JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

Combination of Methodology Building for Multi-Layer FFED Forward Neural Network (MLFF) and Linear Modelling (LM): A Case Study by Biometry Modelling

Muhammad KhairanShazuanJusoff[a], Wan Muhamad Amir W Ahmad[a*], Nor Farid Mohd Noor[b], Nor AzlidaAleng[c], Nur FatihaGhazalli[a], Mohamad Shafiq Mohd Ibrahim[d], Farah Muna Mohamad Ghazali[a],Mohamad Nasarudin Adnan[a], Nurfadhlina Abdul Halim [e]

[a]School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM),
16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
[b]Faculty of Medicine, Universiti Sultan Zainab Abidin (UniSZA), Medical Campus, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia.
[c]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia.
[d]Kulliyyah of Dentistry, International Islamic University Malaysia, IIUM Kuantan Campus, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia.
[e]Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800, Nilai, Negeri Sembilan, Malaysia.
*__Correspondence to:__ Wan Muhamad Amir W Ahmad, School of Dental Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10.

Background: The purpose of this study is to develop and illustrate an optimum variable selection approach using Multiple Linear Regression (MLR) and validation using Multi-Layer Feed Forward Neural Network (MLFF), also known as Multilayer Perceptron Neural Network (MLP) while taking bootstrapping into account. All of the factors specified will be evaluated to determine if they have a meaningful association.
Objective: The goal of this study was to create a new method in making predictions for oral health data by combining a few statistical techniques. This would make the model more accurate.
Material and Methods:A set of medical data that consists 30 observations was used to develop the methodology. The data descriptions of variables being used in this retrospective research including Creatinine, Fasting Blood Sugar, Haemoglobin A1C, and Urea were evaluated using advanced computational statistical modelling approaches. The medical data are used to test the R syntax that is developed in this study. The statistics for each sample are calculated using a model that incorporates bootstrapping and multiple linear regression techniques.
Results: The statistical strategy which combines Bootstrap, MLFF and MLR is better than regular statistical method being used for this type of data. The hybrid model technique's accuracy was increased when the data was separated into training and testing dataset. Four variables are taken into account in this case: fasting blood sugar, creatinine, haemoglobin A1C, and urea. Fasting Blood Sugar (β_1: 0.461889; $p< 0.05$), Creatinine (β_3: 0.029761; $p< 0.05$) and Urea (β_2:-0.454766; $p< 0.05$)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452
are all significant, whereas MLFF and MLRhave the predicted mean square error (PMSE) values of 0.01226 and 0.3531 respectively.

Conclusion: The value of PMSE is mainly used to diagnose the performance of MLR and MLFF. The MLFF is used to determine how close predicted values are to real data, while the model that has low PMSE value indicates that the model is accurate. The R syntax for combining Bootstrap, MLR, and MLFF is also included in this research article. In a word, this research establishes the superiority of the hybrid model method.

Keywords: Creatinine, Fasting Blood Sugar, Haemoglobin A1C, Urea, Multiple linear regression, R square, Predicted Mean Square Error

INTRODUCTION

Multi-Layer Feed Forward Neural Network (MLFF) has been broadly used in medical science. MLFF is a type of mathematical or computational model that imitates the way the human brain learns by using rules deduced from data patterns to build hidden layers of logic for analysis [1]. In MLFF, there exists an interconnection of perception where the flow of information and calculations is unidirectional, which is from the input to the output data. MLFF aims to evaluate independent variables towards dependent variables from a linear regression point of view. A good methodology plays a very important role in data analysis because it leads to good and precise decision-making. Thus, the fundamental issue of preciseness and accuracy of calculation must be highlighted to harmonized between theory and applied programming based.

Multiple linear regression (MLR) is a well-known technique that can be found in a wide range of sectors, as well as health-related ones. It is typically used in measuring the relationship between two or more variables and a single continuous output variable. MLR fits a straight line to the data points using regression coefficient (b_1) as the slope, and the *y*-axis intercept (b_0). This relationship between dependent (Y) variable and independent (X) variable is best described as the equation $Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$ [2]. A previous study in 2012 deployed 40 simulation data, each having one dependent factor (Y) and four explanatory factors (X_1, X_2, X_3 and X_4). In this previous study, the analysis revealed that R values for variables X_1 to X_4 were 0.933, 0.911, 0.725, and 0.148 respectively. X_1, X_2, and X_3 were statistically significant variables (p<0.001) but X_4 was not (p = 0.36)[3].

The bootstrap technique uses random sampling to produce random samples of the same size as the original sample, each of which yields their own value of parameter of interest, such as mean. The phrase "with replacement" means that any observation can be picked as the sample several times in a bootstrap sample. It is important because in the absence of replacement, resampling would only produce a random rotation of the original data, with many statistics like the mean staying unaffected [4]. To get a better idea of how much the estimator varies, it is necessary to run the calculation more than once [5].

Haemoglobin A1C (HBA1C) is used as an indicator to determine long-term glycaemic control. Due to its ability as an indicator for the mean blood glucose level, HBA1C predicts the risk of diabetic discomfort improvement in patients with diabetes. Purohit and colleagues studied the connection between the level of HBA1C and serum lipid profile to examine the importance of HBA1C in terms of

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

dyslipidaemia marker and potential risk of cardiovascular disease. This study result shows a significant relationship between HBA1C-triglyceride (TG), HBA1C-low density lipoprotein cholesterol (LDL-C), HBA1C-total cholesterol (TC), HBA1C-Tc/HDL proportion, and HBA1C-non HDL cholesterol. Therefore, using HBA1C to determine whether a person has dyslipidaemia was found to be accurate, according to the study [6].

The urea cycle is among the most important processes in the human body. Cellular proteins are constantly being degraded and synthesised in all living things. Nearly seventy-five percent of freed amino acids are used again. Extra amino acids are quickly degraded due to the fact that they are not stored. Amino acid nitrogen waste is used to make urea. The kidneys excrete urea as a water-soluble particle. Uraemia is a condition marked by elevated blood urea levels and is associated with a number of metabolic illnesses [7-8]. Laranjinha in the year 2016 conducted a study to analyse the relationship between urea and dyslipidaemia-peripheral neuropathy [9]. Patients with elevated urea had abnormal ratings of peripheral diabetic neuropathy, according to this study. In addition, there was also a study done on adults with type 2 diabetes to see how their lipid profiles correlated with their renal functions. Creatinine was one of the predictor variables for HBA1C, the dependent variable in this research. In this previous study, it was found that a strong correlation between HBA1C and creatinine exist with the significance value of $p=0.017$. This previous study also concluded that creatinine levels can be used to guide the treatment of patients with type 2 diabetes [9].

MATERIALS AND METHODS

This section contains the specifics of the materials and methods that were used in the case study.

Data Collection

A trial with 30 observers was simulated for this study. There are data descriptions in Table 1 for this research.

Table 1. Data Description of research variables

Variables	Code	Descriptions
HBA1C	Y	Hemoglobin A1C
FBS	X_1	Fasting blood sugar
UREA	X_2	Urea Reading
CREATININE	X_3	Creatinine Reading

Study Design, Bootstrap and Modelling of Computational Biometry

This study builds a multilayer perceptron (MLP) with multiple linear regression (MLR) using a retrospective approach and advanced computational statistical modelling methodologies. This developed methodology was based on the testing and training dataset, MSE-predicted and model evaluation. The study was approved by the Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/21080565). A random sample of the population is chosen, and then the bootstrap procedure begins, which entails calculating the sample statistics.

An artificial population is created using a large number of substitution samples, which are then replicated numerous times following several replications of the original sample. Random sampling results in samples that are different from the original sample in the case of replacement sampling. When a sample is

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

drawn with replacement, the bootstrap calculates statistics for each sample and is used to draw samples with replacement [6-7]. The result for the model is displayed in Table 2. The MLR is fitted through the R Software. The complete step by step method is given Figure 1.1. R-Studio software was used to analyse the data, which included MLP and linear regression syntax. There are three models that are used in the advanced technique: the bootstrap, MLP and MLR. During the modelling process, the training data will be used, and the testing data will be used for testing purposes. This section uses a series of linear regression models to evaluate the connection between haemoglobin A1C and the specified explanatory variables.

$$Haemoglobin\ A1C = \beta_0 + \beta_1(FBS) + \beta_2(UREA) + \beta_3(C$$

where $\beta_0 ... \beta_3$ are parameters, and Y is the value of Haemoglobin A1C

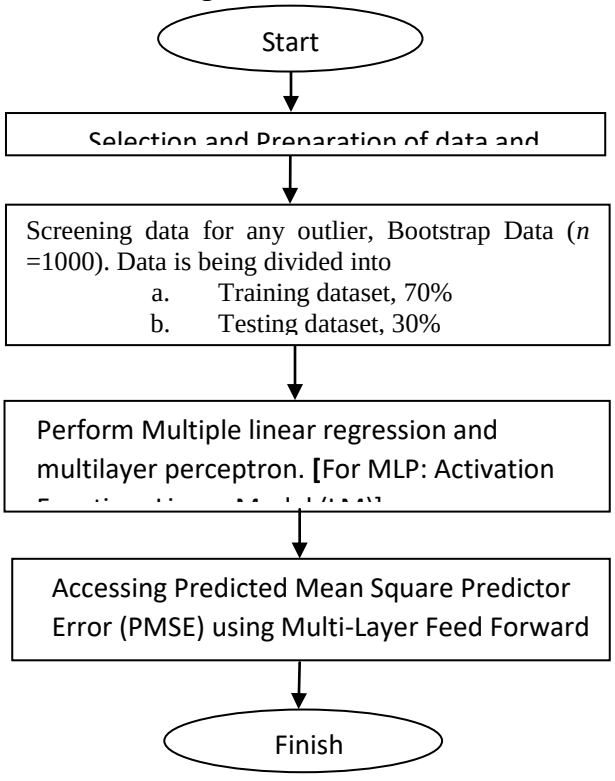


Figure 1. Flowchart of the Propose Analysis

Figure 1 shows a diagram to illustrate this process. One of the study's strengths is that it investigates a model that takes clinically relevant variables into account. Following the data preparation, a bootstrapping technique will be developed. The bootstrap method generates a sample the same size as the original sample, but each observation is repeated multiple times and others are discarded [5-6].

1.3. The syntax in R

```
Input =("
FBS   UREA  CREATININE  HBA1C
     7     6      97     8
     6     5     129     8
     8     5      83     7
    12     6     124    11
     5     6     111     7
     6     4     113     7
     7     5      99     7
     6     7      87     6
    10     8     125     8
     8     8     123     6
     5     5      94     6
     8     6     101     8
     5     8     149     9
     5     4      98     6
    11     5     107     9
    12     5     106    10
    12     3      63    10
    10     6      64     7
    17     5      85    12
     6     6      94     6
     7     6     121     9
     8     3      77     8
     5     4     106     8
     8     5     133     8
     6     5      66     6
     4     4      91     6
     9    10     168     7
    10     7     152     9
    15     9     150    11
     8     4      91     9

")
data =
read.table(textConnection(Input),head
er=TRUE)
```

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

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

#/install the neuralnet package/
if(!require(neuralnet)){install.packa
ges("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*nr
ow(bootdata)))
train_data                             <-
as.data.frame(data_scaled[index,])
test_data                              <-
as.data.frame(data_scaled[-index,])

#////////////////////////////////////
////////////////////////////////////
/
# Print Data
#print(train_data)
#print(test_data )

#/Build the network
#KSe  3  hidden  layers  have  3  and  2
neurons respectfully
#Input layer = 2
#Output layer = 1/

n = names(bootdata)
f    =    as.formula(paste("HBA1C    ~",
paste(n[!n  %in%  "HBA1C"],  collapse  =
" + ")))
nn                                     =
neuralnet(f,data=train_data,hidden=c(
2),linear.output=T)
plot(nn)
options(warn=-1)

#/30 percent of the available data to
do this:
#using   only   the   first   2   columns
representing the input variables
#of  the  network  and  1  is  the  output
for NN/
predicted                              <-
compute(nn,test_data[,1:4])

#/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$HBA1C   -
predicted$net.result)^2)/nrow(test_da
ta)
MSE.net

#////////////////////////////////////
////////////////////////////////////
/

Model                                  <-
lm(HBA1C~FBS+UREA+CREATININE,
data=bootdata)  # build the model
summary(Model)

data$PredictedHBA1C <- predict(Model,
data)
distPred <- predict(Model, data)
preds <- predict(Model, data)
modelEval <- cbind(data$HBA1C, preds)
colnames(modelEval)                    <-
c('Actual','Predicted')
modelEval <- as.data.frame(modelEval)
print (modelEval)
test <- data[-index,]
predict_lm <- predict(Model,test)
```

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

```
MSE.lm <- sum((predict_lm - test$HBA1C)^2)/nrow(test)
MSE.lm

#/Printing the Value of MSE for Linear Model and Neural Network/
print(paste(MSE.lm,MSE.net))

# Finished
```

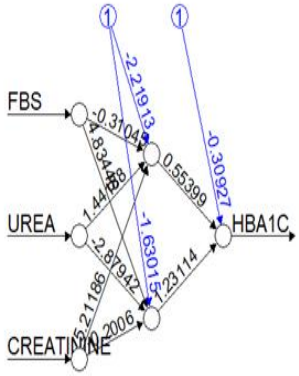
RESULTS

The purpose of this research is to look into the performance of a Multilayer Perceptron (MLP) that is based on an activation function: linear model. This MLP consider both the training and testing datasets.

The Regression Modeling and Multilayer Perceptron Results

Table 2 shows the results of Multi-Layer Feed Forward Neural Network (MLFF), where Haemoglobin A1C act as dependent variable and the study's output. The goodness of the model is being measure through mean square error (MSE) and through the model evaluation. Through the model evaluation the actual data will be compared to the predicted data. Remarkably similar data will have a small spread. In contrast, wildly dissimilar data will have a large distance between them. In this instance, it speaks to the accuracy and reliability of our forecast data. The training to testing data split is in the ratio of 70:30, meaning that 70% of the data is available for modeling while another 30% for testing. Table 2 summarises the results of the MLR.

Below is the model that obtained through the developed syntax in R.




```
> #/Printing the Value of MSE for Linear Model and Neural Network/
> print(paste(MSE.lm,MSE.net))
[1] "0.353132759574172 0.0122671713894197"
```

Figure 2. The architecture of the best (M.L.P.) model with three input variables, one hidden layer, and one output node

Variable	Estimate	Std. Error	t-Value	Pr(>|t|)
(Intercept)	3.54305	0.122114	29.01	<2e-16
FBS	0.46189	0.008767	52.69	<2e-16
Urea	-0.45477	0.021073	-21.58	<2e-16
Creatinine	0.02976	0.001337	22.25	<2e-16

Table 2. Result of Multiple Linear Regression with combining the bootstrap method training and testing dataset

Multiple Linear Regression was applied
Significant at the level of the 0.05

The model evaluation can be derived from the forecast value in this case. The prediction's accuracy will be determined by a comparison of the observed and predicted values. The testing dataset will be used to assess the model constructed using the training data set. When

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

comparing actual and predicted data, the difference between them will be determined using distance prediction. The R syntax includes a method for evaluating models that can be used to evaluate subsequent methods. The "Actual" and "Predicted" values obtained using the proposed methodology are shown in Table 3.

Table 3. The "Actual and "Predicted" value of HBA1C from Multi-Layer Feed Forward Neural Network

Actual	Predicted
8	6.9345
8	7.8798
7	7.4345
11	10.0475
7	6.4274
7	7.8583
7	7.4488
6	5.7202
8	8.2440
6	7.2607
8	6.9345

Table 3 displays the “Actual” and “Predicted” values of HBA1C based on the proposed model. According to the findings, “Actual” and “Predicted” were statistically indistinguishable. This demonstrates the proposed model's superiority.

DISCUSSION

The proposed method, which is extremely useful for estimating event probabilities, was successfully applied. We obtained and harmonised a hybrid method that produced an extremely accurate and reliable model. The findings revealed that Urea, Fasting Blood Sugar, and Creatinine are the most critical factors affecting Haemoglobin A1C.

The purpose of this paper was to discuss the development of methodologies for MLR. The primary focus of this paper is the development of methods for MLR. The study's primary goal was to create, test, and validate a regression model. This project's main goal was to combine bootstrapping and MLR techniques for developing and implementing medical statistics strategies. Clinical expert opinion is incorporated during the variable selection process. The bootstrap method uses the initial data set to create a “mega” file at the start of the operation. The bootstrap procedure, on the other hand, generates a large number of file substitutions. It's also important to note that the bootstrap method also develops and saves statistical samples. Finally, the bootstrap technique goes through the same steps repeatedly, often hundreds of thousands of times before it stops. Application and methodology can be seamlessly integrated with the R syntax algorithm. First and foremost, it is necessary to consult an expert when choosing variables. In the following steps, the bootstrap procedure will be performed to the data. At this point, 70% of the bootstrap data will be regarded as training dataset, while 30% will be regarded as testing dataset. We will use data from the training dataset to build and test the model. A successful model will exhibit the lowest mean absolute deviation possible. The following formula was calculated using syntax and actual and predicted values. The study's findings aided the decision-maker in obtaining the best outcome possible. By combining statistical formulations, R syntax, and the multiple linear regression package, we were able to produce highly successful linear modelling. Selecting appropriate input parameters, data cleaning for linear modelling,

Volume 13, No. 1, 2022, p. 302-309
https://publishoa.com
ISSN: 1309-3452

and standardising it are the most challenging tasks.

CONCLUSION

Bootstrap, multiple linear regression, and multi-layer feed forward neural network methods was combined in this research. The syntax for this method was written using R software in a manner where the researcher may fully understand the concept. The dependent variable in this study is hypertension.At the same time, Fasting Blood Sugar, Urea, and Creatinine are the independent variables. According to the model, variables emerged as the most critical variables. The more accurate a model is, the greater its R squared value, according to regression theory. Additionally, the test's fit can be evaluated using the smallest predictive value generated from the obtained model. As a result, a hybrid model can assist us in determining the method's utility and relative contribution to the outcome. According to the combination technique strategy proposed in this study by using R software, it demonstrates that regression modelling outperforms with a Mean Square Error of 0.3531 for predictive modelling and 0.01226 for multilayer perceptron modelling. This study concludes that the hybrid model technique proposed in the study outperforms other conventional methods since the obtained value of PMSE is very low.

CONSENT

Both the patient's privacy and medical condition were protected.

ETHICAL APPROVAL

The Human Research Ethics Committee of Universiti Sains Malaysia reviewed and approved the study.

REFERENCES

1. Fashoto, S. G., Adeyeye, M., Owolabi, O. and Odim, M., 2015. Modelling of The Feed Forward Neural Network with Its Application in Medical Diagnosis. *International Journal of Advances in Engineering & Technology*, *8*(4), p.507.
2. Schober, P., Vetter, T. R. 2021. Linear Regression in Medical Research. Anesthesia and analgesia, 132(1), 108–109.
3. Mukhopadhyay, J., Kanjilal, S., Biswas, M. 2012. Diabetic dyslipidemia-Priorities and targets in India. *Med Update.* 20;155-9.
4. Campbell, M., J. 2006. Statistics at Square Two: Understanding Modern Statistical Applications in Medicine, *2*(3), 109-116.
5. Walters, S., J., Campbell, M., J. 2004. The Use of Bootstrap Methods for Analysing Health-Related Quality of Life Outcomes (Particularly The SF-36). *Health and Quality of Life Outcomes,2*(1), pp.1-19.
6. Purohit, K., Rathore, H. S., & Köhler-Rollefson, I. 2017. Increased risk of type 2 diabetes mellitus in the Maru Raika community of Rajasthan: A cross-sectional study. *International Journal of Diabetes in Developing Countries*, *37*(4), 494-501.
7. Topçiu-Shufta, V. & Haxhibeqiri, V. 2019. Lipid Disorders in Uremia, Cellular Metabolism and Related Disorders. doi: 10.5772/intechopen.90043. https://www.intechopen.com/chapters/70012
8. Rao, M. U., Siddhartan, S., Sowmya, R., Sathivel, A., Zin, T. and Raj, N. B. 2021. Assessment of Malaysian University undergraduate's knowledge and awareness on metabolic syndrome and condition related to it. *Research Journal of Pharmacy and Technology*, *14(*4), pp.1893-1898.

9. Laranjinha, I., Matias, P., Mateus, S., Aguiar, F., Pereira, P., Santos, M., P., … & Campos, L. 2016. Diabetic kidney disease: Is there a non-albuminuric phenotype in type 2 diabetic patients? *Nefrologia (English Edition), 36*(5), 503-509.
10. Rizal A. M., Raj, N. B., Sowmya, R., Siddhartan, S., Zaleha, M. I. 2020. Weight Reducing Interventions for Overweight and Obese Employees. *European Journal of Molecular Clinical Medicine, 7*(11):835-844.
11. Fan, S., Tang, Q. L., Lin, Y. J., Chen, W. L., Li, J. S., Huang, Z. Q., Yang, Z. H., Wang, Y. Y., Zhang, D. M., Wang, H. J., Dias-Ribeiro, E., Cai, Q. & Wang, L. 2011. A review of clinical and histological parameters associated with contralateral nect metastases in oral squamous cell carcinoma. Int J Oral Sci, 3(4), 180-191.
12. Frithiof, L. & Wersall, J. (1967). Virus-like particles in papillomas of the human oral cavity. Archives of Virology, 21(1), 31-44.
13. Selvaraj, S., Naing, N. N., Wan-Arfah, N., Karobari, M. I., Marya, A. & Prasadh, S. 2022. Development and Validation of Oral Health Knowledge, Attitude and Behavior Questionnaire among Indian Adults. *Medicina, 58*(1), p.68.
14. Selvaraj, S., Naing, N. N., Wan-Arfah, N. & Prasadh, S. 2021. Confirmatory Factor Analysis of Knowledge, Attitude and Behavior Questionnaire among Indian Adults. *Journal of personalized medicine, 11*(4), p.320.