Using Machine Learning to Predict Diabetes

Terence Lam
Faculty of Science and Technology
University of Canberra
Canberra, Australia
u3206488@uni.canberra.edu.au

Uyen Nguyen
Faculty of Science and Technology
University of Canberra
Canberra, Australia
u3206201@uni.canberra.edu.au

***Abstract*—Diabetes is a prevalent issue in nowadays societies, and it is beyond serious to neglect the problem. As in 2019, there are an estimated 1.5 million deaths directly related to diabetes. The number gives us an idea of how severe the problem is. This paper compares the performance of Logistic Regression, Support Vector Machine and K-Nearest Neighbours.**

***Keywords*— Diabetes, Logistic Regression, K-Nearest Neighbours, Support Vector Machine, F1-Score, Receiver Operating Characteristic curve, Precision**

I. Introduction

The number of Diabetes patients have risen to an excessive amount of roughly 463 million adults in 2019. This has set an alarm to the medical world since Diabetes can affect other organs as well as the whole immune system of the patient body. For this reason, the prevention of diabetes onset at earlier stages is closely investigated by many medical practitioners. [3]

To gain the outcome, early prediction of diabetes by using machine learning has been proposed as one of the most effective methods besides other physical, chemical as well as biological therapies. Machine learning is a method in data science in which the machine will learn from its experience with multiple datasets to put forward the outcome with high accuracy.

The dataset experimented is the Pima Indians Diabetes Database. This report will compare three machine learning algorithms – Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN). The result will then be evaluated with the F1-Score, Receiver Operating Characteristic (ROC) and Precision.

II. Literature review

A. Logistic Regression

Logistic Regression is one of the regression analysis models in supervised machine learning. Logistic Regression is used to discover the relationship between two or more independent (features) variables with a binary dependent (label) variable. The value of the label will either be 0 or 1, symbolising that whether the occurrence of an event is true or false. [8][9]

As mentioned in [9], the slope coefficient of exponentiated logistic regression can be easily interpreted as an odds ratio, which gives logistic regression an advantage over other regression models and thus it is widely used by medical researchers.

B. Support Vector Machine

Support Vector Machine (SVM) is a linear model for solving classification and regression problems. SVM is capable for handling linear and nonlinear non-separable data.

SVM was used in previous diabetes diagnostic prediction research and was able to achieve a decent percentage of accuracy. As shown in [10], SVM was able to achieve an acceptable accuracy of 66.25% in the Pima Indians Diabetes Database.

C. K-Nearest Neighbours

K-Nearest Neighbours (KNN) algorithm is a lazy learning machine learning algorithm that is easy-to-implement. It categorises a sample based on the class of its neighbours within a certain distance. KNN can be used for classification and regression problems.

As shown by [7], KNN has been previously implemented and evaluated with different combinations and variations of KNN using the Pima Indians Diabetes Database. One of the variations used was amalgam KNN. The amalgam KNN method pre-processes the dataset by removing noise, applying k-means and replacing missing values. This technique ultimately improves the overall quality of the data minded and shortens the time of mining. When k = 1, the amalgam KNN method was able to achieve a high accuracy of 95.57%.

D. F1-Score

F1-Score is an evaluation measure of a model that finds the harmonic mean of the combination of precision and sensitivity.

$$F1 = \frac{(\text{Precision} \cdot \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})} \tag{1}$$

E. Receiver Operating Charateristic curve

Receiver Operating Characteristic (ROC) curve represents the connection between sensitivity and specificity. As stated in [5], sensitivity and specificity are very popular in clinical test, which is similar to what we had experimented in this research.

Sensitivity is used to represent the number of true positives classified out of the overall actual positives (sum of true positives and false negatives); Specificity is used to represent the number of true negatives classified out of the overall actual negatives (sum of false positives and true negatives). The calculation of sensitivity and specificity are:

$$Sensitivity = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \tag{2}$$

And

$$Specificity = \frac{\text{True Negative}}{(\text{False Positive} + \text{True Negative})} \tag{3}$$

As stated in [4], in medical research, ROC curve is capable of considering event status and marker value within a fixed time. However, there are no follow-up analyses after a patient is diagnosed with no disease, but disease can develop later and goes undetected. Kamarudin et al. suggested that a time-dependent ROC curve is a more appropriate approach to continuously detect existing disease and potential disease.

F. Precision

Precision (also known as positive predictive value) is a form of evaluation strategy that is very well known in the machine learning field including medical research. The purpose of precision is to find out the rate of true positives classified among the overall number of positives (the sum of true positives and false positives) classified. Or in other words, to find how many positives are true positives.

$$Precision = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})} \quad (4)$$

III. Implementation Strategies

A. Reasons for choosing the models

1) Logistic Regression

Logistic Regression is easy-to-implement and has been popular in the medical classification for a long time, which makes the model more reliable in the dataset we experiment. Logistic Regression usually returns a good accuracy in simple datasets. Logistic Regression is less likely to be overtrained in low dimensional datasets.

2) Support Vector Machine

SVM uses less memory space as it only requires a small number of training points and implements a faster training process and returns a good accuracy score at the same time. Although training with SVM is time-consuming, the Pima Indians Diabetes Database is considered a small dataset, which makes training faster compared to other datasets. SVM is easy-to-implement in this dataset, with the indication of higher accuracy compared to logistic regression and KNN.

3) K-Nearest Neighbours

KNN is a non-parametric and lazy model, which means the data distribution and the training session of the dataset are disregarded. The training phase of KNN does not require a large amount of time. KNN is capable to handle non-linear data. Since there are irrelevant data pints in the dataset, the KNN model can benefits from the reduction of the size of the dataset, which makes the entire training process faster.

B. Methodology

1) Dataset

The Diabetes dataset found is the Pima Indians Diabetes Database from the National Institute of Diabetes and Digestive and Kidney Diseases. The Diabetes dataset consists of 768 records and nine attributes i.e. pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome (label).

2) Pre-processing

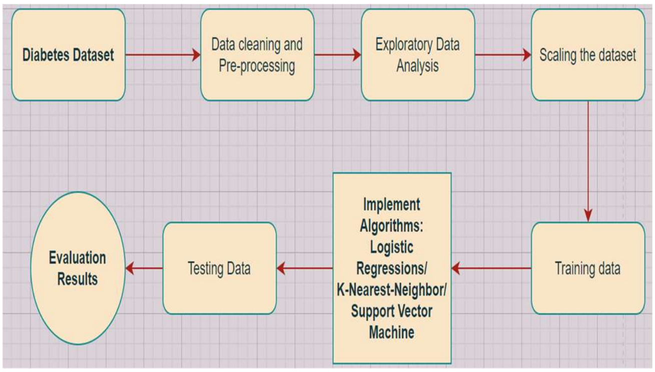
The dataset is witnessed to have a large amount of ‘0’ values. Therefore, it is indispensable that these ‘0’ values have to be replaced in the dataset. By understanding so, ‘0’ have been filled with the mean values in the attributes Glucose, Blood Pressure, Skin Thickness, BMI. Meanwhile, the attribute of Insulin has been modified by replacing ‘0’ with the median value of the column. The decision has been proceeded by looking at the model distribution of each variable in the dataset. Afterward, the Diabetes Pedigree Function was dropped since it does not have a good correlation and a high variance within the dataset, hence proved to be less correlated with the outcome.

Tackling the outliners in the dataset is also an area which we need to focus on. After delving into each attribute distribution by using box whisker plots, the decision to get rid of the outliners had been made in the variables such as Pregnancies, Blood Pressure, Skin Thickness, Insulin and BMI. As for Glucose, this variable shows no outliers to handle.

3) Procedure

The procedures of the research are:

1. Processing and scaling the model before proceeding the model fitting process.
2. The dataset is splitting into 80:20 ratio for data training and testing.
3. Standardize before fitting the models.
4. Experimenting the dataset with logistic regression, SVM and KNN and record their accuracy score
5. Evaluating the score with evaluation strategies i.e. precision, ROC curve and F1-score
6. Tuning the model with GridSearchCV and choose the best parameters.
7. Comparing and analyzing the results


Diabetes Dataset
Data cleaning and Pre-processing
Exploratory Data Analysis
Scaling the dataset
Training data
Implement Algorithms: Logistic Regressions/ K-Nearest-Neighbor/ Support Vector Machine
Testing Data
Evaluation Results

IV. Exploratory Data Analysis

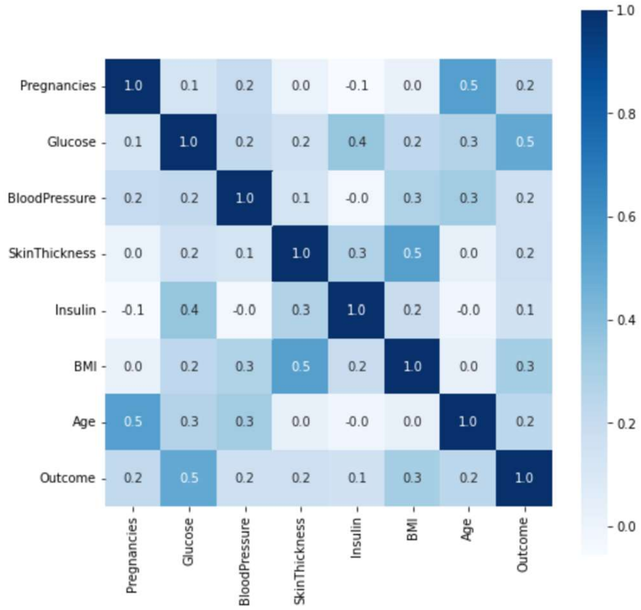


Figure 1

As can be seen from the figure, all of the variables have a relative correlation with the outcome. Out of the results, Glucose and BMI have the most outstanding correlation outcomes with 0.5 and 0.3 respectively.

1. Target variable

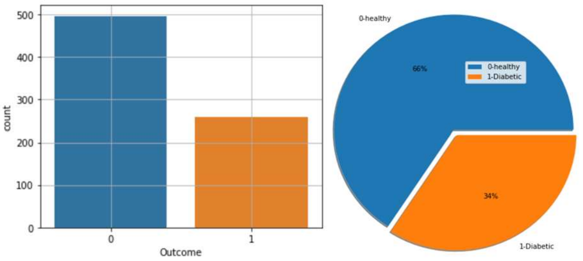


Figure 2: Outcome proportions

As illustrated by figure 2, the percentage of healthy women is predominant that of diabetic women. The percentages for both outcome labels are 66% and 34% for healthy and diabetic women correspondingly, which indicates a non-balanced distribution.

2. Pregnancies and outcome

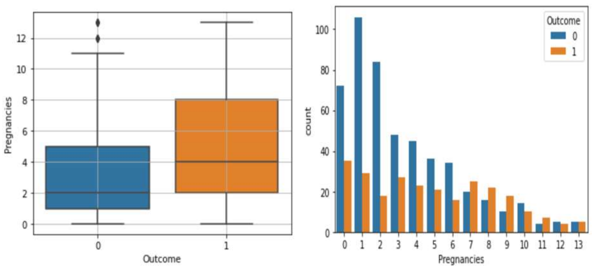


Figure 3: The Box-whisker plot for Pregnancies-Outcome and Pregnancies distribution

The number of pregnancy of people having diabetes is significantly higher than that of people who are non-diabetic. As illustrated in the pregnancy distribution, we can see that the diabetic level increases in contrast with the healthy rate as coming to a higher number of pregnancy. There are a few outliers in this plot after the number of 12 times of pregnancy.

3. Glucose and Outcome

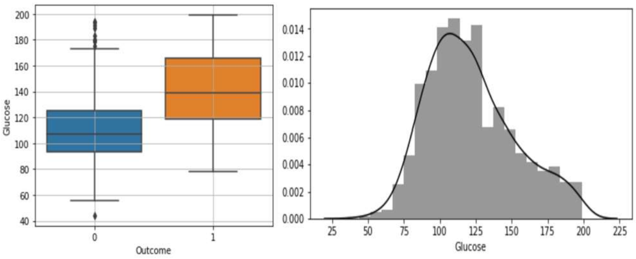


Figure 4: The Box-whisker plot for Glucose-Outcome and Glucose distribution

The higher the Glucose number is, the more the woman will be involved in Diabetes. There are also several outliners occurring from the Glucose level from 175 onwards. As witnessed in the distribution, the model is normally distributed with most of women having the Glucose level from 85 to 150.

4. Blood pressure and Outcome

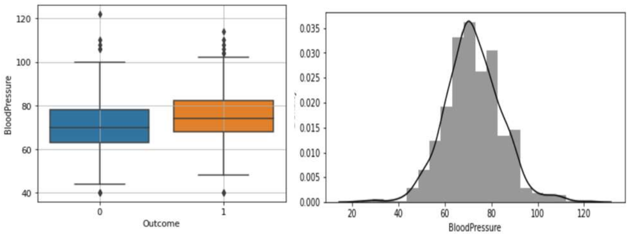


Figure 5: The Box-whisker plot for Blood Pressure-Outcome and Blood Pressure distribution

In contrast with the significant differences witnessed in Pregnancies number and Glucose level, we can perceive that there is just a slight difference between the outcome variables. Overall, the blood pressure of people having diabetes is higher than that of healthy people. Turning now to the Blood Pressure distribution, the blood pressure model is normally distributed with the common range from 60 to 80.

5. Skin Thickness and Outcome

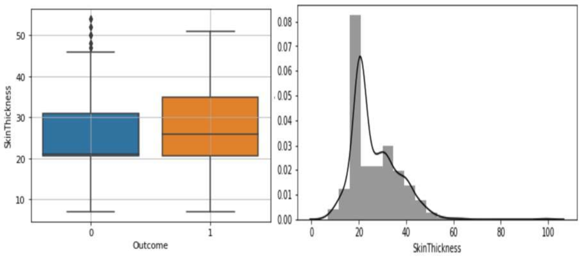


Figure 6: The Box-whisker plot for Skin Thickness-Outcome and Skin Thickness distribution

As is illustrated in the box-whisker plot, the higher the skin thickness size, there will be a considerably higher level of diabetic women. In this case, the outliners are still outstanding after the thickness size after 47.The model is also normal distributed within the range from

20 to 40. There is a considerable number of outliners as seen in the box plot.

6. Insulin and Outcome

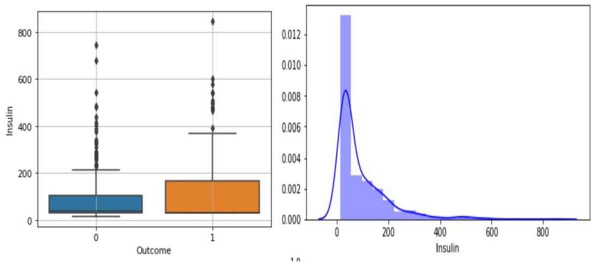


Figure 7: The Box-whisker plot for Insulin-Outcome and Insulin distribution

There is just a slight difference between the Insulin of non-diabetic and diabetic women. This subtle difference still reveals a higher insulin level in diabetic patients compared to that of non-diabetic patients. Due to a surge in a value near 0, the distribution of Insulin is skewed to the right. Significantly, there are numerous outliners in the box plot.

7. BMI and Outcome

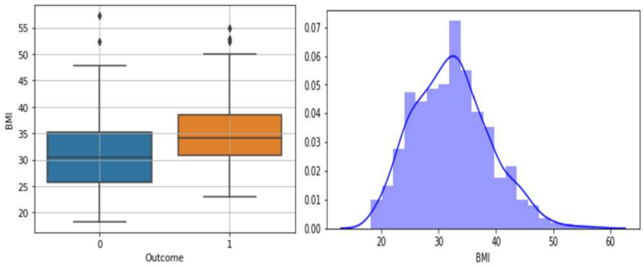


Figure 8: The Box-whisker plot for BMI-Outcome and BMI distribution

As is demonstrated in the above figure, the BMI level of diabetic women is obviously higher than that of non-diabetic women. It is witnessed that there are few outliners in the plot. In the distribution graph, it shows that most of the points will be allocated from 25 to 40 in the BMI level distribution.

8. Age and Outcome

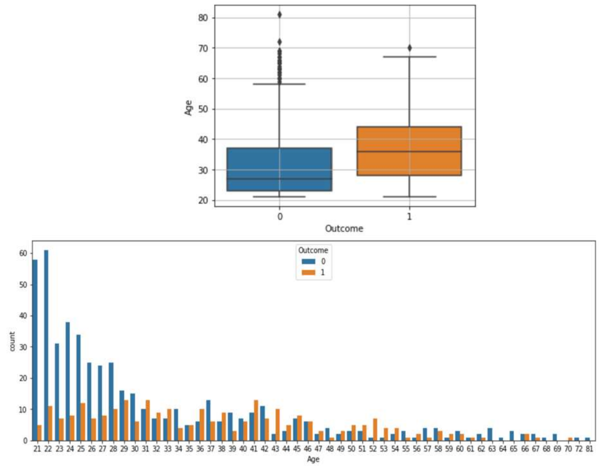


Figure 9: The Box-whisker plot for Age-Outcome and Age distribution

Considering age differences, as the age grows higher, there will also be a higher probability of getting diabetic as shown in the figure 9. As for the model age range, this thesis is also proved as the level of diabetic patient escalates in compatibility with the growing number of age.

V. Evaluation and Results

Model 1: Logistic Regression

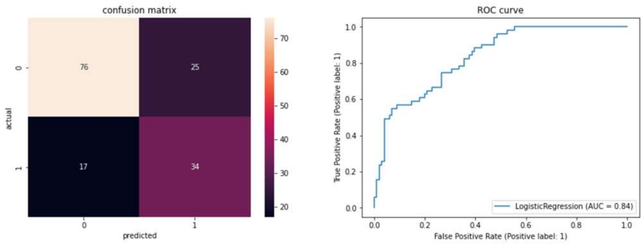


Figure 10: Logistic regression Confusion matrix and AUC

In this model, we can see that the false positive and false negative of the model are 25 and 17 comparatively. To upgrade the accuracy level of the model, these values should be decreased.

- Accuracy: 0.7237
- Precision: 0.74
- Recall: 0.72
- F1-score: 0.72
- AUC: 0.84

To improve the performance of the model, we will apply GridSearchCV with the best parameter.

Applying hypermeter tuning:

The model performance has been improved due to the decrease in the false positive (9) and false negative (22), hence the prediction ability of the model has been upgraded.

- Accuracy: 0.7649
- Precision: 0.79
- Recall: 0.80
- F1-score: 0.80
- AUC-score: 0.7397

Depending on the tuning results of the evaluation metrics, we can conclude that the new parameter (C=10) has made a considerable improvement in the logistic regression performance.

Model 2: K-Nearest-Neighbor

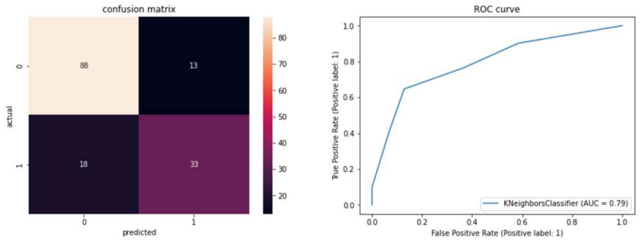


Figure 11: KNN Confusion matrix and AUC

As for this model, we can realize that there is a considerable of false results which is estimated of 31 (False negative + False positive). To optimize the prediction mechanism of the model, these values should be minimized as much as possible.

- Accuracy: 0.7960
- Precision: 0.79
- Recall: 0.80
- F1-score: 0.80
- AUC: 0.79

To improve the performance of the model, we will apply GridSearchCV with the best parameter.

Applying hypermeter tuning:

Due to an ineffectiveness of the model, we believe that the tuning method did not return a promising result with the best parameter of {'metric': 'euclidean', 'n_neighbors': 25}.

As illustrated below, all the evaluation metrics are predicted with lower values compared to before-tuning results.

- Accuracy: 0.7549
- Precision: 0.74
- Recall: 0.75
- F1-score: 0.75
- AUC-score: 0.6857

As stated above, this is not a good choice to optimize the performance of the whole model since the false predictions have been added up to 38.

Model 3: Support Vector Machine

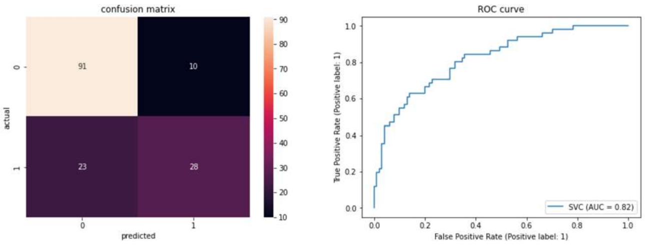


Figure 12: SVM Confusion matrix and AUC

In this model, we can see that the false positive and false negative of the model are 25 and 17 comparatively. To upgrade the accuracy level of the model, these values should be decreased.

Based on the SVM matrix, it can figure out that there is a substantial number of false values which is counted of 33 values in total (False negative + False positive). If these values and be minimized along with the better metrics evaluation, the model will be improved.

- Accuracy: 0.7829
- Precision: 0.78
- Recall: 0.78
- F1-score: 0.78
- AUC: 0.82

To improve the performance of the model, we will apply GridSearchCV with the best parameter.

Applying hypermeter tuning:

As for SVM tuning with the best parameter of {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}, the false negative and the false positive have been observed with a decrease with 8 and 22 respectively. Unfortunately, the evaluation results are illustrated with expected outcomes.

- Accuracy: 0.7516
- Precision: 0.80
- Recall: 0.80
- F1-score: 0.80
- AUC: 0.7447

As can be seen from the evaluation results, SVM has not been improved based on the tuning mechanism. Therefore, this tuning improvement method will not be implemented for the overall model performance regardless of the decrease in the false prediction value.

VI. Evaluation Table

In this section, we will compile all the evaluation results in one table for comparison among the algorithms.

Model	Accuracy	F1-score	AUC-score
Logistic Regression	0.7237	0.72	0.84
KNN	0.7960	0.80	0.79
SVM	0.7829	0.78	0.82
Tuning LR	0.7649	0.80	0.7397
Tuning KNN	0.7549	0.75	0.6857
Tuning SVM	0.7516	0.80	0.7447

Conclusion: KNN, SVM and tuning LR have the best evaluation results. To be specifically, the accuracy levels of the models are 0.79, 0.78 and 0.76 for KNN, SVM and tuning LR respectively. Overall, the best model to predict the diabetes is KNN with the accuracy level of up to 79%.

VII. Contribution and Suggestion

1. Contribution

The research has put forward multiple evaluation results of different models including logistic regression, K-nearest-neighbor, and Support Vector Machine. Besides, the tuning mechanism is also tested for these models to assess the effectiveness of the whole models. Based on these findings, people can evaluate their overall health by applying their variables statistics such as Insulin, BMI, Blood Pressure, Age, Pregnancy times, Skin Thickness, Glucose levels on the predictive model. On the world, woman have a high possibility of being diabetic throughout their pregnant process. This will be a general assessment for them to prevent the disease from the onset stage.

This study contribution should be taken into consideration for further research in the same field of biological relevant issues. We believe the model can be applicable for other diseases to be identified by other relevant datasets at the early stage.

2. Suggestion

In this project, we have mainly experimented on three models i.e. logistic regression, SVM, KNN. Three evaluation strategies i.e. accuracy, ROC curve - AUC, F1-score. Due to lack of experience in machine learning research, we could not go further into the research.

We suggest that we can experiment more learning models and evaluation strategies on the same dataset and potentially receive a better score and findings from other models and strategies. We can also try out different variations of models and evaluation strategies we used in this project i.e. amalgam KNN, time-dependent ROC curve.

Moreover, as we experimented on a relatively small dataset in this project, we will implement the same concepts into other datasets e.g. a larger dataset. We will analyze the performance for each model and strategy and compare them to the dataset we did in this project.

REFERENCES

[1] apoooooorv (2021) *EDA 📊 and Hyperparameter tuning 📈* [computer software], version 2, apoooooorv, Kaggle (online), accessed 29 October 2021

[2] Arnab Das (2021) *Diabetes prediction (LR, SVC, KNN, DT, RF, XGB)* [computer software], version 6, Arnab Das, Kaggle (online), accessed 29 October 2021

[3] International Diabetes Federation (2020) *Diabetes facts & figures*, International Diabetes Federation website, accessed 26 October 2021

[4] Kamarudin AN, Cox T, Kolamunnage-Dona R (2017) 'Time-dependent ROC curve analysis in medical research: current methods and applications', ProQuest, 17(1), accessed 27 October 2021

[5] Lalkhen AG, McCluskey A (2008) 'Clinical tests: sensitivity and specificity', Science Direct, 8(6):221–223, accessed 27 October 2021

[6] Loukas S (26 May 2020) 'How and why to Standardize your data: A python tutorial', *Towards Data Science*, accessed 29 October 2021.

[7] NirmalaDevi M, Appavu S, Swathi UV (2013) 'An amalgam KNN to predict Diabetes Mellitus', IEEE, 691-695

[8] Schober P and Vetter TR (2021) 'Logistic Regression in Medical Research', Anesthesia and analgesia, 132(2):365–366, accessed 26 October 2021

[9] Vetter TR and Schober P (2018) 'Regression: The Apple Does Not Fall Far From the Tree', *PubMed*, 127(1):277-283, accessed 26 October 2021

[10] Viloria A, Herazo-Beltran Y, Cabrera D, Pineda OB (2020) 'Diabetes Diagnostic Prediction Using Vector Support Machines', Science Direct, 170:376-381, accessed 26 October 2021