Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Diabetes Prediction using Machine Learning Ensemble Model

Ong Yee Hang[1], Virgiyanti Wiwied[1,*], Rosly Rosaida[1]

[1] Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

ARTICLE INFO

Article history:
Received 9 May 2023
Received in revised form 7 September 2023
Accepted 1 December 2023
Available online 9 January 2024

Keywords:
Diabetes prediction; Machine Learning; AdaBoost; Support Vector Machine (SVM); Ensemble Model

ABSTRACT

Malaysia National Health and Morbidity Survey revealed that one-fifth of Malaysian adults are diagnosed with Diabetes. It exists in different age groups and is hardly discovered especially among youths as the test could only be performed in certain places which require special equipment. It is essential to develop a tool that is capable to generate high accuracy predictions. This research underwent features selection of a secondary dataset which contains seventeen attributes, with no irrelevant data and missing values, and fed it into an AdaBoost with Decision Tree as Base Algorithm Model, Support Vector Machine (SVM), and an ensemble model developed by the machine learning knowledge. The first five most influenced features in the dataset were selected using SelectKBest for each model to conduct training and testing on the dataset and higher accuracy prediction results were achieved. The predictions from the three models were compared and the results from AdaBoost and SVM were combined in the ensemble model. A diabetes prediction prototype was developed to compare the accuracy of the three methods using the observed dataset. This research concludes the ensemble model gives the highest accuracy for Diabetes prediction and might be considered the most suitable method applied in Diabetes prediction tools.

1. Introduction

1.1 Research background

According to Malay Mail [1], the Malaysia Ministry of Health (MOH), National Health and Morbidity Survey (NHMS) 2019 found that there was one (1) out of five (5) Malaysian adults diagnosed with Diabetes, which makes up 3.9 million people aged 18 years and above. This statistic increased significantly from 13.4% in 2015 to 18.3% in 2019. Moreover, roughly about 49% of Diabetics cases in Malaysia had not been diagnosed or discovered by Malaysians themselves. In that survey, the data were collected from more than 32,000 respondents from all the states in Malaysia. It is summarized a few findings where Malaysians need to take serious action where one (1) out of four (4) Malaysian adults was physically inactive; a total of 95% of Malaysian adults failed to consume the correct daily number of vegetables or fruit as recommended by MOH; 50% of Malaysian adults were overweight. Individuals who practice a sedentary lifestyle will lead to a probability of being

* *Corresponding author.*
E-mail address: wiwied.virgiyanti@umt.edu.my

https://doi.org/10.37934/araset.37.1.8298

overweight, or obese, which are the main causes of Diabetes. Around the world, the World Health Organization (WHO) [2], stated that between 2000 and 2019, the mortality rate of diabetes according to age-standardized was 3% increased, and for the countries which having lower to middle income had increased 13& of mortality rate. These findings rang the public health alarm in our society.

Diabetes is a chronic disease, but it is preventable. Early predictions and precautions should be made to prolong human life expectancy by having a healthy body. Diabetes Mellitus as known as Diabetes has always been a concerning issue for our health. Diabetes mellitus is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or even by them both according to American Diabetes Association [3]. The simplest explanation of Diabetes is "sugar in urine". If we leave Diabetes patients with no treatments, eventually it will damage the patients' eyes, kidneys, immune system activation, etc.

This research focused on making early Diabetes prediction models by using the Ensemble Model and comparing the model performance of this model with another two (2) models. The Ensemble Model was implemented by using the knowledge of machine learning (ML). The implementation of ML would assist the model to learn the behavior and study every characteristic of the dataset inserted, and the Ensemble Model would construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions by Dietterich, Thomas. G [4]. Hence, the feeding of datasets with attributes into the models will enable them to generate a better accuracy of prediction results.

Diabetes prediction equipment with high accuracy is usually only available in medical centres. This might cause working adults facing difficulties to pick a suitable time from working days to receive blood glucose tests in medical centres. This may eventually cause individuals to lose their chance on tracking their health conditions and might fail to treat Diabetes at the right moment. Secondly, the Diabetes prediction calculator tools available on the Internet have low accuracy as they have weak counting algorithms which will miscount and display false results on Diabetes predictions. Thus, prediction the accuracy of Diabetes is needed in order for Diabetic patients to obtain improved information about the Diabetes predications and analysis, which will help to reduce the time and cost in monitoring the Diabetes diseases. This paper presents the comparison of different classifiers and an ensemble model on datasets of Diabetes such as AdaBoost with Decision Tree and Support Vector Machine (SVM), in which the complexity and performance aspects are considered in the selection process of the classifiers.

Therefore, in this research the objectives were to examine the current research related to Diabetes prediction using ML models, to evaluate the efficiency of Diabetes prediction using the Ensemble Model, and to design and develop a user-friendly Diabetes prediction prototype by applying the ML Ensemble Model. Lastly, the scopes of this research the dataset of reviews used in this project is the stage Diabetes risk prediction which is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. Secondly, the dataset consists of 17 attributes which are age, gender, polyuria, polydipsia, sudden weight, loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and the target variable "class". Lastly, the machine learning approaches used in this research are AdaBoost with Decision Tree as Base Algorithm and Support Vector Machine (SVM) which will also be implemented into the Ensemble Model.

1.2 Literature review

Alehegn *et al*., [5] analysed by using a dataset which was obtained from UCI and consisted of 768 instances. They employed five (5) algorithm techniques which were Support Vector Machine (SVM), Naïve Net (NN), Decision Stump (DS), AdaBoostM1, and Proposed Method (PM) into this research. In this research, the dataset underwent data pre-processing, split pre-processed into training set and test set, and lastly compared the prediction results made by both sets through the evaluation of percentage on accuracy and error rate. As result, the accuracy of correctly classified on training set prediction of SVM was 88.80%, NN was 88.54%, DS was 83.72%, AdaBoostM1 was 85.68% and PM was 90.36%. On the other hand, the percentage of incorrectly classified for SVM was 11.20%, NN was 11.46%, DS was 16.28%, AdaBoostM1 was 14.32% and PM was 9.64%. Hence, PM owned the highest for percentage of correctly classified and lowest percentage for incorrectly classified. Therefore, we could conclude that PM has the best model performance. In PM, they had composed SVM, NN, DS and AdaBoostM1 as one model which is PM.

On the other hand, Soni, Ankit Narendrakumar [6] examined "Diabetes Dataset for PIMA Indians (PIMA)" which was obtained from UCI and consisted of 768 instances with eight (8) input attributes and an output attribute. Soni, Ankit Narendrakumar [6] employed Naive Bayes (NB), SVM, Multilayer Perceptron (MLP), Decision Tree and Ensemble Method (EM). This research was carried out by using WEKA 3.8 and the "Replace Missing Value" function in WEKA was used to replace incomplete instances in the PIMA dataset. The following were the percentage of accuracy achieved by each algorithm technique where NB was 93.83%, SVM and MLP both were 97.82%, Decision Tree was 96.00%, and EM was 98.55%. Based on this research, we could conclude that EM gained the highest accuracy.

Alternatively, Perveen *et al*., [7] employed the dataset from Canadian Primary Care Sentinel Surveillance Network (CPCSSN) which contained 667907 of records from 2003 until 2013. Perveen *et al*., [7] carried out this research by grouping the records based on age groups which were adolescence (age group of 18-35), middle aged adults (age group of 36-55) and aged more than 55. The algorithm techniques that were employed were AdaBoost, J48 decision tree and bagging. In each technique, the dataset was split into 60% used for allocating the induction of the models and the remaining 40% were used as testing the accuracy of models. As a result, AdaBoost owned the highest percentage of AROC for groups of adolescence and middle age, whereas bagging owned the highest percentage for groups of people aged more than 55. In conclusion, AdaBoost overly performed better.

Hence, from the mentioned related work, for the individual technique, SVM, Decision Tree, and AdaBoost scored a good performance from each research work and lastly, the PM which can also be known as the EM will perform better than the individual technique. Therefore, in this research work, we would include three (3) models which were AdaBoost with Decision Tree as Base Algorithm, SVM and Ensemble Model which was composed of AdaBoost with Decision Tree as Base Algorithm and SVM. The expected outcome of this research should be the same as related work mentioned above, where the Ensemble Model should achieve the highest score.

2. Methodology

2.1 Data sample

In this research, the dataset named "Early-stage Diabetes risk prediction dataset", which was denoted on 7 July 2020 by UC Irvine Machine Learning Respiratory (UCI) was used. This dataset contained 520 instances and 17 attributes which included the target variable named "class". This dataset was collected by distributing survey forms to the patients of the Hospital in Sylhet, Bangladesh, with the approval of a doctor. As presented in Table 1 below, the attribute type for all the attributes were nominal with labelling of "Yes" / "No", except for the attribute named "Age". The attribute type for the "Age" was numeric. Besides, there was no missing value and irrelevant data samples in this dataset too.

Table 1
Attributes and characteristics

Attribute	Attribute Types	Characteristic	Missing Value	Explanation of Attribute
Age	Numeric	Age ranges from 16 - 90	None	The age of patients ranges from 16 years old to 90 years old.
Gender	Nominal (Yes / No)	Male: 63%, Female: 37%	None	The gender of the patients is male or female.
Polyuria	Nominal (Yes / No)	Yes: 50%, No: 50%	None	The experience of abnormal production of excessive urine.
Polydipsia	Nominal (Yes / No)	Yes: 45%, No: 55%	None	The experience of always being thirsty and increasing the amount of water intake.
Sudden weight loss	Nominal (Yes / No)	Yes: 42%, No: 58%	None	The drastically dropping of body weight without any trying.
Weakness	Nominal (Yes / No)	Yes: 59%, No: 41%	None	The feeling of being tired or weak all the time.
Polyphagia	Nominal (Yes / No)	Yes: 46%, No: 5054%	None	The unexplainable hunger even just a short while after eating.
Genital thrush	Nominal (Yes / No)	Yes: 22%, No: 78%	None	The yeast infection that might occur in man patients will cause irritation on the skin, mouth, throat, or genital.
Visual blurring	Nominal (Yes / No)	Yes: 45%, No: 55%	None	The vision of the patient is weakening and causes blurry sight or loss of vision.
Itching	Nominal (Yes / No)	Yes: 49%, No: 51%	None	The itching of the body leads to scratches and irritation.
Irritability	Nominal (Yes / No)	Yes: 49%, No: 51%	None	The uncomfortable feeling and tend to be having mood swings.
Delayed healing	Nominal (Yes / No)	Yes: 46%, No: 54%	None	The speed of healing wounds is much slower than it used to be.
Partial paresis	Nominal (Yes / No)	Yes: 43%, No: 57%	None	The body muscle is experiencing lack of power

				and could not be able to do work.
Muscle stiffness	Nominal (Yes / No)	Yes: 38%, No: 63%	None	The muscle in a certain part of the body is feeling numb and difficult to move.
Alopecia	Nominal (Yes / No)	Yes: 34%, No: 66%	None	The loss of hair from any part of the body, especially the scalp, due to attacks by the immune system on the hair follicles.
Obesity	Nominal (Yes / No)	Yes: 17%, No: 83%	None	The accumulation of excessive fats in the body and who has a body mass index (BMI) in the range of 30 – 34.9 kg/m².
Class	Nominal (Positive / Negative)	Positive: 62%, Negative: 38%	None	The column of recording the patient as "positive" (having Diabetes) and "negative" (do not have Diabetes).

2.2 Method

This experiment consisted of five (5) modules which were data collection, data pre-processing, features selection, classification and predicting method, and lastly evaluation, as shown in Figure 1. The classification and prediction method involved three (3) models. The models were AdaBoost with Decision Tree as Base Algorithm as Model 1, SVM as Model 2, and Ensemble Model as Model 3. The Model 3 was made out of a combination of Model 1 and Model 2 together.

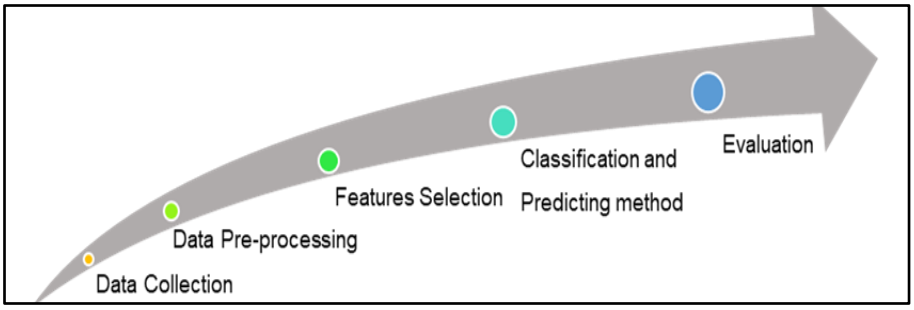


Fig. 1. Proposed model of Diabetes predictions in this research

2.2.1 Data collection

The dataset used in this research study was provided by distributing questionnaires to the patients in Sylhet Diabetes Hospital which were from Bangladesh. This dataset was denoted in UC Irvine Machine Learning Repository (UCI) on 7 July 2020. This dataset consisted of 520 instances with 17 attributes, including the target variable "class". In this dataset, there were no irrelevant attributes and missing values data.

2.2.2 Data pre-processing

This dataset collected a total of 520 instances and 17 attributes. Moreover, this dataset did not contain any missing values or any irrelevant attributes too. Therefore, there were no attributes being removed or imputed in this dataset before inputting the dataset into the feature selection process.

The importance of determining the necessity either to remove or impute any of the attributes in this dataset before moving into feature selection was due to the presence of missing values or irrelevant attributes would contribute a huge effect to the accuracy of results obtained from feature selection.

2.2.3 Feature selection

According to Dutta *et al*. [8], features selection is important as it simplified the dataset and helped to select only the most useful features to decrease the time consumption on training times, improve the accuracy of results and lastly was aiming to prevent the possibility of over-fitting. Hence feature selection was performed before feeding the selected attributes from dataset as input into all the models. The feature selection method applied in this experiment was SelectKBest, which aimed to select only the top five (5) highest k-score ranked features and remove the least relevant attributes based on the sensitivity level of each attribute to the target class in this dataset. The SelectKBest was an example of univariate feature selection, where each feature in the dataset would be tested individually to identify the relationship between the feature and the target variable. The filter-based technique applied by univariate feature selection would be ranked based on the scores. According to Table 2, SelectKBest ranked the features from the highest score to the lowest. The higher the score, the more influenced the features of the target variable. The Table 2 showed the "column" was the position of features in the dataset, whereas the "features", were the attributes in the dataset, and the "score" was the score given by SelectKBest to each attribute. Hence, only data samples of these five (5) features were selected and fed into the models.

Table 2
Results of features selection by SelectKBest

Column	Features	Score
3	Polydipsia	120.7855
2	Polyuria	116.1846
1	Gender	66.1939
4	Sudden Weight Loss	57.7493
12	Partial Paresis	55.3143

After the models in the classification and predicting method received the input of the dataset, the dataset was split into a train set and a test set in each model accordingly. The size of the train set and a test set of each model were adjusted based on the optimal value in each model itself. The train set was used to train the model, whereas the test set was used to identify whether each model was well-trained or left untrained. Lastly, the performance of each model was measured by using a confusion matrix and ROC Curve.

2.2.4 Classification and predicting method

The classification of the dataset was executed by the AdaBoost model with Decision Tree as Base Algorithm, SVM, and Ensemble Model. In the AdaBoost model using Decision Tree as the Base Algorithm model, each attribute had its stump independently and a forest of stumps would be created. For example, according to the attributes in the dataset, all the attributes formed into decision stumps respectively. Figure 2 below showed the stump forming of randomly selected attributes from the dataset, such as the polyuria attribute, itching attribute, and obesity attribute. In AdaBoost, the DS which was made out of the attributes from the dataset were unrelated, they were

independently made. Besides, while processing the classifier, mistakes from the previous stump made would be added to the following stump and the process would be carried on until all the stumps were successfully classified.

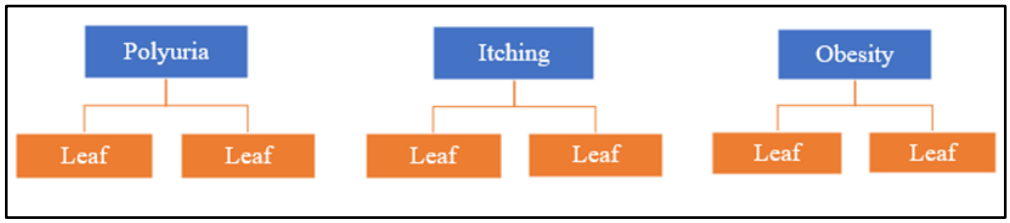


Fig. 2. Example of decision stump

In this research project, the SVM with RBF kernel was applied to predict the selected features from the database of Diabetes. The application of RBF kernel in SVM algorithm would provide the highest 33% accuracy of Diabetes predictions on training data compared to Linear kernel, Polynomial kernel, and Sigmoid kernel, Singh *et al*., [9]. By using the RBF kernel, a hyperplane would be created in a high dimensional space, which do a good separation of classifying data.

2.2.4.1 Model 1: AdaBoost with Decision Tree as Base Algorithm

The AdaBoost with Decision Tree as Base Algorithm created a forest of stumps for predicting Diabetes based on the data of Diabetes patients provided in the dataset. In this ML model, successive models would be added to the failure models. This was to assist the wrongly predicted results obtained from the failure models, it would attempt to make corrections on the errors made by the failure models which was explained by Laila *et al.,* [10]. Firstly, all the attributes were given sample weight (SW). The list of equations applied in this model were listed in Table 3 and the summarized algorithms may refer to Figure 4. For example, in this condition the SW for each attribute was 1/17, where a total of 17 samples had been selected from the features selection process, Alehegn *et al*., [5]. By assigning SW to each attribute, all the attributes were equally important in the beginning of the process, and we would be able to randomly pick any of the attributes to begin the iteration. The Gini Index of each stump would be calculated by first cacalculating the Gini Index for each leaf of the stump, as shown in Equation (1).

$$Gini\ Index\ of\ Each\ Leaf = 1 - (the\ probability\ of\ "Yes")^2 - (the\ probability\ of\ "No")^2 \quad (1)$$

and then obtained the final value of the Gini Index by getting the average value of combining both leaves, as shown as Equation (2).

$$Gini\ Index\ of\ Stump = \left[\frac{Total\ Sample\ of\ Left\ Leaf}{Total\ Sample\ of\ Left\ and\ Right\ Leaf} \times Gini\ Index\ of\ Left\ Leaf\right] + \left[\frac{Total\ Sample\ of\ Right\ Leaf}{Total\ Sample\ of\ Left\ and\ Right\ Leaf} \times Gini\ Index\ of Right\ Leaf\right] \quad (2)$$

This was because the total sample on the left leaf of the stump and right left of the stump were different. Therefore, the Equation (2) would help to get the average Gini Index of the stump. By

comparing all the Gini Index values for all attributes, the stump with lowest value of Gini Index will be assigned as first stump as it has the lowest impurity. Let's take the first stump as Stump A

Next, the Amount of Say, as shown as Equation (3), Stump A will be calculated.

$$Amount\ of\ Say = \frac{1}{2} \times log^{\frac{1-TE}{TE}} \tag{3}$$

The value of Amount of Say obtained will determine the Total Error (TE) of Stump A, Alehegn *et al*., [5]. The graph of TE as shown in Figure 3. Based on Figure 3, when the Amount of Say is large, the TE was small, while when the Amount of Say was small, the Total Error would be nearer to value one (1). The value of TE was from the range of zero (0) to one (1). The smaller the value of TE the better the stump, Alehegn *et al*., [5]. For example, based on the SW, the TE of Stump A was 1/17, Amount of

$$Stump\ A = \frac{1}{2} log^{\frac{1-\frac{1}{17}}{\frac{1}{17}}} = \frac{1}{2} log^{16} = 0.602$$

∴ Based on Figure 3, the value of TF for Stump A was 0.2.
Therefore, by comparing TE and SW of Stump A, we know that Stump A was incorrectly classified. The SW of the Stump A will be increased by calculating its New Sample Weight, as shown as Equation (4).

$$New\ Sample\ Weight = old\ SW\ \times e^{amount\ of\ say} \tag{4}$$

and the SW of the other attribute would be lessened than its old SW, as shown as Equation (5).

$$New\ Sample\ Weight = old\ SW\ \times e^{-amount\ of\ say} \tag{5}$$

The New Sample Weight of all the attributes would be normalized, due to the total amount of all New Sample Weight should be one (1) or rounded off to one (1), as shown as Equation (6).

$$Normalization = \frac{New\ Sample\ Weight\ of\ Particular\ Attribute}{Total\ of\ All\ New\ Sample\ Weight\ from\ All\ Attributes} \tag{6}$$

Table 3
List of AdaBoost with Decision Tree as Base Algorithm equations

Equation	Formula
1	$Gini\ Index\ of\ Each\ Leaf\ = 1 - (the\ probability\ of\ "Yes")^2 - (the\ probability\ of\ "No")^2$
2	$Gini\ Index\ of\ Stump = [\frac{Total\ Sample\ of\ Left\ Leaf}{Total\ Sample\ of\ Left\ and\ Right\ Leaf} \times Gini\ Index\ of\ Left\ Leaf + [\frac{Total\ Sample\ of\ Right\ Leaf}{Total\ Sample\ of\ Left\ and\ Right\ Leaf} \times Gini\ Index\ of Right\ Leaf]$
3	$Amount\ of\ Say = \frac{1}{2} \times log^{\frac{1-TE}{TE}}$
4	$New\ Sample\ Weight = old\ SW\ \times e^{amount\ of\ say}$
5	$New\ Sample\ Weight = old\ SW\ \times e^{-amount\ of\ say}$
6	$Normalization = \frac{New\ Sample\ Weight\ of\ Particular\ Attribute}{Total\ of\ All\ New\ Sample\ Weight\ from\ All\ Attributes}$

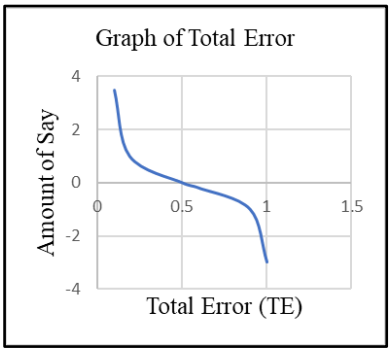


Fig. 3. Graph of TE

Summary algorithm of the model:

① Assigned sample weight (SW) for all the attributes. $SW = \frac{1}{N}$, where N = total number of data on the sample.
② Calculate Gini Index for each of the stump. The stump with lowest Gini Index value will be the first stump in the forest stump.
③ Calculate the Amount of Say of the first stump to determine the Total Error (TE) on the final classification.
④ Calculate New Sample Weight of Stump A.
⑤ Calculate New Sample Weight of the other attributes.
⑥ Normalize the New Sample Weight of all attributes.
⑦ Repeat Step ② to Step ⑥ for the next 17 stumps.

Fig. 4. Summary algorithm of AdaBoost with Decision Tree as Base Algorithm

Lastly, in this research, the decision tree classifier of Model 1 would only use one (1) level as its base algorithm of AdaBoost. The one (1) level decision tree was known as decision stump (DS). Besides, the sequence of DS to be inserted into the calculation was sorted by using Gini Index. Hence, by using the decision tree classifier, I had defined the maximum depth of the decision tree as one (1), and the criterion of the decision tree as "Gini". After defining all these parameters in the decision tree classifier, it is inserted into AdaBoost as the base estimator. In this model, 70% of the inserted dataset was used as a train set, and the remaining 30% was used as a test set.

2.2.4.2 Model 2: Support Vector Machine (SVM)

The SVM was one of the approaches to supervise learning algorithms. According to Kaur, Harleen *et. al.,* [11], where hyperplane of the SVM would help to categorize the data by setting boundaries

for the process of classification and regression problem. In this research, there were more than two features employed in this model, hence these selected features created an infinite dimension. The kernel used in SVM was the Radial Basis Function (RBF) kernel, it helped to identify the position of the hyperplane. The hyperplane would separate the features into different classes. As for non-linear data, the RBF kernel transformed the data into suitable form to perform classification. The kernel function of RBF kernel was defined as in Equation (7).

$$K(x_i, x_u) = e^{-\gamma(x_i - x_u)^2} \tag{7}$$

where gamma (γ) indicated the learnable parameter, and (χ_i, χ_u) indicated the distance point that were nearest to the observation data. In Model 2, the hyper-parameter optimization method had helped to determine the selection of RBF kernels among other kernels. This method used to view the accuracy of predictions on our training set. It helped to sort which kernel of SVM performed the best with the dataset, by giving each kernel a score based on the performance in generating prediction results of each kernel. As shown in Figure 5, the polynomial kernel and the RBF kernel achieved the same score of accuracy. However, when using the test set to evaluate the accuracy of the RBF kernel and the polynomial kernel as shown in Figure 6, the RBF kernel gained a higher accuracy than the polynomial kernel. Hence, it was concluded that RBF kernels would be used, the size of train set, and test set were 25% and 75% respectively.

```
for k in ('linear', 'poly', 'rbf', 'sigmoid'):
    model = SVC(kernel=k)
    model.fit(X_train3, y_train3)
    y_test3 = model.predict(X_train3)
    print(k)
    print(accuracy_score(y_train3, y_test3))
```

```
linear
0.8846153846153846
poly
0.9076923076923077
rbf
0.9076923076923077
sigmoid
0.6743589743589744
```

Fig. 5. Hyper-parameter optimization method

```
#SVM-RBF
X_train3, X_test3, y_train3, y_test3 = train_test_split(X_df, y_df, test_size=0.25,
                                                        random_state=8)  # test size is 0.25, train size is 0.75

SVM = SVC(kernel="rbf", random_state=0)
SVM.fit(X_train3, y_train3)
# accuracy of SVM
target_pred2 = SVM.predict(X_test3)
SVM_train_sc = accuracy_score(y_train3, SVM.predict(X_train3))
SVM_test_sc = accuracy_score(y_test3, target_pred2)

#print(" **A. The Accuracy Result of Train Set** ", ((SVM_train_sc)*100), "%")
print("The Accuracy Result of Test Set (RBF Kernel)", ((SVM_test_sc)*100), "%")
```

```
The Accuracy Result of Test Set (RBF Kernel) 90.76923076923077 %
```

```
#SVM-POLY

X_train3, X_test3, y_train3, y_test3 = train_test_split(X_df, y_df, test_size=0.25,
                                                        random_state=8)  # test size is 0.25, train size is 0.75

SVM = SVC(kernel="poly", random_state=0)
SVM.fit(X_train3, y_train3)
# accuracy of SVM
target_pred2 = SVM.predict(X_test3)
SVM_train_sc = accuracy_score(y_train3, SVM.predict(X_train3))
SVM_test_sc = accuracy_score(y_test3, target_pred2)

#print(" **A. The Accuracy Result of Train Set** ", ((SVM_train_sc)*100), "%")
print("The Accuracy Result of Test Set (Polynomial Kernel)", ((SVM_test_sc)*100), "%")
```

```
The Accuracy Result of Test Set (Polynomial Kernel) 87.6923076923077 %
```

Fig. 6. Comparison of accuracy result on RBF Kernel and Polynomial Kernel

2.2.4.3 Model 3: Ensemble Model

The Ensemble Model was the proposed method in this research project. As shown in Figure 7 below, it consisted of the predicted results generated from AdaBoost Model with Decision Tree as Base Algorithm model (Model 1) and the SVM model (Model 2). In this Ensemble Model, the inputs from Model 1 and Model 2 were labeled as probability p. This was to calculate the unweighted average probability ($\bar{p}$), Semerdjian *et al*., [12]. The $\bar{p}$, would help to determine the decision boundary T where the T firstly would be set to 0.5, and the T then adjusted to obtain a better recall rate, Semerdjian *et al*., [12]. Hence, when the value of $\bar{p}$ is greater than T, the patient would be allocated into the group of diabetic patients and else when $\bar{p}$ was less than T, he or she would be known as a non-diabetic patient.

According to Figure 8, it demonstrated the workflow of the Ensemble Model. The dataset which had gone through the data pre-processing phase and followed by features selection would be then split into two sets of data which were the training set and testing set. In this Ensemble Model, 60% from the original dataset was used for the training set, whereas the remaining 40% was grouped as testing set. The purpose of setting training set was to train the model for capturing the behavior and enhance the performance of the model and testing set was to evaluate the performance of model. After obtaining the result of testing set, it would be evaluated by using confusion matrix and ROC Curve.

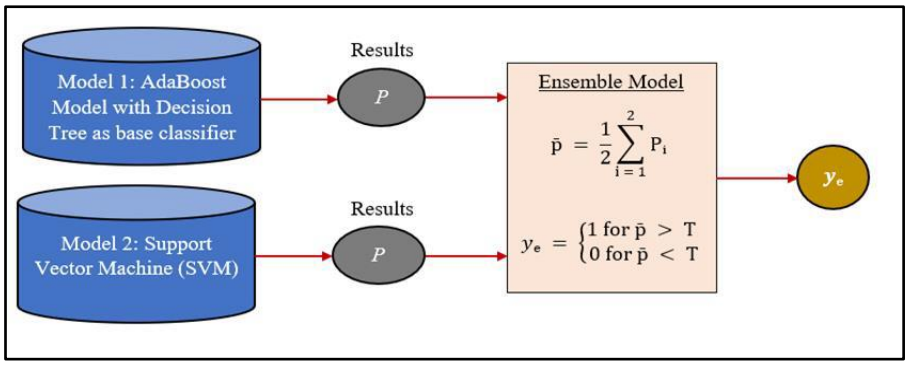


Fig 7. Ensemble Model which consisted of Model 1 and Model 2

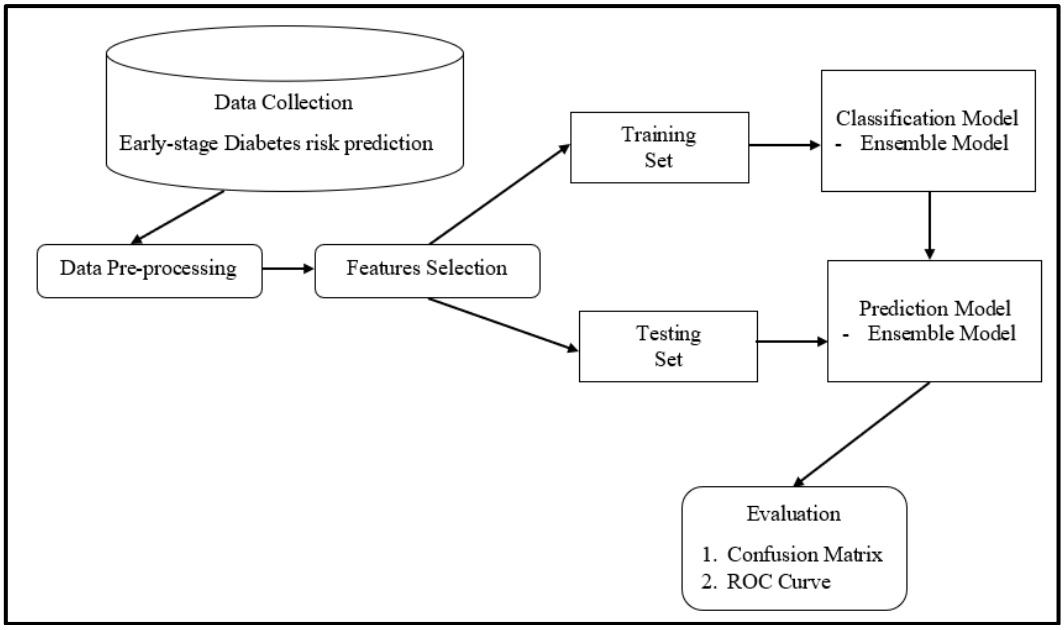


Fig. 8. Workflow of the Ensemble Model

2.2.5 Evaluation

In the evaluation phase, we obtained the results of Diabetes prediction made by applying the AdaBoost Model with Decision Tree as Base Algorithm model, the SVM model and Ensemble Model to undergo evaluation on accuracy. The confusion matrix used to show the prediction of results in the form of a matrix as shown in Figure 9. The values listed in the confusion matrix were used to determine the precision rate and recall rate of all the models.

		Actual Value	
		Positive	**Negative**
Predicted Value	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Fig. 9. Confusion matrix

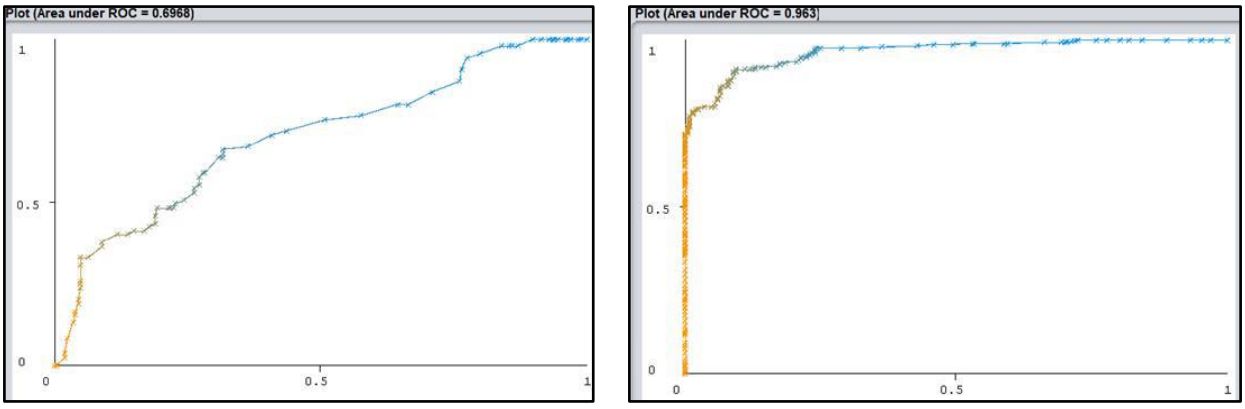
Besides, the Receiver Operating Characteristic Curve (ROC) used to show the relationship between Eq. (8),

$$\frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Negative\ (FN)} \tag{8}$$

and Equation (9),

$$\frac{False\ \ Positive\ (FP)}{False\ Positive\ (FP)\ +\ True\ Negative\ (TN)} \tag{9}$$

by plotting the graph of the ROC curve. The Area Under ROC Curve (AUC) is made to study the performance of Diabetes prediction which had produced by algorithm. In the AUC, the larger the value, which was closer to 1.0, the higher the accuracy of the result. The ideal classification of AUC was between 0.5 to 1.0. For example, by comparing the classification of AUC of the ROC Curve in Figure 10(a) and Figure 10(b), we could see that the prediction result in Figure 10(b) was more accurate than the prediction accuracy in Figure 10(a).



(a) (b)

Fig. 10. The pointer of ROC Curve (a) 0.697 and (b) 0.963

3. Results

3.1 Statement of results

The accuracy of the train set, and test set made to predict the results based on the data samples were measured and recorded in Table 4. The value of the train set for all the models was lower than the test set. Besides, the methods used to evaluate the performance of each model were the confusion matrix and ROC curve. The figures of the confusion matrix and ROC curve of each model are shown in Table 5. The AdaBoost with Decision Tree as Base Algorithm is labelled as Model 1, the SVM is labelled as Model 2, and the Ensemble Model is labelled as Model 3. Besides, the evaluation of the model performance of each model based on the confusion matrix and ROC curve are all listed in Table 6.

By referring to the evaluation of each model in Table 6, Model 3 has the highest accuracy on the test set. Besides, according to Table 6, Model 3 gained the highest accuracy, Model 2 achieved the highest precision rate, and Model 3 obtained the highest recall rate. The accuracy of each model is measured according to the result of the test set. Next, the precision rate is also known as the positive predictive value, which shows the amount of actual positive results that are correctly predicted as positive. On the other hand, the ideal AUC value is between 0.5 and 1.0. By referring to Table 6,

Model 1 and Model 2 have the same AUC, which is 0.95, while Model 3 has the highest value of AUC, which is 0.96.

Table 4
Accuracy of train set and test set of each model

Model	Train Set (%)	Test Set (%)
1	89.29	89.74
2	90.77	90.77
3	90.71	90.87

Table 5
Confusion matrix and ROC Curve of all models

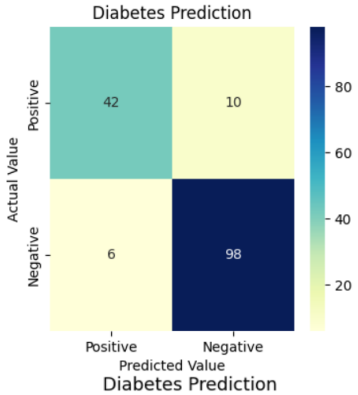
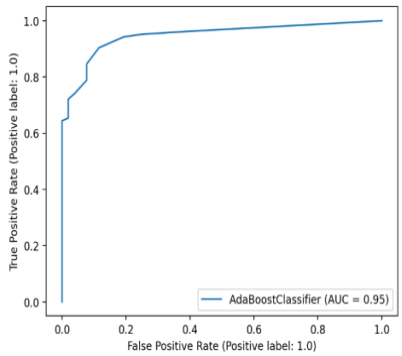
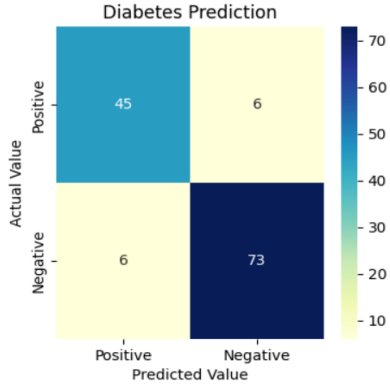
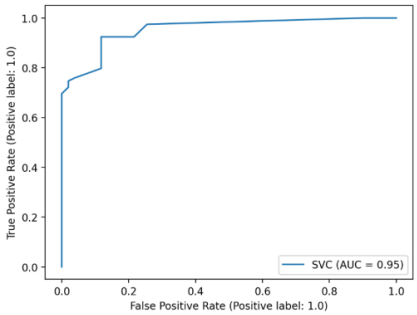
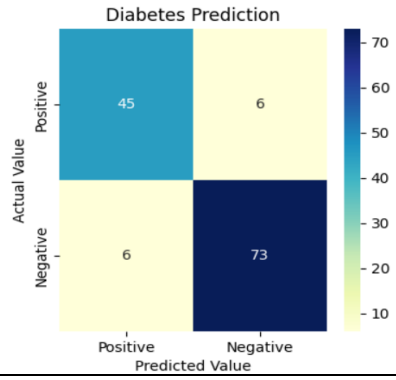
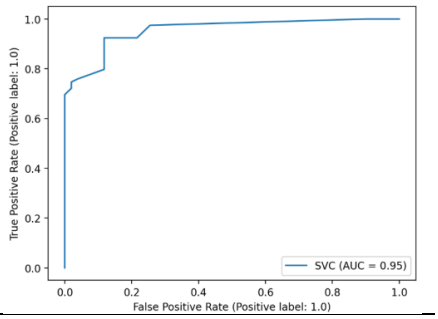
Model	Confusion Matrix	ROC Curve
1		
2		
3		

Table 6
Evaluation of Model 1, Model 2 and Model 3

Model	Confusion Matrix Accuracy (%)	Precision (%)	Recall (%)	ROC Curve Area Under the Curve (AUC)
1	89.74	87.50	80.77	0.95
2	90.77	92.31	76.19	0.95
3	90.87	88.10	89.16	0.96

3.2 Discussions

We could conclude that the results we gained in this experiment where the Ensemble Model (Model 3) is the best model, and this result matched our theoretical concept of the ensemble model. By referring to Table 6, Model 3 obtained the highest rate of accuracy which is 90.87%. I believe that in the future we could gather more high-performance single ML models into this model to help generate a higher accuracy result compared to the results we obtained in this experiment. This is because the main idea of this model is where we are composing few individual algorithm models wishing to produce an optimal model. The combination of these few algorithm models as one will boost the performance level of Ensemble Model compared to those single algorithm models.

Besides, as referring to Alehegn *et al*., [5] the accuracy of correctly classified of PM was 90.36%. The PM introduced by Alehegn *et al*., [5] consisted of SVM, Naïve Net, and DS. Whereas as specified by Soni, Ankit Narendrakumar [6] the accuracy rate of EM is 98.55%. This EM in the research project completed by Soni [6] was a combination of NB, SVM and MLP and DS. From above two results gained from existing research made, Model 3 achieved a better performance than the PM introduced by Alehegn *et al*., [5], and the EM proposed by by Soni [6] had the higher accuracy rate than Model 3.

4. Conclusions

The experiments in this research project proved that Model 3 gained the highest accuracy of prediction results, highest recall rate, the highest amount of Area Under the Curve (AUC) and second highest on the precision rate. As compared to Model 1 and Model 2, Model 3 owned the most highly performed achievements. As a result, it is recommended to implement the concept of Model 3 as a ML backend of the future Diabetes prediction prototype for the use of society. According to Yahyaoui *et al.,* [13], the Ensemble model may achieve higher performance as improvements were made in this model. The prototype of this research project was as displayed in the form of Quick Response (QR) code in Figure 11 below. By scanning the QR code would show the prototype.

However, there are a few improvements that could be made for future work. Firstly, the present Diabetes prediction prototype should provide more choices of models for the users to select, and they could freely choose at least any two (2) of the desired models available in the prototype to be inserted into the Ensemble Model. Secondly, the input of data samples in this prototype should be more flexible. For example, the format of the file could be dataset in the form of .csv, dataset in the form of an image, and enable the users to answer questions embedded in this prototype for generating a prediction result instead of only using the dataset. Thirdly, the prototype should provide another functionality in which the users could view the history of the predictions they made and generate a copy of their results by downloading it in a PDF format. Lastly, the prototype in the future should produce a result that displays the results as “You may have Diabetes, please consult Doctor immediately” or “You are a healthy person” based on the inputs received from users. This will give the users a clear idea of their current health conditions.

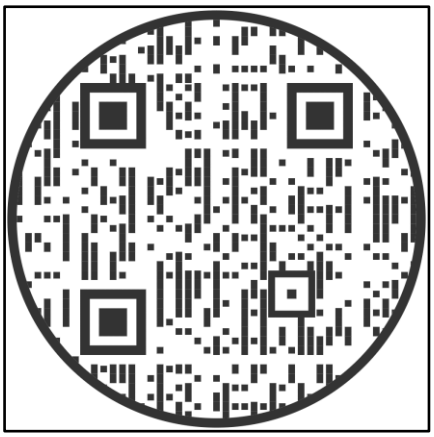


Fig. 11. QR code of the model

Acknowledgement
This research was not funded by any grant.

References
[1] Malay Mail. 2020. "One in five adult Malaysians has diabetes, 2019 National Health and Morbidity Survey shows.", Accessed December, 12, 2022, https://www.malaymail.com/news/malaysia/2020/05/29/one-in-five-adult-malaysians-has-diabetes-2019-national-health-and-morbidit/1870631
[2] 2023. "Diabetes." Who. April 5, 2023. https://www.who.int/news-room/fact-sheets/detail/diabetes.
[3] AD Association. "Introduction: standards of medical care in diabetes-2022." *Diabetes Care* 45, no. 1 (2022): S1-s2. https://doi.org/10.2337/dc22-Sint
[4] Dietterich, Thomas G. "Ensemble methods in machine learning." In *International workshop on multiple classifier systems*, pp. 1-15. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. https://doi.org/10.1007/3-540-45014-9_1
[5] Alehegn, Minyechil, Rahul Joshi, and Preeti Mulay. "Analysis and prediction of diabetes mellitus using machine learning algorithm." *International Journal of Pure and Applied Mathematics* 118, no. 9 (2018): 871-878. https://www.researchgate.net/publication/323278139
[6] Soni, Ankit Narendrakumar. "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques." *Available at SSRN 3642877* (2020). https://doi.org/10.2139/ssrn.3642877
[7] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." *Procedia Computer Science* 82 (2016): 115-121. https://doi.org/10.1016/j.procs.2016.04.016
[8] Dutta, Aishwariya, Md Kamrul Hasan, Mohiuddin Ahmad, Md Abdul Awal, Md Akhtarul Islam, Mehedi Masud, and Hossam Meshref. "Early prediction of diabetes using an ensemble of machine learning models." *International Journal of Environmental Research and Public Health* 19, no. 19 (2022): 12378. https://doi.org/10.3390/ijerph191912378
[9] Singh, Harasis., and Sharma, Richa. "Diabetes Prediction Using Different Kernel SVM Classification Algorithm." *Journal of Emerging Technologies and Innovative Research (JETIR)* 8, no. 3 (2021): 290-295.
[10] Laila, Umm E., Khalid Mahboob, Abdul Wahid Khan, Faheem Khan, and Whangbo Taekeun. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." *Sensors* 22, no. 14 (2022): 5247. https://doi.org/10.3390/s22145247
[11] Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." *Applied computing and informatics* 18, no. 1/2 (2022): 90-100. https://doi.org/10.1016/j.aci.2018.12.004
[12] Semerdjian, John, and Spencer Frank. "An ensemble classifier for predicting the onset of type II diabetes." *arXiv preprint arXiv:1708.07480* (2017). https://arxiv.org/abs/1708.07480
[13] Yahyaoui, Amani, Akhtar Jamil, Jawad Rasheed, and Mirsat Yesiltepe. "A decision support system for diabetes prediction using machine learning and deep learning techniques." In *2019 1st International informatics and software engineering conference (UBMYK)*, pp. 1-4. IEEE, 2019. https://doi.org/10.1109/UBMYK48245.2019.8965556
[14] UCI Machine Learning Repository. n.d. Accessed December 12, 2022, https://archive.ics.uci.edu/ml/
[15] Joshi, Tejas N., Chawan, Pramila M. "Diabetes Prediction Using Machine Learning Techniques." *Int. Journal of Engeneering Research and Application* 8, no. 1 (2018): 09-13. 10.9790/9622-0801020913

[16] Mohammed Alhamid. 2021. "Ensemble Models." Accessed December 12, 2022. https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c
[17] Namugenyi, Christine, Shastri L. Nimmagadda, and Torsten Reiners. "Design of a SWOT analysis model and its evaluation in diverse digital business ecosystem contexts." *Procedia Computer Science* 159 (2019): 1145-1154. https://doi.org/10.1016/j.procs.2019.09.283
[18] Rosly, Rosaida, Mokhairi Makhtar, Mohd Khalid Awang, Mohd Isa Awang, and Mohd Nordin Abdul Rahman. "Analyzing performance of classifiers for medical datasets." *International Journal of Engineering & Technology* 7, no. 2.15 (2018): 136-138. https://doi.org/10.14419/ijet.v7i2.15.11370
[19] Sneha, N., and Tarun Gangil. "Analysis of diabetes mellitus for early prediction using optimal features selection." *Journal of Big data* 6, no. 1 (2019): 1-19. https://doi.org/10.1186/s40537-019-0175-6
[20] Tom Michael Mitchell. Machine Learning, 1st ed. (New York: McGraw-Hill, 1997), 23