A NEW REGRESSION MODELLING APPROACH AND ITS APPLICATION IN BIOSTATISTICS

SUMAIYA ZABIN EUSUFZAI

UNIVERSITI SAINS MALAYSIA
2024

A NEW REGRESSION MODELLING APPROACH AND ITS APPLICATION IN BIOSTATISTICS

By

SUMAIYA ZABIN EUSUFZAI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

July 2024

ACKNOWLEDGEMENT

First and foremost, I am grateful to Allah for the health, protection, and guidance in completing this Ph.D. study. The journey has not been an easy one. However, with His permission, it has been smooth and possible. It is a great honour to have this opportunity to extend my gratitude and appreciation to acknowledge the contributions of the many people who supported me during my Ph.D study. My first debt of gratitude goes to my supervisor, Associate Professor Ts. Dr. Wan Muhamad Amir W Ahmad for his constant guidance, encouragement, patience, continuous support, comments, and endurance during this Ph.D. study. I have benefited enormously from his biostatistics knowledge, experience, and expertise. I am also grateful to him for the opportunities to attend biostatistics workshops before and during my PhD study. I dedicate this thesis to my family members, especially my parents, my husband Dr. Nafij Bin Jamayet, and my daughter Nashmia Binte Nafij for their continuous support and encouragement throughout this study, and also my sibling, whom I value very much for constantly providing me with love, hope, and continuance courage to endure the challenges I have throughout this study. I would also like to thank my senior, Dr. Samiya Riaz, who provided helpful suggestions and encouragement in completing this study.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF FIGURES ..... **vii**

LIST OF TABLES ..... **viii**

LIST OF ABBREVIATIONS ..... **ix**

LIST OF APPENDICES ..... **x**

ABSTRAK ..... **xi**

ABSTRACT ..... **xiii**

CHAPTER 1 INTRODUCTION **1**

1.1 Preview of the chapter 1

1.2 Background of the study – A review of statistical modelling 1

1.3 Research motivation and problem statement 5

1.4 The rationale of the study 8

1.5 Conceptual framework of the research study 10

1.6 Research hypothesis 11

1.7 Research Objectives 11

1.7.1 General Objective 11

1.7.2 Specific Objective 11

1.8 The scope and methodology 11

1.9 The contribution of the study 13

1.10 The limitation of the study 13

1.11 The organization of the thesis 14

CHAPTER 2 LITERATURE REVIEW **16**

2.1 Preview of the chapter 16

2.2 History of regression 16

2.3 Statistical content using multiple logistic regression model in and dental public health modelling 22

2.4 Bootstrapping Approach 26

2.5 Multi-layer feed-forward neural network (MLFFNN) 29

2.5.1 History, Advantages, and Drawbacks of Multi-Layer Feed-Forward Neural Network (MLFFNN) 32

2.5.2 Advantages and drawbacks of using Logistic regression modelling 34

2.6 Advantage of the combination method utilising both logistic regression and neural network models 35

2.7 The implementation of a multilayer-feedforward neural network in public health 37

2.8 The previous study design of the combined method of multilayer feed-forward neural network (MLFFNN) and regression for modelling in public health. 41

2.9 The bootstrapping, regression modelling and MLFFNN in Health Science .. 47

2.10 The specific implementation of the statistical methods 51

2.10.1 Case study I: Dental caries and its risk factors among preschool children 51

2.10.2 Case study II: Knowledge of caries prevention among parents of preschool children 60

2.11 Review of Methodology, Limitations, and Research gap 66

2.12 Concluding remarks 69

CHAPTER 3 METHODOLOGY 71

3.1 Chapter preview 71

3.2 Study area 71

3.3 Study design 72

3.4 Study population 72

3.4.1 Reference population 72

3.4.2 Target population 73

3.4.3 Source population 73

3.4.4 Sampling frame and Sample size calculation 73
3.4.5 Subject criteria 74
3.5 Study period 74
3.6 Ethical consideration 74
3.7 Software used in research 74
3.8 The variables 75
3.8.1 Variables for Case Study I 75
3.8.2 Variables for Case Study II: 76
3.9 The main three elements of the methodology constructed with R syntax 77
3.9.1 Bootstrapping method 77
3.9.2 Multi-layer Feed-forward Neural Network (MLFFNN) 78
3.9.3 Logistic regression model 80
3.9.4 The multiple logistic regression model 81
3.10 Model specification for the case study I 85
3.11 Model specification for the Case Study II 86
3.12 Testing for the significance of the model 86
3.13 The Proposed Methodology 87
3.13.1 Case study 1 88
3.13.2 Case Study II 91
3.14 Summary 95
CHAPTER 4 RESULTS 97
4.1 Chapter Overview 97
4.2 Study case I: Dental caries modelling among preschool children 97
4.2.1 Multiple logistic regression modelling-case I 98
4.2.2 Diagnostic test and model evaluation - Case 1 101
4.2.3 Paired sample t-test 101
4.2.4 Summary Case Study I 102

4.3 Study case II: Modelling on knowledge status of dental caries among parents 103

4.3.1 Multiple logistic regression model- Case II 103

4.3.2 Diagnostic test and model evaluation- Case II 106

4.3.3 Paired sample t-test 106

4.3.4 Summary Case Study II: 107

4.3.5 Concluding remarks: 108

CHAPTER 5 DISCUSSION 109

5.1 Chapter Overview 109

5.2 The development of an integrated multi-layer feed-forward neural network model 109

5.2.1 Discussion on case study I and case study II 112

5.3 Conclusion 118

CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS 119

6.1 Chapter Overview 119

6.2 Summary and Conclusion 119

6.3 Recommendations and Future Directions 121

REFERENCES 123

APPENDICES

LIST OF FIGURES

Page

Figure 1.1 Conceptual framework of the study 10

Figure 3.1 The architecture of Multi-layer Feed-forward Neural Network 80

Figure 3.2 Flowchart of a new proposed logistic regression Model for Case Study I and Case Study II 95

Figure 4.1 The architecture of the multilayer feed-forward neural network (MLFFNN) with five input nodes, two hidden layers, and one output node 101

Figure 4.2 The architecture of the multilayer feed-forward neural network (MLFFNN) with seven input nodes, two hidden layers, and one output node 106

LIST OF TABLES

Page

Table 3.1 Description of data concerning parents and preschool children in Case Study 1 76

Table 3.2 Description of data concerning caries in Case Study II 77

Table 4.1 Result of multiple logistic regression by combining the bootstrap method training and testing data set for case I 100

Table 4.2 *t-test* results of the "Actual" and "Predicted" values of the caries status from the proposed methodology..... 101

Table 4.3 Result of multiple logistic regression by combining the bootstrap method training and testing data set for case II 104

Table 4.4 *t-test* results of the "Actual" and "Predicted" values of the knowledge status from the proposed methodology..... 106

LIST OF ABBREVIATIONS

SOC	Sense of Coherence
AI	artificial Intelligence
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ECC	Early Childhood Caries
GIS	Geographic Information Systems
KAP	Knowledge, Attitude, And Practices
MLFFNN	Multilayer Feed-Forward Neural Network
MLP	Multilayer Perceptron
MSE-F	Mean Square Error For Forecasting
CAR	Number of Dental Caries
COR	Ribenna Drinking Practice Among Children
TEA	Milk Tea
INC	Household Income
PSC	Practice Score Towards Caries Prevention Among Parents
ASC	Attitude Score Towards Caries Prevention Among Parents
BIC	Bicarbonate Drinking Practice
OLR	Ordered Logistic Regression
MLP	Multilayer Perceptron Neural Network
PMSE	Predicted Mean Square Error
ReLU	Rectified Linear Unit
MAD	Mean Absolute Deviance

LIST OF APPENDICES

APPENDIX A	ETHICAL APPROVAL LETTER
APPENDIX B	R-Syntax for Case 1
APPENDIX C	R-Syntax for Case II
APPENDIX D	RESEARCH PUBLICATIONS

PENDEKATAN PEMODELAN REGRESI BAHARU DAN APLIKASINYA DALAM BIOSTATISTIK

ABSTRAK

Penyelidikan ini bertujuan membangunkan metodologi bersepadu yang akan dirumuskan dalam rangka kerja Rangkaian Neural Hadapan Suapan Berbilang Lapisan (MLFFNN) dan regresi logistik. Purata sisihan mutlak dan ralat min kuasa dua digunakan untuk menilai prestasi model bersepadu, dan ia dijadikan kayu ukur untuk menentukan tahap ketepatan dan kecekapan peramalan. Keperluan untuk mendapatkan keputusan yang lebih baik menjadi dorongan untuk meneruskan kajian ini. Objektif kajian ini adalah untuk membangunkan dan mengaplikasikan model bersepadu yang menggabungkan butstrap dan MLFFNN dengan pemodelan regresi logistik (LRM) untuk mencapai ketepatan ramalan dan kebolehtafsiran yang lebih baik. Kaedah bersepadu yang digunakan dalam kajian ini adalah berdasarkan prinsip butstrap, LRM, dan MLFFNN. Ketepatan teknik yang dicadangkan dinilai menggunakan Rangkaian Neural Ralat Kuasa Dua Min (MSE.net), Min Sisihan Mutlak (MAD), dan ketepatan peratusan. Setiap komponen ini bertindak sebagai penanda aras untuk menilai ketepatan dan keberkesanan model yang dicipta. Satu Ujian-t digunakan untuk menilai perbezaan antara nilai sebenar dan nilai ramalan daripada model ini. Analisis data dijalankan menggunakan program R dan SPSS versi 26. Dua kajian kes daripada kesihatan awam pergigian telah digunakan untuk mengesahkan model bersepadu yang baru dibangunkan ini, i) karies gigi di kalangan kanak-kanak prasekolah, dan ii) Kajian kes untuk pengetahuan kesihatan mulut di kalangan ibu kanak-kanak prasekolah. Penggabungan butstrap, MLFFNN, dan regresi logistik dalam pendekatan bersepadu meningkatkan ketepatan parameter anggaran dan menangani hubungan tidak pasti antara pembolehubah bersandar dan bebas. Kajian kes yang memfokuskan kepada

karies gigi dalam kalangan kanak-kanak prasekolah, Min Sisihan Mutlak (MAD) ialah 0.0221126 dan Ralat Min Kuasa Dua Ramalan (PMSE) ialah 0.07909. Ujian-t sampel berpasangan mendedahkan tiada perbezaan yang signifikan antara nilai sebenar dan nilai ramalan, dengan min dan sisihan piawai seperti berikut: Sebenar (Min [SD] = 0.30 [0.483]) dan Diramalkan (Min [SD] = 0.31 [0.373]) ; df = -0.067(9); nilai-p > 0.05. Dalam kajian berkenaan pengetahuan kesihatan mulut di kalangan ibu kanak-kanak prasekolah, MAD ialah 0.05303337, dan PMSE ialah 0.053033. Keputusan daripada ujian-t sampel berpasangan menunjukkan tiada perbezaan yang signifikan antara nilai sebenar dan nilai ramalan, dengan min dan sisihan piawai seperti berikut: Sebenar (Min [SD] = 0.600 [0.940]) dan Diramalkan (Min [SD] = 0.940 [0.030]); df = -2.154(9). Penemuan kajian ini akan menyumbang kepada metodologi penyelidikan epidemiologi, terutamanya pemetaan hubungan, dengan memperkenalkan model bersepadu. Berkaitan MAD, PMSE, dan nilai-p, jelas menunjukkan kedua-dua model menunjukkan ketepatan yang tinggi dalam ramalan hasil. Hasil sintaks yang dihasilkan akan mencadangkan proses membuat keputusan yang lebih tepat dalam pencegahan penyakit.

A NEW REGRESSION MODELLING APPROACH AND ITS APPLICATION IN BIOSTATISTICS

ABSTRACT

This research aims to develop an integrated methodology that will be formulated within a Multilayer Feedforward Neural Network (MLFFNN) framework and logistic regression. The mean absolute deviation and predicted mean square error will be utilised to evaluate the performance of the integrated model, and it serves as a yardstick to determine the accuracy and efficiency of the forecasting that is achieved as a result. The urgency of better significant results serves as a motivation for this study. The objective of this study is to develop and implement an integrated model combining Bootstrap and MLFFNN with logistic regression modelling (LRM) to achieve better prediction accuracy and interpretability. The integrated method used in this study is based on the principles of Bootstrap, LRM, and MLFFNN. The accuracy of the proposed technique is assessed using the Predicted Mean Squared Error Neural Network (PMSE.net), Mean Absolute Deviance (MAD), and the accuracy percentage. Each of these components acts as a benchmark for assessing the precision and effectiveness of the created model. A *t-test* was used to explore the difference between actual and predicted values from the models. Data analysis was conducted using the R program and SPSS version 26. Two case studies from dental public health have been used to validate this newly developed integrated model, i) dental caries among preschool children, and ii) The case study for oral health knowledge among mothers of preschool children. The incorporation of bootstrapping, MLFFNN, and logistic regression in an integrated approach enhances the accuracy of parameter estimation and addresses the uncertain relationship between dependent and independent variables. In the case study focusing on dental caries among preschool children, the Mean

Absolute Deviation (MAD) is 0.0221126 and the Predicted Mean Squared Error (PMSE) is 0.07909. A paired sample t-test reveals no significant difference between the actual and predicted values, with means and standard deviations as follows: Actual (Mean [SD] = 0.30 [0.483]) and Predicted (Mean [SD] = 0.31 [0.373]); df = -0.067(9); *p*-value > 0.05. In the study concerning oral health knowledge among mothers of preschool children, the MAD is 0.05303337, and the PMSE is 0.053033. Results from the paired sample t-test indicate no significant difference between actual and predicted values, with means and standard deviations as follows: Actual (Mean [SD] = 0.600 [0.940]) and Predicted (Mean [SD] = 0.940 [0.030]); df = -2.154(9). This study's findings will considerably contribute to epidemiological methodology research, particularly relationship mapping, by introducing an integrated model. Concerning MAD, PMSE, and *p*-value, these indicate both models showed high accuracy in outcome prediction. The significance of the produced syntax outcome will suggest a more accurate decision-making process in disease prevention.

CHAPTER 1

INTRODUCTION

1.1 Preview of the chapter

This chapter provides a comprehensive summary of the thesis. It begins by elucidating the background of the study and proceeds to address the research problem and the underlying motivation. The subsequent section delves into the rationale behind the study, while the conceptual framework is elucidated through a systematic flowchart depicting the entire investigative process. The research hypothesis is expounded upon in the subsequent segment.

Moving forward, the objectives of the study are categorised into general and specific objectives. The chapter systematically explores the scope of the study, methodology, and contributions in alignment with the defined research objectives. It further acknowledges the limitations inherent in the study, concluding this section with an overview of the organisation of the thesis, and outlining the detailed methodology employed in conducting the analysis.

1.2 Background of the study – A review of statistical modelling

In medical research, Multilayer Feed-Forward Neural Networks (MLFFNNs) have encountered extensive applications. MLFFNN is a perception-based interconnection where computations and data move in one direction, from the input to the output. Logistic regression is a statistical technique specifically tailored for binary classification scenarios, wherein the dependent variable assumes categorical binary values. Employing the logistic function, this method transforms a linear combination of input features into a probability value, indicating the likelihood of an instance belonging to the positive class (Dreiseitl & Ohno-Machado, 2002). In contrast to

logistic regression, MLFFNN, with its multilayer architecture and capacity for non-linear transformations, offers a more versatile approach to regression and classification tasks. While logistic regression is adept at binary classification, MLFFNN extends its applicability to both binary and multiclass classification challenges, showcasing enhanced adaptability to intricate data patterns through its network structure and activation functions.(Warner & Misra, 1996) Good and accurate decision-making is the result of using an excellent approach, which is crucial in data analysis. To bring theory and practical programming into harmony, it is necessary to accentuate the underlying problem of computation precision and accuracy. This will guarantee that the computational capabilities can function effectively and efficiently while offering high-quality results for the objective of the study. Applications of statistics have grown essential in many scientific domains, including biology, corporate management, economics, and education (Sasmita *et al*., 2023; Kunz & Wirtz, 2023; Thapliyal & Nautiyal, 2024; Gofman & Jin, 2024; Stephany & Teutloff, 2024).

Biostatistics, a subfield of applied statistics, uses statistical techniques to infer correlations and make conclusions in the biological sciences (Hamilton & Kingston, 2024). Regression analysis plays a crucial role in scientific experiments, enabling the exploration of relationships between variables, forecasting future values, and identifying factors influencing outcomes (Donnelly *et al*., 2024).In 1805, Gauss and Legendre invented the least-squares approach, which is commonly referred to as regression(Stephen S.M., 1981). When examining the possibility of having tall parents and/or children, Francis Galton noticed that the average height of children “regressed” towards the average height of the population. This led Galton to develop the word “regression” (Krashniak & Lamm, 2021). Subsequently, Karl Pearson found that

regressing tall and short children to the mean height of the population yielded consistent results (Krashniak & Lamm, 2021). The initial investigations that provided the groundwork for the regression approach heavily relied on biostatistics. One such initial implementation was Galton's research on the seed weights of sweat peas. Since Galton's contributions, research has continued, leading to breakthroughs in regression techniques (Krashniak & Lamm, 2021). Nevertheless, the contemporary definition of regression deviates from Galton's and includes the analysis of the relationship between variables to estimate and forecast the average or mean of a population (Soufya & Assari, 2020; Senn, 2011). Regression analysis is a flexible statistical method that may be used in a wide range of areas, including finance, social sciences, physical sciences, chemistry, and health sciences (Sun *et al*., 2023). The main aim of this approach is to predict the value of the dependent variable by evaluating the correlation between the independent variables (Sun *et al*., 2023). Regression analysis is often used by medical researchers to provide precise diagnoses; the two main models used in this process are linear and non-linear regression (Feng *et al*., 2022). In a more complex form, multiple linear regression considers many explanatory factors. However, the association between a categorical variable and one or more categorical independent variables is evaluated using Logistic regression modelling (Feng *et al*., 2022). The integration of a multilayer feedforward neural network (MLFFNN) with regression modelling presents a synergistic approach that capitalises on the respective strengths of both methodologies. MLFFNNs excel in capturing intricate, nonlinear relationships, contributing to enhanced predictive accuracy when compared to conventional regression models. This amalgamation provides a flexible modelling framework capable of adapting to diverse data patterns, thereby accommodating complex relationships that might elude simpler regression models. Additionally,

MLFFNNs automatically extract relevant features, facilitating regression models in identifying crucial variables for improved interpretability. The combination addresses the limitation of regression models in handling nonlinearity, ensuring efficient treatment of complex relationships and resulting in more accurate predictions. Furthermore, the collaborative model demonstrates improved generalisation performance by leveraging the complementary strengths of both MLFFNNs and regression, culminating in enhanced predictive capabilities on unseen data(López-Martín, 2015). Additionally, the robustness of MLFFNNs complements the interpretability offered by regression models and provides a more comprehensive understanding of the influential variables (Dreiseitl & Ohno-Machado, 2002). Previously, a study used multiple Logistic regression modelling(LRM) to examine the prevalence of Human papillomavirus infection together with Oral Squamous Cell Carcinoma (Yaaqob *et al*., 2019). Ahmad *et al.*, (2023) invented and employed an integrated methodology, in a similar context intending to overcome the limitations of the previous study method and improve the accuracy of models by integration approaches. After a while, A novel study conducted by Adnan *et al.* (2023), employed a highly effective combination method integrating the LRM and multilayer perceptron to analyse simulated forensic data, highlighting its remarkable efficacy. However, no study has been found from the current literature review implementing the integrated method of MLFFNN with LRM in the dental public health field. Most of the studies regarding risk factors of dental caries prediction and assessing predictors of oral health knowledge used a single statistical technique. Considering these circumstances, in dental public health, the current study aims to integrate the MLFFNN with LRM for two cases in the field of dental public health. Additionally, bootstrapping is applied

to improve the predictability and accuracy of regression models, offering valuable insights for researchers in terms of validation and prediction.

1.3 Research motivation and problem statement

The development of a completely automated computerised medical diagnostic system continues to be difficult owing to the complexities inherent in medical diagnosis. Advancements in intelligent systems, driven by Artificial Intelligence (AI) approaches, have created opportunities for wider use of computers in medical diagnostics (Fernandes et al., 2020; Mirbabaie et al., 2021; Kumar et al., 2023). The advent of computer-based solutions in the healthcare industry has resulted in the digitisation of all medical records and the widespread use of computers to observe clinical data. New technologies that use deep learning and machine learning to forecast future health events have attracted a lot of investment (Fernandes et al., 2020; Kumar et al., 2023). This demonstrates the increasing interest in enhancing healthcare using predictive analytics. Clinical prediction algorithms used to be able to identify individuals who were at a higher risk of disease. These models make therapeutic choices and provide patient advice based on patient data. However, since the system is dynamic and scalable, healthcare personnel must deal with issues like interruptions and changing duties (Wang et al., 2018; Mirbabaie et al., 2021). The motivation of medical professionals to diagnose diseases can wane due to the complexities of medical data, particularly for those with limited diagnostic knowledge. The diagnostic process is further complicated by time constraints, rapid disease progression, and patient condition variability(Udegbe & Ekesiobi, 2024). However, timely and accurate diagnosis becomes essential for prompt treatment and patient safety(Udegbe & Ekesiobi, 2024). Predictive analytics has become critical in healthcare, significantly

enhancing disease prognosis accuracy and saving time (Alowais et al., 2023). Accurate predictions can save lives, whereas inaccuracies can ruin patient well-being. Therefore, it is necessary to assess and forecast diseases with high accuracy (Badawy *et al*., 2023). Therefore, reliable, accurate and effective techniques for healthcare predictive analysis are needed and have become a prime concern for health policymakers. Besides, In the discipline of dental public health, researchers commonly investigate predictors of oral diseases and other oral health behaviours. In these investigations, conventional LRM were widely used. These models were well-known for their interpretability and simplicity. Nevertheless, these models may encounter difficulties in capturing the intricate, non-linear relationships seen in the extensive data used in dental public health. The use of a neural network variation called MLFFNN allows for the investigation of complex patterns and nuanced correlations within the data for its ability to accurately identify subtle patterns (Amin & Noor, 2024). Considering the constraints of conventional models, the integration of MLFFNN, a neural network variation is applied in combination with regression modelling to get better accuracy in this research.

The integrated LRM and MLFFNN modelling applied in this research will present a departure from traditional approaches by enabling the prediction of having dental caries based on surveys or fundamental information before undergoing a detailed diagnosis by a specialist among preschool children. This innovation holds the potential to substantially reduce the human resources, time, and costs associated with oral health examinations. The capacity of this modelling to classify individuals at high risk allows for precise diagnosis and targeted treatment from specialists, further optimizing resource allocation. Moreover, the predictive model's identification of influential factors impacting dental caries and subsequently oral health knowledge

offers an avenue for caries prevention by effectively controlling these identified factors and underscores its role in optimizing oral health preventive strategies. The objective of this study is to create and implement an integrated model combining MLFFNN with regression modelling using data from two case studies. In the first case study, predictors of dental caries among preschool children and in the second case study, predictors of oral health knowledge among the parents of preschool children will be explored. This study design effectively combines the advantages of both techniques and is then applied to dental public health. This integrated model enhances the prediction accuracy of MLFFNN while simultaneously providing the clear interpretability of LRM. This study presents an innovative method that combines many statistical tools to thoroughly analyse the patterns and changes in diseases across the Malaysian population, with a particular focus on oral health. This integrated model is positioned to provide a possible resolution to the inherent difficulties in oral disease modelling. Furthermore, by using the R statistical software, this research work significantly advances the fields of oral disease modelling and mapping and ensures accurate forecasts and a deeper understanding of the factors that affect health outcomes. Moreover, this prediction paradigm provides improved accuracy and important insights into assessing predictors of oral health outcomes. As a result, it supports informed decision-making in dental public health interventions and policies by enhancing model accuracy and performing disease prediction rapidly. The exploration of newly developed statistical methodologies is not common in the field of dental public health. The investigation of the use of integrated statistical methodologies in assessing risk factors for dental caries and variables associated with oral health knowledge, attitudes, and practices among the Malaysian population is lacking. The current study attempts to bridge the present research gap by constructing

and implementing an integrated model to reduce the functional discrepancy between the two existing techniques of Bootstrapping, MLFFNN and LRM. The effectiveness of this recently developed method will be evaluated by quantifying the Predicted Mean Square Error (PMSE), Mean Absolute Deviance (MAD), and accuracy. Following the coordination of the corresponding routines, these measurements will be retrieved. Furthermore, the suggested integrated model may be used to examine the presence of a "cause-and-effect" relationship and establish connections with many relevant components that contribute to the situation under investigation.

1.4 The rationale of the study

There is a compelling justification for integrating LRM with neural networks in dental public health applications since both methods provide distinct benefits. Neural networks excel at comprehending intricate patterns and traits, making them very proficient in discerning complicated, non-linear relationships within extensive datasets. It is advantageous to use this approach when examining diverse patient data about dental well-being, including medical records, diagnostic images, and lifestyle preferences. Nevertheless, a neural network is a complex mathematical structure with many linked layers, and it may be challenging to grasp how it arrives at a certain conclusion or prediction. The absence of transparency is disconcerting, especially in critical sectors such as healthcare, where the confidence and support of patients and professionals rely significantly on the capacity to understand and evaluate information. The architecture gains a transparent and comprehensible layer that facilitates understanding the decision-making process of the model by using LRM. Through the integration of the neural network's predictive power and logistic regression's interpretability, this method promotes acceptance and confidence

amongst medical professionals and patients. By combining these methodologies, there is the potential to enhance treatment planning and increase the accuracy of diagnoses in oral health applications. The topic of dental public health lacks a distinct research methodology, especially in the areas of method construction and implementation. This research presents three novel methodologies: Bootstrapping, MLFFNN and LRM. In the contemporary period, there is a growing need for highly dependable and accurate models. Numerous advancements in research methodology do not have a robust mechanism for validating models, such as bootstrapping. It is essential to devise strategies to bridge this gap, which may include bringing together current methodologies or enhancing and modifying existing ones. The current scenario identified that the application of integrated statistical methodologies for assessing risk factors for dental caries and variables related to oral health knowledge, attitudes, and practices among the Malaysian population is unexplored. This study endeavours to bridge this gap by developing, validating, and implementing an integrated model consisting of Bootstrapping, MLFFNN, and LRM, in predicting dental caries and oral health knowledge to obtain better accuracy. An integrated model of MLFFNN, bootstrapping, and LRM can potentially address issues of multicollinearity and missing values more effectively than using individual techniques alone. MLFFNNs can manage complex, non-linear relationships and handle imputed missing data, whereas bootstrapping provides stable estimates by averaging over multiple resamples, thus mitigating the effects of multicollinearity. Logistic regression, particularly with some regularization techniques, can further address multicollinearity and improve model interpretability. By combining these approaches, the integrated model leverages the strengths of each method, providing a more robust solution to these common data issues.

1.5 Conceptual framework of the research study

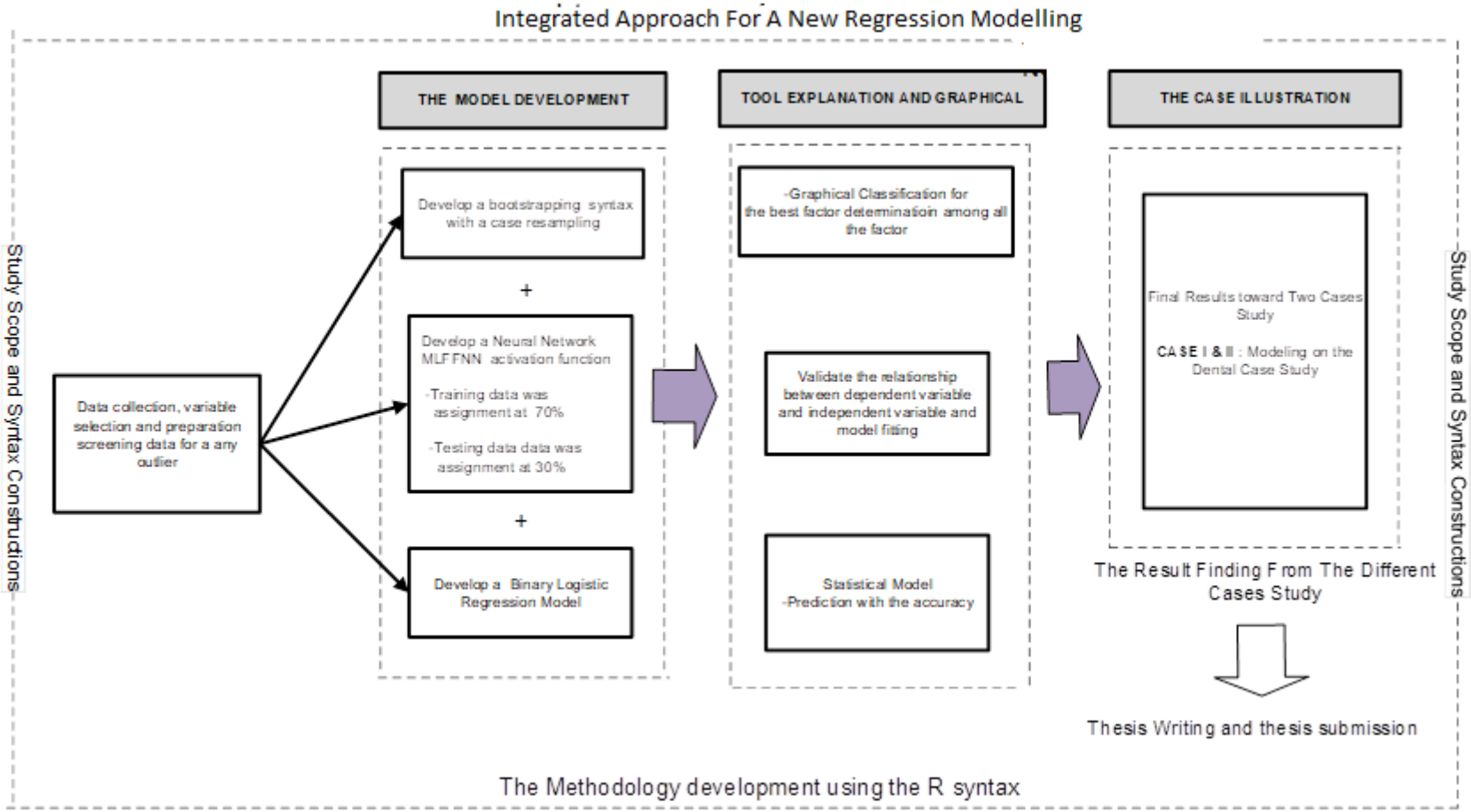

Figure 1.1 Conceptual framework of the study

1.6 Research hypothesis

This research is crucial in advancing the fundamental principles of statistical computing methodology. The hypotheses being tested are as follows:

a. The integrated model is thought to be able to improve the depth and effectiveness of analysis.

b. The study suggests that the integrated model can become a standard in the field of dental caries modelling and oral health knowledge modelling.

1.7 Research Objectives

1.7.1 General Objective

To develop a new dimension of the regression modelling approach and its application in biostatistics.

1.7.2 Specific Objective

1. To elucidate the Bootstrapping, LRM and the MLFFNN for dental public health
2. To design and develop an integration model by taking into consideration Bootstrapping, LRM and the MLFFNN.
3. To implement the newly developed syntax for dental public health

1.8 The scope and methodology

The scope of this research encompasses an in-depth exploration of the factors influencing dental caries and oral health knowledge by employing a novel approach that combines neural network methodology with LRM and Bootstrapping. This study uses a unique strategy that combines Bootsrtrapping, LRM and neural network

methods to investigate the variables that influence dental caries and oral health knowledge in detail. This study rigorously examines a diverse array of factors, such as oral hygiene practices, dietary habits, sociodemographic traits, and access to dental care. The objective is to achieve an extensive understanding of how these factors interact to impact the occurrence of dental caries among preschool children and the extent of oral health knowledge among parents of preschool children. By combining the interpretability of LRM with the non-linear pattern recognition power of neural networks, an integrated statistical model is constructed as part of the process. The datasets were obtained by conducting questionnaires and clinical examinations and were used to train and verify the integrated model. This strategy involves integrating many fundamental statistical methodologies to improve the accuracy of study results. The study involves a unique technique for two separate examples in the field of dental public health. The three components used in model development, MLFFNN, LRM, and data bootstrapping were highlighted in the methodology. The second phase of the strategy is using this comprehensive methodology to examine two separate case studies in the field of dental public health, specifically focusing on dental caries and oral health knowledge. The approach presented notably highlights the use of integrated MLFFNN with LRM with Bootstrapping for both case studies. The examination of both situations entails the assessment of their accuracy and predictivity performance, relying on the acquired findings. The results will not only help identify important predictors of dental caries and oral health knowledge, but they will also give policymakers and healthcare professionals a clear and understandable framework to use when developing focused interventions and educational initiatives to enhance oral health outcomes. Additionally, findings from this study will considerably contribute

to epidemiological methodology research, particularly relationship mapping, by introducing an integrated model.

1.9 The contribution of the study

The contribution of the study may be classified into three main categories. The main emphasis is on developing a complete methodology that can be used to provide predictions for future trends in the field of dental public health. This technique can accurately forecast oral health diseases. This work greatly advances technique development by offering a novel methodology known as integrated LRM and MLFFNN. The second contribution is to enhance the R-syntax for statistical tools. Researchers may develop tailored R-syntax to fit their requirements, hence improving the efficiency of results and facilitating optimal variable selection. Moreover, the syntax offers comprehensive output, aiding researchers in deriving meaningful discoveries, particularly for decision-making. Another notable contribution is the expansion of scientific literature, providing valuable insights for future research initiatives by other researchers. This research work aids inexperienced researchers in comprehending the intricacies of method development and analysis. Furthermore, the study proposes an advanced methodology for modelling, prioritising the attainment of exceptional levels of predictability and accuracy.

1.10 The limitation of the study

The limitations of the study are identified across four components. The initial limitation of the bootstrapping approach is that it only focuses on enhancing parameter accuracy by using case resampling to fine-tune estimated parameters. A second limitation is that the “enter” approach in multiple LRM only allows the inclusion of variables that have been previously established as both clean and clinically significant

from previous research. The third component of the MLFFNN assesses accuracy and predictability by comparing anticipated values from testing data with actual values. It also verifies components identified in regression studies and calculates predictability. The third limitation is software selection, where R-software is solely selected due to its easily approachable interface and unrestricted accessibility, hence restricting the use of all created techniques to this platform. The last limitation is found in the data itself, which only uses secondary sources, which may potentially restrict the ability to reflect on and understand to a lesser extent. The scope is also constrained by the fact that the data used in the two case studies, which investigate dental caries and oral health knowledge among preschoolers in Kelantan, was obtained only from secondary sources.

1.11 The organization of the thesis

The thesis is structured into five chapters, organised as follows:

The first chapter, Introduction, offers a comprehensive examination of the historical context of the study, therefore framing its importance. This chapter provides a more detailed explanation of the study objectives, reasoning, extent, approach, importance, and constraints.

The second chapter, Literature Review, examines and evaluates statistical techniques and variables that may be used. This chapter also examines the statistical approaches used in previous instances of dental public health, with a focus on their limitations.

The third chapter, Methodology, provides detailed information on the techniques and statistical models used, such as MLFFNN, LRM, and Bootstrap. This chapter provides a comprehensive overview of the research design, including the

location, duration, and case studies. Additionally, it presents a flow chart that is based on the suggested statistical logistic modelling.

The fourth chapter, Results, provides comprehensive findings and analyses for each case study.

The fifth chapter, Conclusion, provides a summary of the conclusions derived from the results and delves into the analysis and interpretation parts of model construction. It also offers suggestions for future study improvements.

CHAPTER 2

LITERATURE REVIEW

2.1 Preview of the chapter

Sections 2.2 to section 2.3 offer a comprehensive exploration of Logistic Regression and multiple LRM in the dental public health field. The technique of bootstrapping is detailed in section 2.4. Section 2.5 delves into the history, advantages, and drawbacks of the (MLFFNN). Emphasising the limitations of statistical approaches, section 2.6 underscores the necessity for a combined strategy. The application of an MLFFNN in public health is depicted in section 2.7. Section 2.8 highlights the previous study design of the combined method involving MLFFNN and LRM in public health. Additionally, section 2.10 provides an overview of the specific implementation of the statistical methods. Brief explanations of both case studies involving past statistical methods are outlined in sections 2.10.1 and 2.10.2. Researchgap is provided in section 2.11 The concluding section offers final remarks.

2.2 History of regression

Regression analysis has accumulated a significant amount of information for several decades, moving from simple correlation studies to the application of advanced statistical techniques common in modern research. Since its inception some centuries ago, this analytical approach has seen revolutionary developments that have greatly impacted its methodology and expanded its application across many fields. A succinct synopsis illuminates this path, revealing the gradual improvement and modification of regression analysis as it evolved into a crucial tool for comprehending relationships, forecasting results, and revealing insights in other fields. The ensuing discussion investigates the historical progression of regression analysis.

During 18th Century

Carl Friedrich Gauss and Adrien-Marie Legendre both separately developed the least squares approach in the late 18th century, which is a fundamental part of regression analysis. The technique of least squares was presented by Carl Friedrich Gauss in the early 19th century and independently refined by Adrien-Marie Legendre around the same period. His work established the fundamental principles of Ordinary Least Squares. The approach entails reducing the total of the squared disparities between observed and anticipated values, offering a reliable means of estimating the parameters of a mathematical model(Stephen, 1981).

During 19th Century

Regression analysis originated from the study carried out by Sir Francis Galton, a cousin of Charles Darwin, in the late 19th century. Galton examined the relationship between the heights of parents and their children, using the word "regression" to describe the pattern of extreme values moving toward the average. He made substantial contributions to the field of regression analysis. During his research of characteristic inheritance in the late 19th century, he generated his concept of regression to the mean. Nevertheless, his contributions to regression analysis were not extensively acknowledged throughout that period (Krashniak & Lamm, 2021). The foundation for the regression notion was laid by his seminal study, which was summarised in the 1886 publication "Regression Towards Mediocrity in hereditary stature," which was published in the Journal of the Anthropological Institute. Statistical Galton shared his results and observations in this publication. He invented the phrase "regression towards mediocrity" to explain the statistical phenomenon in which extreme values in a dataset tend to migrate or regress towards the mean or average. This research examines the tendency of extreme results to move closer to the

average, which is a significant milestone in the advancement of statistical analysis (Galton, 1886; Gorroochurn, 2016).

Regression analysis and statistics in general have benefited greatly from Ronald Fisher's contributions, which are factual and provide a succinct assessment of his impact. The evolution of statistical methods was greatly influenced by Fisher's broad competence in genetics, statistics, and evolutionary biology. Fisher's multidisciplinary expertise as a statistician, geneticist, and evolutionary biologist enabled him to provide a distinctive viewpoint to the study of statistics (Esposito, 2016). His statistical work was impacted by his ideas from genetics, and vice versa. Fisher's work on the analysis of variance (ANOVA) was revolutionary. ANOVA is a statistical technique that divides the variability in data into distinct parts, offering a robust tool for comparing averages across many groups. This approach is essential in the process of experimental design and hypothesis testing. "The Statistical Methods for Research Workers" (1925) by Fisher is a renowned and influential work on the subject. Published in 1925, this work provided a comprehensive overview of statistical techniques, including regression analysis(Fisher, 1925). The book underlined the significance of robust statistical techniques and meticulous experimental design in scientific investigation, ultimately exerting a substantial impact on the development of statistical methodology. To guarantee reliable statistical conclusions, Fisher was a strong supporter of meticulous experimental design, placing a strong emphasis on randomization and control. His concepts about experimental design have established a fundamental basis in the investigation of science. Regression analysis and other theoretical underpinnings of statistics continue to benefit from Sir Ronald Fisher's influence on modern research, and statistical practice has adopted his methodologies. In many publications, including his 1922 paper "On the Mathematical Foundations of

Theoretical Statistics" Sir Ronald A. Fisher's significant contributions to statistics have been observed however, it is important to remember that his work was more focused on statistical inference, analysis of variance (ANOVA), and maximum likelihood estimation rather than on the development of Ordinary Least Squares (Williams, 2014). As part of his research on maximum likelihood estimation and likelihood-based inference, Fisher created the likelihood ratio test (Aldrich, 1997). The ratio of the likelihoods under the alternative hypothesis and the null hypothesis serves as the foundation for the test and is known as the likelihood ratio statistic (Aldrich, 1997). The Goodness of Fit of Regression Formulae (1922) and The Theory of Statistical Estimation (1925) are two of Fisher's foundational books that include his contributions to the likelihood ratio test (Rao, 2019). Fisher created the likelihood function as a metric to quantify the level of evidence that the observed data provide for various parameter values in a statistical model (Fisher, 1925; Rao, 2019).

In contrast, Abraham Wald made significant advances in econometrics and statistical decision theory. Among his contributions are sequential analysis and the creation of Wald tests. Although he did not originate the likelihood ratio test, he made substantial advances to statistical theory independently. A major contribution to the theoretical foundations of regression and hypothesis testing came from Wald's work on statistical decision theory and sequential analysis. His 1943 work "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large" has had a big impact (Wald, 1943). The theoretical foundation of regression analysis was significantly shaped by Fisher and Wald's contributions, which also gave statisticians and other academics vital resources for developing and evaluating hypotheses. In addition to advancing regression analysis, Wald's work has

wider ramifications for decision theory and hypothesis testing, strengthening the statistical underpinnings of many scientific fields (Wolfowitz, 1952).

Galton's theoretical model of heredity involved variables with defined means, irregular deviations from it, and an aggregate forming a normal distribution. This model became instrumental in the formulation of regression analysis. Galton's experiments with peas and humans revealed the law of reversion or regression, expressed mathematically as a linear equation. The mean of the offspring deviated from the population mean and returned proportionately to the displacement of their parents (Krashniak & Lamm, 2021).

During 20th Century

Galton's groundbreaking investigation into the hereditary characteristics of sweet peas played a pivotal role in shaping the evolving concept of linear regression. This influence is discernible through a thorough examination of the literature reviews by Sir Francis Galton and Karl Pearson. Following this, the collaborative efforts of Galton and Pearson expanded the practical applications of these initial concepts, leading to the development of more extensive methodologies such as multiple regression and the product-moment correlation coefficient. (Krashniak & Lamm, 2021) Notably, correlation principles are often introduced and clarified in current instructional materials before getting into the intricacies of prediction issues and the real-world applications of linear regression (Barnes, 1994). The field of regression analysis continued to evolve with advancements in statistical theory and computing. The theory of estimate and hypothesis testing was advanced by Jerzy Neyman and Egon Pearson (son of Karl Pearson), strengthening the statistical grounds of regression. Jerzy Neyman and Egon Pearson were influential statisticians who made significant contributions to the development of statistical theory, particularly in the

context of hypothesis testing. The Neyman-Pearson lemma, developed by Jerzy Neyman and Egon Pearson in the early 20th century, laid the foundation for a rigorous approach to hypothesis testing (Neyman & Pearson, 1933). The lemma introduced the concept of likelihood ratios as a basis for making decisions about hypotheses. (Ronald, 2014). The key idea was to compare the likelihood of the observed data under different hypotheses to guide the testing procedure. With the advent of computers, regression analysis became more accessible and applicable to a broader range of fields. Multiple types of regression models, including linear and non-linear forms, were developed. The field expanded to include logistic regression, robust regression, and other specialised techniques (Ronald,2014; Helmreich, 2016)

During 21st Century

Regression modelling is still an essential method in data science and statistics. Advances in computational power, along with the availability of vast amounts of data, have led to the development of more sophisticated regression models and techniques, including machine learning algorithms that incorporate regression principles.

Modern Developments and Applications (21st Century)

In recent times, regression analysis has been extensively used in various fields, including agriculture, economics, finance, biology, social sciences, and machine learning (Storm *et al*., 2020; Lockhart, 2023; Bluwstein *et al.,* 2023) *et al.,* 2023). In addition, machine learning techniques, such as linear regression and logistic regression are widely applied for prediction and modelling in data science (Malekzadeh, 2023).

2.3 Statistical content using multiple logistic regression model in and dental public health modelling

The application of statistical modelling through a multiple logistic regression model is instrumental in the field of dental public health. This method allows for the examination and analysis of various factors simultaneously to understand their influence on specific outcomes or conditions. By utilising multiple LRM, researchers can assess the relationship between multiple independent variables, such as oral health behaviours, fluoridated toothpaste use, dental visits, and social determinants like religion, while considering their collective impact on oral health outcomes. This approach provides a comprehensive understanding of the complex interplay of factors affecting dental public health, facilitating more informed decision-making and targeted interventions (Preisser *et al*., 2012; Javali & Pandit, 2012; Boateng & Abaye, 2019).

A worldwide survey investigated the contribution of multiple factors like oral health behaviours, the use of fluoridated toothpaste, and frequency of dental visits in oral health outcomes. In this study, the multiple regression method has been used to investigate the association between factors contributing to oral health disparities all over the world among all age group populations. The study findings also revealed social factors such as social conditions, environment, economic conditions, and religion play a significant role in influencing oral health inequalities (Chaudhary *et al*., 2024). Numerous studies have employed multiple LRM analyses to investigate the factors influencing oral health outcomes. This analytical approach allows researchers to explore the simultaneous impact of various variables on oral health conditions. These studies often consider factors such as oral health behaviours, utilisation of fluoridated toothpaste, frequency of dental visits, and social determinants like

economic conditions and environmental influences. By employing multiple LRM, researchers can identify and quantify the relative importance of each factor, providing a more nuanced understanding of the complexities involved in determining oral health outcomes. Such studies contribute valuable insights to the field of dental public health, helping to inform preventive measures, intervention strategies, and health policies aimed at improving overall oral health in populations (Stangvaltaite-Mouhat *et al*., 2024). For example, a prevalence study aimed to examine the connection between chewing gum and improved oral health status among US citizens. Employing logistic regression, the study explored the relationships between self-reported chewing gum habits and the oral health status of adults. Study findings indicated that self-reported use of chewing gum by American adults did not show any significant impact on their oral health status (Lu *et al*., 2024).

Recently another study applied a multinomial LRM to assess the association between Sense of Coherence (SOC) scores of Saudi Arabian mothers and the oral health behaviour of children with Special Health Care Needs. Exploratory variables of this study included the type of motherhood, the mother's educational status, and monthly family income whereas SOC was considered as outcome variable. From this study, it was observed that individuals with monthly incomes less than 5000 SAR and those in the 5000-10,000 SAR range were more inclined to score lower on the SOC scale. Intriguingly, in the present study, the educational level of mothers did not appear to be a confounding factor in children's oral health behaviour. Additionally, a noteworthy association was observed between the monthly family income and the frequency of children's consumption of sugary drinks and their brushing habits. Additionally, the findings indicated that children from households with higher income levels were significantly less prone to consuming sweet drinks. These results

underscore the complex interplay of socio-economic factors in influencing children's oral health behaviours, with family income demonstrating a notable association with both dietary habits and oral hygiene practices (Iyer *et al*., 2024).

Currently, another study applied multiple LRM in dental public health among the Japanese population. In this study, the potential relationship between the experienced stressful life events and their association with the occurrence of oral health issues including tooth pain, gum swelling or gum bleeding, and difficulty chewing was explored. The analysis has been performed while accounting for the influence of covariates such as gender, age, and concurrent medical conditions in this study. The utilisation of logistic regression facilitated the estimation of the magnitude and direction of this association, providing valuable insights into the interplay between stress and oral health problems within the studied population (Aoki *et al*., 2023).

Furthermore, a literature review identified a research endeavour investigating the determinants of periodontal disease within the Saudi population. This study was conducted in the field of dental public health, indicating a focused examination of factors influencing periodontal health among male patients in Saudi Arabia. Results from the binary logistic regression analysis from this study data demonstrated a significant association between age, nationality, risk factors, and the type of periodontal disease. According to the formulated model, individuals aged less than twenty-six years were found to be almost seven times more likely to exhibit gingivitis than periodontitis. Furthermore, patients with plaque and calculus had an eight times higher likelihood of developing gingivitis compared to periodontitis. The accuracy of the model was determined from this study to be almost 80%. The logistic regression model employed for predicting periodontal diseases provides valuable insights into