Original Article

A STUDY ON A NEW APPROACH TO HYBRID REGRESSION MODELING: A CASE FOR DIABETES MELLITUS WITH DYSLIPIDAEMIA PATIENTS WHO VISITED HOSPITAL USM

Mohamad Nasarudin Adnan[1], Wan Muhamad Amir W Ahmad[1*], Farah Muna Mohamad Ghazali[1], Hazik Bin Shahzad[1], Noraini Mohamad[1], Norhayati Yusop[1]**, Nor Farid Mohd Noor[2], Nor Azlida Aleng[3]**

[1]School of Dental Sciences, Health Campus. Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.
[2] Faculty of Medicine, Universiti Sultan Zainal Abidin (UniSZA), Medical Campus, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia.
[3]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia.
*Correspondence to: Wan Muhamad Amir W Ahmad. School of Dental Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.
DOI: 10.47750/pnr.2022.13.S01.196

Abstract

Background and Objective: The incidence of type 2 diabetes has been steadily rising over the past few decades, which has contributed significantly to the rise in the prevalence of diabetes (DM). Statistics from the World Health Organization show that more than 422 million adults worldwide had diabetes in 2014, and an ongoing rise in DM prevalence is anticipated. This study aims to create a method that can use to predict and manage diabetes cases in light of the importance of statistical modeling in diabetes. Decision trees and ordinal regression were the two methods used in this study. With some modification and extension, both methods will be harmonized in the R syntax. **Materials and Methods:** In this paper, we developed a method for analyzing decision trees using R syntax and embedding classification predictions. The classification for prediction with accuracy will indicate a successful classification analysis. This study illustrated the development method using diabetes data consisting of one thousand observations. Before further testing, the clinical relevance and significance of each preselected variable will be assessed. The decision tree will be used to evaluate nine variables. The selected variables are body mass index, total cholesterol, diabetes status, glucose reading, high-density lipoprotein, patient height, hip circumference, hypertension status, smoking status, and triglycerides. The classification obtained will be used as an input for the ordinal regression modeling. **Result:** It has been discovered that the status of diabetes can be determined by the level of glucose during fasting, which is consistent with the most recent research that has been published. one variable was chosen and used for the input of the ordinal regression. The suggested variables will apply to the ordered logistic regression, and the developed syntax will be used to assess the goodness of measurement and the significance level is set at a 0.05 level. **Conclusion:** Our proposed method achieves the highest level of forecasting precision possible. The methodology offers a precise evaluation of the fit of the final model. The superior performance of the model resulted in improved outcomes and efficient decision-making management.
Keywords: Diabetes, decision tree analysis, ordinal logistic regression

1.0 INTRODUCTION

Diabetes is a disease characterized by a high level of glucose in the sufferer's blood. It is extremely harmful to the human body. The blood glucose testing method is used to control diabetes which involves measuring blood glucose levels after fasting and after consuming glucose [6]. A fasting blood sugar test is required to diagnose diabetes, and an oral glucose tolerance test (OGTT) is used to monitor a patient's glucose level in the blood which measures the first and second stages of insulin response to glucose [1,2]. Insulin is a hormone that controls or regulates the body's glucose levels. The blood sugar level rises when the body does not produce enough insulin or when the insulin produced does not work effectively due to other factors, resulting in diabetes. After testing the blood glucose level in the patient fasting, 75grams of glucose dissolved in water is given to the patient and another test is conducted at one hour and two hours" intervals. Medical attention will be required if it is too high or too low [3,4,7].

Dyslipidemia, which has taken the place of the more outdated term "hyperlipidemia" is the term used to describe abnormal changes in body composition, particularly in body fat and lipid profiles [5]. Dyslipidemia, which is associated with lower levels of HDL (high-density lipoprotein) cholesterol, higher levels of plasma triglycerides, and higher levels of small dense particles of LDL (low-density lipoprotein) cholesterol, is a significant risk factor for coronary heart disease with diabetes mellitus [8]. Because crucial enzymes and metabolic pathways involved in lipid metabolism are impacted by insulin resistance or deficiency, lipid abnormalities are frequently present in diabetes mellitus [3]. Furthermore, it has been suggested that the lipid particle composition of diabetic dyslipidemia is more atherogenic than that of other forms of dyslipidemia. The implication of this is that even normal lipid levels may be more atherogenic in diabetics than in non-diabetics [9].

2.0 MATERIAL AND METHODS

Decision Tree

Decision Trees are a form of Supervised Machine Learning in which data is continuously partitioned according to a specific parameter. The tree is comprised of two entities: decision nodes and leaves. Decision nodes split the data into decision leaves representing decisions and results. According to Mesari and Ebalj (2016), a decision tree is a great and effective tool for categorizing data, producing predictions, and assisting with decision-making in sequential choice issues. This approach has been widely utilized in a variety of fields. In the medical field, for instance, the decision-maker is frequently confronted with a sequential decision problem involving options that, depending on chance, result in different outcomes. Figure 1 shows that decision trees are the best way to display this type of information graphically. It has aided stakeholders in intuitively understanding problems and making better decisions, as well as assisting in how decisions can be made and what may occur. A decision tree (Fig.1) is made up of three types of nodes: (a) decision nodes, (b) chance nodes, and (c) endpoint/terminal nodes. [14]. In this study, a data item x is a vector of d attribute values with an optional class label y. We refer to the collection of attributes as A(set of attributes) = $A_1, A_2, ..., A_d$. Thus, we can characterize x as $\{x_1, x_2, ..., x_d\}$, where $x_1 \in A1$, $x_2 \in A2$, ..., $x_d \in A_d$. Let Y (domain of class values)= $\{y_1, y_2, ..., y_m\}$ be the set of class labels. Each training item x is mapped to a class value y where $y \in Y$. X is the total set of training data (set of training data). Using a splitting rule called S, the data set X is divided into a set of subsets collectively known as X_S; that is, $X_S = \{X_1, X_2, ..., X_k\}$ where $\cup_i X_i = \mathbf{X}$. A decision tree is a rooted tree in which the root represents the full data set, and each set of parent nodes corresponds to an (X_s) partitioning of the parent's data set. The number of items in X_i that belong to class y_j is $|\boldsymbol{X}_{ij}|$. The likelihood that a chosen participant in a X_i is of class y_j is $p_{ij} = |X_{ij}| / |X_i|$

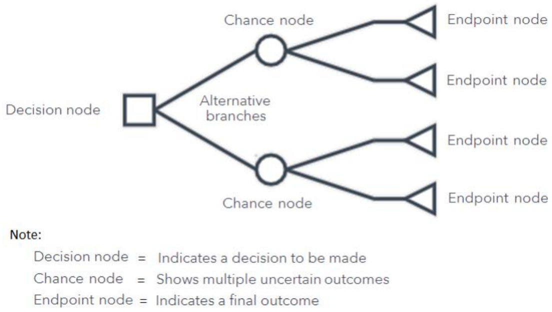

Figure 1: Decision trees are graphical models for describing sequential decision problems

Conditional Inference Trees (ctree)

The developed syntax employs the R package partykit's conditional inference trees (*ctree*). Hothorn et al. (2006) proposed a conditional inference tree or *ctree*. *ctree* separates the process of splitting into two distinct steps. First, the variable to split is determined by measuring the relationship between each covariate and the outcome of interest. After determining the splitting variable, the optimal split point for that variable is calculated. *ctree* employs formal statistical inference procedures at every stage of the splitting process. The coefficient measures the relationship between each covariate and the outcome in a regression model. A node is split only if sufficient evidence exists to reject the global null hypothesis, which states that none of the covariates have a one-to-one relationship with the outcome. If the global null hypothesis is rejected, the covariate with the strongest association to the outcome of interest is chosen as a candidate for splitting. No variable is selected for splitting if the minimal p-value is greater than the multiplicity-adjusted significance threshold and the node is designated as a terminal node. After the initial split, subsequent inference occurs inside subgroups, i.e., conditional on subgroup membership; Ctree-based splitting decisions on marginal regression models. The word "conditional" refers to this. Nine selected variables body mass index, total cholesterol, glucose reading, high-density lipoprotein, patient height, hip circumference, hypertension status, smoking status, and triglycerides were tested for the decision tree classification.

Bootstrap

Bootstrap starts by randomly selecting a sample from the population, after which sample statistics are computed. The bootstrap creates a pseudo population by substituting samples after several iterations of the initial samples. Samples obtained from random sampling with substitution differ from those obtained from the original sample. As each sample is drawn using replacement, the bootstrap calculates statistics for each sample.

Ordinal Regression

Ordinal regression can be useful when analyzing the relationship between one or more independent variables and a categorical dependent variable (one with more than two categories). When there are more than two categories in the response variable, ordinal logistic regression (OLR) is the type of logistic regression analysis that is used. We need an ordinal-scaled dependent variable in ordinal logistic regression. In this study, the diabetic reading is on an ordinal scale, and we will use the result from the decision tree classification to fit the ordinal regression. The maximum likelihood method will be used to estimate the regression parameter's value. The model for ordinal is given by

$$y_i^* = x_i\beta + \varepsilon_i \tag{1}$$

The dependent variable, however, is categorized, so we must use:

$$C_x(x) = \ln\left[\frac{P(Y \le j \mid x)}{P(Y > j \mid x)}\right]$$

and $\ln\left(\frac{\sum pr(event)}{1-\sum pr(event)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 +, \ldots, +\beta_k X_k$. It can be summarized as

$$\ln\left(\frac{P(Y \le j \mid x)}{1-P(Y \le j \mid x)}\right) = \alpha_j + \beta_i X_k \quad i = 1...k, \ j = 1, 2, ..., p-1 \tag{2}$$

where

α_j = called threshold or intercept

β_i = Parameter in the model

X_{i1} = Set of factors or independent variables.

Equation (2) above is an ordinal logistic model for *k* predictors with the *p*-1 levels response variable (Adeleke and Adepoju, 2010).

DATA AND R SYNTAX

The research was conducted at Hospital Universiti Sains Malaysia (USM) in Kubang Kerian, Kelantan, Malaysia. A total of 1000 patients took part in this study. The data summary for the selected variable in the analysis is described in Table 1.1.

Table 1.1: Data description of research variables

Code-variables	Explanation of user Variables
BMI	Body Mass Index
Choltot	Total of Cholesterol
Diab	Diabetes Status
	0 = Normal, 1= Borderline and 2 = Diabetes
Glucose	Glucose Reading
Hdl	High-Density Lipoprotein
Height	Patient Height

Hip	Hip circumference
Hyper	Hypertension Status 0 = Normal, 1= Borderline and 2 = Hypertension
Smoke	Smoking Status
Trig	Triglycerides

The statistical analysis was done with a method that was built into the R Studio software package. This study combines three statistical methodologies into a single syntax to get the best research results.

Bootstrapping Ordinal Regression Modeling with Decision Tree Analysis Using R Syntax

```
#install.packages("party")
library(party)
#install.packages("partykit")
library(partykit)
#install.packages("caret")
library(caret)
#install.packages("class")
library(class)
#install.packages("tree")
library(tree)
#install.packages("rpart")
library(rpart)
#install.packages("rpart.plot")
library(rpart.plot)
#install.packages("caTools")
library(caTools)

Input =("
bmi choltot diab glucose hdl height hip hyper smoke trig
37  209  2  114  37  178  130  2  1  168
31  175  1  109  34  181  107  1  2  332
27  228  3  153  33  183  105  2  1  304
26  194  1  92  50  178  98  0  1  81
23  156  1  94  48  175  100  0  2  98
⋮             ⋮            ⋮             ⋮
22  179  1  91  43  164  96  2  2  131
29  210  1  106  56  173  104  2  2  147
19  98  2  113  52  170  90  0  2  70
20  203  1  91  57  161  89  1  1  82
33  219  1  99  54  167  107  0  2  102
")
data = read.table(textConnection(Input),header=TRUE)
head(data)

# STEP 1: Modeling The Decision Tree Analysis
#To Convert Factors To Numbers
data$diab<-as.factor(data$diab)

#Scenario For Decision Tree Using-The Whole Data
dtm <- ctree(diab~hyper+ bmi+choltot+glucose+hdl+height+hip+smoke+trig, data=data)
plot(dtm)

# Calculating The Prediction For The Test
pred = predict(dtm, data[,-5])
confusionMatrix(pred,as.factor(data$diab))

##########################################################################
#Whole Scenario- Prediction for the Data
table_mat <- table(data$diab ,pred)
table_mat
```

```
# Calculating the Accuracy For The Test
accuracy_test <- sum(diag(table_mat)) / sum(table_mat)
print(paste("Accuracy for Test is ", accuracy_test))

################################################################################
# STEP 2: Modeling Ordinal Model
#if(!require(MASS)){install.packages("MASS")}
library(MASS)
#if(!require(ordinal)){install.packages("ordinal")}
library(ordinal)
#if(!require(erer)){install.packages("erer")}
library(erer)

#Performing Bootstrap For 1000
mydata <- rbind.data.frame(data, stringsAsFactors = FALSE)
iboot <- sample(1:nrow(mydata),size=10000, replace = TRUE)
Bootdata <- mydata[iboot,]

data$diab<-factor(data$diab)
model  <- clm(diab~hyper+ bmi+choltot+glucose+hdl+height+hip+smoke+trig, data = Bootdata)
options(warn=-1)
summary(model)
options(warn=-1)
anova(model, type="II")

################################################################################
## Fit Ordered Logit Model And Store Results 'M'
m <- polr(diab~hyper+ bmi+choltot+glucose+hdl+height+hip+smoke+trig, data = Bootdata, Hess = TRUE)
## view a summary of the model
summary(m)

## Store Table
(ctable <- coef(summary(m)))

## Calculate and Store P Values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
## combined table
(ctable <- cbind(ctable, `p value` = p))

(ci <- confint(m))  # default method gives profiled CIs
## Calculating For The Odds Ratios
exp(coef(m))

## Calculating For The OR and CI
exp(cbind(OR = coef(m), ci))

################################################################################
```

RESULTS

Decision Tree Result

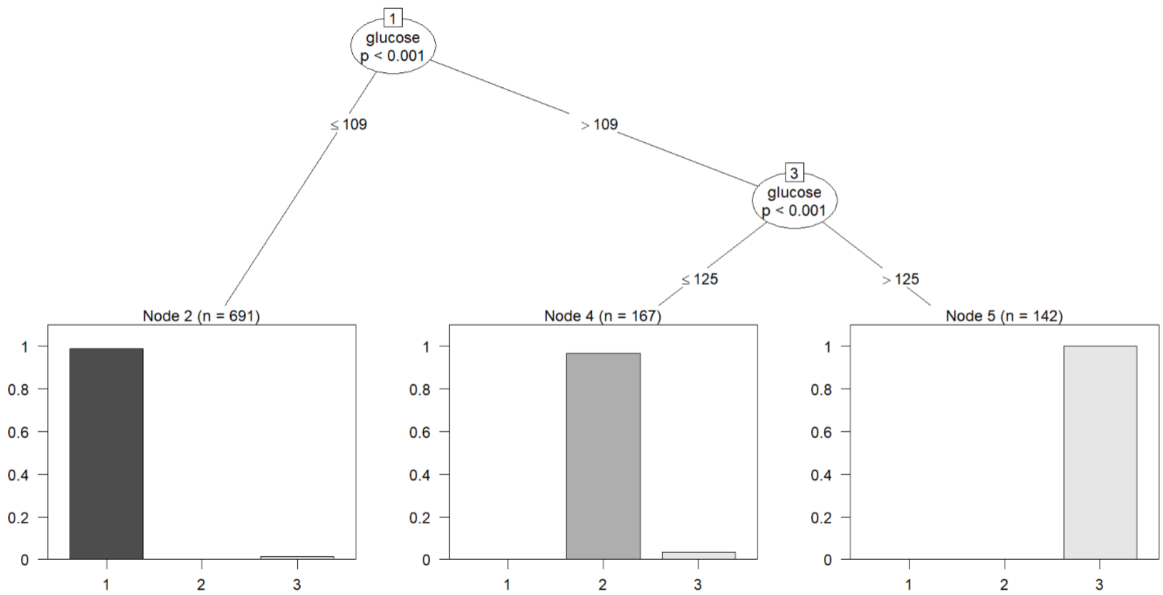

Figure 2: The Decision Tree Analysis

The outcome of the constructed decision tree analysis is depicted in Figure 2, and the only variable that played a role in determining this categorization was glucose ($p < 0.05$). The ordinal regression modeling will make use of this variable as a feeder variable for the ordinal regression modeling. The next step is to predict the outcome using decision tree analysis. This is to ensure that the selected variables are significant to contribute to the diabetes factor.

Prediction Using the Result of Decision Tree Result

Table 1.2: Confusion matrix and Statistics

		Reference		
		Normal	Borderline	Diabetes
Prediction	Normal	682	0	0
	Borderline	0	161	0
	Diabetes	9	6	142

Table 1.2 show the results of the prediction of decision tree analysis. The accuracy of the prediction can be assessed through the value obtained from the confusion matrix and statistics. The true value will be determined at the point of prediction and reference. Table 1.3 shows the summary of statistics for decision tree analysis.

Table 1.3: Overall Statistics for Decision Tree Analysis

Statistics	Value
Accuracy	: 0.985
95% CI	: (0.9754, 0.9916)
P-Value	: < 2.2e-16
Kappa	: 0.9687

Table 1.4: The Summary of Statistics by Class

	Class :1	Class :2	Class :3
Sensitivity	1.0000	1.0000	0.9045
Specificity	0.9717	0.9928	1.0000
Pos Pred Value	0.9870	0.9641	1.0000
Neg Pred Value	1.0000	1.0000	0.9825

Prevalence	0.6820	0.1610	0.1570
Detection Rate	0.6820	0.1610	0.1420
Detection Prevalence	0.6910	0.1670	0.1420
Balanced Accuracy	0.9858	0.9964	0.9522

According to Table 1.4, the accuracy rate is 98.5%, with a positive predictive value for Class 1 at 98.7 %, Class 2 is 96.41%, and Class 2 is 100.0%. This table contains various other statistical summaries.

Table 1.5: Parameter estimate on the ordinal logistic regression model.

Response	**Coefficient**	**St. Error**	**t-value**	**Sig.**	**Odds Ratio**	**95% Confidence Interval**	
						Lower	Upper
Cut1	44.2267	0.9404	47.0264	0.0000	-	-	-
Cut2	49.4803	1.0504	47.1036	0.0000	-	-	-
Glucose	0.4043	0.0088	46.5022	< 2.2e-16*	1.4942	1.4697	1.5201

** Significant at the level of 0.05*
Ordinal logistics regression was applied

Table 1.5 summaries the outcomes of an ordinal regression model that incorporates inputs from a decision tree analysis. The outcomes of an ordinal regression model that incorporates inputs from a decision tree analysis are summarized in Table 1.5. It was discovered that the glucose factor is crucial to the diabetic condition. The glucose factor [β_1 = 0.4043; p <0.05, CI: 95% (1.4697, 1.5201)] found to be very significant for the classification. This factor's standard deviation is low and acceptable. Regarding Table 1.3, there are two distinct thresholds designated as cut 1 and cut 2. This threshold can be used to determine the predicted probabilities that a patient with a given set of characteristics belongs to a specific category. The proposed ordered logistic regression models for the various cut-off points must be distinct and represented by a separate equation; thus, the formulations for the first and second categories become (the estimated model): The proposed ordinal regression is given as

Ordinal logistic regression for Cut 1

$$\text{Logit}(P(Y \leq 1) = 44.2267 + (0.4043 \times \text{Glucose level}) \quad (1)$$

Ordinal logistic regression for Cut 2

$$\text{Logit}(P(Y \leq 2) = 49.4803 + (0.4043 \times \text{Glucose level}) \quad (2)$$

Table 1.5 displays the ordered results of logistic regression for the significant variables selected from the decision tree modeling. This ordinal regression model takes into account only one factor that modifies the class odds ratio, namely glucose levels. As glucose levels rise, diabetes prevalence increases.

DISCUSSION

The discussion of this suggested model may be divided into two phases: first, the discussion focuses on the development of the technique, and second, the discussion focuses on the findings, which contribute to the study's findings. The hybrid model, comprised of an ordinal regression model, is constructed based on the model assumption of parallel lines for all corresponding outcome categories**.** The suggested methodology for an ordered logistic regression model with decision tree analysis requires the premise that the response variable is ordinal and the bootstrap technique's presence increases accuracy. This proposed method can serve as an alternate way to ordinal regression modeling. This proposed method can serve as an alternate way to ordinal regression modeling.

This proposed model will be discussed in two parts: first, the evolution of the method will be discussed, followed by a discussion of the results and their implications for the overall findings of the study. The hybrid model, which comprises an ordinal regression model, is based on the premise that each outcome category is represented by parallel lines. The methodology given for an ordered logistic regression model with decision tree analysis involves the assumption that the response variable is ordinal, and the existence of the bootstrap procedure enhances accuracy. This approach offers an alternative to ordinal regression modeling. Predictive modeling for forecasting the development of diabetes is crucial for public health professionals as healthcare costs continue to rise due to the prevalence of non-communicable chronic illnesses. This research will aid medical professionals in spotting patients who are at risk for developing diabetes. The suggested approach helps achieve the best level of predicting accuracy. This paper's objective was to discuss the evolution of ordinal regression approaches.
The main objective of the project was to build, test, and evaluate a decision tree, bootstrap, and ordinal regression combination for generating and using medical statistic strategies. The bootstrap approach creates an enormous dataset at first. The decision tree approach provides a precise evaluation of the variables that must be carefully

chosen for the final model in this investigation. Alternatively, researchers might utilize decision trees to determine which data points are most important when developing an ordinal regression model and also to test a clinical hypothesis. Decision tree inference of which variables to use in making a forecast proves the usefulness of this approach to health care planning. This strategy leads to an accuracy rate of 98.5%, with a positive predictive value for Class 1 at 98.7 %, Class 2 is 96.41%, and Class 2 is 100.0%. The most significant variables obtained from the decision tree classification were glucose levels, while the other variables is being excluded from the classification. This method can be utilized as an alternative to ordered regression modeling in situations where the selection of acceptable variables was based on a computational study that predicted the significance of the independent variable that should be selected for the final model. This technique simplifies the most difficult aspect of any research project, which is choosing the right input parameters. To assess the effectiveness of the established approach, the predictive model is applied to actual data and its output is compared to the actual data. The results helped the person making the decision get the best possible results. Throughout the previous decade, several researchers have studied the risk factors for diabetes using a variety of statistical methods, including logistic regression, correlation, and decision trees. By combining a decision tree with ordered regression analysis, we may improve the clinical application of risk variables. In a study in Qatar by AlKaabi, comparable results through the random forest and logistic regression analysis linked age and physical activity, consumption of fruits and vegetables, and history of diabetes as essential predictors of diabetes. In a longitudinal study, Dimitriadis showed a significant association of diabetes with risk factors of age, gender, and blood glucose levels. Using decision tree analysis, more conclusive, detailed, and reliable results can be obtained.

The proposed method and obtained outcomes illustrate the superiority of the employed modeling technique. The goal of this article is to create a hybrid approach that combines ordered logistic regression, decision trees, and bootstrapping. Future researchers will be able to replicate the processes because the syntax offered in R is meant to be easily understandable. In addition, this method can assist policymakers and health professionals in establishing a new program by improving existing preventive measures. Using bootstrap, decision trees, and ordered logistic regression, the developed predictive models can create accurate and robust diagnostic parameters. Using a decision tree to choose variables, an ordered regression may learn from the data and validate its prediction of diabetes risk. These models can be adapted to develop preventive strategies.

CONCLUSION

The R syntax synchronizes and harmonizes the application with the concept of a developed methodology-based approach. This research output demonstrates the findings and their application to diabetes risk factors. The decision tree ranks the importance of key risk factors.

The model's sensitivity and specificity were both greater than 75%. We can conclude that our proposed method produces excellent results with the highest possible forecasting precision. The method approach provides an accurate assessment of the final model's fit. The model's superior performance resulted in improved outcomes and effective decision-making management. This information will hopefully aid clinicians in managing and educating patients about diabetes-related risk factors. The R syntax algorithm integrates the idea of a methodology-based approach to finding risk factors in managing diabetic patients. The parameters of glucose levels contributed significantly to a patient's overall diabetes status, demonstrating that the methodology and findings were developed successfully.

Acknowledgments: The authors would like to thank Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2022/STG06/USM/02/10 and Universiti Sains Malaysia (USM).

REFERENCES

1. Almdal T, Scharling H, Jensen JS, Vestergaard H. The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: A population-based study of 13,000 men and women with 20 years of follow-up. Arch Intern Med 2004;164:1422-6.
2. Bagheri B, Radmard N, Faghani-Makrani A, Rasouli M. Serum creatinine and occurrence and severity of coronary artery disease. Med Arch 2019;73:154-6.
3. Bamanikar SA, Baravkar D, Chandanwale S, Dharwadkar A, Paranjape S. Study of cervical cytology and its correlation with clinical and histopathological findings. Clin Cancer Invest J 2016;5:403-8.
4. Chahil TJ, Ginsberg HN. Diabetic dyslipidemia. Endocrinol Metab Clin North Am 2006;35:491-510.
5. Farmer JA. Diabetic dyslipidemia and atherosclerosis: Evidence from clinical trials. Curr Diab Rep 2008;8:71-7.
6. Sghaireen MG, Alduraywish AA, Srivastava KC, Shrivastava D, Patil SR, Al Habib S, et al. Comparative evaluation of dental implant failure among healthy and well-controlled diabetic patients – A 3-year retrospective study. Int J Environ Res Public Health 2020;17:5253.
7. Solano, M. P., & Goldberg, R. B. (2006). Management of dyslipidemia in diabetes. Cardiology in Review, 14(3), 125–135. Retrieved from https://doi.org/10.1097/01.crd.0000188034.76283.5e
8. Taskinen, M. R. (2002). Diabetic dyslipidemia. Atherosclerosis Supplements, 3(1), 47-51.

9. VinodMahato, R., Gyawali, P., Raut, P. P., Regmi, P., Singh, K. P., Pandeya, D. R., & Gyawali, P. (2011). Association between glycaemic control and serum lipid profile in type 2 diabetic patients: Glycated haemoglobin as a dual biomarker. Biomedical Research (0970-938X), 22(3).