Predicting Type 2 Diabetes: A Machine Learning Approach

Lim Zi Hao, Mafas Raheem, Seetha Letchumy

Abstract*: Diabetes is a well-known common disease among people around the world. Diabetes causes many anomalies in the body and results in the patients to become under a long term medication. Detecting diabetes has been done via hectic medical tests and causes a delay for the patients to get to know their test results. However, data mining and machine learning approaches are in the frontline supporting the health care domain to make effective predictions in this regard. This paper elaborates about predicting Type 2 Diabetes Mellitus using classification models. A suitable secondary dataset was used to build classification models and the more suitable model was selected via the valid performance measures. In this line, the Random Forest, Support Vector Machine, Naïve Bayes and Artificial Neural Network models were built. Based on the performance measures, Random Forest has been identified as the more suitable classifier with the accuracy of 90%, the recall and precision value of 0.90.

***Keywords*: diabetes prediction, machine learning, predictive models, optimization, model tuning.**

I. INTRODUCTION

Diabetes is a condition which is created when the blood glucose level of a human is beyond the tolerance level. The human body generates glucose from the food intake and the glucose beyond the required level in the blood is controlled by the insulin produced by the human body. Diabetic condition arises when the human body cannot produce sufficient amount of insulin (Type 1 Diabetes) or does not make use of insulin well (Type 2 Diabetes). These conditions seem very common among people nowadays due to their lifestyle, food habit and sometimes due to genetics. Nearly 10% of people do suffer from the Type 1 Diabetes Mellitus (T1DM), yet people get affected regardless of their age and surprisingly diagnosed in children and young adults [1]. These blood glucose level of these people should be maintained with adequate amount of insulin. However, several research works have taken place to identify the risk factors of T1DM, though, it is found that genetic code is a serious factor which would increase the risk of developing the TIDM. However, it is impossible to prevent TIDM as the factors are unclear.

Nearly 90% cases are identified as Type 2 Diabetes Mellitus (T2DM) which is yet another common type of diabetic condition [1]. It is mostly diagnosed in older adults, but at present the children, adolescents and young adults are getting affected. The increasing levels of obesity, physical inactivity and improper diet are being identified as the major causes of this diabetic condition. Diabetes can cause serious problems such as eye sight issues, kidney failures and nerves system malfunction, heart diseases, stroke and bone-related issues. Having said that people used to follow the conventional methods to get themselves diagnosed for diabetes regardless of their educational background. It has been realised that type 2 diabetes is a serious problem and most people are less aware of it and suffer due to late diagnosis. One must consult a specialist medical officer to get diagnosed whether he/she is affected by diabetes, and expected to wait for a day or more to get the test report, thus a costly effort too. Subsequently, medical check-ups are less convenient and sometimes less reliable due to the facilities available in the labs. In this line, an efficient application and implementation of technology in terms of data analytics and machine learning would be more useful in diagnosing diabetes on time. Therefore, the project is aimed at building an effective machine-learning based predictive model to diagnose diabetes on time and to respond towards the unfavourable medical condition with hope. Further, the results of the predictive model would reveal the significant factors which affect diabetes, thus would help the practitioners and patients to handle it more effectively. As the on-time diagnosis of diabetes is essential, this project would take a prominent place in the health care industry and would enable the healthcare practitioners to make timely decisions.

II. LITERATURE REVIEW

Numerous pieces of literature concerning T2DM were reviewed to obtain relevant knowledge of the previous research. In particular, the causes of T2DM and similar models built in the past were explored and assembled in this section.

A. T2DM – risk factors and medical tests

A range of lifestyle factors such as physical inactivity, sedentary lifestyle, smoking habit and alcohol consumption have a great impact as the cause of T2DM. Further, obesity has also been identified as one the most important risk factor of T2DM and proved from several substantial studies [2], as it resists the secretion of sufficient amount of insulin and leads to the progression of the disease. Subsequently, during the pregnancy period, Gestational Diabetes Mellitus (GDM) is developed which is a condition of glucose intolerance.

Revised Manuscript Received on June 22, 2020.
***** Correspondence Author
Lim Zi Hao, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia. E-mail: lzhao1880@gmail.com
Mafas Raheem*, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia. E-mail: rmafas@gmail.com
Seetha Letchumy, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia. E-mail: seetha@apu.edu.my

There are approximately 7% of all pregnancies complicated by GDM, thus accounted as more than 200,000 cases in a year [4]. The risk of developing T2DM is high for the women with GDM compared to others with normoglycemic pregnancy [5]. This proves that the development of T2DM is supported more by pregnancy.

However, medical practitioners who treat these T2DM patients recommend some medical tests to perform an effective diagnosis. The A1C test, also known as haemoglobin A1C, HbA1C, glycated haemoglobin, or glycohemoglobin test is a blood test done to measure the level of blood glucose, also known as blood sugar, in the human body over the past 3 months [6]. The fasting blood glucose test is another blood test done after 8 hours of fasting [7]. It is normally done after an overnight fasting where a blood sample will be taken and measured like the A1C test. A Random Blood Glucose Test is a test similar to this test without fasting, but it is less accurate due to the possibility of the ingested food before the test which would affect the blood glucose levels.

The Oral Glucose test is a lab test done to examine how the body moves sugar from the blood to tissues [8]. A sample of blood is taken and then the patient will be asked to ingest a certain amount of glucose. Blood samples will usually be taken on the 30 to 60-minutes mark and the whole test may take up to 3 hours. The blood samples are then compared to measure how the body tolerates glucose.

As stated above, the diagnostic process of T2DM seems very hectic and sometimes it can give wrong results too. However, the patients and medical practitioners do rely on these kinds of tidy medical tests to diagnose the T2DM. But, with the involvement of the data analytics, the researchers interested in the healthcare domain started building predictive models which would support the stakeholders to perform early detection of the T2DM. This would be more beneficial rather than diagnosing it after being affected. In this line, similar predictive models were gathered and presented that were built in the past by a significant number of researchers.

B. Predictive Models – past researches

Researchers started working on this aspect to come up with machine learning-based predictive models which could predict the target as to whether a patient would be affected by diabetic or not. In the year 2013, Xue-Hui Meng et al. from China built predictive models for diabetes or prediabetes based on risk factors by comparing three data mining models such as Logistic Regression, ANN and Decision Tree. The Logistic Regression model obtained 76.13%, 79.59% and 72.74%, for accuracy, sensitivity and specificity respectively, the ANN model obtained 73.23%, 82.18% and 64.49% for accuracy, sensitivity and specificity respectively; and the Decision Tree model obtained 77.87%, 80.68% and 75.13% for accuracy, sensitivity and specificity respectively, which was the best among those three models [9].

In the year 2014, an enhanced J48 classification model was proposed to predict the diabetic conditions by two researchers named Gaganjot Kaur and Amit Chhabra [10]. The WEKA was used as the data mining API of MATLAB for building the J48 classifier with the accuracy of 99.87%. However, this was not supported by the sensitivity and specificity values.

In the year 2015, a proposal was made for a quicker and more efficient technique for diagnosing diabetes by using J48 Decision Tree and Naïve Bayes algorithms [11]. The Decision Tree model achieved 76.96% accuracy with 62.34% sensitivity and the Naïve Bayes model achieved a 79.56% accuracy with 69.84% sensitivity. The data used to build these models had imbalanced class and no action was taken to balance the class.

Some works of literature were obtained about building predictive models for the prediction of diabetes. However, the past research works did not mention the hyperparameters tuning with proper optimisation aspects involved in the model building process. This research has extremely incorporated the proper preprocessing and optimisation aspects to build more effective predictive models with better accuracy, precision and recall.

III. MATERIALS AND METHODS

The application of machine learning algorithms in the domain of healthcare has been significantly increasing from the recent past. The health practitioners immensely rely on the support rendered by the data mining and predictive modelling in predicting the diseases.

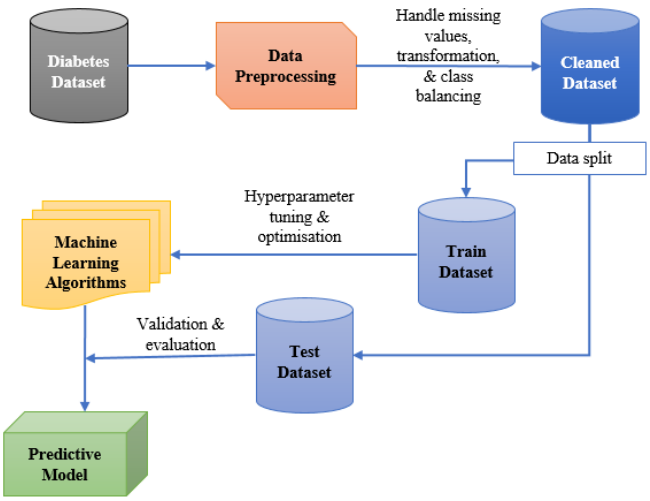


Fig. 1.Block diagram of predictive model building

The data mining and model building in this project has been planned as depicted in Fig. 1.

A. Dataset

The Pima Indian Diabetes dataset from the National Institute of Diabetes and Digestive Kidney Diseases was chosen for this project along with the features as given in Table-I. All patients were identified as at least 21 years old of Pima Indian heritage. The dataset was used to build a more effective machine learning predictive model that could predict whether a patient has diabetes or not based on certain diagnostic measurements found in the dataset. The dataset contains 768 rows, 8 features and 1 target variable.

Table-I: Features of the Dataset

Features	Data Type	Description
Pregnancies	Integer	Number of times pregnant
Glucose	Integer	Plasma glucose concentration after 2 hours in an oral glucose tolerance test.
Blood Pressure	Integer	Diastolic blood pressure (mm Hg)
Skin Thickness	Integer	Triceps skin fold thickness (mm)
Insulin	Integer	Insulin levels after 2 hours in an oral glucose tolerance test. (mu U/ml)
BMI	Decimal	Body mass index (BMI) is a measure of body fat based on height and weight (weight in kg/ (height in m)2)
Diabetes Pedigree Function	Decimal	Diabetes pedigree function is a function that scores the likelihood of diabetes based on family history
Age	Integer	Age (years)
Outcome	Integer	Target variable of the project. Class variable (0 or 1)

B. Pre-processing

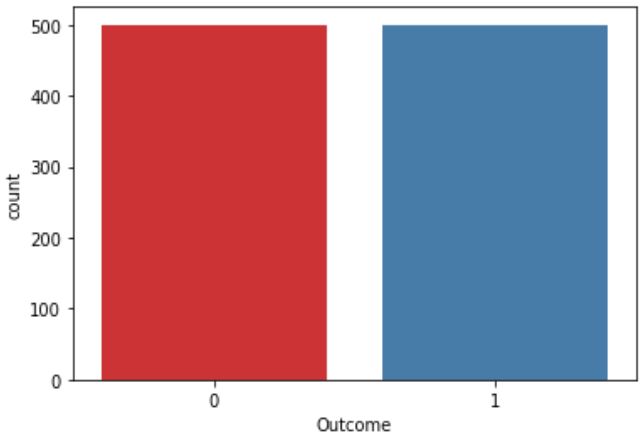


Fig. 2.Class balancing

Pre-processing has been recognised as the most important and time-consuming process in any data analytics project. Similarly, a significant effort was taken in this project to get the data more suitable for the analytics pipeline, especially handling missing values, data transformation and class balancing. The class balancing was done via the oversampling technique using the resample package, thus bringing both the classes' counts equal as shown in Fig. 2.

Further, the dataset did not show any missing values but identified as many zero values in most of the significant variables. The zero values were then replaced by null values except for pregnancies (possible to be zero) and Outcome (where zero meaning negative). Then the null values were replaced with the median of the respective variables. Subsequently, the dataset except the target variable showed significant skewness, thus a log transformation was done to normalize the dataset to avoid the biases in the results as shown in Fig. 3.

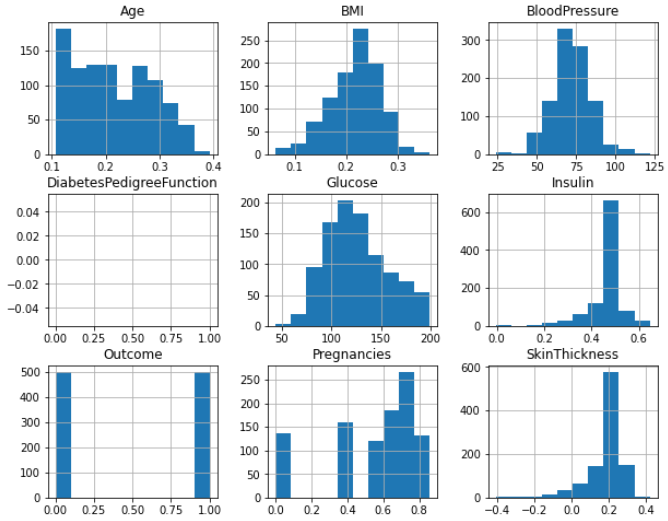


Fig. 3.Histograms after normalization

C. Hyperparameter Tuning

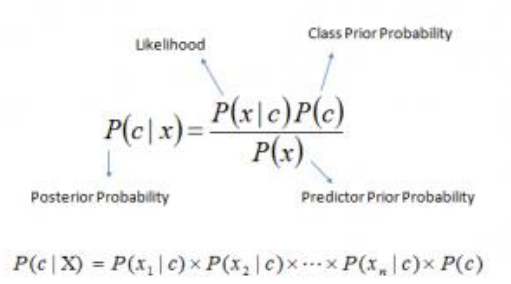
Hyperparameters are properties that exist in every machine learning algorithm that governs the training process [12]. Depending on the set hyperparameters, the machine learning models can perform differently even though the dataset is the same. The grid and random search along with the cross-validation was selectively used as the hyperparameter optimization techniques in this research. This would reduce the possibility of producing over fitted machine learning predictive models with the best accuracy levels. The grid search is the most basic hyperparameter optimization technique where a finite set of values is specified for each parameter and the Cartesian product of the set is evaluated [13]. The random search is an alternative of grid search techniques, where random samples of the model configuration are evaluated until the specified parameters are selected [14].

D. Predictive Models

The literature revealed that several machine learning algorithms were used to build predictive models in this regard. However, certain algorithms were widely selected by many researchers to build predictive models where they obtained different performance measures. Subsequently, it could be noted that the models were built with minimal tuning and optimisation aspects which are expected to add more accuracy in predicting the existence of diabetes. In this line, the following machine learning algorithms were chosen to build a more effective predictive model on Python platform.

E. Naïve Bayes (NB)

Naïve Bayes is a classification algorithm based on Bayes Theorem assuming that the predictors are independent [15]. It is a probabilistic algorithm used to build the baseline predictive model by taking into consideration of the posterior probability values. Bayes Theorem affords a mathematical way to calculate the posterior probability value of P(c|x) from P(c), P(x) and P(x|c) using the Equation 1.



Eq. 1. Naïve Bayes Equation [15]

P(c|x) → posterior probability of the target class c, given predictor x.
P(c) → prior probability of the target class c.
P(x|c) → likelihood probability of predictor with given target class c.
P(x) → prior probability of the predictor x.

F. Random Forest (RF)

Random Forest is a classifier that consists of multiple decision trees that functions as an ensemble algorithm [16]. It uses a huge number of decision trees to vote on the outcome as depicted in Fig. 4. The main reason for RF to be more effective is that the large number of trees defend each other from their discrete errors. A Random Search Optimizer is used in this classifier to prune the Decision Trees of the Random Forest.

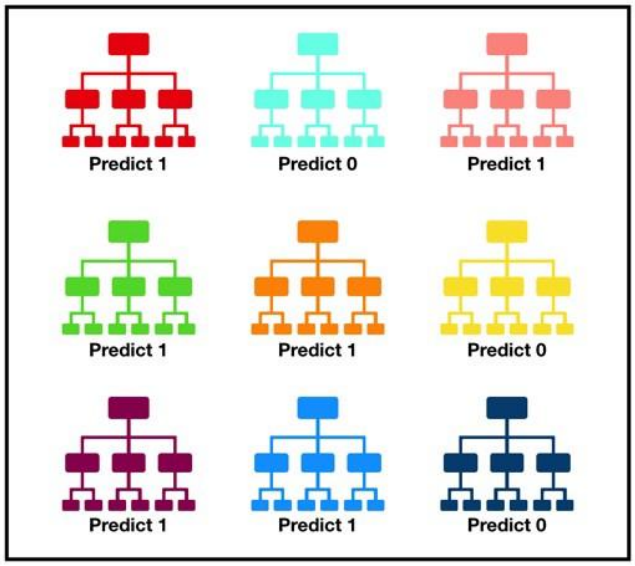


Fig. 4. Visualization of a RF Model Prediction [16]

G. Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm which can be used to build a classification model. In the SVM the features are plotted on an n-dimensional space (where n=number of features) and a hyperplane that separates the classes as depicted in Fig. 5 [17]. Grid Search was applied to find the best hyperparameter for the classifier to build the model.

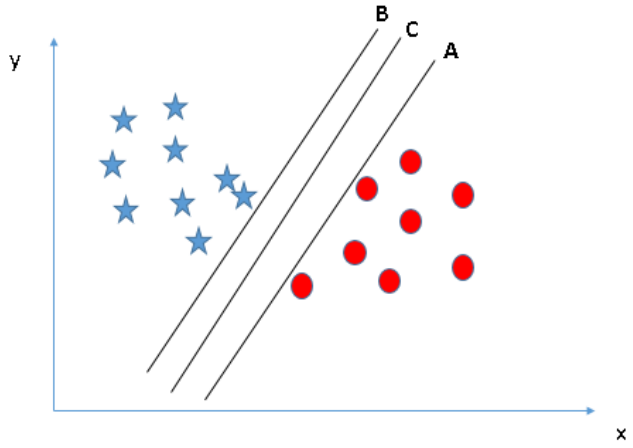


Fig. 5. Visualization of a SVM Model Prediction [17]

H. Artificial Neural Network (ANN)

ANN is one of the popular machine learning algorithms used to build predictive models in various domains. Those are brain-inspired systems that follow and replicate how a human brain learns. Those are made up of layers that consists of neurons that are connected from one layer to another as depicted in Fig. 6 [18]. Random Search was used in the hyperparameter tuning to find the best hyperparameter for the classifier to build the model.

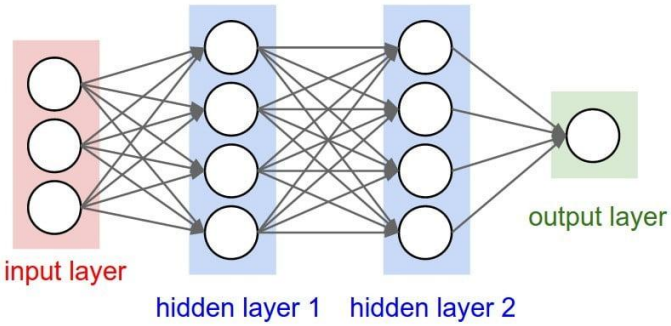


Fig. 6. Visualization of an ANN Model Prediction [18]

I. Model evaluation

The predictive models were validated while being built using random and grid search methods along with cross-validation techniques. Further, the models were evaluated using the performance measures such as accuracy, precision and recall (sensitivity) and the scores of the respective measures were obtained using the sklearn.metrics package.

IV. RESULTS AND DISCUSSION

The output results from the respective predictive models were recorded and compared, where the RF model obtained the best accuracy (89.6%) value than the other models as depicted in Fig. 7. The performance measures including the recall and the precision of the models were tabulated in Table-II to support the selection of the best model for this problem.

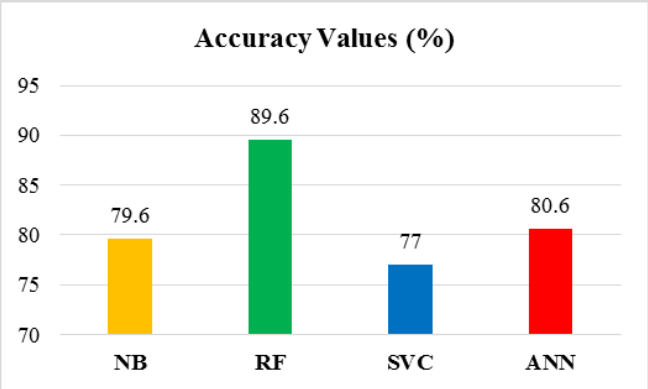


Fig. 7.Accuracy values of the predictive models

Table-II: Output result of the predictive models

	NB	RF	SVM	ANN
Accuracy (%)	79.6	89.6	77.0	80.6
Sensitivity/Recall	0.8	0.9	0.77	0.81
Precision	0.8	0.9	0.77	0.81

The models were inconsistent in predicting the new target with significant accuracy values. The relevant performance measures were taken to compare and choose the most suitable model. The models did very well where the RF model was the best with 89.6% accuracy and 0.9 recall on the diabetic target prediction. Similarly, the other models achieved significant accuracy, recall and precision values. The sensitivity/recall is one of the most important performance measures as it reflects the right prediction of the target

Table-III: Comparison with previous models

Models	Research	Accuracy (%)	Recall
Logistic Regression	Xue Hui-Meng et al., 2011	76.0	0.80
ANN		73.0	0.82
Decision tree		78.0	0.81
Improved J48 Decision Tree	Kaur et al., 2014	99.0	-
Naïve Bayes	Iyer et al., 2015	80.0	0.70
J48 Decision tree		77.0	0.62
Naïve Bayes	**Current Research**	**79.6**	**0.80**
Random Forest		**89.6**	**0.90**
SVM		**77.0**	**0.77**
ANN		**80.6**	**0.81**

In general, the models used in the current research achieved better accuracy and recall values than the previous models as detailed in Table-III. The NB model accuracy is nearly the same where the recall is better than the previous models. Similarly, the accuracy of ANN from the current research is 7% better than the one built in the year 2011. The Naïve Bayes and SVM algorithm used in this research sits at the lower end of the accuracy spectrum.

V. CONCLUSION

Setting a solution platform with the aid of the technology advancements especially by using a predictive model is very beneficial for the healthcare domain. This project had produced four classifications/predictive models that achieved reasonable results among which the RF model was selected as the best predictive model for this problem with approximately 90% accuracy. Proper data preprocessing and optimisation techniques supported well in building a more effective predictive model with better accuracy than the past.

There can still be improvements made on these models especially on the optimization of the ANN as many options are available to build a more effective classification model using the different number of hidden layers and neurons. Also, deep learning can be implemented using Tensor flow and Keras to see whether a predictive model could be built with more than 90% accuracy.

REFERENCES

1. International Diabetes Federation, “International Diabetes Federation - Type 2 diabetes”, 2019. Accessed on July 1, 2019. [Online]. Available: https://www.idf.org/aboutdiabetes/type-2-diabetes.html
2. Walley, A. J., Blakemore, A. I., Froguel, P., “Genetics of obesity and the prediction of risk for health”, Human Molecular Genetics, Vol. 15, No. 2, pp. R124-R130, 2006.
3. Belkina, A. C., Denis, G. V., “Obesity genes and insulin resistance”, Curr Opin Endocrinol Diabetes Obes., Vol. 17, No. 5, pp. 472-477, 2010.
4. Diabetes Care, “Gestational Diabetes Mellitus”, Diabetes Care, Vol, 25, No. 1, pp. S94-S96, 2002.
5. Bellamy, L., Casas, J. P., Hingorani, A. D., Williams, D., “Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis”, The Lancet, Vol. 373, No. 9677, pp. 1773-1779, 2009.
6. The National Institute of Diabetes and Digestive and Kidney Diseases, “The A1C Test & Diabetes NIDDK”, 2018. Accessed on January 30, 2020. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/tests-diagnosis/a1c-test.
7. Villines, Z., “Fasting blood sugar: Normal levels and testing”, 2019. Accessed on January 30, 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/317466.
8. Medline Plus, “Glucose tolerance test - non-pregnant: MedlinePlus Medical Encyclopedia”, 2019. Accessed on January 30, 2020. [Online]. Available: https://medlineplus.gov/ency/article/003466.htm.
9. Xue-Hui, M., Yi-Xiang, H., Dong-Ping, R., Qiu, Z., Qing, L., “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors”, The Kaohsiung Journal of Medical Science, Vol. 29, No. 2, pp. 93-99. 2013.
10. Kaur, G., Chhabra, A., “Improved J48 Classification Algorithm for the Prediction of Diabetes”, International Journal of Computer Applications, Vol. 98, No. 22, pp. 13-17. 2014.
11. Iyer, A., Jeyalatha, S., Sumbaly, R., “Diagnosis of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 5, No. 1, pp. 1-14. 2015.
12. Feurer, M., Hutter, F., Kotthoff, L., Vanschoren, J., “Automated Machine Learning”, 1st ed. Cham: Springer. 2019.
13. Scikit-Learn Devs, “Tuning the hyper-parameters of an estimator — scikit-learn 0.22.2 documentation”, 2019. Accessed on January 30, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html.
14. Bergstra, J., Bengio, Y., “Random Search for Hyper-Parameter Optimization”, Journal of Machine Learning Research, Vol. 13, pp. 281-305. 2012.
15. Ray, S., “6 Easy Steps to Learn Naive Bayes Algorithm”, 2017. Accessed on February 20, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.
16. Yiu, T., “Understanding Random Forest - Towards Data Science”, 2019. Accessed on February 20, 2020. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

17. Ray, S., “Understanding Support Vector Machines(SVM) algorithm”, 2017. Accessed on February 20, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.
18. Dormehl, L., “What is an artificial neural network? Here's everything you need to know”, 2019. Accessed on February 21, 2020. [Online]. Available: https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/.

AUTHORS PROFILE

Lim Zi Hao – A Computer Science graduate specialized in Data Analytics from the Asia Pacific University of Technology & Innovation, Malaysia. He is experienced in data analytics and web development.

Mr. Mafas Raheem - Data Scientist | Business Analyst Mafas is an academic specialized in the field of Data Science & Business Analytics with nearly 15 years of academic & industry experience. He holds an MSc in Data Science & Business Analytics and a Master of Business Administration degree and reading his PhD in the area of Machine Learning/Text analytics. Currently, he works as an academic at the Asia Pacific University of Technology & Innovation, Malaysia. His research areas are business intelligence, visual analytics, predictive analytics, text analytics & sentiment analysis in various domains. He has published a significant number of journal articles in the area of data analytics and machine learning.

Seetha Letchumy M. Belaidan is a Lecturer at Asia Pacific University of Technology and Innovation (APU) Malaysia. She obtained her Master in Computer Science from USM. She is passionate about teaching and has over 20 years of teaching experience. Her research interests include database technologies, data analytics and visualization.