International Journal of Global Optimization and Its Application

Vol. 1, No. 2, June 2022, pp.126-134 e-ISSN: 2948-4030

https://doi.org/10.56225/ijgoia.v1i2.22

Article

Predictive Data Mining Approaches for Diabetes Mellitus Type II Disease

Shahira Ibrahim [1] and Siti Shaliza Mohd Khairi [1,*]

[1] Department of Statistics and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia; shahira.ibrahim@tmsk.uitm.edu.my

***** Correspondence: shalizakhairi@tmsk.uitm.edu.my

Citations: Ibrahim, S. & Khairi, S.S.M. (2022). Predictive Data Mining Approaches for Diabetes Mellitus Type II Disease. *International Journal of Global Optimization and Its Application, 1*(2), 126-134. https://doi.org/10.56225/ijgoia.v1i2.22

Academic Editor: Liew Pay Jun.

Received: 23 March 2022 Accepted: 8 June 2022 Published: 30 June 2022

Abstract: Diabetes is among the major public health problem especially in developing countries which cause by abnormal insulin secretion in human body. It is a common disease that can led to several health complications and mortality. In Malaysia, most of the cases are categorized as Diabetes Mellitus (DM) Type II. Patients with diabetes increases from year to year due to unhealthy lifestyles e.g. smoking, overweight and hypertension. Therefore, this study meant to identify the influential factors that may contribute to DM Type II by comparing the performance of different data mining approaches. Between April 2017 and November 2018, 684 patients from a public clinic participated in this retrospective cross-sectional study. Four predictive models involved in the study are Logistic Regression, Decision Tree, Naïve Bayes, and Artificial Neural Network (ANN). The error measures (Average Squared Error and Misclassification Rate) with ROC Index are used to evaluate the performance of the models. Results show that the performance of Logistic Regression-Stepwise outperformed to other predictive models with classification accurateness of 73% and able to predict positive outcome (Y=1) correctly by 90%. The significant inputs that affect DM Type II prediction (Y=1) are Hypertension and Glycated Hemoglobin (HbA1c) given the Root Mean Squared Error (RMSE) of model is 0.424. The importance of study may be able to contribute in improving the strategies and planning on diabetes diseases in Malaysia.

Keywords: data mining; diabetes mellitus; Naïve Bayes; artificial neural network; logistic regression; decision tree

1. Introduction

Diabetes Mellitus (DM) is one of the growing chronic diseases and had become a global health problem. In 2014, the number of people with diabetes has risen to 422 million worldwide (World Health Organization, 2018). Furthermore, diabetes spreading rapidly especially in developing countries (Hussein et al., 2015). Malaysia is one of the emerging and developing countries that grow in both social and economic terms. With this sturdy progress, the lifestyle and dietary patterns among citizens have changed to commit with

current needs. Still, everyone craves a healthy and happy life because health is the greatest wealth. In 2018, according to Executive Chairman of National Diabetes Institute (NADI), about 2.5 million adults in Malaysia aged 18 years old and over were diagnosed with diabetes and based on the histories show that most of the patients are not aware they have diabetes. Besides, diabetes is also a "silent killer" among patients. Thus, this situation is alarming the government especially the Health of Ministry Malaysia.

Diabetes Mellitus is divided into two types; Type I and Type II. DM Type I is caused by genetic factors and some sort of environment factors which more likely to attack children and teenagers. DM Type I is also known as adolescent diabetes. When the pancreas does not produce insulin then DM Type I will occur. Meanwhile, DM Type II usually suffered by adults and senior citizens because it is closely related to genetic predisposition. Instead of an unknown trigger, DM Type II also involved with the genetic tendency factors such as high blood pressure and obesity (Okwechime et al., 2015) 10.1371/journal.pone.0145781.

This study focuses on patients with DM Type II because most of the cases happen in Malaysia involved DM Type II. In addition, there are numerous studies in Malaysia focuses on knowledge, practice and attitude towards diabetes. However, there are few studies regarding significant inputs that contribute to DM Type II using predictive models. Efficient predictive model is important in predicting diabetes, to prevent and control the occurrence of diabetes. Therefore, this study focuses to predict DM Type II category among adults in the urban area by using data mining approaches with several influential factors to be considered.

2. Materials and Methods

Data mining approaches is popular in big dataset. The characteristics of these approaches help in building an efficient model for prediction. This study will implement Logistic Regression, Pruning – Decision Tree, Artificial Neural Network and Naïve Bayes approaches. Logistic regression is a regression model that suitable for modelling binary target (1=diabetes, 0=pre-diabetes). It allows the estimation of probability of an event occurs where the target variable containing Y=1 for probability of success (p) and Y=0 for probability of failure ($1-p$). For logistic regression, the curve is built using the natural logarithm of the "odds" of the target variable (diabetes or pre- diabetes). The best-fit line for logistic regression, is obtained from maximum likelihood estimation. The log-odds (logit) of the binary logistic regression model can be written as:

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \mathrm{B}_0 + \mathrm{B}_1\mathrm{X}_1 + \cdots + \mathrm{B}_\mathrm{k}\mathrm{X}_\mathrm{k} \tag{1}$$

In a mathematical expression, logistic function and can be expressed as:

$$f(z) = \frac{\mathrm{p}}{1-\mathrm{e}^{\mathrm{z}}} \tag{2}$$

Where $z = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$. As for this study, coefficient in logistic model referring to coefficient of the risk factors correlated with DM Type II disease (Wah, 2006). Logistic Regression does not require a linear relationship between the target and inputs, normally distributed error terms (residuals) and homoscedasticity is also not required. However, other assumptions still apply which is no multicollinearity. According to Alin (2010), multicollinearity refers to the relationship among two or more inputs. This assumption can be tested by checking the Variance Inflation Factor (VIF) statistic and tolerance statistics

A decision tree is a top-down approach which involves partitioning the data into subgroups that contain cases with similar classes (homogeneous). A decision tree is started with the root node, containing all training data used to grow the tree. The root node has *n* children and rules that specifies which records go to which child. The rule is based on the most important input variables selected by the splitting algorithm. The nodes that ultimately get used are at the ends of their branches, with no children also known as leaves nodes (Esmaily et al., 2018). The most common tree algorithm that had been used widely is Chi-Square Automatic Interaction Detector (CHAID), Classification and Regression Tree (CART), and C5.0. However, in this study, one splitting algorithm (Gini) and two pruning algorithms (CHAID, CART) were used. When splitting, we want diversity in the parent node to be greater than summed diversities in child nodes. Many different criteria may be used to evaluate potential splits.

However, in this study, chi-square test using log-worth value and Gini are used for splitting criterion since the target variables are categorical (Y=1 or Y=0). The purity of the target variable in the children is used to calculate a potential split. A good split is split with high number of purities in the children and produces nodes with same size (Esmaily et al., 2018). Gini is calculated by adding all the squares of the

proportions of the classes and it is called perfectly pure if Gini has a score of 1. The Gini index at a node D is given by:

$$GINI(D) = 1 - \sum_{i=1}^{m} P_i \quad (3)$$

The value measures how likely or unlikely a split is for chi-square test. Then, the best split of chi-square using log- worth value is determined by calculating the chi-square statistic of association between the binary target variables and all potential splits of each competing input variables. The split with the highest log-worth for each input is determined by using equation (4), and the highest log-worth is chosen as the best split.

$$\text{logworth} = -\log 10(\text{p} - \text{value}) \quad (4)$$

On the other hands, human brain consists of many neural cells that process information. Similar as human brain, an Artificial Neural Network (ANN) comprises artificial neurons and relations between them (Kaur & Wasan, 2006; Wah et al., 2011). An artificial neuron takes its inputs and produces an output. The overall behaviour is called the node's activation function. The combination function and the transfer function are the two different parts in the activation function. A single value that gets from the combination of the inputs using the combination function is passed to the transfer function to produce an output (Dreiseitl & Ohno-Machado, 2002). Each input has its own weight. The strength of the inputs depends on the value of weight of an artificial neuron. The computation of the neuron will be different depends on the weights. However, to obtain the output for specific inputs can be achieved by the process of adjusting the weight known as training or learning. The combination function typically uses a set of weights assigned to each of the inputs. The products of each input with its weight are added together also called the weighted sum.

In most data mining tools, the weighted sum is a default. Meanwhile, a mathematical illustration of the association between the inputs and the outputs is shown by the transfer function. There are some transfer functions that have been commonly used to produce outputs which are the threshold, linear, logistic, and hyperbolic tangent functions. Moreover, the functions used are depends on the target variables (diabetes or pre-diabetes). Thus, this study will use the sigmoid function. Nowadays, multilayer perceptron is one of the types of neural network that are broadly used in medicine. The structure of a typical neural network (Figure 1) consists of the first layer, the input layer where the data enters the network. The second layer known as the hidden layer, comprised of artificial neurons and, an output layer, a layer that combines results summarized by the artificial neurons. The input layer will standardize all inputs to have similar input ranges. The hidden layer contains the non-linear activation functions, and all units in the input layer are completely connected to every unit in the second layer. The transfer function is applied after computing the weighted sum for the items in the second layer. The number of the hidden layer for a neural network can be more than one but, usually, one hidden layer is enough.

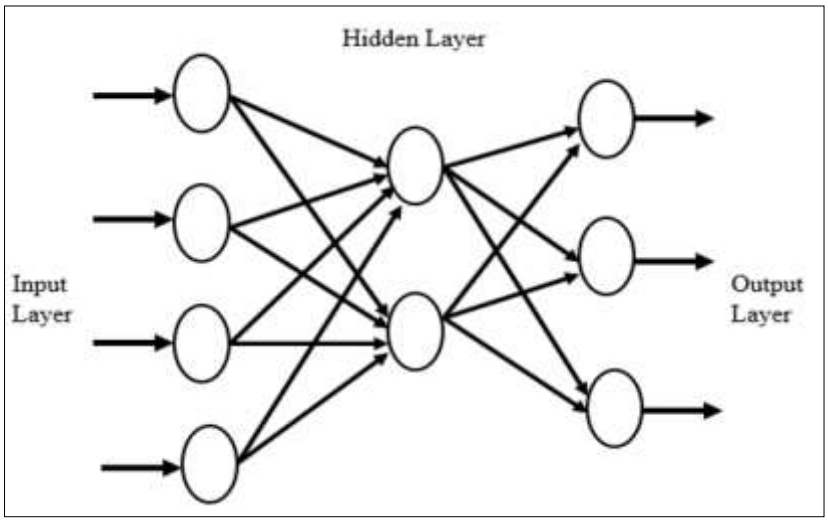

Figure 1. A typical neural network diagram

Sources: Kazemnejad et al. (2010)

Lastly, Naïve Bayes also named as Bayesian Network because Bayes Classifier uses a probabilistic framework for solving classification problems. Most people acknowledge it as Bayes theorem which provides a way of computing a posterior probability $P(Y|X)$ from $P(Y)$, $P(X)$ and $P(X|Y)$. The equation for Bayes theorem (Anitha & Sridevi, 2019), as stated in equation (5).

$$P(Y|X) = \frac{\mathrm{P(X|Y)P(Y)}}{P(X)} \tag{5}$$

where $P(Y|X)$ = The posterior probability of target (Y) given predictor for diabetes (X), $P(Y)$ = The prior probability of target variables, $P(X|Y)$ = The likelihood which is the probability of predictor diabetes given target and $P(X)$ = The prior probability of predictor diabetes. Data mining process involving 6 major steps from data collection until model deployment.

Step 1: Data Acquisition - This study used secondary data collected from a public clinic in Malaysia. The calculated sample size for this study is 384 patients. However, a total number of 684 patients are obtained from the clinic. The population of this study comprised of the patients who have undergone treatment at the clinic. The data are collected from April 2017 to December 2018 which comprised of 7 inputs (Age, Gender, Smoking Status, Cardiovascular Risk, BMI, HbA1c and Family History) with Diabetic Category as the target of the study (refer to Table 1).

Step 2: Data Understanding - This step begins with identifying the type of data used either quantitative or qualitative and determine the data measurement level used (nominal, ordinal, interval and ratio). In this step, we also explore and identify data quality problems such as missing values, outliers and recoding errors.

Table 1. Description of target and inputs.

Variables	Role	Measurement Level	Description
Diabetic Category	Target	Binary	It is divided into 2 groups: Blood glucose level ≥ 7.0 mmol/litre considered as diabetes (Y=1). Blood glucose level between 6.1 mmol/litre and < 7.0 mmol/litre considered as pre-diabetes (Y=0) (Meng et al., 2013).
Age	Input	Interval	Age in years
Gender	Input	Binary	Coded as: 1=Male; 0=Female
Smoking	Input	Binary	Patient with smoking history (1=Yes,0=No)
Cardiovascular Risk	Input	Binary	Patient with heart disease (1=Yes; 0=No)
BMI	Input	Interval	Measured in kg/m2
HbA1c	Input	Interval	Measured in percentage
Hypertension	Input	Binary	Patient with high blood pressure (1=Yes; 0=No)
Family History	Input	Binary	Patient with diabetes family history (1=Yes; 0=No)

Step 3: Data Preparation - The analysis will be performed using SAS Enterprise Miner Workstation 14. Data preparation step covers all activities in preparing the data such as cleaning, transformation and modifying before modelling. Some of the activities at these steps are data audit, identifying missing, incorrect, and extreme values, data selection, and restructuring of data in the form required for analysis. As for the analysis purpose, data are divided into 70% training data and 30% validation data. Table 2 shows the existence of the missing values and outliers among interval input. Missing values will be imputed by using mean and outlier is removed to produce stable parameter.

Step 4: Modelling - Four predictive models are used to predict DM Type II disease. The approaches include are Logistic Regression, Pruning Decision Tree, Artificial Neural Network and Naïve Bayes. Approaches are chosen based on literature review from past studies on data mining models.

Table 2. Interval input summary.

Variable	Missing	N	Minimum	Maximum	Skewness	Kurtosis
Age	0	684	22	87	-0.1331	0.048
BMI	0	684	30.02	674.62	11.344	140.003
HbA1c	17	667	4.9	15.4	1.1178	1.368

Step 5: Model Assessment and Comparison - The ability of an estimated response model to predict diabetes can be assessed using accuracy measures. The tools from which various accuracy measures are derived include ROC chart and some statistics such as average squared error, misclassification rate and ROC index.

3. Results

The secondary data collected from a public clinic consists of 684 observations with 17 missing values for input HbA1c (2.49%). Majority of the patients (96.5%) does not have family history with diabetes while about 60% of the patients are non-smoking. A total of 356 male (52%) patients been diagnosed with diabetes and pre-diabetes disease. The chi-square test was used to test the association between Diabetic Category, Y and the categorical inputs. Hence, result shows that Gender (Chi-square=3.2668, p=0.0707) and Family History (Chi-square=1.4294, p=0.2319) have no statistical significance, while the other two inputs (Hypertension and Smoking) showed statistically significant association between the two groups, at 5% significant level. On the other hand, input BMI is transformed using power transformation to normalize the input distribution for better model performance. According to Chen & Deo (2004), power transformation is used to improve the normal approximation and ameliorates the sample effect.

Table 3. Model assessment for logistic regression.

	Logistic Regression (Enter)	**Logistic Regression (Stepwise)**
Accuracy	0.7470	0.7485
Sensitivity	0.8986	0.9058
Specificity	0.3636	0.3636
Misclassification Rate	0.2745	0.2696
Average Squared Error	0.1783	0.1799
ROC Index	0.7610	0.7610

Table 3 illustrates the validation results on performance measures for Logistic Regression (LR) models. The misclassification rate for LR Stepwise is slightly lower compared to LR Enter with accuracy 74.85%. Therefore, LR (Stepwise) is a better model. LR Stepwise can 90.58% predict positive category correctly.

Table 4. Analysis of maximum likelihood estimates (LR stepwise).

Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Exp(Est)**
Intercept	1	-4.1799	0.7315	32.65	<.0001	0.015
Hypertension	1	0.3928	0.123	10.2	0.0014	1.481
HbA1c	1	0.6335	0.0926	46.81	<.0001	1.884

Table 4 shows the coefficient, *B* that determines the risk factors correlated with diabetes (Y=1) and the corresponding odds ratios. For input Hypertension (B=0.3928, odds ratio= 1.481), the odds of having diabetes is higher by 1.481 times for patients with no hypertension as compared to patients with hypertension. Meanwhile, for HbA1c (B=0.6335, odds ratio= 1.884), odds of having diabetes is increased by 88.4% for one unit increase in percentage of HbA1c among patients. On the other hands, results for Decision Tree (DT) are tabulated in Table 5. It shows the misclassification rate, DT (Gini) and DT (CART) has lower rate as compared to DT (CHAID). However, DT (Gini) has ASE value compared to DT (CART). Therefore, Decision Tree (CART) is the best model for Decision Tree approach.

Table 5. Model assessment for decision tree.

Decision Tree (Gini)		Decision Tree (CART)	Decision Tree (CHAID)
Accuracy	0.7288	0.7288	0.7222
Sensitivity	0.9348	0.9348	0.9420
Specificity	0.2727	0.2727	0.2424
Misclassification Rate	0.2790	0.2790	0.2840
Average Squared Error	0.1890	0.1878	0.1900
ROC Index	0.7230	0.7230	0.7140

For Artificial Neural Network (ANN) models, six ANN models are used with different hidden nodes for model comparison based on model complexity. The output shows that the misclassification rate values for model ANN_3 (3 hidden nodes) is the smallest (27.90%) with sensitivity (89.13%) and precision (74.55%). Meanwhile, ANN_2 has highest ROC Index which is 78.40%. Hence, ANN_3 is the best model among other ANN models. Next, is to choose the best model among the best to predict DM Type II disease. Models LR (Stepwise), DT (CART), ANN_3 and Naïve Bayes are run for best model comparison as per illustrated in Figure 2.

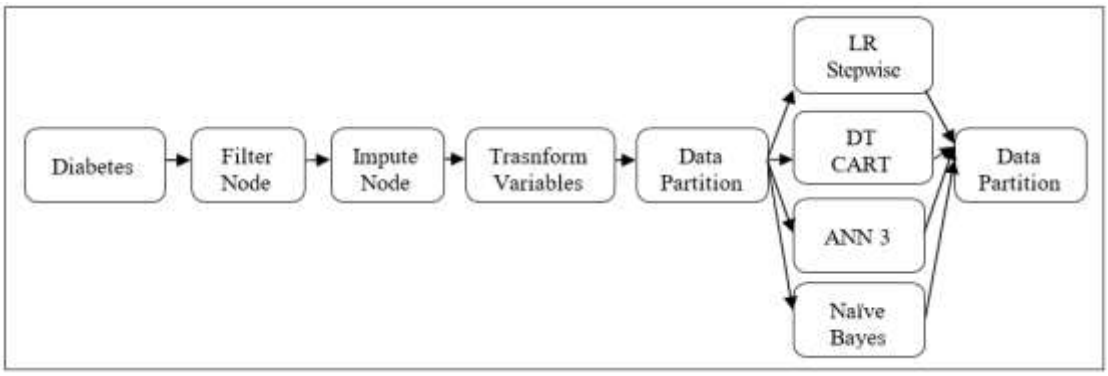

Figure 2. Best model comparison.

3.1. Best Model Comparison

Often the area under the Receiver Operating Characteristics (ROC) curve is used to determine the quality of classification models predicts the classes best (Esmaily et al., 2018). Range value area under the curve (AUC) for a good model should be between 0.5 to 1. Greater AUC value will lead to a better model.

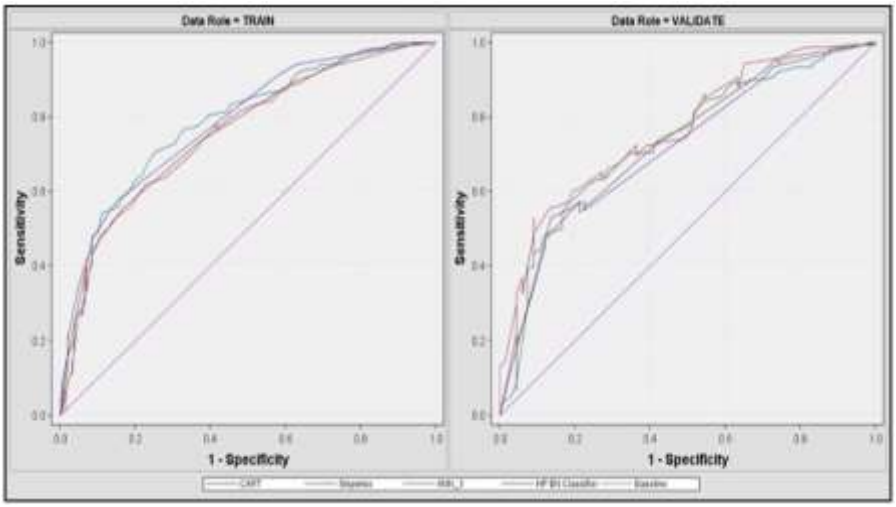

Figure 3. ROC charts for best model comparison.

Figure 3 shows the ROC Chart of the four chosen models. According to the AUC it could be said that prediction of Logistic Regression (Stepwise) is better than other models. In fact, the curves which climb quickly toward the top-right meaning the model correctly predicted the cases. Hence, Logistic Regression (Stepwise) is the best model to predict DM Type II Disease. Summary of model comparison (Table 6) for LR (Stepwise), DT (CART), ANN_3 and Naïve Bayes which also shows that Logistic Regression (Stepwise) is the best model since it has smallest misclassification rate (0.26961) meaning that the model is high in accuracy with smallest ASE and highest ROC Index. Given that, the Root Mean Squared Error (RMSE) value for LR (Stepwise) also indicate a good model (RMSE= 0.424).

Table 6. Summary of best model comparison.

	Logistic Regression Stepwise	**Decision Tree (CART)**	**Artificial Neural Network 3**	**Naïve Bayes**
Accuracy	0.7485	0.7288	0.7455	0.7293
Sensitivity	0.9058	0.8913	0.8913	0.9565
Specificity	0.3636	0.3636	0.3636	0.2576
Misclassification Rate	0.2696	0.2794	0.2794	0.2696
Average Squared Error	0.1799	0.1878	0.1888	0.1902
ROC Index	0.7610	0.7220	0.734	0.729

3.2. Model Scoring

Logistic Regression (Stepwise) is selected as the best model in predicting DM Type II disease. Model scoring is performed to assess the accurateness and efficiency of the selected model (Steyerberg et al., 2001). Another dataset is used for the model scoring purpose. The dataset used are secondary data which obtained from a public hospital in Klang Valley. Output for model scoring are as tabulated in Table 7. The probability is calculated by using the following formula:

$$\hat{p} = \frac{1}{1 + e^{-m}}$$

Where $m = \log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = -4.1799 + 0.3928$; hypertension + 0.6335 HbA1c

This study used ten observation of future data as testbeds.

Table 7. Summary of model scoring.

No	**Hypertension**	**HbA1c**	**Predicted: Y=1**	**Predicted: Y=0**	**Probability**	**Prediction for Y**
1	1	9.3	0.7889	0.2111	0.7889	1
2	0	10.2	0.9355	0.0645	0.9355	1
3	1	6.3	0.3585	0.6415	0.6415	0
4	1	12.4	0.9638	0.0362	0.9638	1
5	1	5.7	0.2765	0.7235	0.7235	0
6	0	7.7	0.7485	0.2515	0.7485	1
7	0	6.3	0.5507	0.4493	0.5507	1
8	1	6.9	0.4497	0.5503	0.5503	0
9	1	7.5	0.5444	0.4556	0.5444	1
10	1	10.6	0.8949	0.1051	0.8949	1

Table 7 displays a patient will be predicted for having diabetes (Y=1) when the probability value is more than and equal to 0.5 and the pre-diabetes (Y=0) will be diagnosed when the person has probability value less than 0.5. The prediction for observation 3, 6 and 8 are to have pre-diabetes would be incorrectly prediction because the probability value is greater than 0.5 for all observation. Meanwhile, for observation

1, 2, 4, 5, 7, 9 and 10 are correctly predicted to have diabetes. Thus, the prediction error rate is 0.3 (30%) and 70% of accuracy.

4. Conclusions

In conclusions, this study has successfully predicted the DM Type II. The performance of Logistic Regression-Stepwise outperformed to other predictive models with classification accurateness of 73% and able to predict positive outcome (Y=1) correctly by 90%. The significant inputs that affect DM Type II prediction (Y=1) are Hypertension and Glycated Hemoglobin (HbA1c) given the Root Mean Squared Error (RMSE) of model is 0.424. The importance of study may be able to contribute in improving the strategies and planning on diabetes diseases in Malaysia.

Author Contributions: Conceptualization, S.I., and S.S.M.K.; methodology, S.I., and S.S.M.K.; software, S.S.M.K.; validation, S.S.M.K.; formal analysis, S.I., and S.S.M.K.; investigation, S.I., and S.S.M.K.; resources, S.I.; data curation, S.S.M.K.; writing—original draft preparation, S.I., and S.S.M.K.; writing—review and editing, S.I., and S.S.M.K.; visualization, S.I.; supervision, S.S.M.K.; project administration, S.S.M.K.; funding acquisition, S.S.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express their gratitude to Ministry of Health (MOH), record unit from one of public clinic in Kuala Lumpur, Malaysia for providing data of diabetes. Also special thanks to the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor for the financial support provided.

Conflicts of Interest: The authors declare no conflict of interest.

References

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374.

Anitha, S., & Sridevi, N. (2019). Heart disease prediction using data mining techniques. *Journal of Analysis and Computation*, *8*(2), 48–55.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359.

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of Research in Health Sciences*, *18*(2), 412.

Hussein, Z., Taher, S. W., Singh, H. K. G., & Swee, W. C. S. (2015). Diabetes care in Malaysia: problems, new models, and solutions. *Annals of Global Health*, *81*(6), 851–862.

Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, *2*(2), 194–200.

Kazemnejad, A., Batvandi, Z., & Faradmal, J. (2010). Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes. *EMHJ-Eastern Mediterranean Health Journal, 16 (6), 615-620, 2010*. https://doi.org/10.5829/idosi.wasj.2013.23.11.1119

Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, *29*(2), 93–99. https://doi.org/10.1016/j.kjms.2012.08.016.

Okwechime, I. O., Roberson, S., & Odoi, A. (2015). Prevalence and predictors of pre-diabetes and diabetes among adults 18 years or older in Florida: a multinomial logistic modeling approach. *PloS One*, *10*(12), 1–17. https://doi.org/10.1371/journal.pone.0145781

Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781.

Wah, Y. B. (2006). Some applications of data mining. *National Statistics*.

Wah, Y. B., Ismail, N. H., & Fong, S. (2011). Predicting car purchase intent using data mining approach. *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, *3*, 1994–1999.

World Health Organization. (2018). *Diabetes*. available at https://www.who