Atomic Spectroscopy
Article

A Performance Of Robust Hybrid Methodology: A Forensic Crime Case Study

Mohamad Nasarudin Adnan[a], Wan Muhamad Amir W Ahmad[a]*, Nor Azlida Aleng[b], Nor Farid Mohd Noor[c], Norsamsu Arni Samsudin[a], Mohamad Shafiq Mohd Ibrahim[d], Noor Maizura Mohamad Noor[b], Ruhaya Hasan[a]
[a]School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
[b]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia
[c]Faculty of Medicine, Universiti Sultan Zainal Abidin (UniSZA), Medical Campus, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia
[d]Kuliyyah of Dentistry, International Islamic University Malaysia, IIUM Kuantan Campus, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia

ABSTRACT: Background: This study aims to showcase an effective technique for variable selection using established Multiple Linear Regression (MLR) models. Furthermore, the study aims to validate these selected variables using multilayer feed-forward neural network (MLFFNN) models and enhance the analysis by incorporating contour plots and surface plots as visual tools. Initially, all chosen variables will undergo the bootstrap methodology to assess their significance and screen for relationships. Objective: The primary aim of this study is to create, standardize, and validate a hybrid model that integrates multiple linear regression, multilayer feed-forward neural networks, surface plot methodology, and contour plots. The model will be implemented using the R software, which offers a comprehensive modeling approach and various diagnostic tools. These diagnostic tools will assist researchers in accurately interpreting the results and obtaining optimized outcomes. Material and Methods: Approximately 200 simulated data points were utilized to establish the methodology for this study. Advanced computational statistical modeling techniques were employed to assess the data characteristics of various variables in this retrospective analysis. These variables encompassed aspects such as total victim count, gender, age, marital status, presence of adults and children in the household, burglary and sexual victimization, and victim reporting. The case study was devised and executed using the R-Studio program and corresponding syntax. Results: The statistical analysis revealed that regression modeling outperforms the R-squared and mean-square error tests in most scenarios. The researchers observed that when the data was divided into two sets for training and testing, the hybrid model approach exhibited significantly superior predictive capabilities for the experimental outcome. To determine the validity of the variables, the well-established bootstrap-integrated MLR approach was employed. In this case, seven characteristics were considered: gender (β1: 2.0882e+00; $p < 0.05$), age (β2: -5.8824e-02; $p < 0.05$), marital status (β3: 3.8235e-01; $p < 0.05$), adult in the household (β4 1.6176e+00; $p < 0.05$), burglary victim (β5: 2.5588e+00; $p < 0.05$), the sexual victim (β6: -2.3529e-01; $p < 0.05$), and victim's report (β7: -3.2353e-01; $p < 0.05$). The linear model in this scenario yielded a predicted mean square error (PMSE) of 4.261. While the predicted mean square error for the neural network is 0.19099. Conclusion: The main aim of this research is to develop and thoroughly assess a hybrid approach that combines bootstrapping, multilayer feedforward neural network, multiple linear regression, and contour plot and surface plot techniques. The methodology involves creating and presenting R syntax to ensure researchers have a comprehensive understanding of the approach. The statistical analysis conducted using R software in this study demonstrates that linear regression modeling surpasses other methods in terms of accuracy measures and the Mean Square Error value. As a result, the findings of the study strongly support the superiority of the hybrid model technique, contributing to a deeper understanding of its significant impact on the outcomes in this particular case.

Keywords: Criminal Case, Multiple Linear Regression (MLR), Multilayer Feed-Forward Neural Network (MLFFNN), Contour Plot, and Surface Plot

INTRODUCTION

Regression modeling involves the integration of two phenomena, represented by equations incorporating dependent and independent variables. One such phenomenon is the Applied Linear Regression Model (ALRM) [1,7]. This model illustrates how the dependent variable reacts to variations in the independent variables, while also determining the association between

the dependent variable and the independent variable, irrespective of whether the relationship is linear or nonlinear [5,12]. Multiple linear regression (MLR) is a widely recognized and extensively utilized technique, particularly in the medical field [1,18]. It is employed to model the relationship between multiple independent variables and a continuous dependent variable. MLR is applicable when there are two or more independent variables and a single dependent variable. In this study, MLP (multilayer perceptron) is an extension of simple linear regression that is employed to achieve improved results [6,11]. For instance, Cohen et al. (2003) utilized MLR in a study to identify the factors influencing nurses' attitudes toward patients' families. Due to the violation of the normality assumption, robust standard errors were employed in the regression analysis. The findings indicated no issues, and the variables exhibited variance inflation factor (VIF) values ranging from 1.00 to 1.36 [5].

Neural networks serve as modeling tools for neurophysiology and artificial intelligence [13]. However, they are also utilized as statistical models in the medical field to describe and draw inferences from data, particularly using multilayer perceptron (MLP) models [10,19]. Many epidemiological studies lack sufficient information regarding the statistical properties of covariates, such as distribution normality, colinearity, and the unknown link function between variables [10,11,13,20]. In this regard, MLFFNN offers a potential solution, as three-layer perceptron networks are theoretically capable of universal approximation [10]. Furthermore, some studies argue that MLP models can match or even outperform classical statistical methods in terms of goodness of fit, estimation, and prediction [12,18]. The objective of this study is to analyze and construct an Applied Linear Regression Model (ALRM) and an MLFFNN model to investigate factors associated with the total number of crime cases.

Neural networks have emerged as valuable tools in the field of crime analysis, enabling the modeling and prediction of crime cases. The application of neural network techniques allows for the identification of patterns and relationships within crime data, resulting in improved comprehension and predictive capabilities. One of the key advantages of neural networks in crime analysis is their ability to handle complex and nonlinear relationships between variables. They can effectively capture intricate patterns and dependencies within crime data that may not be easily discernible using traditional statistical methods. These networks learn from historical crime data, incorporating various factors such as time, location, demographics, and other relevant variables, to construct predictive models. Once trained, the neural network can forecast future crime occurrences and identify areas or populations that may be at higher risk. This empowers law enforcement agencies and policymakers to gain valuable insights for enhancing crime prevention strategies, optimizing resource allocation, and implementing targeted interventions. By leveraging neural networks in crime analysis, a proactive approach can be taken toward crime reduction and community safety [15,16,20,21].

In the study conducted by Liu et al. (2019), a novel ensemble learning approach was employed, combining neural networks with linear regression, to predict crime rates. This hybrid model demonstrated improved accuracy compared to standalone models [15]. Similarly, Raza et al. (2019) developed a hybrid model for crime hotspot prediction by integrating neural networks with ridge regression. This approach utilized spatial features to enhance the accuracy of predictions, leading to more accurate identification of crime hotspots [20]. Another study by Gao et al. (2020) proposed a hybrid model that combined neural networks with linear regression and gradient-boosting decision trees for crime prediction. The effectiveness of this hybrid approach was demonstrated, showcasing improved prediction accuracy and hotspot identification in crime analysis tasks. These studies serve as examples of successful integration between neural networks and linear regression techniques in crime analysis. The hybrid models effectively leverage the strengths of both approaches, resulting in enhanced understanding and prediction of crime patterns [9,15,20].

In the study conducted by Mohler et al. (2011), a self-exciting point process model based on neural networks was employed to analyze crime data and predict crime hotspots [16]. Liu et al. (2019) utilized neural networks to incorporate social media data into crime rate predictions. This approach enabled the identification of shifting hotspots and changing crime patterns [15]. Mohler (2011) conducted research utilizing convolutional neural networks (CNNs) to predict spatiotemporal crime incidents. The study demonstrated the efficacy of deep learning methods in crime analysis [16]. Furthermore, Wang et al. (2019) explored the use of deep learning techniques, specifically long short-term memory (LSTM) neural networks, to forecast crime rates in Los Angeles. This study showcased the potential of neural networks in crime prediction tasks. These studies exemplify the application of neural networks in various aspects of crime analysis, including crime hotspot prediction, incorporation of social media data, spatiotemporal crime incident prediction, and crime rate forecasting. The utilization of neural networks in these studies highlights their effectiveness and potential in enhancing crime analysis and prediction capabilities [21].

Surface plots can be utilized in crime analysis to visually represent the density or intensity of crime incidents across a geographic area. This three-dimensional visualization portrays crime patterns, where the x and y axes correspond to spatial coordinates like longitude and latitude, while the z-axis represents the frequency or intensity of crime incidents. By examining the surface plot, areas with high crime rates or hotspots can be identified, facilitating resource allocation and targeted policing strategies. On the other hand, contour plots are effective tools for mapping crime hotspots and revealing spatial patterns of criminal activity [2,9]. These plots consist of contour lines that connect areas with similar crime incident densities or frequencies. By analyzing the contour plot, regions with higher or lower concentrations of crime incidents can be identified. This visualization technique enhances the understanding of crime distribution, enabling law enforcement agencies to prioritize patrol areas, allocate resources, and implement preventive measures. The integration of surface plots and contour plots into crime analysis empowers researchers and law enforcement agencies to acquire valuable insights into the spatial patterns and distribution of criminal activity. By utilizing these visualizations, a deeper comprehension of crime dynamics can be achieved, hotspots can be identified, and effective crime prevention strategies can be developed [15,16,20].

Recent studies have reaffirmed the significance of contour plots and surface plots in the field of forensic sciences. These visualizations play a crucial role in uncovering patterns, trends, and relationships within health-related data, thereby informing forensic practices and interventions. Moreover, contour plots and surface plots serve as effective communication tools, enabling researchers to convey their findings to various stakeholders, including healthcare professionals, policymakers, and patients. As technology advances and more sophisticated data visualization tools become available, contour plots and surface plots are increasingly accessible and versatile in the realm of health sciences. They empower researchers to delve into complex datasets, identify associations, and generate hypotheses for further exploration. As the field of health sciences continues to progress, these visualization techniques will undeniably remain indispensable for data analysis, contributing to advancements in healthcare and promoting evidence-based decision-making [3,14].

MATERIALS AND METHODS

Data Collection

The focus of this research was to analyze simulation data derived from a trial that included a total of 200 participants acting as observers. Detailed information about the research variables can be found in Table 1, which provides a comprehensive description of the data used in the study.

Table 1. Data Description of the research variables

Code	Variables	Descriptions
Y	Total victim	Number of times respondents were victimized in a year
X_1	Gender	Gender 1 = Male 2 = Female
X_2	Age	Age in years
X_3	Marital	Marital status 1 = Married, 2 = Divorced, 3 = Separated 4 = Other
X_4	Adult	Number of adults in the household
X_5	Burglary	Number of times victimized by burglary
X_6	Sexual	A number of times victimized by the sexual offense.
X_7	Report	Number of victimizations reported to the police

Study Design

In this research, a sophisticated computational statistical modeling approach called a multilayer feed-forward neural network with ordinal regression is employed. This methodology effectively addresses complex statistical problems and manages large datasets by integrating statistical theory, algorithms, and computational power. To facilitate the model's development, the data is randomly divided into testing and training datasets. The methodology considers various factors, including the testing and training datasets, MSE-predicted values, and the accuracy of the mean absolute deviance. Computational statistical modeling entails utilizing computational techniques and algorithms to analyze intricate statistical relationships within data, enabling statistical analysis, predictive modeling, and data-driven inferences. To ensure the protection of patient privacy and medical information, ethical approval has been obtained from the Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/16050184).

Modeling of Computational Biometry

This study employs an advanced methodology that goes beyond a single approach, instead utilizing a combination model that integrates multiple statistical techniques. The methodology incorporates bootstrap resampling, which is a method that repeatedly samples from the available data to estimate the uncertainty of statistical estimates. This approach provides a robust framework for analysis. Furthermore, the methodology integrates multilayer feed-forward neural networks, which are artificial neural networks with multiple layers of interconnected nodes. This allows the model to capture complex relationships and patterns within the data. The methodology used in this study follows a two-phase process that resembles the common practice of dividing data into training and testing sets. The first phase, known as the modeling phase, primarily focuses on developing and training the multilayer neural networks. This involves constructing the neural networks using the training data and iteratively adjusting the network's weights and biases to minimize the discrepancy between predicted and actual outcomes. By leveraging the training data, the model learns the underlying patterns and relationships necessary for making accurate predictions.

Upon completing the initial step, the second phase of this study is dedicated to validating the developed models. In this stage, multiple linear regression models are employed to investigate the relationship between the total number of victims and the selected explanatory variables. These models allow for the assessment of the association between the total number of victims and various factors while considering the numerical nature of the outcome variable. The research aims to gain insights into the impact and significance of the explanatory variables on the total number of victims, thereby providing a comprehensive understanding of the factors contributing to this forensic situation. The methodology employed in this study surpasses the utilization of multiple linear regression techniques alone. It incorporates additional components such as bootstrap resampling, multilayer feed-forward neural networks, linear regression, surface plots, and contour plot approaches. By integrating these diverse statistical techniques within a two-phase process, the research strives to uncover meaningful

relationships and patterns within the data. Ultimately, these findings contribute to an enhanced understanding and management of the case study.

a) Bootstrap

The bootstrap method is a resampling technique commonly used in statistical analysis. It starts by randomly selecting a sample from a population. From this sample, various statistics are computed. This initial sample is used to create a "pseudo-population" by generating multiple subsamples with replacements, each replicating the original sample. Due to the random selection process, the subsamples may differ from the initial sample [19]. Throughout the bootstrap process, statistics of interest are calculated for each subsample. These statistics can include means, medians, variances, and confidence intervals, among others. By examining the distribution of these statistics across the subsamples, valuable insights can be gained about the overall population. The bootstrap method is a powerful tool for estimating uncertainty in statistical estimates and making reliable inferences about the population. In this study, the researchers used the R software, a widely used programming language, and environment for statistical computing and graphics, to fit a multiple linear regression model. This model is suitable for analyzing continuous dependent variables. By employing R, the researchers were able to implement the multiple linear regression and analyze the relationship between the dependent variable (likely the number of victim levels) and the selected explanatory variables. The flexibility and functionality of R facilitated the execution of the multiple linear regression and provided a comprehensive platform for data analysis in this study [1,8,14].

b) Multiple Linear Regression

Linear regression is a statistical technique used to model and examine the relationship between two variables: a response variable and one or more predictor variables. It provides insights into how the explanatory variables influence the response variable and can also predict the value of the response variable when the predictor variables change [12]. Let's denote the response variable as '*y*' (dependent variable), which has a linear association with '*n*' explanatory variables (independent variables) represented as $\chi_1, \chi_2, \chi_3, ..., \chi_n$. The linear regression model can be defined as follows: $y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + ... + \beta_n\chi_n + \varepsilon$. In this equation, *y* represents the dependent variable, χ_n represents the independent variable, β_0 is the intercept, β_n represents the regression coefficients, and ε represents the error term or statistical error [1,12]. The regression equation not only allows us to predict the value of the dependent variable, *y*, but also provides insights into the direction and magnitude of the impact of the independent variable, *x*, on *y*. Let's consider an experiment with a sample of *n* observations and *k* independent variables. We assessed the homoscedasticity, normality, and linearity of the multiple linear regression (MLinearR) models [1,2,12]. We developed R-syntax using Bootstrap, MLinearR, and multi-layer perceptron approaches in R-Studio. The data was divided into two categories: training data for modeling and testing data for validation. Linear regression models were fitted to investigate the relationship between the total number of cases and the selected explanatory variables.

$$Y = \beta_0 + \beta_1(Gender) + \beta_2(Age) + \beta_3(Marital) + \beta_4(Adult) + \beta_5(Burglary) + \beta_6(Sexual) + \beta_7(\mathrm{Re}\,port) \quad (1)$$

where $\beta_1, \beta_2, ..., \beta_0$ are the parameters, and *Y* is the number of times respondents were victimized in a year. In the current study, the parameters of a regression model are determined using the Maximum Likelihood Estimator (MLE) technique R. A. Fisher proposed the method of MLE in 1922 as a means of point estimation. Maximum Likelihood Estimation (MLE) is a method for identifying the function that most likely explains observed data [12].

c) Multilayer Feed-Forward Neural Network (MLFNN)

In this study, we will be utilizing the multilayer feed-forward neural network (MLFNN) procedure. MLFNN is a commonly used artificial neural network that consists of three main layers: the input layer, the hidden layer, and the output layer [5,10]. In the context of this research, the dependent variable has three distinct classifications. Therefore, the output node for this analysis remains constant, but with three classifications. The MLFNN model, represented by Equation (1), incorporates N input nodes, H hidden nodes, and one output node [12]. For this specific analysis, eight variables are considered as inputs or independent variables: gender, age, marital status, social class, presence of adults and children in the household, burglary and sexual victimization, victim reporting, and household location. On the other hand, the output or dependent variable is defined by the total victim count, which is measured on a continuous scale.

Multilayer Feed-Forward Neural Network (MLFNN) With Two Hidden Layers Approach

Artificial Neural Networks (ANN) are computational models inspired by the structure and functionality of biological neural networks, commonly known as neural networks (NNs). Among the various types of neural networks, the Multilayer Feed-Forward (MLFF) architecture is utilized in this study [12]. It consists of one or more layers situated between the input, hidden, and output layers. Since this study focuses on a single dependent variable, the output node in the MLFF model is set to one. Figure 1.1 illustrates the MLFF model with N input nodes, H hidden nodes, and one output node [13]. The values of

the hidden node h_j, j =1…3 are given by $h_j = g_1\left(\sum_{j=1}^{3} v_{ji} x_i + E_1\right)$ where v_{ji} the output *weight,* E_1 is the bias. The values of the hidden node n_j, j =1…3 are given by $n_j = g_2\left(\sum_{j=1}^{3} v_{ji} h_i + E_2\right)$ where v_{ji} the output *weight.* ,E_2 is the bias. The values of the hidden node Y_j, j =1,2 are given by where $Y_i = g_3\left(\sum_{j=1}^{3} v_{ji} n_i + E_3\right)$ v_{ji} the output *weight, E_3* is the bias.

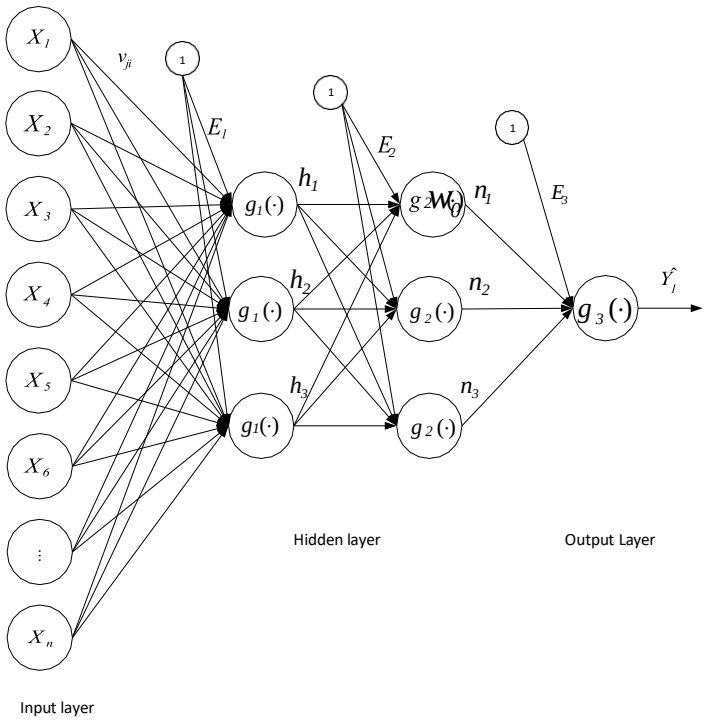


Figure 1.1 The general architecture of the MLFNN with two hidden layers, N input nodes, and one output node.

The variable chosen through the MLFNN procedure will serve as the input for the multiple linear regression. Multiple linear regression expands on simple linear regression by incorporating multiple explanatory variables instead of just one.

d) The Contour Plot and Surface Plot

Contour plots and surface plots are widely utilized visualization techniques across different fields to examine intricate data. These plots offer visual depictions of data correlations, patterns, and trends, significantly augmenting the intuitive comprehension of the fundamental information [4,17].

Contour Plot

A contour plot is a visualization technique that uses contour lines to connect points with equal values, providing a two-dimensional representation. It effectively demonstrates the fluctuations and gradients of a specific variable within a given space, enabling the identification of regions with similar values. Similarly, graphs are commonly used to visually depict functions involving two variables, such as $z = f(\chi, y)$. Alternatively, contour diagrams or contour maps can be employed as an alternative visualization method. These diagrams essentially act as "topographical maps" for the graphs representing $z = f(\chi, y)$. In this context, a topographical map refers to a two-dimensional representation of three-dimensional terrain, achieved by utilizing level curves or contours that indicate points with equal elevation. By analyzing these contours, valuable insights into the structure and variations of the function across different input values can be obtained [17].

Surface Plot

A surface plot is a visualization technique in three dimensions that illustrates the connection between two independent variables and a dependent variable. It presents a continuous surface where the height or color intensity corresponds to the value of the dependent variable at different combinations of the independent variables. Surface plots enable the visualization of intricate relationships and interactions among variables. Typically, the two independent variables are represented on the x1 and x2-axes, while the dependent variable is displayed on the y-axis [3,4].

$$y = f(\chi_1, \chi_2) + \varepsilon \qquad (1)$$

The values of the independent variables, χ_1 and χ_2, are of utmost importance in determining the response variable, *y*. This dependent variable, *y*, is influenced not only by the values of χ_1 and χ_2 but also by an experimental error term denoted as ε. The error term ε accounts for various sources of variation, including measurement errors in *y* and other factors not considered by the function *f*. In statistical terms, the error term ε is assumed to follow a normal distribution with a mean of zero and a variance of σ^2. To approximate the function *f* accurately, researchers commonly start by considering a low-order polynomial within specific regions of interest. When the response can be represented by a linear relationship with the independent variables, a first-order model is usually employed [6,9].

A First-Order Model

A first-order model involving two independent variables can be mathematically expressed as follows:

$$y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \varepsilon \qquad (2)$$

To achieve optimal outcomes in polynomial approximation, it is essential to utilize an appropriate experimental design for data collection. Afterward, the Method of Least Squares is employed to estimate the parameters in the polynomials based on the collected data. The resulting fitted surface is then utilized for conducting response surface analysis [18]. Response surface designs are specific types of designs tailored for fitting response surfaces. By employing these methodologies, the objective of studying response surface methodology (RSM) can be effectively accomplished. Additionally, contour plots and surface plots play a crucial role in visually representing intricate information. These visualization techniques contribute to a deeper understanding of the data and facilitate informed decision-making processes [4,17].

e) The Hybrid Method

The comprehensive step-by-step process of the proposed hybrid method is illustrated in the flowchart and can be summarized as follows:

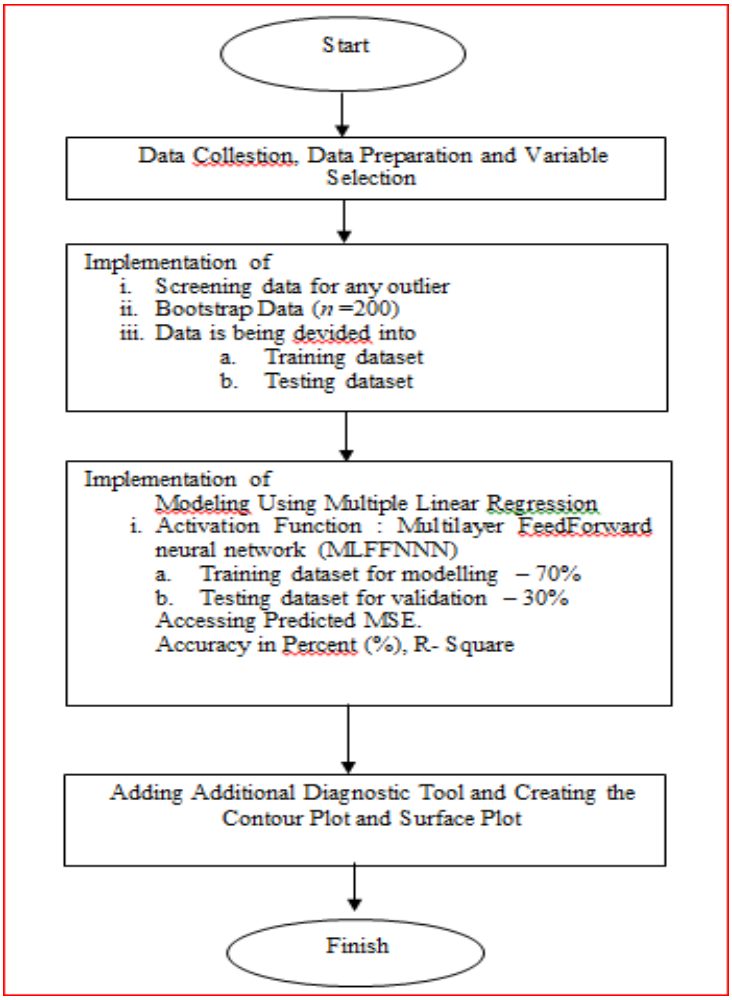


Figure 1.2 Flowchart of the proposed statistical ordinal modeling

Figure 1.2 in the diagram illustrates the intricate research process. In this study, great attention is devoted to the meticulous collection and preparation of data, as well as the selection of variables. Variable selection is a crucial step that involves identifying and choosing independent variables that have a significant impact on the dependent variable. By carefully selecting relevant variables, the model can effectively capture the most influential factors that drive the desired outcome.

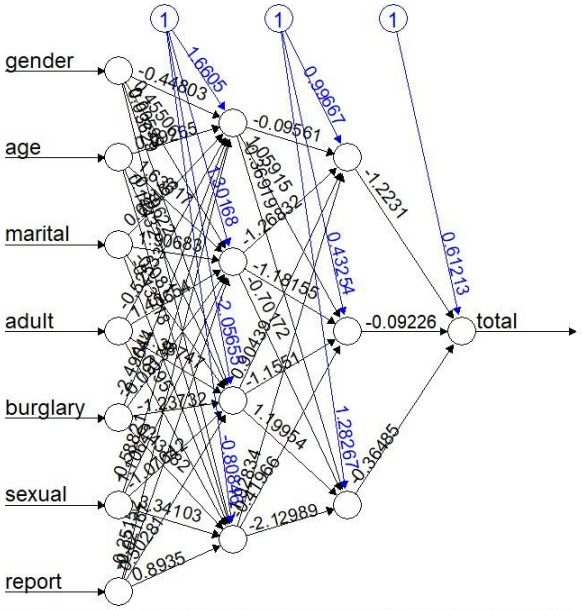
After thorough data preparation, a robust analysis is constructed using the bootstrapping method. The bootstrapping technique involves generating a new sample, equal in size to the original, by repeatedly selecting observations from the original sample [1, 18]. This approach allows for the possibility of selecting the same observation multiple times while discarding those not chosen in the bootstrap sample [8]. Following the bootstrapping process, the data is divided into separate training and testing datasets. The training dataset is used for modeling purposes, while the testing dataset is utilized for validation. Subsequently, the data is modeled using ordered logistic regression with multi-output. This modeling technique aims to obtain an inferential model that provides valuable insights. The developed syntax not only facilitates the modeling process but also offers the flexibility to utilize surface plots and contour plots, which play a crucial role in precise decision-making. These options are implemented in the final stages of the programming, enhancing the analytical capabilities of the study.

RESULTS

The main aim of this study is to evaluate the efficacy of a Multilayer Feed-Forward Neural Network (MLFFNN) using the ordered logistic model as its activation function. This evaluation encompasses both the training and testing datasets. The MLFFNN algorithm is utilized to determine the most suitable model for multiple linear regression by identifying clinically significant variables that minimize the MSE.net. Additionally, the inclusion of contour plots and surface plots provides valuable supplementary information about the characteristics of the variables under investigation. These visual representations contribute to a comprehensive understanding of the data and enhance the interpretability of the findings.

The Result of MLFFN Modeling

The results of the ordered logistic regression analysis on the training dataset, focusing on hypertension status as the dependent variable, are presented in Table 2. The MSE.net value, calculated to be 0.190999, indicates the level of accuracy in the distribution of the available data. A lower MSE.net value suggests a more effective analysis, indicating a closer alignment between the predicted and actual data and a smaller variance. Conversely, a larger variance implies significant discrepancies among dissimilar data points. To ensure the robustness of our predictions, a 70:30 train-to-test split was implemented, with 70% of the data allocated for modeling and the remaining 30% for testing. This division allows us to showcase the accuracy and reliability of our predicted data. Table 2 provides a summary of the results obtained from the ordered logistic regression analysis. Furthermore, the specific model with the best MLFFNN architecture is presented below, representing the most optimal approach for this study.




```
> MSE.lm <- sum((predict_lm - test$total)^2)/nrow(test)
> MSE.lm
[1] 4.261134
>
> #/Printing the Value of MSE for Linear Model and Neural Network/
> print(paste(MSE.lm,MSE.net))
[1] "4.26113419897977 0.190999755935112"
```

Figure 1.1 The architecture of the best (MLFFNN) model with five input variables, two hidden layers, and three output nodes (Obtained model)

Table 2. Result of Multiple Logistic Regression by combining the bootstrap method training and testing dataset

Variable	Estimate	Std. Error	t-Value	P-Value
(Intercept)	-5.2647e+00	1.7685e-14	-2.9769e+14	< 2.2e-16 ***
Gender	2.0882e+00	7.9009e-15	2.6430e+14	< 2.2e-16 ***
Age	-5.8824e-02	9.2992e-17	-6.3257e+14	< 2.2e-16 ***
Marital	3.8235e-01	7.6542e-16	4.9953e+14	< 2.2e-16 ***
Adult	1.6176e+00	3.0580e-15	5.2899e+14	< 2.2e-16 ***
Burglary	2.5588e+00	5.1502e-15	4.9684e+14	< 2.2e-16 ***
Sexual	-2.3529e-01	-2.3529e-01	-6.4889e+13	< 2.2e-16 ***
Report	-3.2353e-01	-3.2353e-01	-9.6773e+13	< 2.2e-16 ***

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

R^2 : 98.16%

By the equation, the model is given as follows :

Total victim = -5.2647e + 2.0882 * *Gender* + -5.8824e-02 * *Age* + 0.38235 * *Marital* + 1.6176 * *Adult* + 2.5588 * *Burglary* + -0.23529 * *Sexual* + -0.32353 * *Report*

Table 2 presents a summary of the detailed output, showcasing the results derived from the integrated linear regression model. To validate the chosen variables in the model, the established bootstrap method is utilized in this section. This technique enables the evaluation of the variables' significance and impact on the regression model. In this study, a total of seven variables were selected for analysis, namely gender (β_1 : 2.0882e+00; $p < 0.05$), age (β_2 : -5.8824e-02; $p < 0.05$), marital status (β_3 : 3.8235e-01; $p < 0.05$), adult in the household (β_4 : 1.6176e+00; $p < 0.05$), burglary victim (β_5 : 2.5588e+00; $p < 0.05$), sexual victim (β_6 : -2.3529e-01; $p < 0.05$), and victim's report (β_7 : -3.2353e-01; $p < 0.05$). By employing the bootstrap method, we assessed the significance and impact of these variables on the multiple linear regression model. The obtained results confirm the statistical significance of these factors concerning the outcome variable.

Model evaluation of the model

The model's evaluation relies on the forecast value, comparing the actual and predicted values to determine accuracy. The testing data set is utilized to assess the model constructed from the training data set. Distance prediction is employed to compare the predicted and actual data. The R syntax offers a model assessment approach that can be applied to evaluate subsequent methods. Table 3 displays the "Actual" and "Predicted" values obtained from the proposed methodology.

Table 3. The "Actual" and "Predicted" values obtained through the proposed methodology

Actual	Predicted
3	3.00
3	3.00
0	0.13
0	0.16
1	1.00
0	0.02
3	3.00
2	1.97
0	0.00
0	0.00
2	2.00
1	1.01
0	0.00
6	5.90
3	2.90

The variance between the "Actual" and "Predicted" values is negligible. Additionally, a paired sample t-test indicates that there are no statistically significant differences.

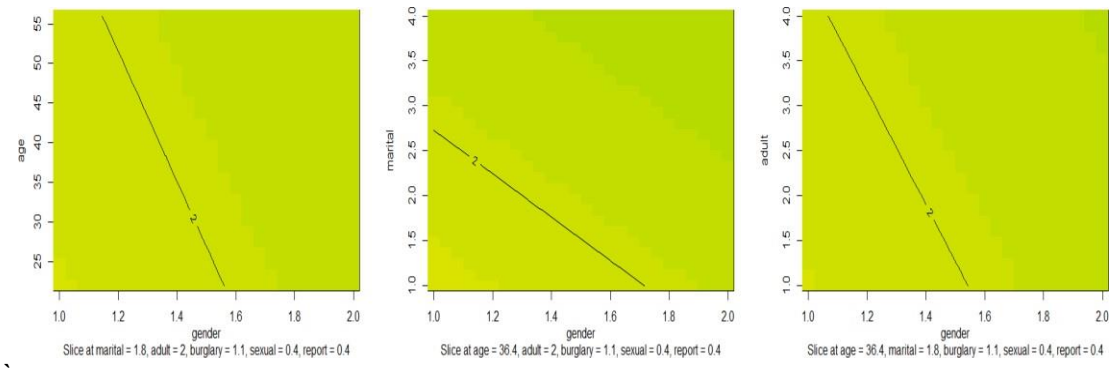
Table 4 Comparison of the "Actual Data" with "Predicted Data"

Paired Samples Test	
Variables	Mean (SD)
Actual	1.74(1.790)
Predicted	1.88(1.75)
T Statistics (d.f)	-3.37 (18)
p-value	0.331
Paired Sample Correlation (ρ)	0.998
p-value	0.000*

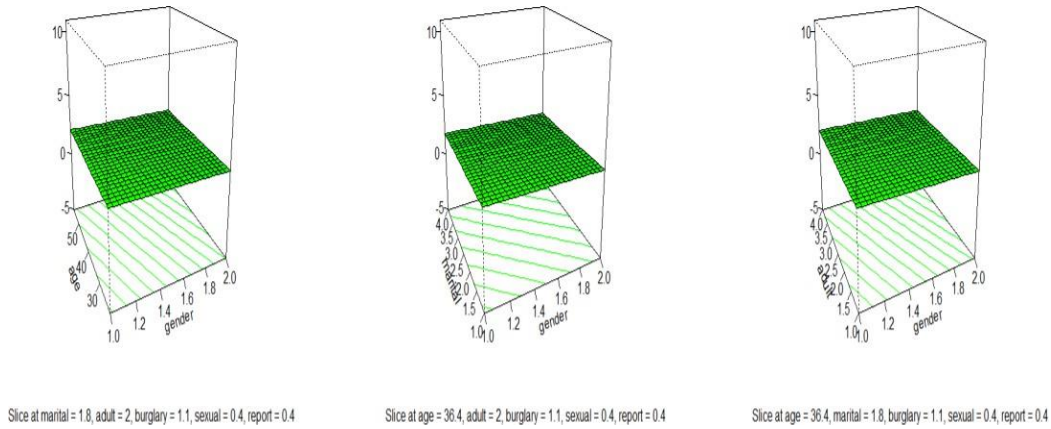
**significant at 0.05*
Paired samples T-Test was applied
Assumptions normality is fulfilled

The "Actual" and "Predicted" values are presented in Table 4, reflecting the outcomes of the proposed model. Based on the results, there were no statistically significant distinctions between the terms "actual" and "predicted" This highlights the superior performance of the proposed model.

Figure 1.3 showcases a contour and a surface plot that illustrates the relationship between various variables, denoted as a(i) a(ii) to u(i) u(ii). However, for detailed analysis, the selected contour and surface plot will be discussed further. In Figure 1.3 d(i) d(ii), a contour and surface plot demonstrates the relationship between the total number of victims, burglary, and gender. The plot indicates that as the incidence of burglary increases, the total number of victims also tends to increase. Additionally, it suggests that females are more likely to be victims in this scenario, although the contour line shows a gradual but significant increase. Figure 1.3 i(i) i(ii) presents a contour and surface plot depicting the relationship between the total number of victims, burglary, and age. The plot reveals that as the occurrence of burglary increases, the total number of victims also tends to rise. Moreover, it suggests that age factors are more likely to influence victimization in this context, with the contour line showing a gradual yet significant increase. Figure 1.3 m(i) m(ii) displays a contour and a surface plot that depicts the relationship between the total number of victims, burglary, and marital status. The plot reveals that as the occurrence of burglary increases, the total number of victims also tends to increase. Moreover, it suggests that marital status significantly influences victimization in this scenario, as indicated by the gradual yet meaningful rise in the contour line. In Figure 1.3 p(i) p(ii), a contour and surface plot is presented, illustrating the relationship between the total number of victims, burglary, and the presence of adults. The plot demonstrates that as the incidence of burglary increases, the total number of victims also tends to rise. Additionally, it suggests that the presence of adults significantly influences victimization in this context, as evidenced by the gradual but meaningful increase in the contour line. Figure 1.3 t(i) t(ii), exhibits a contour and surface plot that depicts the intricate relationship among the total number of victims, reports, and burglaries. The plot provides valuable insights into the dynamics observed when burglary incidents escalate, leading to a corresponding increase in the total number of victims. However, a notable divergence is observed as the number of victims reported to the police shows a decrease, which is indicated by the region situated in the lower left corner of the plot. This intriguing finding highlights the significance of considering the interplay between reports and burglary in understanding victimization patterns. The plot underscores the influential role of both variables in shaping the total number of victims, as evidenced by the discernible upward trajectory depicted by the contour line. This gradual yet meaningful increase further emphasizes the importance of factors such as reporting behavior and the occurrence of a burglary in the overall victimization dynamics within this specific context.

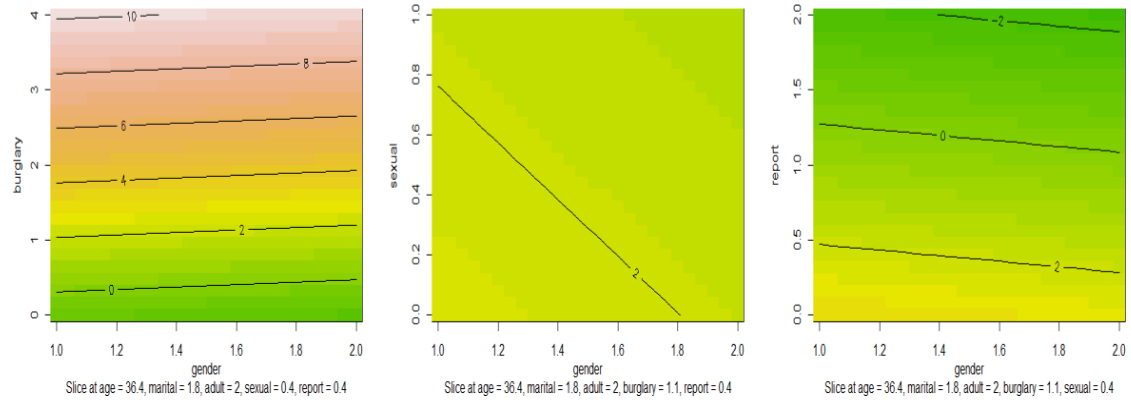


a(i) **b(i)** **c(i)**

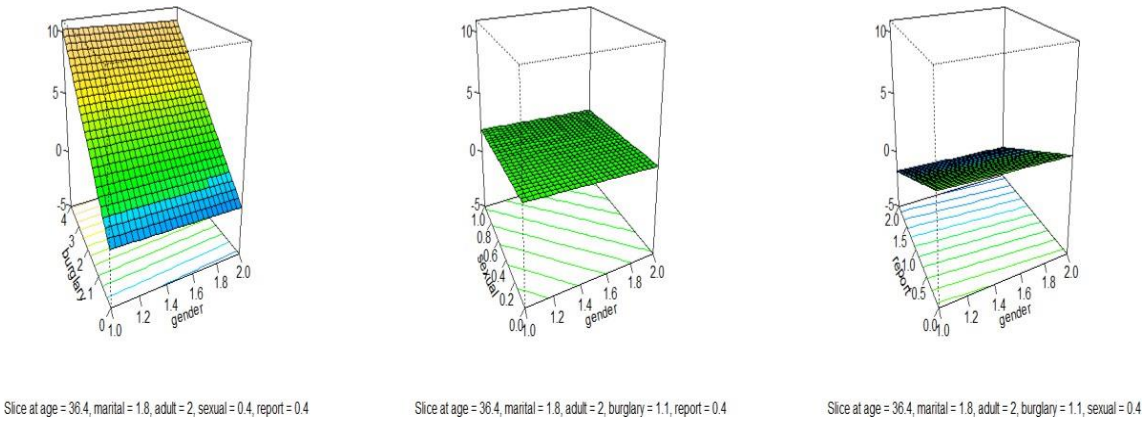


a(ii) **b(ii)** **c(ii)**

Figure 1.3 Contour and surface plot a(i)-a(ii) Total victim versus age and gender, b(i)-b(ii) Total victim versus marital and gender, c(i)-c(ii) Total victim versus adult and gender

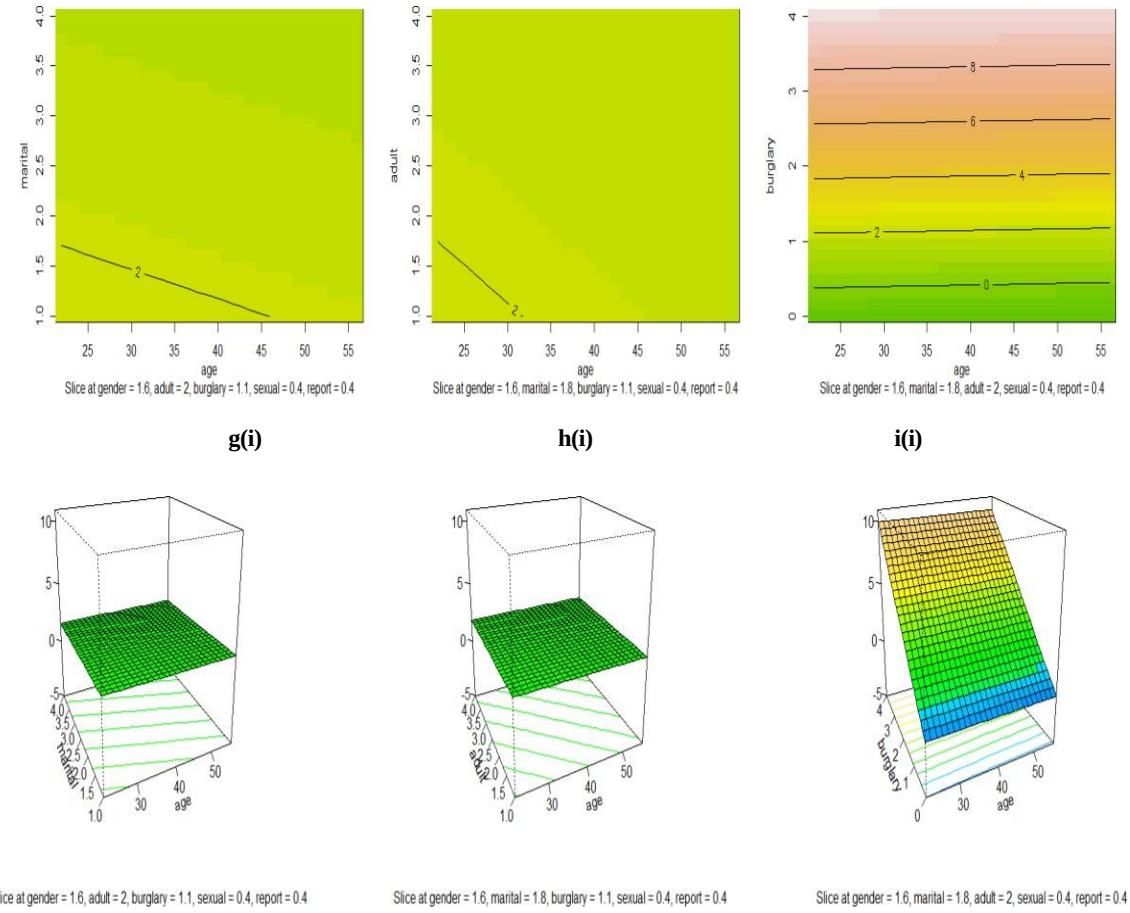


d(i) e(i) f(i)



d(ii) e(ii) f(ii)

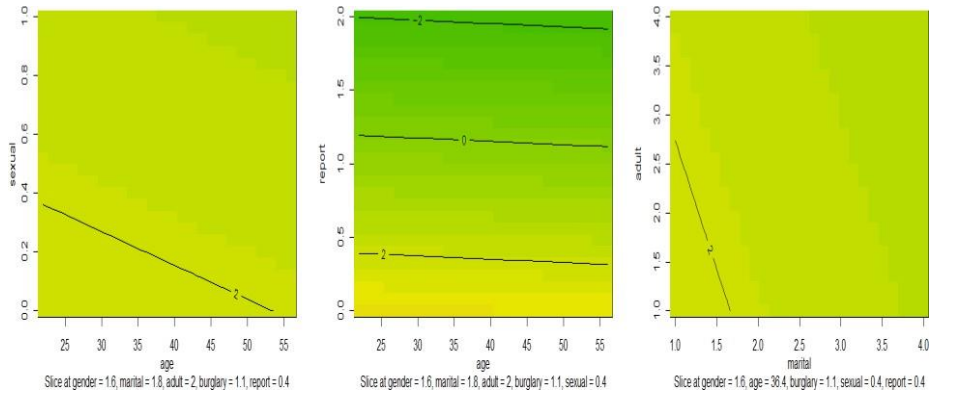
Figure 1.3 Contour and surface plot d(i)-d(ii) Total victim versus buglary and gender, e(i)-e(ii) Total victim versus sexual and gender, f(i)-f(ii) Total victim versus report and gender



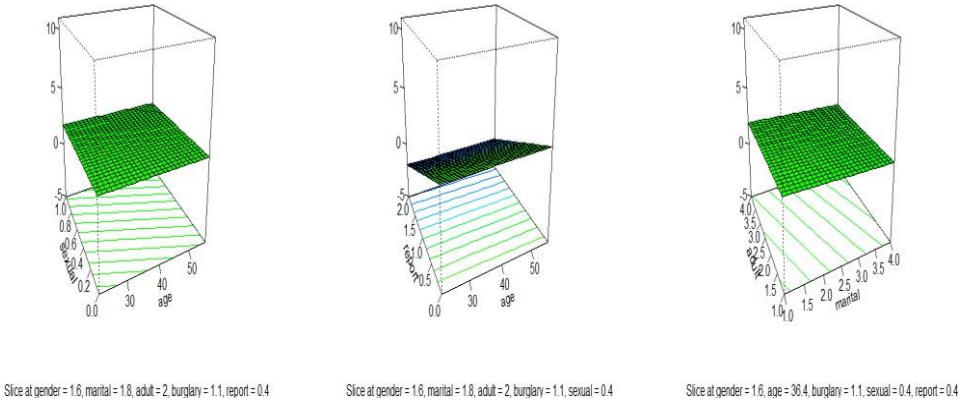
g(i) **h(i)** **i(i)**

g(ii) **h(ii)** **i(ii)**

Figure 1.3 Contour and surface plot g(i)-g(ii) Total victim versus marital and age, h(i)-h(ii) Total victim versus adult and age, i(i)-i(ii) Total victim versus burglary and age

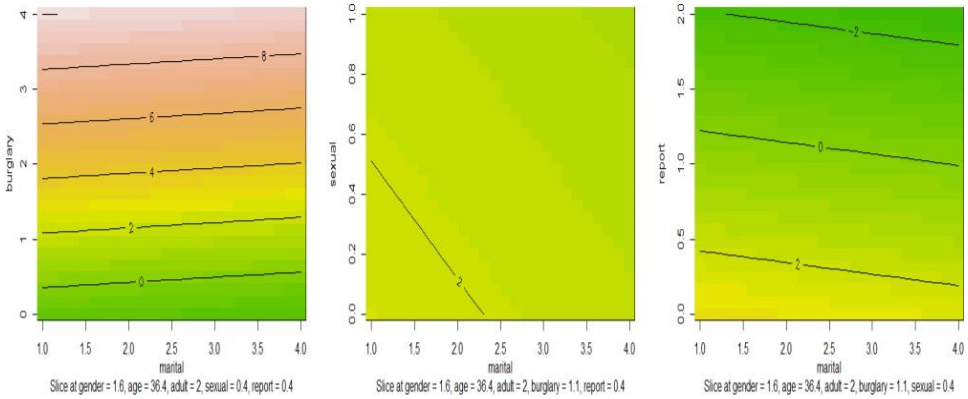


j(i) k(i) l(i)

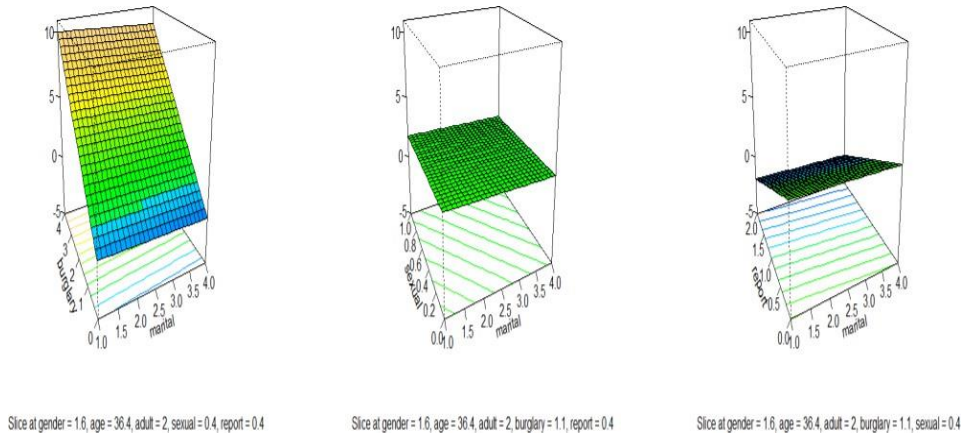


j(ii) k(ii) l(ii)

Figure 1.3 Contour and surface plot j(i)-j(ii) Total victim versus sexual and age, k(i)-k(ii) Total victim versus report and age, l(i)-l(ii) Total victim versus adult and marital

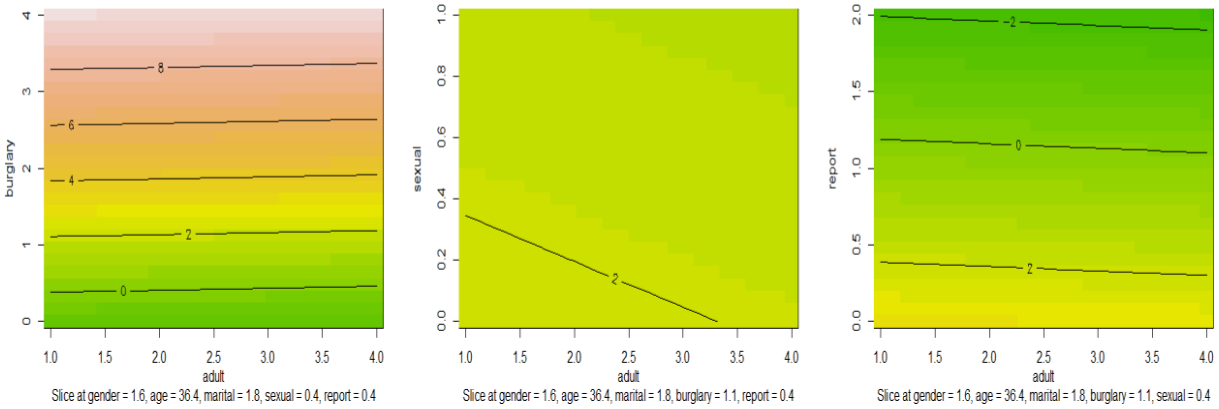


m(i) n(i) o(i)

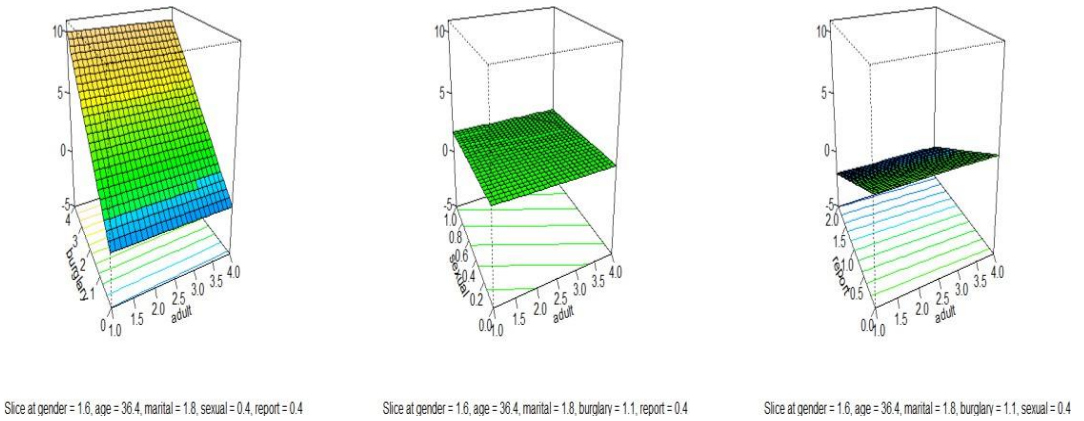


m(ii) n(ii) o(ii)

Figure 1.3 Contour and surface plot m(i)-m(ii) Total victim versus burglary and marital, n(i)-n(ii) Total victim versus sexual and marital,
o(i)-o(ii) Total victim versus report and marital

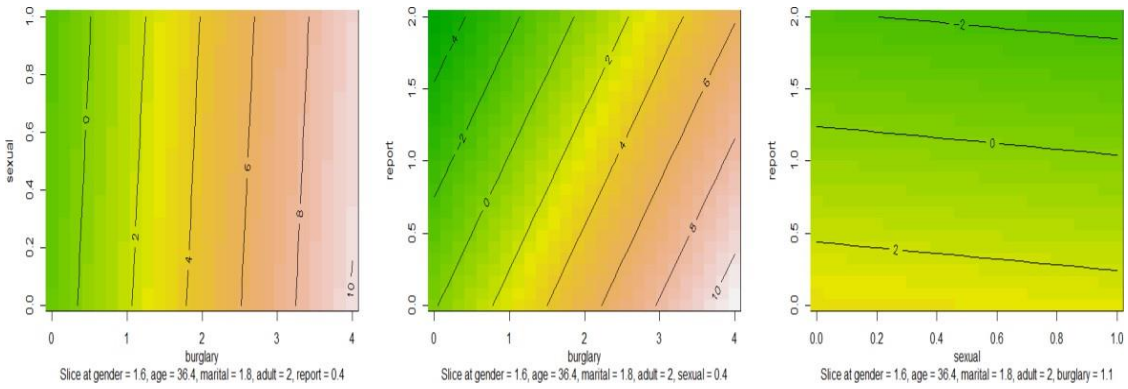


p(i) **q(i)** **r(i)**

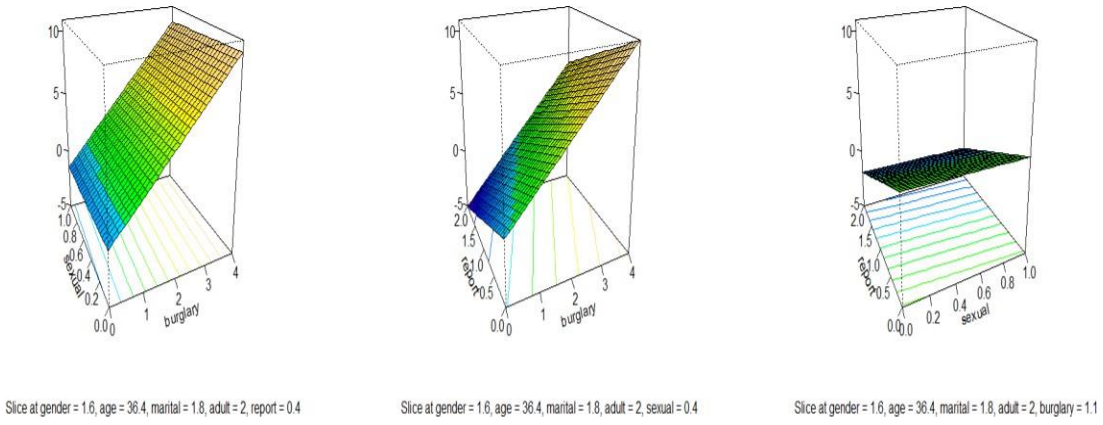


p(ii) **q(ii)** **r(ii)**

Figure 1.3 Contour and surface plot p(i)-p(ii) Total victim versus burglary and adult, q(i)-q(ii) Total victim versus sexual and adult, r(i)-r(ii) Total victim versus report and adult



s(i) **t(i)** **u(i)**



s(ii) **t(ii)** **u(ii)**

Figure 1.3 Contour and surface plot s(i)-s(ii) Total victim versus sexual and burglary, t(i)-t(ii) Total victim versus report and burglary, u(i)-u(ii) Total victim versus report and sexual

DISCUSSION

By utilizing the proposed method, we achieved an accurate estimation of the dependent variable's predicted value. The hybrid method we discovered and harmonized allowed for the creation of an exceptionally precise and reliable model. The analysis involved nine variables: gender, age, marital status, adult in the household, burglary victim, sexual victim, and victim's report, all of which were found to be significant. The main objective of this study was to develop, test, and validate a regression model using multiple linear regression (MLR) methodologies with Multilayer Feed-Forward Neural Networks (MLFNN). Expert advice was incorporated into the variable selection process. The bootstrap method was employed, initially generating a “mega” file using the original dataset. Subsequently, the bootstrap procedure produced numerous replacement files. Statistical samples were generated and saved through the bootstrap method. This process was iteratively repeated, often thousands of times, until the data was ready for analysis. This proposed methodology can be integrated into an application concept using the R syntax algorithm. The first step in this approach is to consult with an expert for variable selection. Then, the dataset undergoes the bootstrap procedure. The training and testing data are kept separate. The R syntax algorithm links the application to the methodology based on the proposed method. The initial phase involves selecting variables with the guidance of an expert. Subsequently, the bootstrap procedure is applied to the data. In this stage, 70% of the bootstrap data is designated as training data, while 30% is allocated for testing purposes. The model is built and tested using the training dataset. A successful model is determined by the lowest mean absolute deviation. The developed syntax calculates the difference between the actual and predicted values. The study findings assist decision-makers in achieving optimal outcomes. Selecting appropriate input parameters, preparing data for linear modeling, and standardizing it are identified as the most challenging tasks. The study demonstrates that employing statistical formulations, performing computations in R syntax, and utilizing the multiple linear regression package lead to highly successful linear modeling. The integration of statistical formulations, computation using R syntax, and the implementation of the linear regression package played a significant role in achieving successful outcomes in the linear regression modeling approach. However, it is important to acknowledge that certain tasks in this process present specific challenges. The selection of suitable input parameters, data preparation for linear regression modeling, and data standardization are identified as particularly demanding and complex tasks encountered throughout the process. Furthermore, this study introduces a diagnostic tool for evaluating model fitting, which includes the provision of contour and surface plots for each variable. This approach aids researchers in visually interpreting and understanding the optimized output derived from the results, thereby enhancing their ability to gain meaningful insights.

CONCLUSION

The primary objective of this study is to make a significant contribution to the field by introducing innovative hybrid methods that combine bootstrapping, multilayer neural networks, and multiple linear regression. These methods offer a unique approach to analyzing complex phenomena. To ensure clarity and practicality, we have developed an R syntax that provides a detailed implementation process. Our research focuses on investigating the dependent variable of total victims, aiming to understand the factors that influence its occurrence by examining various independent variables. These independent variables have been identified as significant factors contributing to the number of total victims. By utilizing hybrid methods and conducting a comprehensive analysis of the relationship between the dependent variable (total victims) and the mentioned independent variables, our goal is to uncover valuable insights that can inform stakeholders, decision-making processes, and interventions. This research represents a significant advancement in our understanding of the complex dynamics involved in forensic cases, specifically the frequency of victimization and its associated risk factors. Additionally, the model's adequacy can be assessed by utilizing the predictive value obtained from the model with the lowest error. The incorporation of neural networks, contour plots, and surface plots has provided comprehensive information to assist researchers in optimizing results and gaining valuable insights. Employing the hybrid model, this approach allows for the evaluation of the model's effectiveness and its contribution to the outcomes. The statistical analysis conducted in this R-

based study demonstrates the superiority of regression modeling compared to other techniques, as evidenced by a mean square error of 0.1909. The study's findings firmly establish the exceptional performance and superiority of the proposed hybrid model technique over alternative methods.

ACKNOWLEDGMENT

The authors express their gratitude to Universiti Sains Malaysia (USM) for their support in funding this study through the Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS/1/2022/STG06/USM/02/10).

REFERENCES

1. Ahmad, W.M.A.W., Nawi, M.A.A., Aleng, N.A., Ibrahim, S.M. 2016. An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach (SAS). Journal of Modern Applied Statistical Methods. 5(2).
2. Ahmad, W.M.A.W., Aleng, N.A., Dasril, Y., Halim, N.A., Hasan, R., Ali, Z., Baharum, A., Zakaria, S. 2017. Modified Nonlinear Model for Exponential Growth Method and Its Application in Biostatistics Using SAS. International Journal of Multidisciplinary Research and Modern Education, 3(1), 89-94.
3. Ahmad, W.M.A.W., Ibrahim, S.M., Mokhtar, K., Aleng, N.A., Rahim, H.A. 2016. Simple response surface methodology using RSREG (SAS). Journal of Modern Applied Statistical Methods, 15(1), 855-867.
4. Anderson, M.J. Whitcomb, P.J. 2004. Design Solutions from Concept Through Manufacture: Response Surface Methods for Process Optimization. Desktop Engineering.
5. Cohen, J., Cohen, P., West, S. G., Aiken, L. S. 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
6. Douglas, E.A. 2021. Introduction to Linear Regression Analysis. John Wiley & Sons.
7. Draper, N. R., Smith, H. 1992. Applied Regression Analysis, Second Edition. John Wiley & Sons, Inc.
8. Efron, B., Tibshirani, R. 1985. The Bootstrap Method for Assessing Statistical Accuracy. Behaviormetrika, 12, 1-35.
9. Gao, Z. 2020. A Hybrid Neural Network Model with Linear Regression and Gradient Boosting Decision Tree for Crime Prediction. Applied Sciences, 10(7), 2301.
10. Gardner, M.W., Dorling, S.R. 1998. Artificial Neural Networks (The Multilayer Perceptron) – Review of Applications in The Atmospheric Sciences. Atmospheric Environment, 32(14), 2627-2636.
11. Gaudart, J., Guillaume, B. 2004. Performance Comparison of Multi-Layer Perceptron and Linear Regression for Epidemiological Data. Computational Statistics & Data Analysis, 547-570.
12. Ghazali, F.M.M., Ahmad, W.M.A.W., Srivastava, K.C., Shrivastava, D., Noor, N.F. M., Akbar, N.A.N., Aleng, N.A., Alam, M.K. 2021. A Study of Creatinine Level among Patients with Dyslipidemia and Type 2 Diabetes Mellitus using Multilayer Perceptron and Multiple Linear Regression. Journal of Pharmaceutical and Bioallied Sciences, 13(Suppl 1), S795-S800.
13. Hornik, K., Stinchcomber, M., White, H. 1989. The Universal Approximation Power of Multilayer Feedforward Networks. Neural Networks, 2, 359-366.
14. Kim, Y. M., Im, J. 2019. Frequency domain bootstrap for ratio statistics under long-range dependence. Journal of the Korean Statistical Society, 48(4), 547-560.
15. Liu, L. 2019. Crime Rate Prediction Utilizing Ensemble Learning with Neural Networks. IEEE Access, 7, 71312-71320.
16. Mohler, G. 2011. Self-Exciting Point Process Modeling of Crime. Journal of the American Statistical Association, 106(493), 100-108.
17. Montgomery, D.C. 2005. Design and analysis of experiments: Response surface method and designs. New Jersey: John Wiley and Sons, Inc.
18. Ngo, T.H.D., La Puente, C.A. 2012. The Steps to Follow in a Multiple Regression Analysis. SAS Global Forum 2012: Statistics and Data Analysis. Paper 333-2012, 1-12.
19. Noor, N.F.M., Ahmad, W.M.A.W., Nawi, M.A.A., Ghazali, F.M.M., Aleng, N.A., Shaari, R., Harun, A.M., Abas, R. 2021. Prediction the best predictor for urea reading among diabetic patients using artificial neural networks (ANNS) models. Sapporo Medical Journal, 55(1), 1-7.
20. Raza, H. 2019. Crime Hotspot Prediction Using Spatial Features and Neural Network with Ridge Regression. Future Generation Computer Systems, 91, 126-136.
21. Wang, Z. 2019. Forecasting Crime Rate with Deep Learning Techniques: A Study of Los Angeles. ISPRS International Journal of Geo-Information, 8(7), 308.

Appendix

#First, import the tidyverse and neuralnet packages.

```
if(!require(tidyverse)){install.packages("tidyverse")}
library(tidyverse)
if(!require(neuralnet)){install.packages("neuralnet")}
library(neuralnet)
if(!require(dplyr)){install.packages("dplyr")}
library(dplyr)
```

##Dataset for Biometry Modeling Study/##

```
Input =("
```

```
bmi hdl height hyper smoke trig diabc diab
26 50 178 0 1 181 1 Normal
23 48 175 0 2 198 1 Normal
28 54 182 2 2 145 1 Normal
27 47 173 0 2 196 1 Normal
25 41 170 0 3 261 1 Normal
⋮  ⋮  ⋮
27 38 164 1 1 137 2 Borderline
31 48 172 1 3 111 2 Borderline
37 43 172 1 1 199 2 Borderline
27 95 177 1 2 180 2 Borderline
32 33 182 1 2 263 2 Borderline
26 37 178 1 1 179 2 Borderline
28 34 186 1 1 145 2 Borderline
⋮  ⋮  ⋮
29 41 177 2 2 214 3 Diabetes
32 45 175 2 2 100 3 Diabetes
33 35 178 2 2 148 3 Diabetes
28 55 176 2 2 169 3 Diabetes
31 35 170 2 2 235 3 Diabetes
22 52 187 2 2 166 3 Diabetes
28 40 169 2 1 204 3 Diabetes
24 46 163 2 1 138 3 Diabetes
23 44 187 1 2 108 3 Diabetes
 ")
data = read.table(textConnection(Input),header=TRUE)
mydata <- rbind.data.frame(data, stringsAsFactors = FALSE)
iboot <- sample(1:nrow(mydata),size=1000, replace = TRUE)
Bootdata <- mydata[iboot,]

#MultiLayer MLFFNN
##Install the Neuralnet Package
if(!require(neuralnet)){install.packages("neuralnet")}
library("neuralnet")

#Determine the Training and Testing of the Dataset
#70% for Training and 30% For Testing
index = sample(1:nrow(data),round(0.70*nrow(data)))
Training <- as.data.frame(data[index,])
Testing <- as.data.frame(data[-index,])

##Plotting the Architecture of MLFFNN Neural Network
nn <- neuralnet(diab~bmi+hdl+height+hyper+smoke+trig, data=Training,
            hidden=c(4,3),linear.output = F, stepmax = 1000000)
plot(nn)
options(warn=-1)
nn1 <- neuralnet(diabc~bmi+hdl+height+hyper+smoke+trig,data=Training,
              hidden=c(4,3),act.fct = "logistic",
              linear.output = FALSE, stepmax = 1000000)
plot(nn1)
nn1$result.matrix

##Testing The Accuracy of The Model-Predicted Results
##Predicted Results Are Compared To The Actual Results
Temp_test <- subset(Testing, select = c("bmi","hdl","height","hyper","smoke","trig"))
head(Temp_test)
nn1.results <- compute(nn1, Temp_test)

##Results
results <- data.frame(actual = Testing$diabc,
                  prediction = nn1.results$net.result)

##Use The Predicted Mean Squared Error NN (MSE-forecasts the Network)
##As a Measure of How Far the Predictions Are From The Real Data
predicted <- compute(nn1,Testing[,1:6])
MSE.net <- sum((Testing$diabc - predicted$net.result)^2)/nrow(Testing)
```

```
##Printing the Predicted Mean Square Error
MSE.net

##Neural Network Parameter Output
library(neuralnet)

nn1 <- neuralnet(diabc~bmi+hdl+height+hyper+smoke+trig,data=Training,
                 hidden=c(4,3),act.fct = "logistic",
                 linear.output = FALSE, stepmax = 1000000)
nn1$result.matrix
results <- data.frame(actual=Testing$diabc,prediction=nn1.results$net.result)

predicted1=results$prediction*abs(diff(range(data$diabc)))+min(data$diabc)

##Print(Predicted)
actual1=results$actual*abs(diff(range(data$diabc)))+min(data$diabc)

##Print(Actual1)
deviation= ((actual1-predicted1))
##Print(deviation)

##Mean Absolute Deviance
value=abs(mean(deviation))
print(value)
accuracy_in_percent=(1-((value)/137))*100
accuracy_in_percent

#Modeling Ordinal Model
#Build Ordinal Logistic Regression Model
if(!require(MASS)){install.packages("MASS")}
library("MASS")

polr(as.factor(diab)~bmi+hdl+height+hyper+smoke+trig,data=Bootdata,
    Hess=TRUE, method=c("logistic"))
m<-polr(as.factor(diab)~bmi+hdl+height+hyper+smoke+trig,data=Bootdata,
       Hess=TRUE, method=c("logistic"))
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))

## Odds Ratios
exp(coef(m))
if(!require(rsm)){install.packages("rsm")}
library(rsm)
first <- rsm(diabc~FO(bmi, hdl, height, hyper, smoke, trig), data=data)
summary(first)

par(mfrow = c(2,3))       # 2 x 3 pictures on one plot
contour(
  first,           # Our model
  ~ bmi+hdl+height+hyper+smoke + trig,   # 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
  ~ bmi+hdl+height+hyper+smoke + trig,   # A formula to obtain the 6 possible graphs
  col = topo.colors(100), # Color palette
```

```
  contours = "colors"     # Include contours with the same color palette
)

if(!require(party)){install.packages("party")}
library(party)
if(!require(partykit)){install.packages("partykit")}
library(partykit)

if(!require(caret)){install.packages("caret")}
library(caret)

if(!require(tree)){install.packages("tree")}
library(tree)

if(!require(caTools)){install.packages("caTools")}
library(caTools)

#To Convert Factors To Numbers
data$diab<-as.factor(data$diab)

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

# calculating The Prediction For The Test
pred = predict(dtm, data[,-6])
confusionMatrix(pred,as.factor(data$diab))
```