab, and M. Guptac, "Role of K-nearest neighbour in detection of Diabetes Mellitus," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 373–376, 202.

[9] I. H. Sarker, M. F. Faruque, H. Alqahtani, and A. Kalim, "K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 26, pp. 1–9, 2020.

[10] B. Premamayudu, K. Muralikrishna, and K. Pramodh, "Diabetes Prediction Using Machine Learning KNN -Algorithm Technique," *Int. J. Innov. Sci. Res. Technol.*, vol. 7, no. 5, pp. 941–944, 2022.

[11] R. Garcia-Carretero, L. Vigil-Medina, I. Mora-Jimenez, C. Soguero-Ruiz, O. Barquero-Perez, and J. Ramos-Lopez, "Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population," *Med Biol Eng Comput*, vol. 58, no. 5, pp. 991–1002, 2020.

[12] K. VijiyaKumar, B. Lavanya, I. Nirmala, and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," in *IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–5.

[13] T. Ooka, H. Johno, K. Nakamoto, Y. Yoda, H. Yokomichi, and Z. Yamagata, "Random forest

approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan," *BMJ Nutr. Prev. Heal.*, vol. 0, pp. 1–9, 2021.
[14] V. A. Maksimenko *et al.*, "Exploratory study on classifcation of diabetes mellitus through a combined Random Forest Classifer," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 105, pp. 1–14, 2021.
[15] O. Daanouni, B. Cherradi, and A. Tmiri, "Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 2020, pp. 1–5.
[16] L. Breiman, "Random forests," *Mach. Learn.*, pp. 5–32, 2001.

Biography of all authors

Picture	Biography	Authorship contribution
	'Ayuni binti Zamri is a final year student in Bachelor of Computer Sciences at UiTM Perak, Tapah Campus. Her research interest in on Artificial Intelligences and Machine Learning	Drafting article, running experiment, design the research work
	Mohd Faaizie Darmawan received his PhD from Universiti Teknologi Malaysia, Malaysia. He also obtained his Bachelor in Computer Science (Industrial Computing) from the same university. Currently, he is a senior lecturer at Faculty of Computer & Mathematical Sciences at Universiti Teknologi MARA, Perak Branch, Tapah Campus, Malaysia. His research of interest includes Artificial Intelligence, Machine Learning and Image Processing.	Advising the research, editing and revise the article
	Shahirah Mohamed Yatin is a lecturer at Faculty of Computer & Mathematical Sciences at Universiti Teknologi MARA, Perak Branch, Tapah Campus, Malaysia. His research of interest includes Artificial Intelligence, Evolutionary Algorithm and Internet of Things (IoT)	Advising the research and conclusions
	Ahmad Firdaus distinctively received his PhD from University of Malaya, Malaysia. He also obtained his Masters of Computer Science (Networking) from University Teknologi Mara, Malaysia. He is currently a senior lecturer at the Faculty of Computing at Universiti Malaysia Pahang, Malaysia. His area of research includes Mobile Security, Artificial Intelligence, Blockchain and Intrusion Detection System.	Data collection and advising the research

	Mohd Zamri Osman currently works at the Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang as a senior lecturer. His research interest including Artificial Neural Network, Computer Graphics and Artificial Intelligence. His current project for now is 'Hybrid Feature for Automatic Race Identification'.	Data analysis and interpretation