Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

MODELING CHOLESTEROL LEVELS IN PATIENTS WITH DYSLIPIDEMIA AND TYPE 2 DIABETES MELLITUS USING AN INTEGRATED STATISTICAL METHOD

Wan Muhamad Amir W Ahmad[1*], Nor Azlida Aleng[2], Nurfadhlina Abdul Halim[3], Nor Farid Mohd Noor[4], Mohamad Shafiq Mohd Ibrahim[5], Nur Fatiha Ghazalli[1], Mohamad Nasarudin Adnan[1], and Farah Muna Mohamad Ghazali[1]

[1]School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM)
16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
[2]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia
[3]Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM)
Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
[4]Faculty of Medicine, Universiti Sultan Zainal Abidin (UniSZA) Medical Campus
Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia
[5]Kuliyyah of Dentistry, International Islamic University Malaysia, IIUM Kuantan Campus
Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia
*Corresponding author: Wan Muhamad Amir W Ahmad. Email: wmamir@usm.my

ABSTRACT

Background: Cholesterol levels in the blood, comprising both LDL and HDL cholesterol, can lead to artery plaque formation and potential blockages. Researchers are studying cholesterol levels in individuals with dyslipidemia and type 2 diabetes mellitus. **Objective:** The objective of this paper is to utilize the developed methodology to model the factor associated with the total cholesterol status in patients with dyslipidemia and type 2 diabetes mellitus. This undertaking has the potential to improve the prediction of total cholesterol levels among the analyzed patients by integrating comprehensive supplementary data from a statistical standpoint. **Material and Methods:** The data was collected from Hospital Universiti Sains Malaysia (Hospital USM), using statistical modelling techniques to evaluate data descriptions of numerous variables, including height, total cholesterol, triglyceride levels, low-density lipoprotein (LDL) levels, high-density lipoprotein (HDL) levels, and alkaline phosphatase (ALP) levels. The developed method was implemented and evaluated using the R-Studio, employing a neural network model with bootstrapping method and response surface methodology. **Results:** Our proposed method showed superior accuracy when dividing data into training and testing sets, offering a more precise prediction. The neural network's mean square error was approximately 0.021, demonstrating high precision. **Conclusion:** In this study, the proposed model demonstrates the method's capability for the improvement of research methodology. The outcome suggests that the methodology established for this investigation is capable of producing favourable results. The study's final analysis demonstrates that the model technique created for research is preferable.

Keywords: *Contour plot, multilayer feed-forward neural network (MLFNN), response surface methodology (RSM), surface plot*

Introduction

In 2008, the World Health Organization (WHO) identified four non-communicable diseases (NCD) as the leading causes of mortality worldwide. The aforementioned illnesses comprise chronic respiratory disease, cancer, cardiovascular disease (CVD), and diabetes, as indicated by the source [5]. According to the World Health Organization (WHO), it is projected that within the next decade, there will be a 17% increase in mortality rates attributed to non-communicable diseases (NCDs). The regions with the highest estimated rates are Africa, with approximately 27%, and the Eastern Mediterranean, with approximately 25%. In addition, despite having a normal body mass index (BMI) according to global criteria, individuals from South Asian backgrounds have a greater likelihood of exhibiting risk factors for cardiovascular disease, developing type-2 diabetes, and experiencing an earlier onset of CVD [7,12].

Fortunately, through appropriate interventions aimed at reducing the influence of risk factors, it is possible to prevent over 80% of cases of heart disease, stroke, and type 2 diabetes, as well as almost one-third of cancer cases [2]. Worldwide, CVD is the most prevalent health issue. According to international reports, almost 48% of men and 29% of women are predicted to die from CVD in developed nations between 1990 to 2020 [1]. Gerstein et al., (2008) research indicate that there is a link between cholesterol and triglyceride levels and cardiovascular disease in individuals with type 2 diabetes mellitus [6].

Research has demonstrated that Asians have a greater prevalence of lipid abnormalities compared to non-Asians [8,9]. A combination of low levels of HDL cholesterol and high concentrations of triglycerides has been linked to an increased risk of cardiovascular disease and is referred to as atherogenic dyslipidemia [3,11]. Several factors such as insulin deficiency or resistance, adipocytokines, and hyperglycemia may contribute to the changes in lipid metabolism seen in diabetic patients [10, 14], although the exact mechanism underlying the development of hypertriglyceridemia is reasonably well understood. Nonetheless, many aspects of the pathophysiology and consequences of diabetes-related dyslipidemia are not yet fully understood [15].

Dyslipidemia is a common characteristic of diabetes and is marked by an increase or decrease in the concentration of lipoproteins in the blood and may even serve as a precursor for cardiovascular disease [2,4]. High levels of triglycerides and low levels of high-density lipoprotein (HDL) cholesterol are key features of both dyslipidemia and type 2 diabetes mellitus, and they can be present many years before the onset of clinically significant hyperglycemia [1,10]. Patients with type 2 diabetes have a heightened cardiovascular risk before the onset of biochemical hyperglycemia. During this time, metabolic syndrome, which includes obesity, insulin resistance, hypertension,

and dyslipidemia, is often observed [13]. Recent research suggests that low HDL cholesterol is an independent risk factor for the development of diabetes as well as cardiovascular disease [14].

Materials and Methods

Data Collection

The present study analyzed information from patients treated at the outpatient clinic of Hospital Universiti Sains Malaysia (USM). A total of 122 patients were included in this investigation, and Table 1 offers an overview of the variables chosen for the study, along with information about the data collected.

Table 1: Data description of the selected blood profile

Variable	Code	Description
Total Cholesterol	Y	Reading of Total Cholesterol
Triglycerides	X1	Triglycerides reading
LDL	X2	Low-Density lipoprotein (LDL)
HDL	X3	High-Density Lipoprotein (HDL)
ALP	X4	Alkaline Phosphatase

Study Design and The Methodology

This study is a methodology based which is on computational statistical modeling. The methodology begins with the response surface methodology (RSM) to determine the best model which can represent the case, either the first-order model or the second-order model. After the model recognition process, the next procedure is to apply the multilayer feed-forward neural network toward the selected model either the first-order model or the second-order model. This investigation employed a multilayer feed-forward neural network (MLFNN) that was trained and tested on a specific dataset, with MSE used for prediction. The Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/16050184) approved the study, and patient privacy and medical information were safeguarded.

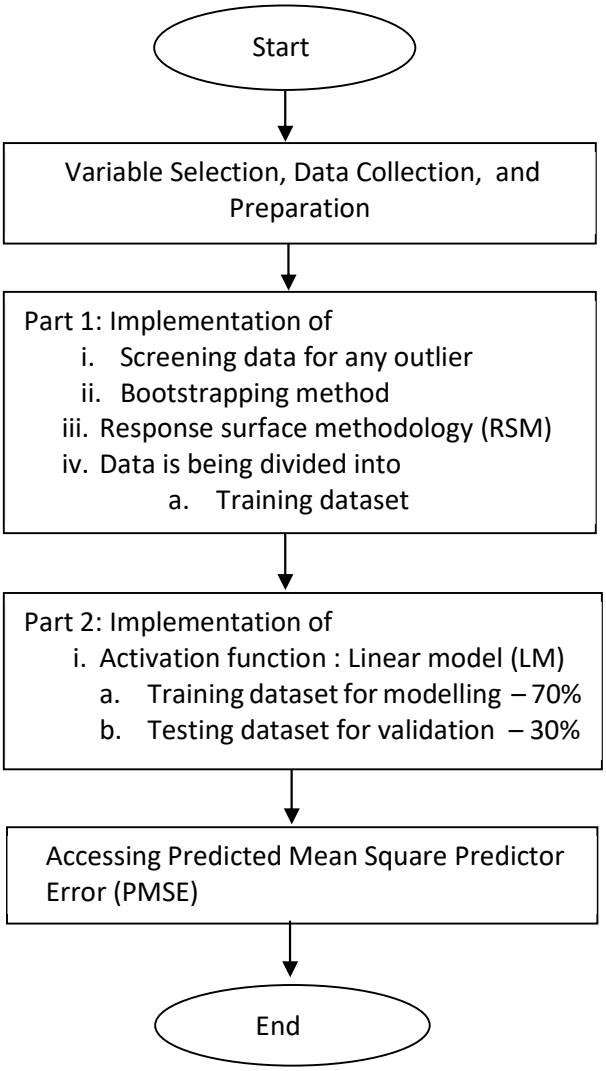

Figure 1: Flowchart of the proposed statistical method

Figure 1 provides a visual representation of this procedure. The diagram highlights the study's notable features, which include an optimal selection model for modeling, model validation, and factor characteristics demonstrated through the contour and surface plots. The syntax created for the investigation is presented in the appendix.

Results

The purpose of this study is to evaluate the effectiveness of a newly created technique that employs response surface methodology and a multilayer feed-forward neural network, which is based on the linear model activation function. This section presents the results of the analysis, including the findings regarding the effectiveness of the newly developed procedure.

Response Surface Methodology (RSM)

The characteristic of total cholesterol can be seen clearly versus all the studied variables (refer to Figure 2 and Figure 3.

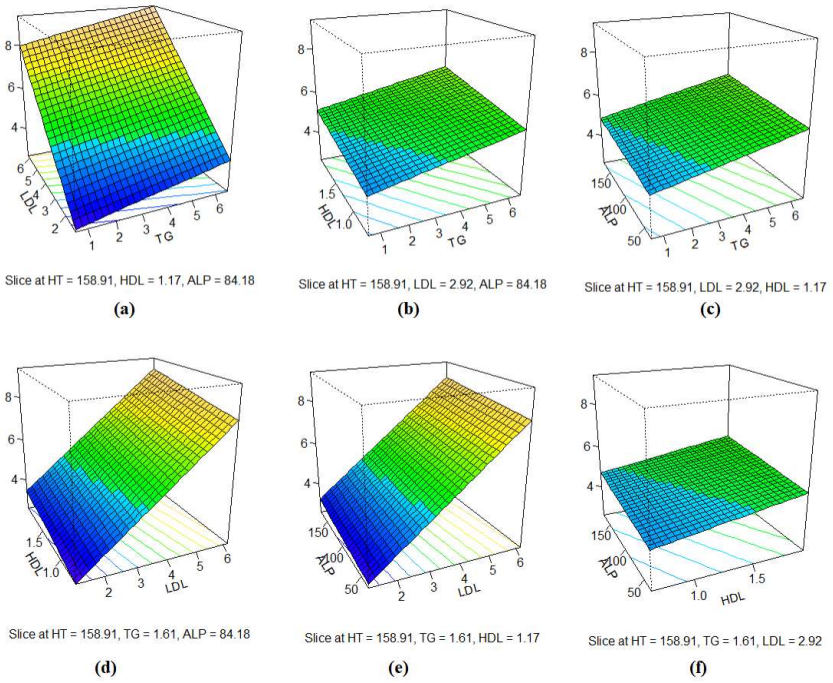

Figure 2: The surface plot for the total cholesterol with the selected variables

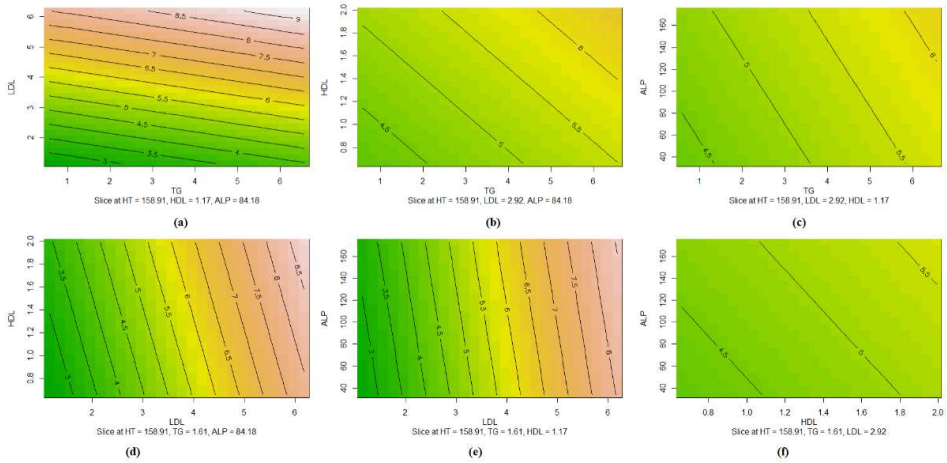

Figure 3: The contour plot for the total cholesterol with the selected variables

(a) *The plot of the total cholesterol versus HDL and ALP.*

The graph illustrates the relationship between total cholesterol, HDL, and ALP.

The contour and surface plots in Figures 2(a) and 3(a) indicate that the highest total cholesterol value is obtained at high HDL and ALP levels. This region is also located in the upper right corner of the plot.

(b) *The plot of the total cholesterol versus LDL and ALP.*

The graph illustrates the relationship between total cholesterol, LDL, and ALP.

The contour and surface plots in Figures 2(b) and 3(b) indicate that the highest total cholesterol value is obtained at high LDL and ALP levels. This region is also located in the upper right corner of the plot.

(c) *The plot of the total cholesterol versus HDL and LDL.*
The graph depicts the connection between total cholesterol, HDL, and LDL.
The contour and surface plots in Figures 2(c) and 3(c) show that the highest value of response total cholesterol is obtained at high levels of HDL and LDL. This area is also located in the plot's upper right corner.

(d) *The plot of the total cholesterol versus triglycerides and ALP.*
The graph shows the relationship between total cholesterol, triglycerides, and ALP.
The contour and surface plots in Figures 2(d) and 3(d) show that at high levels of triglycerides and alkaline phosphatase, the highest value of response total cholesterol is obtained. This area can also be found in the plot's right upper corner.

(e) *The plot of the total cholesterol versus triglycerides and HDL.*
Figures 2(e) and 3(e) exhibit contour and surface plots, respectively, indicating that the maximum value of the response variable, total cholesterol, is achieved at elevated levels of triglycerides and high-density lipoprotein. The aforementioned region is also situated in the upper right-hand corner of the plot.

(f) *The plot of the total cholesterol versus triglycerides and LDL.*
Figure 2(f) and Figure 3(f) show the contour and surface plot indicating that the highest value of response total cholesterol is obtained when the reading of triglycerides and LDL is high. This area appears at the right upper corner of the plot.

Multilayer Feed-Forward Neural Network (MLFNN)
Below is the result of the MLFNN obtained from the analysis.

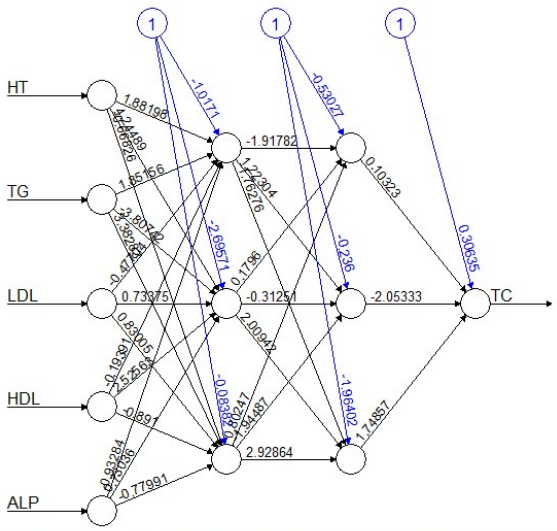

Figure 4: The architecture of the best (MLFFNN) model with five input variables, one hidden layer, and one output node

Figure 4 presents the outcomes of the MLFNN, with total cholesterol serving as the dependent variable of interest. The predicted neural network has a mean square error (MSE) of 0.01489982. The extremely small predicted value indicates that the developed model is highly accurate. This speaks to the reliability and precision of our projected data. The MLFNN utilized a 70:30 split, with 70% of the data used for modelling and 30% for testing.

Evaluation of the Developed Model

In this instance, the forecast value will be used to evaluate the model. The degree of precision of this prognostication shall be ascertained through a comparative analysis between the actual result and the projected outcome. The training dataset is used to create a model, which will then be tested on the testing dataset. Using the linear activation and the selected variable, a sample of the predicted and actual values was given in Table 2. The distance prediction will be used to find the gap between the observed and forecasted values. Using IBM SPSS and the paired sample T-test, the value predicted by the developed method will be evaluated. This is done to ensure that the predicted value and actual value are identical. Table 2 presents the actual and predicted values obtained through the proposed methodology. Based on Table 2, there is a minimal discrepancy between the actual and predicted values. A paired sample t-test revealed no significant statistical differences between the two sets of values. Table 2 provides an overview of the observed and forecasted values generated by the suggested model.

Table 2: The Actual and predicted value for using the developed methodology

Actual	Predicted	Actual	Predicted

4.09	4.10537	4.60	4.67395
6.10	6.13809	4.03	4.13169
3.90	3.91070	5.29	5.38395
3.60	3.55051	6.10	6.21320
:	:	:	:
4.54	4.63173	3.83	3.93079
5.65	5.72515	6.22	5.96353

The forecasted sample was subjected to a paired sample T-Test, with the outcome reported in Table 3.

Table 3: Summary of mean differences for the "Actual" and "Predicted" values.

Mean (SD)				
Variables	Actual	Predicted	T-Statistics (d.f)	*p*-value
Data	4.7437(1.179)	4.7505(1.103)	-0.186 (121)	0.852

Paired sample T- Test was applied

The proposed model's findings indicate that the null hypothesis should not be rejected, as the p-value exceeds 0.05. This suggests that there is no significant difference between the mean values of the actual (Mean (SD) = 4.7437(1.179)) and predicted (Mean (SD) = 4.7505(1.103)) datasets. This demonstrates the superior quality of the model.

Discussion

The proposed approach was implemented effectively, providing significant benefits for identifying models through statistical inferences. The combined method, which aimed to achieve harmony between two different processes, was successful in producing an accurate and reliable model as well as for statistical inferences. The advantage of using the suggested approach is that it has the potential to increase the amount of information obtained from data analysis; this, in turn, will be advantageous for decision-making as well as other improvement-related purposes. The study found that triglyceride levels, low-density lipoprotein, high-density lipoprotein, and alkaline phosphatase are the most significant factors affecting total cholesterol readings. This paper is centered around creating methodologies for MLFNN utilizing response surface methodology in combination with testing and validation procedures.

Our second focus is on the clinical experts' perceptions, which were considered during the variable selection phase of the study. After selecting the variable, the bootstrap procedure was applied to generate a "mega" file from the seed data this is to produce a large number of data and lead to the high accuracy of parameter estimates. After that, the dataset is subjected to the bootstrap. The response surface methodology was then used to model the data that had been generated. The method is to identify the most effective model to use in the case. The R syntax algorithm employed

in this study enables the application to incorporate the methodology concept. In the following procedure, the data will be split into training and testing sets for use in the MLFFNN process. 70% of the bootstrap data will be allocated to the training dataset, with the remaining 30% designated as the testing dataset. Data from the training set will be utilized to construct and evaluate the model. The ideal model will possess the smallest mean square error for the neural network. Syntax was utilized to compute the following formula based on actual and predicted values. The R syntax algorithm establishes a connection between the proposed methodology by its single calculation. The findings of the study assisted the decision-maker in coming to the best possible conclusion. Incorporating statistical formulations, computation using R syntax, and the RSM and MLFNN package resulted in highly successful modelling. The most challenging tasks are developing the R-syntax, which includes the bootstrap, RSM, and MLFFNN, as well as choosing appropriate input parameters for testing the developed method and standardizing it.

Conclusions

The present investigation involved the development of a comprehensive methodology through the integration of bootstrapping, response surface methodology (RSM), and multilayer feedforward neural network (MLFFNN) techniques. The R syntax of the methodology was devised to facilitate a comprehensive understanding of the illustration by the researcher. The study's dependent variable is the aggregate cholesterol level, while its independent variables comprise height, triglyceride levels, low-density lipoprotein (LDL) levels, high-density lipoprotein (HDL) levels, and alkaline phosphatase (ALP) levels. The analysis showed that the primary factors were the most critical factors in the first-order model. The accuracy of a model is believed to be directly proportional to its R-squared value, based on the principles of regression theory. Moreover, the result acquired through the employment of the MLFFNN approach can be viewed as a reflection of a high-quality model, given its ability to produce the lowest error value. The conclusion of the study says that a proposed combined method is better than other single methods.

Acknowledgment: The authors express gratitude to the Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme, which was awarded under Project Code: FRGS/1/2022/STG06/USM/02/10, as well as to Universiti Sains Malaysia (USM).

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

References

[1] Abbasi, A., Corpeleijn, E., Gansevoort, RT. (2013). Role of HDL cholesterol and estimates of HDL particle composition in the future development of type 2 diabetes in the general population: the PREVEND study. *J Clin Endocrinol Metab*.

[2] Alwan, A. (2008). Action plan for the global strategy for the prevention and control of noncommunicable diseases. *Report World Health Organization*.

[3] Amarenco, P., Labreuche, J., Touboul, PJ. (2008). High-density lipoprotein-cholesterol and risk of stroke and carotid atherosclerosis: A systematic review. *Atherosclerosis*.

[4] Chapman, MJ., Ginsberg, HN., Amarenco, P. (2011). Triglyceride-rich lipoproteins and high-density lipoprotein cholesterol in patients at high risk of cardiovascular disease: evidence and guidance for management. *Eur Heart J*.

[5] Fakhrzadeh, H., Tabatabaei, Ozra., (2012). Dyslipidemia and Cardiovascular Disease. *Tehran University of Medical Sciences*.

[6] Gerstein, HC., Miller, ME. (2008). Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med*.

[7] Joshi, P., Islam, S., Pais, P., Reddy, S., Dorairaj, P., Kazmi, K. (2007). Risk factors for early myocardial infarction in South Asians compared with individuals in other countries. *JAMA*.

[8] Karthikeyan, G., Teo, KK., Islam, S., McQueen, MJ., Pais, P., Wang, X. (2009). Lipid profile, plasma apolipoproteins, and risk of a first myocardial infarction among Asians: An analysis from the INTERHEART Study. *J Am Coll Cardiol*.

[9] Labreuche, J., Touboul, PJ., Amarenco, P. (2009). Plasma triglyceride levels and risk of stroke and carotid atherosclerosis: A systematic review of the epidemiological studies. *Atherosclerosis*.

[10] Livingstone, SJ., Looker, HC., Hothersall, EJ. (2012).Risk of cardiovascular disease and total mortalityin adults with type 1 diabetes: Scottish registrylinkage study. *PLoS Med*.

[11] McBride, PE. (2007). Triglycerides and risk for coronary heart disease. *J Am Med Assoc*.

[12] Mohan, V., Sandeep, S., Deepa, R., Shah, B., Varghese, C. (2007). Epidemiology of type 2 diabetes: Indian scenario. *Indian J Med Res*.

[13] Schofield, J.D., Liu, Y., Rao-Balakrishna, P. (2016). Diabetes Dyslipidemia. *Springer Link*, 203–219.

[14] Taskinen, MR. (2003). Diabetic dyslipidemia: from basic research to clinical practice. *Diabetologia*.

[15] Verges, B. (2015). Pathophysiology of diabetic dyslipidemia: where are we? *Diabetologia*. 58(9).

Appendix

The syntax in R

#/Dataset for Biometry: Statistical Modeling Study #

```
Input =("
HT TC TG LDL HDL ALP
157 4.09 1.11 2.59 1.00 54
157 4.09 1.11 2.59 1.00 54
```

```
158 4.74 2.74 2.70 .79 84
160 4.64 .70 2.84 1.45 89
:          :          :
175 4.46 1.36 2.84 1.00 111
175 1.96 1.04 1.96 1.51 40
176 4.91 4.26 2.12 .85 86
")
data = read.table(textConnection(Input),header=TRUE)

#/Performing the bootstrap for 1000: case resampling procedure /#
mydata <- rbind.data.frame(data, stringsAsFactors = FALSE)
        iboot <- sample(1:nrow(mydata),size=1000, replace = TRUE)
                bootdata <- mydata[iboot,]

#/Performing Response Surface Methodology /#
        if(!require(rsm)){install.packages("rsm")}
         library(rsm)
              first <- rsm(TC~FO(HT,TG,LDL,HDL, ALP), data=data)
summary(first)

par(mfrow = c(2,3))              # 2 × 3 pictures on one plot
contour(
 first,          # Our model
 ~ TG+LDL+HDL+ALP,       # A formula to obtain the 6 possible graphs
 image = TRUE,           # If image = TRUE, apply color to each contour
)


par(mfrow = c(2,3))             # 2 x 3 pictures on one plot
persp(
 first,          # Our model
 ~ TG+LDL+HDL+ALP,       # A formula to obtain the 6 possible graphs
 col = topo.colors(100),       # Color palette
 contours = "colors"           # Include contours with the same color palette
)

#/Install the Neuralnet Package/#
        if(!require(neuralnet)){install.packages("neuralnet")}
                library("neuralnet")
#/Checking for the Missing Values/#
```

```
        apply(bootdata, 2, function(x) sum(is.na(x)))


#/Scaling the Data for Normalization/#
#/Method (Usually Called Feature Scaling) to Get All the Scaled Data/#
#/In the Range [0,1]/#
  max_data <- apply(bootdata, 2, max)
min_data <- apply(bootdata, 2, min)
        data_scaled <- scale(bootdata, center = min_data, scale = max_data - min_data)

#/Randomly Split the Data into 70:30/#
#/70 Percent of the Data at Our Disposal to Train the Network/#
#/30 Percent to Test the Network/#
  index = sample(1:nrow(bootdata),round(0.70*nrow(bootdata)))
     train_data <- as.data.frame(data_scaled[index,])
       test_data <- as.data.frame(data_scaled[-index,])

#/Building the Neural Network/#
#/There are 3 Hidden Layers Have 3 and 2 Neurons Respectfully/#
#Input  = 10/# and Output  = 1/#

n = names(bootdata)
 f = as.formula(paste("TC ~", paste(n[!n %in% "TC"], collapse = " + ")))
   nn = neuralnet(f,data=train_data,hidden=c(3,3),linear.output=T)
     plot(nn)
         options(warn=-1)

#/30 Percent of the Available Data to do the prediction of the data
        predicted <- compute(nn,test_data[,1:5])

#/Use the Mean Squared Error NN (MSE forecasts the network) as a Measure of How Far
#Away Our Predictions Are From The Real Data/
        MSE.net <- sum((test_data$TC-predicted$net.result)^2)/nrow(test_data)
                MSE.net

colnames(modelEval) <- c('Actual','Predicted')
        modelEval <- as.data.frame(modelEval)
          print (modelEval)

#/ finished/#
```