A STATISTICAL COMPARISON BETWEEN GENETIC ALGORITHM AND LOGISTIC REGRESSION FOR A CLINICAL STUDY

Aththanayake Mukaweti Sahabandu Mudiyanselage

Chathuri Malee Aththanayake

(179054E)

Dissertation submitted in partial fulfilment of the requirements for the degree of
Master of Science in Business Statistics

Department of Mathematics

University of Moratuwa

Sri Lanka

December 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other media. I retain the right to use this content in whole or part in future works (such as articles or books)

………………………… ……….......

Signature Date

A.M.S.M.C.M. Aththanayake

The above candidate has carried out research for the Masters dissertation under our supervision.

………………………… ……….......

Prof. W. B. Daundasekera Date

Department of Mathematics

Faculty of Science

University of Peradeniya

Peradeniya

…………………………… ……………….

Dr P.M. Edirisinghe Date

Department of Mathematics

Faculty of Engineering

University of Moratuwa

Moratuwa

ACKNOWLEDGEMENT

My dissertation work of M.Sc./Diploma in Business Statistics could not have been accomplished without the support and cooperation of all those who encouraged me during my project to make it successful.

First of all I would like to convey my sincere gratitude to my supervisor Prof. W. B. Daundasekera, Department of Mathematics, Faculty of Science, University of Peradeniya for his continues guidance, inspiration and advices throughout this study. I would further like to extend my gratitude to my internal supervisor Dr. P.M. Edirisinghe, Senior Lecturer, Department of Mathematics, Faculty of Engineering, University of Moratuwa for his guidance throughout this work. Furthermore, Prof. T. S. G. Peiris, Senior Professor in Applied Statistics and Former Head of the Department, Department of Mathematics, Faculty of Engineering, University of Moratuwa for the guidance and support provided as the Course Coordinator of the Degree Program of M.Sc in Business Statistics.

Moreover, I gratefully acknowledge Prof. Noordeen, Faculty of Medicine, University of Peradeniya for giving data for this study. Finally, I would like to thank non-academic staff members of the Department of Mathematics, University of Moratuwa for their invaluable help and support.

At last but not least, I would like to thank my parents, siblings and my colleagues, who were always there with me in the difficulties. It was their unshakable faith in me that has always helped to proceed further.

A. M. S. M. C. M. Aththanayake

ABSTRACT

A Statistical Comparison between Genetic Algorithm and Logistic Regression for a Clinical Study

Identifying a combination of variables causing infections or infectious diseases is one of the main tasks in clinical models in medicine. Forward and backward variable selection techniques in Logistic Regression (LR) are widely used in such situations, where it assumes linearity of independent variables and the absence of multi-collinearity. More often, the observed data do not satisfy these assumptions and thus, LR is not applicable. Hence, the Genetic Algorithm (GA), which does not depend on pre-defined assumptions, has proven to be better under such circumstances. By evaluating prediction rates of LR and GA techniques, the objective of this study was to perform binary LR and GA to reduce the number of variables on a sample of clinical data and compare the goodness of fit statistics to identify the better variable reduction method. Three models were built using 40 independent variables (3 non-categorical and 37 categorical) for a sample of 497 observations collected from suspected respiratory syncytial virus (RSV) infected children under 5 years of age, who were hospitalized to the Kegalle Base Hospital from May 2016 to July 2018. The binary dependent variable indicates whether the suspected child is infected with RSV positive or negative. Log-likelihood and Area Under Curve (AUC) represent the fitness functions of two GAs. The goodness of fits on the three models was compared using statistical measurements: -2log-likelihood, Psudo R-square values, Correctly Classified Percentage, Specificity, and Sensitivity. Results shown that Log-likelihood GA produces better goodness of fit measurements compared to other the two methods. However, LR reduces 40 variables into 8 with lower number of iterations while two GAs reduced into 17 variables to predict the status of RSV infection. This study suggests that the LR has a better prediction power with the most associated combination of variables. However, GA indicated better in analysing when the predefined assumptions were not satisfied and solving high dimensional classification problems in a large or complex searching space in the background of the study.

Keywords: Clinical Data, Fitness Function, Genetic Algorithm, Logistic Regression

TABLE OF CONTENTS

DECLARATION i

ACKNOWLEDGEMENT ii

ABSTRACT iii

TABLE OF CONTENTS iv

LIST OF FIGURES vii

LIST OF TABLES viii

LIST OF ABBREVIATIONS ix

CHAPTER 1 1

1.1 RSV Disease Burden 1

1.2 RSV in Sri Lanka 1

1.3 Importance of Predictions 2

1.4 Significance of the Study 3

1.5 Research Objectives 4

1.6 Organization of the Thesis 4

CHAPTER 2 5

2.1 Logistic Regression Analysis to Forecast the Response Variable 5

2.2 Logistic Regression Model Selection Using Genetic Algorithm 7

2.2.1 Fitness Function: Statistical Techniques 8

2.2.2 Fitness Function: Machine Learning Technique 8

2.3 Regression Model Selection Using GA 10

2.4 Time Series Forecasting Using GA 10

2.5 Discriminant Analysis Using GA 11

2.6 Summary 11

CHAPTER 3 12

3.1 Secondary Data 12

3.2 Demographic and Climate Data Analysis 12

3.3 Methods of Analysis.. 13

3.4 Descriptive Statistics for Categorical Variables... 13

3.4.1 Contingency Tables (Cross Tabulations) for the Dichotomous Dependent Variable.. 13

3.4.2 Chi-Square Statistic 13

3.4.3 Kruskal Wallis Test.. 14

3.5 Binary Logistic Regression 14

3.5.1 Logistic Regression Parameter Estimation 15

3.5.2 Strategies in Model Selection 16

3.5.3 Summary Measures of Goodness-of-Fit 17

3.6 Genetic Algorithm.. 20

3.6.1 Selection.. 22

3.6.2 Crossover 22

3.6.3 Mutation... 23

3.6.4 A GA for a Variable Reduction in Binary Logistic Regression 23

3.7 Software Packages to Perform Logistic Regression and GA 24

CHAPTER 4 25

4.1 Exploratory Data Analysis of the Dependent Variable RSV 25

4.2 Exploratory Data Analysis of the Explanatory Variables 26

4.2.1 Non Categorical Explanatory Variables 26

4.2.2 Categorical Variables.. 27

4.3 Presence of Correlations with RSV Prevalence and Two General Variables Residence and Ethnicity.. 31

4.4 Seasonality of the Number of RSV Positive Children 33

4.5 Analysis of the Number of RSV Positive Children under Climate Data 34

CHAPTER 5 36

5.1 Variable Selection Procedure 36

5.2 Statistical Approach: Binary Logistic Model 36

5.2.1 Validate the Assumptions 36

5.2.2 Binary Logistic Regression Model Fitting 38

5.3 Approach of the Genetic Algorithm 45

5.3.1 AUC Genetic Algorithm (AUC-GA) 46

5.3.2 Log-likelihood Genetic Algorithm (LL-GA) 50

CHAPTER 6 52

CHAPTER 7 56

7.1 Conclusion 56

7.2 Recommendations 57

7.3 Future Work 57

APPENDIX A: Description of Variables 63

APPENDIX B: Relationship between Independent Variables 66

APPENDIX C: Categorical Variables Coding 70

APPENDIX D: R Code for AUC-GA 72

APPENDIX E: R Code for LL-GA 74

LIST OF FIGURES

Figure 3.1: AUC-ROC Curve 18

Figure 3.2: Flow Chart of a GA 21

Figure 4.1 : RSV Positive Vs Negative... 25

Figure 4.2: Histograms for the Non Categorical Variables 26

Figure 4.3: Pie Chart for the Variable Residence 31

Figure 4.4: Pie Chart for the Variable Ethnicity 32

Figure 4.5: Time Series Plot for Number of RSV Positive Children......................... 33

Figure 4.6: Relationship between Rainfall Data and RSV (+) Counts 34

Figure 5.1:ROC Curve for the Binary Logistic Model 44

Figure 5.2: Adapted GA.. 45

Figure 5.3: Level of Contribution of Explanatory Variables to Predict the Status of RSV.. 47

Figure 5.4: Convergence of Fitness Value on AUC-GA method 47

Figure 5.5: Dependency of each Explanatory Variable 48

Figure 5.6: Convergence of Fitness Value on LL-GA Method 50

Figure 6.1: Reduced Variables to Predict the Status of RSV..................................... 55

LIST OF TABLES

Table 3.1: Sub Categories of Independent Categorical Variables 12

Table 4.1: Frequency Table for RSV 25

Table 4.2: Summary of Non-Categorical Independent Variables 26

Table 4.3: Summary of Chi-Square Test for General Variables 27

Table 4.4: Summary of Chi-Square Test for Symptom Variables 28

Table 4.5: Summary of Chi-Square Test for Risk Factors 29

Table 4.6: Summary for the Dependency of Variables which has an Association with RSV 30

Table 4.7: Summary on Type of Residence in Kegalle District 31

Table 4.8: Summary of RSV Positive Children According to Month 33

Table 5.1:Multicollinearity among Independent Variables 37

Table 5.2: Summary of Goodness-fit Statistics for Six Variable Reduction Methods 38

Table 5.3: Dependent Variable Encoding 40

Table 5.4: Classification Table 40

Table 5.5: Variables in the Equation 40

Table 5.6: Omnibus Tests of Model Coefficients 41

Table 5.7: Model Summary 41

Table 5.8: Hosmer and Lemeshow Test 41

Table 5.9: Variables in the Equation 42

Table 5.10: Classification Table 43

Table 5.11: Statistics on the Confusion Matrix 43

Table 5.12: Area Under the Curve 44

Table 5.13: Final Step of the AUC-GA Method 46

Table 5.14: Fitted Function Value for Several Models 48

Table 5.15: Summary of LL-GA 50

Table 6.1: Summary Statistics 52

Table 6.2: Reduced Variables by Three Methods 54

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ARIMA	Auto-Regressive Integrated Moving Average
ARTI	Acute Respiratory Tract Infections
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristics Curve
BIC	Bayesian Information Criterion
DIB	Difficulty in Breathing
FDA	Fisher Discriminant Analysis
GA	Genetic Algorithm
GARI	Genetic Algorithm Rainfall Intensity
GARS	Genetic Algorithm for Regressors Selection
GARST	Genetic Algorithm for Regressors Selection with the Transformation
LSM	Least Square Method
ROC	Receiver Operating Characteristics Curve
RSV	Respiratory Syncytial Virus
SIC	Schwarz Information Criterion
SOB	Shortness of Breath
VIF	Variance Inflation Factor

CHAPTER 1

INTRODUCTION

This chapter focuses on brief background information about the Respiratory Syncytial Virus (RSV) in Sri Lanka. Furthermore, an introduction to the binary logistic regression model on clinical data and applications of Genetic Algorithm (GA) to reduce the number of variables which describe the status of RSV are discussed.

1.1 RSV Disease Burden

Acute Respiratory Tract Infections (ARTI) is an identified cause for morbidity and mortality of children in developed and developing countries (Cockburn & Assaad, 1973). Across the world, RSV is believed to be the most predominant viral causing ARTI in children under two years old (Weber et al., 1998). Nearly 45% of ARTI infected children under the age of 6 months lost their lives due to the cause of RSV. Also, it can be observed that across the world, 3.2 million children are admitted to hospitals out of which 59,000 died due to various types of new RSV (Shi et al., 2017).

1.2 RSV in Sri Lanka

It is identified that RSV is the most common cause of viral ARTI among Sri Lankan children, which causes a higher number of childhood hospitalizations. Based on the results of a study conducted on children who were admitted to Kegalle General Hospital with symptoms of viral ARTI, RSV was identified as the leading cause of infection with a prevalence of 29.5% (Muthulingama et al., 2014). From the studies conducted at the Anuradhapura General Hospital in the years 2013 and 2014, it is noticeable that more RSV cases were hospitalized from May to July. Also, it can be noticed that in the same years, RSV cases were hospitalized from December to January at Gampola Teaching Hospital (Jayaweera et al., 2016). In a cross-sectional study conducted on patients who are 82.3% of children from a tertiary care hospital in Southern Province of Sri Lanka, RSV was identified as a common viral etiology with a prevalence of 4.2%. RSV activity was found in the peak from May to July in the years 2013 to 2015 (Jayaweera et al., 2016)

1.3 Importance of Predictions

Prediction models are implemented in the medical field in the purpose of decision making to diagnose a disease and to continue or discontinue appropriate treatments and patients' lifestyles. In the healthcare industry, the capability to collect and analysis of large scale data is essential to prevent diseases. Suitable descriptive and inferential statistical approaches were applied to the study of these types of large scale data sets (F.R. & S., 2009). Logistic regression technique is an essential technique that can be used in the healthcare industry. The significance of regression models in the field of clinical practices depends on the simplicity of the model. To reach this goal, the number of variables should be reduced by applying variable reduction methods. The power of predictivity of the model depends on the method which is used to reduce the number of variables in the predictive model. There are some defined methods, such as forward selection method, backward selection method, and elimination methods are described in statistics when selecting the best model.

To apply statistical approaches, data has to follow specific assumptions, such as normality, heteroscedasticity, and stationary. Certain transformations or derivatives can be used to satisfy such assumptions. Accuracy of the original data would no longer exist when taking transformations and derivatives.

Genetic Algorithm (GA) is an alternative approach to the variable selection problem, which was introduced by John Holland in 1975, as suggested by Charles Darwin. GA is one of the evolutionary methods to use as a search method in optimization. GA represents the process of natural selection. The highest fitted individuals are chosen for the reproduction stage to produce offspring of the next generation. A GA is frequently used in the statistical analysis, such as regression models, logistic regression models, time series analysis, discriminant analysis, outliers' detection, graphical model selection, and clustering.

In regression analysis, the Least Square Method (LSM) is used to find the regression coefficients. In the LSM, the objective is to minimize the sum of the errors of the squares. The estimators no longer satisfy these same properties when the LSM is applied to a model with a binary outcome. Maximum Likelihood Function provides the foundation for estimating the parameters of a logistic regression model.

In statistics, a higher number of independent variables with a larger number of individuals, calculating regression coefficients is an NP-Hard problem that takes more computational time. In regression analysis or logistic regression analysis, model selection is performed using stepwise, forward, and backward methods. GA can be used to solve a variety of optimization problems when objectives function is discontinuous, non-differentiable, stochastic, or highly nonlinear. It is an efficient tool to provide suitable near-optimal solutions in a short amount of time.

Furthermore, a GA was used to identify the factors which were highly affected by RSV. A comparison between the adequacy of the logistic regression and GA was made by reporting Log-likelihood value, pseudo-R-square values, and Hosmer-Lemeshow test of the models' respect to the selected number of individual variables. Finally, the best model is selected and used for future predictions.

1.4 Significance of the Study

Statistical approaches are used to extract trends, patterns, relationships, and useful information from a set of existing data. Generally, high-dimension data sets to be considered in the statistical analysis to obtain accurate results. However, when the data set is too large, most of the assumptions which should be satisfied to apply statistical techniques are violated. Therefore, alternative methods would be considered to obtain reliable predictions. GA is one of the alternative algorithms that can be used to have accurate predictions.

GA is widely used in business, scientific, and engineering disciplines. GA searches from a population of points, not a single point. Therefore, the optimization will terminate at the local minimum/maximum. Most of the optimization algorithms assume the objective function to be differentiable, but GA does not. GA is a simple algorithm that can be understood and applied very easily. Besides, most of the real-world scenarios consist of multiple objectives, and GA is an algorithm which supports for the multi-objective optimization. GA uses probabilistic transition rules instead of deterministic rules. Predictions made through GA are accurate. Therefore, responsible authorities can easily take the necessary actions to answer the identified problems and issues according to the predictions obtained by the GA.

1.5 Research Objectives

On the view of the above explanation, the objectives of this study are to:

I. develop a logistic regression model to identify the risk factors for RSV,

II. apply GA into logistic regression modelling to find the combinations of features that produce best-performing predictive models of a virus,

III. carry-out a comparison between the logistic regression model and GA approach to identify the best variable reduction method.

1.6 Organization of the Thesis

In the first chapter, an introduction to this study and a summary background is provided briefly. The second chapter examines the current literature in the field of clinical researches through two aspects called binary logistic regression and the GA approach. Moreover, fitness functions for the GA are discussed clearly in this chapter. Chapter 3 discusses two modelling methodologies called binary logistic regression and the GA approach, which are considered in this study. Different fitness functions for the GA are discussed, and the implementation of each model is described. The fourth chapter provides the preliminary analysis, which gives descriptive statistics of the study. Chapter 5 includes the results of the fitted binary logistic model and implemented GA findings. Moreover, the discussion of the study is also summarized in this chapter. The last chapter states the conclusion drawn from the study and suggests possible directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Logistic Regression Analysis to Forecast the Response Variable

Like all regression analyses, logistic regression is a predictive analysis. The relationship between independent variables and one dependent binary variable is described by the Logistic Regression Analysis. In the literature review, it was observed that a large number of predictive models were built in order to detect the status of the response variable.

Sze & Wong, (2007) used binary logistic regression concepts to determine the associations between the probability of fatality or severe injury and all supportive factors. Hosmer–Lemeshow test (Equation 2.1) and logistic regression diagnostics are used to verify the goodness of fit of the fitted model:

$$H = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{N_g \pi_g (1-\pi_g)} \sim \chi^2_{N-2} \qquad 2.1$$

$n =$ Number of groups

$O_g =$ Observed number of events in group g

$E_g =$Expected number of events in group g

$N_g =$ Total number of observations in group g

$\pi_g =$ Average predicted probability in group g

χ^2_{N-2} : Chi-square with N-2 degree of freedom

Although Sze & Wong (2007)evaluated only the Hosmer-Lemeshow test to find the best-fitted model, Abedin et al.,(2016) have evaluated Hosmer-Lemeshow test statistics as well as another two statistics called Cox & Snell R^2 and Negelkerke R^2. Equations of the aforementioned two goodness-of-fit measurements are stated in equation number 2.2 and 2.3:

$$R^2_{Cox\,\&\,Snell} = 1 - (L_0 \,/\, L_M)^{2/n} \qquad 2.2$$

$$R^2_{\text{Nagelkerke}} = \frac{1 - (L_0 \,/\, L_M)^{2/n}}{1 - (L_0)^{2/n}} \qquad 2.3$$

L_0 = likelihood function value for a model with no predictors.

L_M = likelihood for the model being estimated.

n = sample size.

Moreover, Chauhan et al., (2016) fitted a binary logistic model to forecast whether existing metro users have moved from buses or new alternatives to public transport shifting. Al-Ghamdi, (2002) used logistic regression techniques to estimate the influence factors on accident severity. It is found that logistic regression is a guaranteed tool in providing meaningful interpretations that can be used for future safety improvements.

Similarly, another study on modelling dichotomous outcomes is conducted by Bagley et al., in 2001. Validation analysis, regression diagnostics, and goodness-of-fit measurements were performed to pay greater attention to the use of logistic regression models. Turner et al., (2000) and Abdulqader & Saeed, (2017) used the concept of binary logistic regression in clinical data. Turner et al., modelled the data using the Wald method and likelihood estimation (Equation 2.4) are taken to evaluate the performance of the model. Abdulqader & Saeed, applied a variable selection method and selected the best subsets of explanatory variables to detect Hepatitis disease. Furthermore, the logistic regression modelling technique was used for categorizing people into two different groups, which are influenced and non-influenced by viral Hepatitis disease. Results found that the proportion of visible correctly classified into 98%, which implies a higher capacity of the model for classification.

$$L(\beta) = \prod_{i=1}^{N} [exp\left(\sum_{k=0}^{k} \beta_k X_{ik}\right)]^{Y_i} \left(1 + exp(\sum_{k=0}^{k} \beta_k X_{ik})\right)^{-1} \qquad 2.4$$

N = Sample Size

β_k = Coefficient of the kth independent variable

X = Independent Variable

A study was conducted by Bujang et al., (2018) to select a satisfactory sample size for fitting a logistic regression model. In this study, Bujang et al., proposed a sample size guideline for logistic regression based on observational studies with a large population. Large population size with observational studies that involve logistic regression in the analysis were identified by considering a sample size of 500. Another rule of thumb as $n = 100 + 50i$, where i refers to a number of independent variables in the final model, is recommended in this study.

Midi et al., (2010) and Bergtold et al., (2018) discussed the assumptions of logistic regressions. Midi et al., mentioned that multicollinearity is not uncommon when there are a large number of covariates in the model. Omit one of the correlated variables without increasing sample size is the option given in this study when multicollinearity exists. Furthermore, it was given that examining the correlation matrix is not sufficient to detect the multicollinearity among variables and VIF, condition indices and variance proportions also needed to evaluate for better diagnostics. However, conducting a number of simulations at the presence of multicollinearity, it was found that relatively small sample size can negatively affect the quality of the parameter estimates (Bergtold et al., 2018).

2.2 Logistic Regression Model Selection Using Genetic Algorithm

GA has been applied to statistics in numerous ways. In every situation, the fitness function is identified, whether to maximize it or minimize the negation of its function. Statistical techniques and machine learning techniques are used to improve fitness functions when applying GAs to find optimum solutions. However, there is a tendency to use machine learning techniques in big data analysis to develop the fitness function in GA.

Wu & Norat, (2017) use a GA to predict Medicare payments to physical therapists. In this study, 40,662 observations were gathered under 24 independent variables. The tournament selection method is used with a tournament of size 10 to select a temporary population of parents of the same size as the regular population. Crossover and mutation probabilities are fixed as 0.9 and 0.2, respectively, until stopping criteria is reached (200 generations). The performance of developed canonical GA and a self-adaptive GA is compared with logistic regression. Results show that both

GA approaches are competitive with logistic regression with the canonical GA consistently outperforming logistic regression.

2.2.1 Fitness Function: Statistical Techniques

Bhattacharyya & Pendharkar, in 1998, applied GA to direct marketing response models to evaluate classification accuracy. Fitness of the GA was measured by Hosmer-Lemeshow goodness of fit statistic (Bhattacharyya & Pendharkar, 1998), which calculates the Pearson Chi-square. Furthermore, Sukono et al., in 2014 conducted an analysis of credit scoring for cooperative of financial services is performed using a logistic regression model, which is estimated by GAs using statistical software SPSS 17.0 and Matlab 7.0. Log-likelihood value (see Equation 2.5) is taken as the fitness function of this approach:

$$\ln(L(\beta)) = \sum_{i=1}^{N} \left\{ Y_i \ \sum_{k=0}^{k} \beta_k X_{ik} - ln\left(1 + e^{\sum_{k=0}^{k} \beta_k X_{ik}}\right) \right\} \qquad 2.5$$

2.2.2 Fitness Function: Machine Learning Technique

In medical researches, the GA is frequently used for the detection of diseases. Gayou et al.,(2008) have conducted a study on GA for the variable reduction in Logistic Regression in radiotherapy treatments. The clinical, biological, physiological input factors and dosimetric variables were found from the treatment plan of the lung DVH of each 200 patients. Altogether 17 variables were used for this study and two terms included in the fitness function; one that benefits the overall ability of the prediction power of the model and one that rewards the statistical significance of the variables which were included in the model. The area under the ROC curve (AUC) and the Spearman rank correlation coefficient was the tested fitness criteria. It is mentioned that the two stopping criteria for the GA have reached the defined maximum number of generations, and the minimum penalty has not changed over 50 generations. According to the view of these explanations, the study found that GA provides a trustworthy and efficient way to select significant factors in logistic regression. In 2014 Johnson et al., also did an analysis to recognize the risk factors for Alzheimer's disease. A developed version of AUC (Equation 2.6) is used as the fitness function to predict the progression of the disease by using GA with logistic regression:

$$\text{Fitness Function} = AUC + \rho - \frac{\rho l}{n}, \qquad 2.6$$

where l represents the number of active variables for a particular genome, AUC is the area under the ROC curve, n is the total number of features, and ρ gives the compromise factor. The aim of this study was to identify a combination of risk factors rather than any one alone, which affects the disease. An application has been developed for clinical settings in this study, and different search parameters were tested to achieve the convergence of the fitness function of GA to reach an optimal solution. It is found that optimal values for the parameters are:

Population:50

Mutation Rate: 10%

Crossover rate: 90%

The number of generations: 300.

Finally, the study indicates that variables that were selected frequently by GA might be more important as part of the algorithm to predict the development of the disease. Selected variables were used to fit the logistic regression model to predict the progression of the disease.

Furthermore, Vinterbo & Ohno-Machado,(1999) used a GA to select variables in logistic regression. In this study, the performance of the model is measured by the area under the ROC curve and results were compared by models built with three different variable selection methods (backward, forward, and stepwise). Finally, it was observed that the improvement in the ability of classification yielded by the GA variable selection method was statistically significant ($p < 0.02$).

Another research was conducted by Zhang et al., (2018) to select variables or features in the logistic regression model using a GA. In this paper, 103 independent variables (51 categorical and 52 numerical) with more than 100 observations were generated to discuss the stepwise approach, which is highly used as a variable reduction method, is commonly catch in local optima. According to the study, the subset approach could search the entire space of the covariate pattern, but the set of solutions can be extremely large with a large number of variables, which is the case in clinical studies. Therefore, as a solution, the GA technique by considering AUC as the fitness function is introduced. The analysis is carried out by R.

In addition to logistic regression, the GA is used in many statistical concepts such as regression analysis, time series analysis, and discriminant analysis.

2.3 Regression Model Selection Using GA

A GA is used in regression analysis to fit a model when the data is high-dimensional. Minerva & Paterlini, (2002) and Minghua et al., (2017) have applied GA to linear regression. Minerva & Paterlini measured the optimality by the AIC, BIC, and SIC criteria. But Minghua et al., improved GA by combining two statistical criteria AIC and BIC. It is found that improved GA by Minghua et al., is superior with respect to quality and convergence rate over to classical regression analysis. However, the model built as a GA for Regressors Selection (GARS) was presented by Minerva & Paterlini in 2002. In 2010, their research was developed into a further stage by measuring the performance of two GA for Regressors Selection (GARS) and Regressors Selection with the Transformation of the Regressors (GARST). AIC, BIC, and SIC criteria were used to evaluate the model adequacy, and the results show that GARST is better compared to GARS.

Another study was conducted by Manouchehrian et al., (2013) to apply GA for the selection of best transformed of the independent variables in the regression model to the prediction of strength. Multiple regression analysis (MLR) was performed on a model. It is found that GA models are more precise than MLR models.

2.4 Time Series Forecasting Using GA

GA is used to optimize time series forecasting in order to predict future values. Al-Douri et al., (2018) used GA to implement Auto-Regressive Integrated Moving Average (ARIMA) and forecast error rate. The study implies that GA based on the ARIMA model reflects better forecasting results. Furthermore, Cooper, (2010) used GA to predict rainfall intensity. Results imply that the proposed GA Rainfall Intensity (GARI) model can be used to forecast the rainfall intensity with the lowest mean-squared error between measured and predicted measurements.

2.5 Discriminant Analysis Using GA

GA is used as a key variable identification for classification techniques. Chiang & Pell, (2004) carried out a study to incorporate GA with Fisher Discriminant Analysis (FDA), which was the fitness function for key variable identification. It has been found that GA/FDA correctly gathers the key variables for some particular case studies. In one study, it is found that the classification is incorrect. However, GA/FDA determines that the operating conditions which were related to the case study are different as well as the key variables for the corresponding change. Moreover, Cateni et al., (2010) conducted a study to select variables through GA for classification. Here, the aim is achieved through the process of the selection based on the performance of evaluating a combination of possible variables used to train a decision tree. The proposed approach by different initialization and fitness functions of the GA is performed in this study.

2.6 Summary

Most of the clinical data were analyzed using binary logistic regression. But binary logistic regression approach decreases its accuracy of prediction when the number of independent variables increases and the assumptions are violated. Researchers used evolutionary approaches to classify this type of large scale data sets. When defining the GA, several types of fitness functions were used by previous studies such as MLE, AIC, BIC, and AUC. A comparison between logistic regression and GA with the fitness function was developed using statistical or machine learning concepts, which were conducted separately. Researchers found that evolutionary approaches give better classification than binary logistic regression.

However, a comparison between logistic regression, GA with statistical perspective fitness function and machine learning approached fitness function is not performed in the literature review.

CHAPTER 3

MATERIAL AND METHODS

3.1 Secondary Data

This study is carried out on a collection of a primary clinical experiment conducted in the area of Kegalle district, by the Faculty of Medicine, University of Peradeniya. Specimens of 502 nasopharyngeal aspirates (NPA) were collected from children above the age of 1 month and less than 5 years, who were hospitalized at the Kegalle Hospital, Sri Lanka, between May 2016 and July 2018. The children had a history of ARTI of less than 4 days and with recurrent ARTI were included in the study. A detailed pre-tested questionnaire was used to extract demographic data from patients. Furthermore, demographic and climate data have been collected from the Department of Census and Statistics to identify the monthly virus disease diagnosis.

Forty independent variables, including 3 non-categorical variables (Age in months, Fever Days, Number of People at Home) and 37 categorical variables (Appendix A), were identified from suspected viral ARTI pediatric patients in order to predict the status of the RSV, whether positive or negative. Thirty seven categorical variables were considered under following subcategories.

Table 3.1: Sub Categories of Independent Categorical Variables

Sub Category	Number of Variables
General Variables	5
Symptom Variables	25
Risk Factors	7

3.2 Demographic and Climate Data Analysis

Demographic data, including the disease diagnosis, symptoms, age, gender, ethnicity, residency, and risk factors, were extracted using a questionnaire. The population data of Kegalle district was collected from reports of the Department of Census and Statistics of Sri Lanka to identify the association between prevalence of RSV among several ethnicities and residences. Moreover, monthly average measurements of rainfall are collected from the Metrological Department to check the prevalence of RSV among pediatric patients during the studied period.

3.3 Methods of Analysis

A preliminary descriptive analysis is performed to identify the associated explanatory variables with the status of RSV. Furthermore, a detailed analysis is conducted on identified associate variables. In addition to this, the two variables, Ethnicity, and Residence are analyzed separately with their population data. The number of RSV positive patients and the rainfall measurements are used to identify the pattern between the monthly spread of RSV.

Initially, this study developed a binary logistic regression model using SPSS statistical software to identify the risk factors for RSV. In the second step, two GAs were developed for variable selection using two different fitness functions. In the third step, obtained model adequacy statistics are compared for the above three methods. Finally, the most associated risk factors on RSV are identified according to the better goodness of fit values.

3.4 Descriptive Statistics for Categorical Variables

3.4.1 Contingency Tables (Cross Tabulations) for the Dichotomous Dependent Variable

The contingency table is one of the most common methods to summarize categorical data. The interest in the contingency table is to identify whether there exists a relationship or an association between two or more variables.

3.4.2 Chi-Square Statistic

Karl Pearson proposed the Chi-Square test statistic (equation 3.1) to check the association between two categorical variables. The Pearson chi-squared statistic for testing H_0 where,

H_0: There is no association between X and Y
Vs.
H_1: There is an association between X and Y

$$\chi 2 = \sum \frac{(\mathrm{nij}-\mu\mathrm{ij})^2}{\mu\mathrm{ij}} \sim \chi^2_{df} \qquad 3.1$$

μ_{ij} = Expected frequency of the row category i and column category j

n_{ij} = Observed frequency of the row category i and column category j

3.4.3 Kruskal Wallis Test

Kruskal–Wallis test is a non-parametric test which is applied to check the association between a categorical variable and a continuous variable. It is widely used when the continuous variable does not follow a normal distribution. The null hypothesis for Kruskal Wallis Test assumes that the samples (groups) are from identical populations (Hoffman, 2019). The corresponding test statistic for testing H_0 is given by 3.2:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \sim \chi^2_{k-1}(\propto) \qquad 3.2$$

where, R_j is the rank sum of the jth group and n_j is size of the jth group.

3.5 Binary Logistic Regression

Binary Logistic Regression is a statistical method for modelling binary independent variables. The success or failure is the outcome of the response variable. There are pre-defined assumptions that are needed to be satisfied to apply binary logistic regression.

Assumptions

1. The dependent variable should be on the scale of dichotomous.
2. One or more independent variables (continuous or categorical) should be included.
3. The independence of observations should be satisfied.
4. The linearity of continuous predictors and log-odds.
5. Requires the presence of little or no multicollinearity among the independent variables.

The binary logistic regression model can be fitted if the above assumptions are satisfied. If there is a quantitative single independent variable X and a dichotomous response variable (Y), then "success" probability at value x is denoted by $\pi(x)$ in binary logistic regression. The logistic regression model has a linear format for the logit of this probability,

$$logit[\pi(x)] = log\left[\frac{\pi(x)\,)}{1 - \pi(x)}\right] = \alpha + \beta x \qquad 3.3$$

The probability π(x) is obtained by the exponential function of equation 3.3. (Agresti, 2007).

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} \qquad 3.4$$

In general,

$$\pi(x_i) = \frac{exp(\sum_{k=0}^{k} \beta_k X_{ik})}{1 + exp(\sum_{k=0}^{k} \beta_k X_{ik})} \; ; i = 1,2,3, \ldots\ldots, K \qquad 3.5$$

3.5.1 Logistic Regression Parameter Estimation

The parameters of the logistic regression models are estimated using the maximum likelihood estimation method. Suppose that there are k independent variables $X_1, \ldots, X_k$, and the model parameters, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$, then the conditional density function of the dichotomous variable Y follows a Bernoulli distribution:

$$f(y|\beta) = \prod_{i=1}^{N} \pi(x_i)^{\,Y_i}(1 - \pi(x_i)\,)^{1-Y_i} \;\; ; \; Y_i = 0, 1 \qquad 3.6$$

The likelihood function can be written as: (Feelders et al., 2000)

$$L(\beta) = \prod_{i=1}^{N} \pi(x_i)^{Y_i}\big(1 - \pi(x_i)\big)^{1-Y_i} \;\; ; \; Y_i = 0,1 \qquad 3.7$$

Substituting 3.5 into 3.7 and takes the natural logarithm. :

$$\ln(L(\beta)) = \ln\left[\prod_{i=1}^{N} [exp\left(\sum_{k=0}^{k} \beta_k X_{ik}\right)]^{Y_i} \; (1 + exp(\sum_{k=0}^{k} \beta_k X_{ik}))^{-1}\right]$$

$$\ln(L(\beta)) = \sum_{i=1}^{N} \left\{Y_i \;\; \sum_{k=0}^{k} \beta_k X_{ik} - ln\left(1 + e^{\sum_{k=0}^{k} \beta_k X_{ik}}\right)\right\} \qquad 3.8$$

Equation 3.8 gives the log-likelihood function, and the parameters are estimated by maximizing it. But, it cannot be solved algebraically since partial derivatives are not in closed forms. Numerical methods are used to obtain the maximum likelihood estimates for the model represented by the equation 3.8. After estimating the parameters, the significance of parameter estimates is checked using Wald type t-test and likelihood ratio test.

3.5.2 Strategies in Model Selection

Stepwise Variable Selection Algorithms

Stepwise methods are algorithms that determine either to include or exclude variables from the model to improve the predictability. In SPSS, there are six different variable selection methods for binary logistic models, where the score test for selection and likelihood ratio test for deletion of covariates. Moreover, the Wald test is also used for both entry and removal of covariates. Variable selection methods in SPSS are as follows:

1. Forward - Conditional: Variables entered to the model based on the significance of the score statistic. Removing an estimator is based on the probability of a likelihood-ratio statistic using conditional parameter estimation.
2. Forward - Likelihood Ratio: Variables entered to the model based on the significance of the score statistic. Removing an estimator is based on the probability of a likelihood-ratio statistic using the maximum likelihood estimation.
3. Forward - Wald: Variables entered to the model based on the significance of the score statistic. Removing an estimator is based on the probability of the Wald statistic.
4. Backward Elimination - Conditional probability of the likelihood-ratio statistic using conditional parameter estimates is used to remove an estimator from the model.

5. Backward Elimination - Likelihood Ratio: Probability of the likelihood-ratio statistic using the maximum partial likelihood estimates is considered to remove an estimator from the model.
6. Backward Elimination - Wald: Removing an estimator is based on the probability of the Wald statistic.

3.5.3 Summary Measures of Goodness-of-Fit

The goodness of fit value provides an overall indication of any fitted model.

The Hosmer-Lemeshow Test

Hosmer-Lemeshow Test ((Hosmer et al., 1988)) is produced to check the goodness of fit for logistic regression models.

Classification Tables

The classification table summarizes the power of predictivity of a binary regression model. The binary outcome y with a prediction of whether $y = 0$ or 1 is classified by the table of classification. The prediction is $y\hat{} = 1$ when $\widehat{\pi_\iota} > \pi_0$ and $\hat{y} = 0$ when $\widehat{\pi_\iota} \leq \pi_0$, for some cut-off π_0. Generally, the cut-off is to take π_0 = 0.50. (Agresti, 2007).

Sensitivity and Specificity are two important summary statistics that use to measure the predictive power. The two measurements aforementioned in clinical trials are explained under AUC-ROC Curve.

AUC-ROC Curves

AUC-ROC (Area Under the Receiver Operating Characteristics) is a measurement of performance to classify at various thresholds settings. It is developed under the concept of the confusion matrix, specificity, and sensitivity.

The confusion matrix is used to summarize the classification of the sample.

		Actual	
		True	False
Predicted	True	True Positive	False Positive
	False	False Negative	True Negative

Sensitivity tells what percentage of true counts correctly identified.

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \qquad 3.9$$

Specificity tells what percentage of false counts correctly identified.

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \qquad 3.10$$

ROC Graph is designed by plotting the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity).

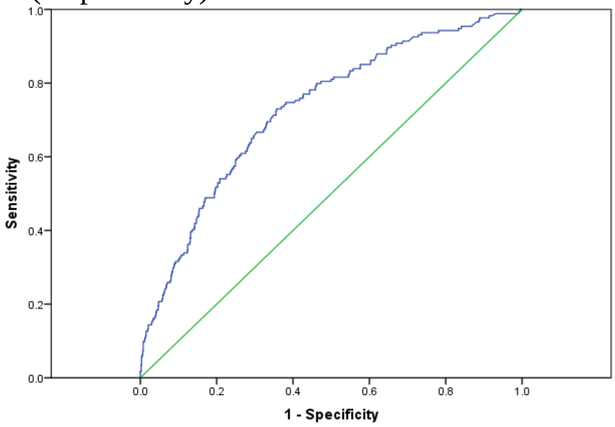


Figure 3.1: AUC-ROC Curve

AUC - ROC varies between zero and one, interprets the ability of classification in the model. Higher AUC is a better classification model's performance (Hosmer et al., 1988).

In general,

if $AUC = 0.5$: No discrimination among classifiers

if $0.7 \leq AUC < 0.8$: Acceptable discrimination among classifiers

if $0.8 \leq AUC < 0.9$: Excellent discrimination among classifiers

if $AUC \geq 0.9$: Outstanding discrimination among classifiers

In practice, it is unusual to observe AUC higher than 0.9.

Pseudo R-Square Values

Both Cox & Snell R-squared and Nagelkerke R-squared values are referred to as pseudo-R-squared values and they indicate the percentage of the variance of the dependent variable explained by the model.

Pseudo-R-squared values are given by:

1. Cox & Snell R-Squared

$$R^2_{C\&S} = 1 - (L_0 \,/\, L_M)^{2/n} \qquad 3.11$$

2. Nagelkerke R-squared

$$R^2_{Nage} = \frac{1 - (L_0 \,/\, L_M)^{2/n}}{1 - (L_0)^{2/n}} \qquad 3.12$$

L_0 = likelihood function value for a model with no predictors.

L_M = likelihood for the model being estimated.

n = sample size.

3.6 Genetic Algorithm

Genetic algorithm introduced by John Holland in 1960 based on the concept of Darwin’s theory of evolution, and later it is extended by David E. Goldberg in 1989. GA is a heuristic optimization technique that imitates the evaluation of biological process and finds a near global optimum for an objective function without any assumptions. In the GA, an initial population is created. "Parents" are selected, and "Children" are produced. The best children from the previous generation are kept to create the next generation. Development of the other operators depends on the design.

In GA, points in the search space are defined by their genes and these points are denoted by chromosomes. Each chromosome has a fitness value that represents the objective function of the GA. Darwin's theory of evolution mentioned that the reproduction, crossover, and mutation are the processes, which maintain the survival of an organism.

Select a large number of different chromosomes at the beginning of the search. Then their fitness values are evaluated. In the initial generation, evolution is mimicked by randomly mating two chromosomes to develop two off springs and randomly mutate variables from these off springs. Chromosomes which have a higher fitness value have a greater possibility to choose for the process of mating and hence, have a higher possibility of producing off springs. Figure 3.2 simply illustrates the process of GA:

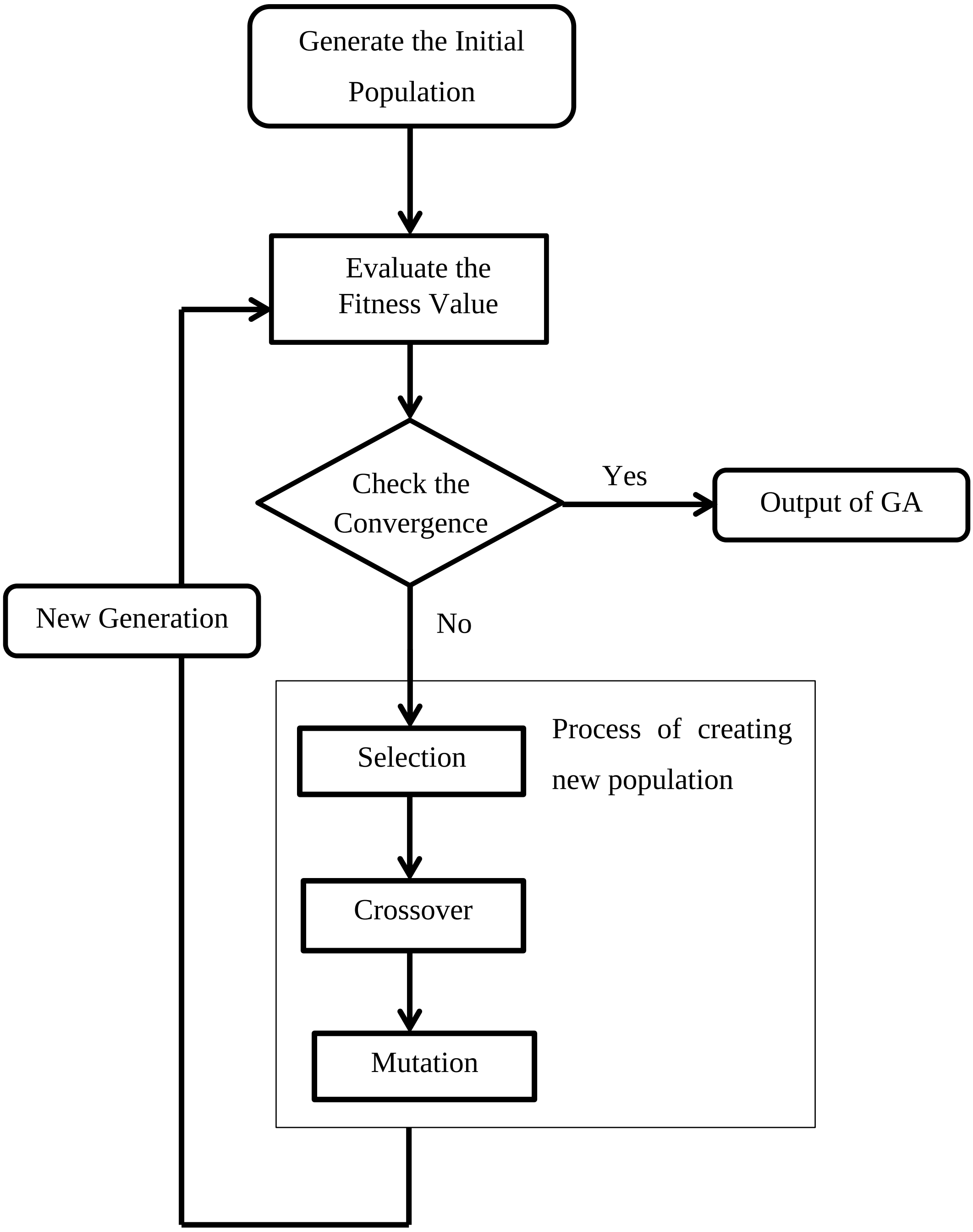

Generate the Initial Population
Evaluate the Fitness Value
Check the Convergence
Yes
Output of GA
No
New Generation
Process of creating new population
Selection
Crossover
Mutation

Figure 3.2: Flow Chart of a GA

3.6.1 Selection

The selection or reproduction is the first operator applied to a population in a GA. Re-production chooses good strings in a population and creates a pool of mating. Average strings are chosen from the current population to make copies of them and added them into the pool of mating. The proportionate selection operator is the most commonly used reproduction operator. In this operator, a string from current population is chosen with probability corresponding to the fitness of the string. Accordingly, the i^{th} string from the population is chosen with probability corresponding to f_i . The size of the population is generally remained fixed in a GA. Therefore, the cumulative probability must be one for all strings in the population. Hence, $f_j / \sum_{j=1}^{N} f_j$ is the probability of selecting i^{th} string, where N is the population size. (Iquebal et al., 2012)

3.6.2 Crossover

The function of crossover is applied after the pool of mating string. Here, randomly two strings are picked, and a part of the strings is interchanged. In the crossover operator of single point, both strings are crop at an arbitrary place, and a right-side section of both strings are exchanged among themselves. Then two new strings are created. It is simply shown below:

$$\begin{array}{lcc|ccc} \text{Parent 1} & 0 & 0 & 0 & 0 & 0 \\ \text{Parent 2} & 1 & 1 & 1 & 1 & 1 \end{array} \quad \Longrightarrow \quad \begin{array}{lcc|cccl} \text{Parent 1} & 0 & 0 & 1 & 1 & 1 & \text{Child 1} \\ \text{Parent 2} & 1 & 1 & 0 & 0 & 0 & \text{Child 2} \end{array}$$

Construction of substrings from a parent string depends on the site. Random site is usually chosen when the knowledge of an appropriate site is unknown. In a random site, a combination of good substrings may not be constructed but more copies in the next pool of mating will be generated by the reproduction operator. Two random sites are chosen in a two-point crossover operator. Likewise, the creation of an operator for multi-point crossover can be extended. Ultimately, it is known as the operator of uniform crossover. For binary strings, 0.5 of probability is taken to choose a parent in a uniform crossover. Searching the parameter space is the main intention of the crossover operator. Operator search is not extensive in the single-point crossover; however, the maximum information is maintained from parent to child. Whereas, search is pervasive in the uniform crossover, but in here, minimum information is maintained from parent to child (Iquebal et al., 2012).

3.6.3 Mutation

The mutation operator is used for the search aspect of GA. It changes from 0 to 1 and conversely with a small mutation probability p_m.

0 0 1 0 0 ➔ 0 0 0 0 0

In this example, the third value of the gene has changed from 1 to 0 by creating a new solution. The process of mutation is essential to maintain variety in the population. The insertion of mutation gives some probability of turning that 1 into 0. Moreover, the mutation is useful to improve a local solution.

One generation of GA is completed after applying the process of reproduction, crossover, and mutation to the current population. The best fitted strings were chosen by the reproduction operator, and the crossover operator recombines good substrings. Through the mutation operator, a better string is created. From the reproduction operator, existing bad strings will be deleted, and good strings will be highlighted in the next generation. Problem-specific operators are applied to achieve faster convergence of GAs in various types of real-world problems. However, three operators mentioned here are the fundamental operations of a GA (Iquebal et al., 2012).

3.6.4 A GA for a Variable Reduction in Binary Logistic Regression

GAs have the ability to search larger solution space and filter them in order to reach the optimal combination of characteristics and solutions which might not find in a lifetime. An application of this study is to apply GA as an alternative method to binary logistic regression technique. In other words, two different fitness functions, likelihood function as in equation (3.8), and Area Under ROC curve (In section 3.5.3) are used in GA to select the best combination of predictive variables.

For the variable selection in GA, chromosome stands for every possible variable subset. The fitness of a chromosome is evaluated with two criteria which are defined previously. For the binary logistic regression, fitness function can be taken as a goodness of fit of the model. Therefore, likelihood function, AIC, BIC, Deviance, and AUC can be considered as a fitness function for the GA.

3.7 Software Packages to Perform Logistic Regression and GA

IBM® SPSS® Statistics 22 statistical software program was used to perform Binary Logistic Regression. Different R packages were used to develop GA. Trevino & Falciani, (2006) developed a R package called "GALGO," which is a popular tool in bioinformatics to provide efficient variable selection strategy. Another package called "GA" is developed by Scrucca, in 2013. "GA" and "GALGO" were used in this study to perform log likelihood and AUC functions respectively.

CHAPTER 4

PRELIMINARY ANALYSIS

In this chapter, explanatory data analysis is performed for all three non-categorical variables: Age in months, Fever Days, and Number of People at Home. The normality of these three variables is investigated and performed statistical analysis to find the association between the dependent variable and the aforementioned three variables. The association between the dependent variable and all the other categorical variables are examined using the chi-square test, and hence, the related variables are identified.

Moreover, certain variables like Residence and Ethnicity were analyzed with population data collected from the Department of Census to evaluate the presence of correlation with RSV prevalence.

Also, a time series graphical analysis is performed to check the presence of seasonality patterns in the number of monthly RSV infected children. The association between rainfall data and the number of RSV infected children is also analyzed using monthly rainfall measurements.

4.1 Exploratory Data Analysis of the Dependent Variable RSV

Table 4.1: Frequency Table for RSV

RSV				
		Frequency	Percent	Valid Percent
Valid	Negative	335	67.4	67.4
	Positive	162	32.6	32.6
	Total	497	100.0	100.0

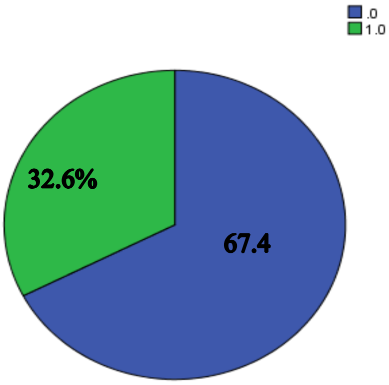


Figure 4.1 : RSV Positive Vs Negative

It can be noticed that there are 335 children who are not infected from RSV among a total of 497, which means approximately only one-third of children are infected from RSV.

4.2 Exploratory Data Analysis of the Explanatory Variables

There are 40 independent variables considered in this study, where three of them are non-categorical, and the rest of the variables are categorical. Preliminary analysis for all 40 variables was conducted and summarized in tables 4.2, 4.3, and 4.4.

4.2.1 Non Categorical Explanatory Variables

Table 4.2: Summary of Non-Categorical Independent Variables

No	Independent Variable	Minimum	Maximum	Mean	Std. Deviation	Kolmogorov-Smirnov Teat Statistic p-Value	Kruskal-Wallis Test Statistic p-value
1	Age in months	1	66	12.95	11.874	0.194 0.000	12.049 0.001
2	Number of People at Home	1	11	4.592	1.295	0.218 0.000	7.956 0.005
3	No. of Fever Days	1	8	3.165	1.383	0.189 0.000	25.995 0.000

Table 4.2 shows that the average age of a child is 12.95 months, the average fever days of a child is close to three days, and the mean number of members in a family is approximately 5. Furthermore, Figure 4.2 indicates that all three continuous variables do not follow a normal distribution since the p-value for the Kolmogorov-Smirnov test is less than 5% significant level. Kruskal-Wallis test is used to check the association between RSV and each non-categorical variable.

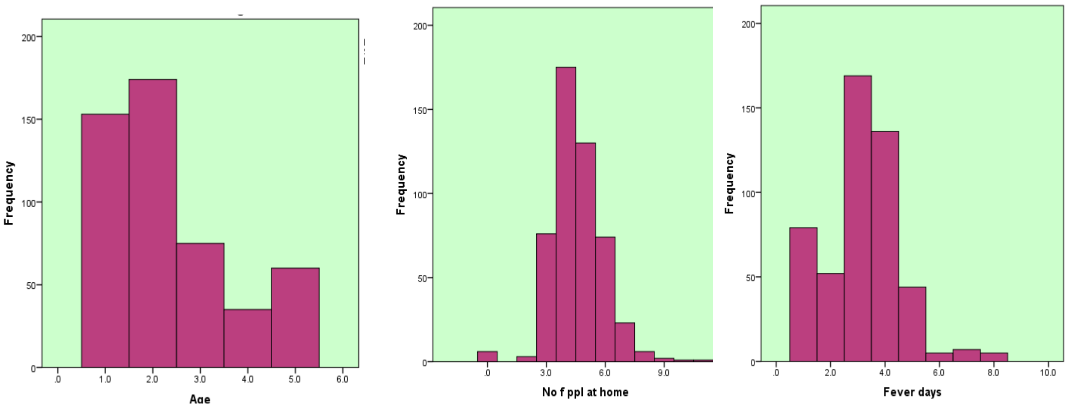


Figure 4.2: Histograms for the Non Categorical Variables

H_0: There is no association between the status of RSV and Independent Variable
H_1: There is an association between the status of RSV and Independent Variable

According to Table 4.2, the p-values of the chi-square test are below 0.05. Therefore, the null hypothesis is rejected at 5% significant level. It can be concluded that, when considered individually, all three non-categorical variables show a relationship between the dependent variable RSV.

4.2.2 Categorical Variables

Exploratory Data Analysis for the categorical variables is performed under three subcategories mentioned in Table 3.1.

a. General Variables

Under general variables, Gender, Residence, Ethnicity, High Dependency Care, and Required Intensive Care are considered.

Table 4.3: Summary of Chi-Square Test for General Variables

No	Independent Variable	Chi-Square	p-value
1	Gender	0.487	0.485
2	Residence	0.018	0.991
3	Ethnicity	3.123	0.210
4	High Dependency Care	4.715	0.030
5	Required Intensive Care	-	-

H_0: There is no association between the status of RSV and Independent Variable
H_1: There is an association between the status of RSV and Independent Variable

Table 4.3 exhibits that only one general variable (High Dependency Care) is statistically significant since other p-values of the chi-square tests are greater than 5% significant level. Therefore, it can be identified that there is an association between RSV and the variable High Dependency Care at 5% significant level.

b. Variables: Indication of Symptoms

Data are collected under 25 symptom variables, and the analysis has been conducted in order to identify the symptom variables which are associated with the status of RSV.

Table 4.4: Summary of Chi-Square Test for Symptom Variables

No	Independent Variable	Chi-Square	p-value
1	Fever	0.870	0.351
2	Cold	2.526	0.112
3	Headache	7.830	0.005
4	Cough	10.286	0.001
5	Sputum	1.038	0.308
6	Vomiting	0.029	0.864
7	Runny Nose	1.031	0.310
8	DIB	1.221	0.269
9	Dyspnoea	10.082	0.001
10	Conjunctivitis	6.975	0.008
11	Tachypnoea	9.674	0.002
12	Sore Throat	3.687	0.055
13	Sinus Congestion	1.950	0.163
14	Stuffiness	4.933	0.026
15	Chills	1.486	0.223
16	Diarrhea	7.603	0.006
17	Fatigue	5.836	0.016
18	Body Aches	2.305	0.129
19	Loose Stool	2.177	0.140
20	Wheezing	1.684	0.194
21	Nasal Block	0.219	0.640
22	SOB	0.002	0.969
23	Nasal Congestion	0.340	0.560
24	Colour of Nasal Secretion	0.256	0.613
25	Severe Hydration	6.241	0.012

H_0: There is no association between the status of RSV and Independent Variable
H_1: There is an association between the status of RSV and Independent Variable

Table 4.4 shows that p-values of the chi-square tests are greater than 0.05 significant level for all the symptom independent variables except Headache, Cough, Dyspnoea, Conjunctivitis, Tachypnoea, Stuffiness, Diarrhea, Fatigue and Severe Hydration. Therefore, there is a significant dependence between RSV and the latter variables; Headache, Cough, Dyspnea, Conjunctivitis, Tachypnoea, Stuffiness, Diarrhea, Fatigue, and Severe Hydration, at 5% significant level. Since their p-values are greater than 0.05, it can be concluded that the rest of the variables do not have a significant relationship with RSV.

c. Variables: Indication of Risk Factors

   Association of risk factors with the status of RSV is summarized to identify the factors which support to spread this virus.

Table 4.5: Summary of Chi-Square Test for Risk Factors

No	Independent Variable	Chi-Square	p-value
1	Daycare Preschool	0.701	0.402
2	Family History of Asthma /	0.466	0.495
3	Anyone Smokes at Home	0.231	0.630
4	Possibility of Inhaling Toxic /	6.670	0.010
5	Did Patient Travel Recently	0.752	0.386
6	Allergic to Pollen / Dust	0.434	0.510
7	Pets at Home	11.345	0.001

H_0: There is no association between the status of RSV and Independent Variable
H_1: There is an association between the status of RSV and Independent Variable

Table 4.5 provides information about the association between RSV and the risk factor among independent variables. Resulted p-values for the chi-square tests are greater than 0.05 in all the symptom variables except the variables; Pets at Home and Possibility of Inhaling Toxic/Fumes. Therefore, there is an association between RSV and the variables; Pets at Home and Possibility of Inhaling Toxic/Fumes at 5% significant level. Since all the other p-values are greater than 0.05, it can be concluded that the rest of all the risk factors do not indicate a relationship with RSV.

Finally, fifteen independent variables (Age (in months), Number of People at Home, Fever Days, High Dependency Care, Headache, Cough, Dyspnea, Conjunctivitis, Tachypnea, Stuffiness, Diarrhea, Fatigue, Severe Hydration, Possibility of Inhaling Toxic/Fumes and Pets at Home) indicate an association with dependent variable RSV. Inter-dependency between the aforementioned fifteen independent variables were summarized in Table 4.6:

Table 4.6: Summary for the Dependency of Variables which has an Association with RSV

Variable	Age in months	Fever Days	Headache	Cough	Dyspnoea	Conjunctivitis	Tachypnoea	Stuffiness	Diarrhoea	Fatigue	High Dependency Care	Severe Hydration	Number of People at Home	Possibility of Inhaling Toxic/Fumes	Pets at Home
Age in months															
Fever Days															
Headache	✓														
Cough	✓		✓												
Dyspnoea	✓	✓													
Conjunctivitis			✓	✓											
Tachypnoea					✓										
Stuffiness	✓				✓		✓								
Diarrhoea					✓		✓								
Fatigue			✓												
High Dependency					✓		✓								
Severe Hydration					✓		✓		✓		✓				
Number of People at Home				✓	✓	✓	✓		✓						
Possibility of Inhaling Toxic/Fumes			✓	✓						✓					
Pets at Home		✓			✓	✓	✓			✓	✓		✓	✓	

In Table 4.6, checkmarks highlight p-values which are less than 5% significant level (Appendix B). It can be concluded that there are 36 paired, significant relationships between independent variables.

4.3 Presence of Correlations with RSV Prevalence and Two General Variables Residence and Ethnicity

Residence and Ethnicity have not indicated a significant association with dependent variable RSV. But previous researchers found that there is an association between Residence and Ethnicity with RSV. Therefore, further analysis has been conducted using population data collected from the Department of Census.

1. Residence:

Mahikul et al., (2019) and G.M. et al., (2013) have done studies about household dynamics on RSV infection and found that there was a relationship between residence type and status of RSV.

The population under the type of residence based on geographical background in 11 divisional secretariats at Kegalle is analyzed and a summary is given in Table 4.7:

Table 4.7: Summary on Type of Residence in Kegalle District

	Urban	Rural	Estate
Population	16,809	806,632	60,104
Admitted to Hospital	219	269	9
RSV Positive from Admitted	72	87	3

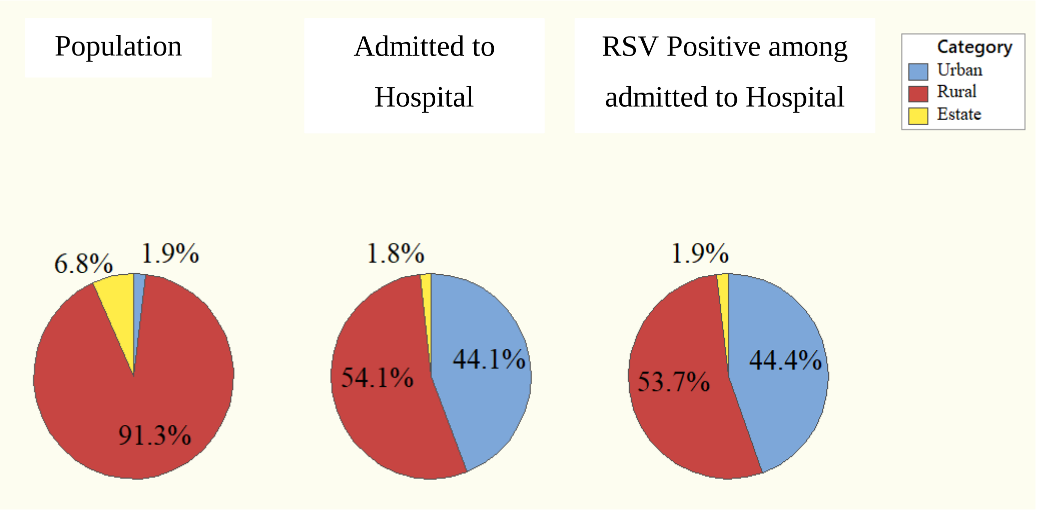


Figure 4.3: Pie Chart for the Variable Residence

According to Figure 4.3, people who are living in an urban area have a high tendency of infecting RSV, while in rural areas have a lower tendency of infecting RSV. A Chi-square test is performed to check whether there is an association between the size of the population and the number of admitted RSV infected children with respect to three different residence types. The calculated p-value is lower than the chosen 5% significant level. Therefore, it can be concluded that there is a dependency between residence type and RSV infected.

Furthermore, the Chi-square test statistic is performed to check the association between children who were admitted to hospital and RSV positive from admitted. Corresponding p-value is 0.995 which is greater than 5% significant level. Hence, it can be concluded that there is no significant association between children who were admitted to hospital and RSV positive from admitted with respective to childrens' Residence type.

2. Ethnicity

Kassem et al., (2019) have made a comparison between ethnicity groups for RSV and found that Ethnicity is a dependent factor for diagnosing RSV positive. Therefore, the analysis for the explanatory variable Ethnicity is developed using population data in Kegalle district.

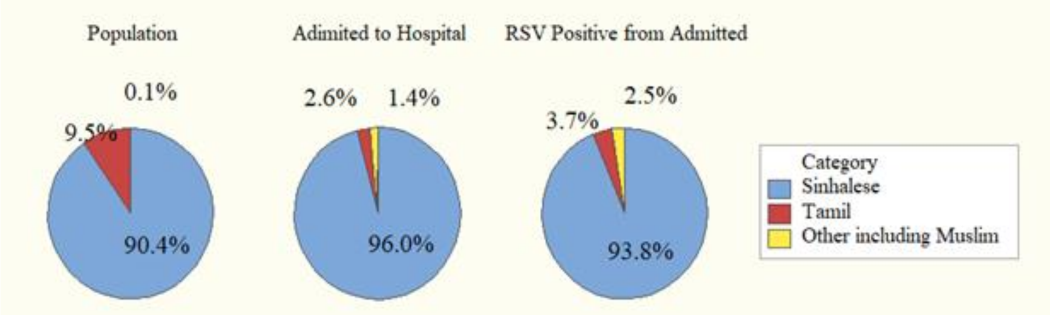


Figure 4.4: Pie Chart for the Variable Ethnicity

According to Figure 4.4, it can be observed that during the study period, there were 9.5% Tamil population in Kegalle district. Percentage of 2.6 of Tamil children were admitted to Kegalle hospital as inpatients. Out of the hospitalized children, 3.7% were RSV infected. In the contingency table, some cells have an expected count of less than five. Therefore, Fisher's Exact test is performed to check the association between the size of the population and the admitted number of RSV infected children with respect to three different Ethnicities. The calculated p-value is 0.229, which is higher than the chosen 5% significant level. Therefore, it can be concluded that there is no significant association between type of Ethnicity and getting RSV infected.

Furthermore, the Chi-square test statistic is performed to check the association between children who were admitted to hospital and RSV positive from admitted. Corresponding p-value is 0.501, which is greater than 5% significant level. Hence, it can be concluded that there is no significant association between children who were admitted to hospital and RSV positive from admitted with respect to three different Ethnicities.

4.4 Seasonality of the Number of RSV Positive Children

Table 4.8 summarizes the number of RSV positive children in each month during the period the study was conducted:

Table 4.8: Summary of RSV Positive Children According to Month

	2016								2017												2018						
Month	May	June	July	August	September	October	November	December	January	February	March	April	May	June	July	August	September	October	November	December	January	February	March	April	May	June	July
RSV (+)	2	8	8	9	3	6	8	2	1	3	9	14	14	11	7	1	6	2	1	0	0	2	2	0	18	20	7

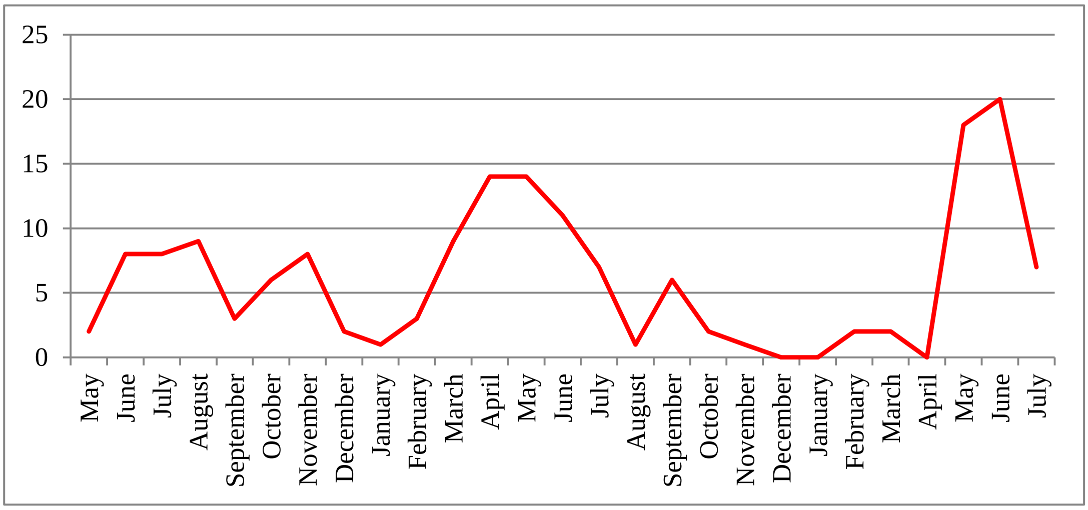


Figure 4.5: Time Series Plot for Number of RSV Positive Children

Figure 4.5 shows that the number of RSV positive children against the month. It can be seen that there is seasonality in the number of RSV positive children in Kegalle district during the investigated time frame. Peaks can be seen from April to June. Therefore, it can be identified that RSV is spreading during the period from April to June.

4.5 Analysis of the Number of RSV Positive Children under Climate Data

Figure 4.6 illustrates the pattern of rainfall (mm) and the number of RSV positive children in Kegalle district during the investigated period:

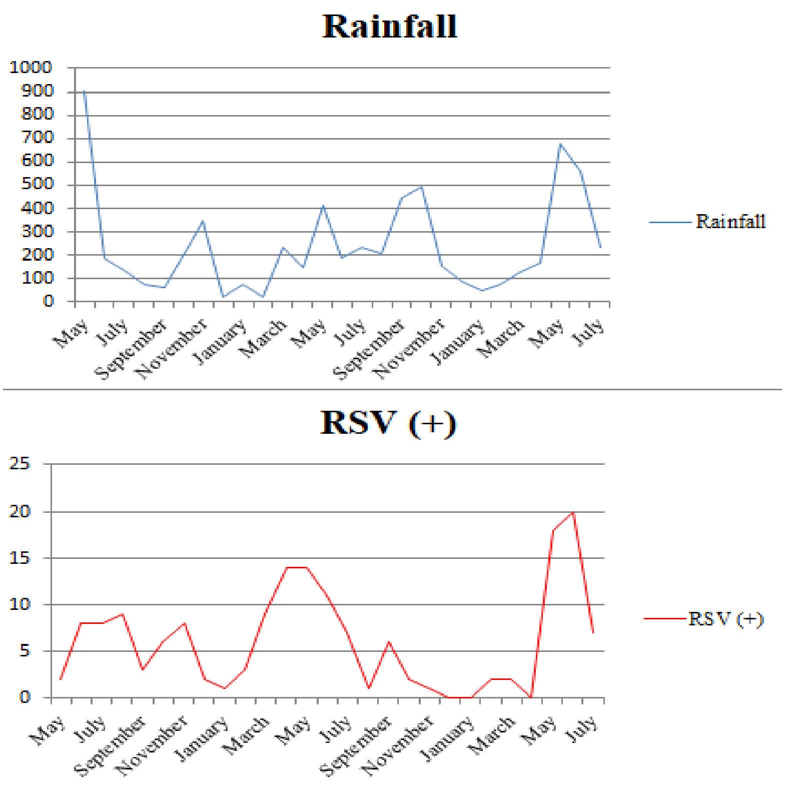


Figure 4.6: Relationship between Rainfall Data and RSV (+) Counts

It can be observed that during the month of May the amount of rainfall has risen rapidly and by the beginning of the month June, it has reached the peak. By the end of July, the amount of rainfall has fallen down below average. Interestingly, the same pattern can be noticed in the number of RSV positive children from May to July. Furthermore, this pattern continues during September to December. However, when compared to these two periods, it can be identified that the number of infected children was significantly low during the season of low rainfall.

Furthermore, Spearman's Correlation between monthly RSV infected children and rainfall is 0.419, and the corresponding p-value is 0.029. It indicates that there is a positive relationship between monthly RSV infected children and rainfall.

In conclusion, the prevalence of RSV depends on demographic factor Residence as well as the rainfall of the region (in sections 4.3 and 4.5).

CHAPTER 5

SELECTION OF VARIABLES USING LOGISTIC REGRESSION AND GA

5.1 Variable Selection Procedure

The variable reduction is an essential step in model specification. In clinical research, identifying risk factors for a specific disease is the primary objective. In this study, six variable selection methods in binary logistic regression model is used to determine the significant explanatory variables. Then, two different GAs were developed and applied to the data set in order to reduce the number of explanatory variables. Finally, the goodness of fit measurements was summarized for the above three methods and comparison is done to identify the most accurate way to predict the status of RSV. Defined significant variables and their effect on the dependent variable are discussed in summary.

5.2 Statistical Approach: Binary Logistic Model

5.2.1 Validate the Assumptions

The dependent variable of the study is RSV with two outcomes (Positive and Negative) and 40 independent variables are included in the set of considered data. The observations are independent in this study since the occurrence of one observation does not provide information on the other observation. Linearity between the continuous predictors and the log odds and multicollinearity among independent variables is checked as follows.

Linearity between continuous predictors and log-odds

Box-Tidwell test can be used to test the linearity between continuous predictors and log odds. Interaction terms between the continuous predictors and their logs were included to the model and such an interaction is significant, then it can be identified that the assumption has been violated. Results indicated that p-values for the three interaction terms were less than 0.05. Therefore, the interaction terms were not significant in this study thus; linearity between continuous predictors and log odds assumption is satisfied.

Multicollinearity among Independent Variables

Section 3.5 stated the assumptions that to be satisfied to apply logistic regression technique. VIF values were calculated to check the multicollinearity among independent variables.

Table 5.1:Multicollinearity among Independent Variables

Variable	VIF
Gender	1.056
Age in months	1.328
Fever days	1.120
Fever	1.194
Cold	1.187
Headache	2.040
Cough	1.292
Sputum	1.184
Vomiting	1.306
Runny Nose	1.122
DIB	1.429
Dyspnoea	1.729
Conjunctivitis	1.387
Tachypnea	1.758
Sore Throat	1.837
Sinus Congestion	1.000
Stuffiness	1.101
Chills	1.486
Diarrhea	1.248
Fatigue	1.304
Body Aches	1.442
Loose Stool	1.101
Wheezing	1.166
Nasal Block	1.142
SOB	1.286
Nasal Congestion	1.106
Residence	1.209
Ethnicity	1.113
Day Care/ Pre School	1.220
Colour of Nasal Secretion	1.000
Required Intensive Care	-
High Dependency Care	1.189
Severe Hydration	1.000
Number of People at Home	1.203
Family History of Asthma	1.333
Anyone Smokes at Home	1.297
Possibility of Inhaling Toxic /	1.138
Did Patient Travel Recently	1.122
Allergic Pollen/Dust	1.000
Pets at Home	1.172

Table 4.6 indicated some interdependencies among independent variables. However, Table 5.1 indicates that all the VIF values are greater than 1 and less than 10. It means that all the independent variables are not correlated with other independent variables. Therefore, the assumption is satisfied to apply logistic regression.

Since all the assumptions were satisfied for the set of collected data, binary logistic regression is applied in order to identify the factors which are highly affected to the status of RSV.

5.2.2 Binary Logistic Regression Model Fitting

The logistic framework is designed for analyzing the coefficients of categorical explanatory variables or a combination of categorical and continuous variables when the dependent binary variable is coded 0 or 1.

Logistic regression is performed using SPSS. There are six different variable reduction methods available in SPSS, as described in Section 3.5.2. Entry probability for the stepwise method is defined as 0.05 and the removal probability is set to 0.1. Stepwise algorithm is iterated until one of the defined two stopping criteria is met. Two stopping criteria were:

1. Maximum number of iterations is reached - 30 iterations
2. Minimum change of the parameter estimation – 0.01

Six goodness of fit statistics methods were summarized in the Table 5.2:

Table 5.2: Summary of Goodness-fit Statistics for Six Variable Reduction Methods

Method	-2 Log likelihood	Cox & Snell R-squared	Nagelkerke R Square	Correctly Classified Percentage	AUC	Iteration No	Number of Significant Variables
Forward Wald	535.674	0.169	0.235	71.0%	0.754	8	8
Forward Likelihood Ratio	535..674	0.169	0.235	71.0%	0.754	8	8
Forward Conditional	535.674	0.169	0.235	71.0%	0.754	8	8
Backward Wald	524.900	0.187	0.260	74.8%	0.770	27	9
Backward Likelihood Ratio	515.003	0.203	0.282	75.1%	0.778	27	9
Backward Conditional	515.003	0.203	0.282	75.1%	0.778	27	9

The Section 3.5.1 explains that to have better predictions for a binary logistic regression model, Log-likelihood value should be maximized. Therefore the statistic ‘-2 Log-likelihood’ should be minimized in a logistic regression model. Furthermore, according to Section 3.5.3, Cox & Snell R-squared and Nagelkerke R Squared statistics give the percentage of the variance of the dependent variable explained by the model. These two statistics should be maximized to have accurate predictions. Moreover, it is explained in Section 3.5.3 that the value of AUC and correctly classified percentage should be maximized to have better predictions.

Table 5.2 summarizes that best goodness fit statistics were given by the Backward Conditional and Backward Likelihood Ratio methods in 27th step. The number of variables reduced into 13 and only 9 of them were statistically significant at 5% significant level (Appendix C). However, these two methods meet the stopping condition in the 27th step, and thus terminate. Hence, the exact optimal solution cannot be determined and the current solution is considered as the near optimal solution within the stipulated stopping condition. Also, it was observed that the solution could not be improved further.

Forward Wald, Forward Likelihood Ration and Forward Conditional methods give the same goodness of fit measurements in the 8th step. Moreover, the parameter estimation change is less than 0.01 in the aforementioned three methods and all the variables in the 8th step are statistically significant at 5% significant level.

Although all three backward methods give better goodness of fit measurements to reduce to 13 variables at 27th step, only 9 of them are statistically significant at 5% significant level and output indicates that the final solution cannot be found. Conversely, all three forward methods give the optimum solution at the 8th step and all the variables at the optimum level are statistically significant at 5% significant level. Since all three forward methods give the same value for each statistic, a stepwise explanation to reduce the number of explanatory variables is performed by the Forward Wald method.

Table 5.3: Dependent Variable Encoding

Original Value	Internal Value
RSV Negative	0
RSV Positive	1

Status of the dependent variable is encoded according to Table 5.3

Status of all 37 independent variables is encoded. (Appendix D)

Table 5.4: Classification Table

Step 0	Observed		Predicted		
			RSV		Percentage Correct
			Negative	Positive	
	RSV	Negative	335	0	100.0
		Positive	162	0	00.0
	Overall Percentage				67.4
a. Constant is included in the model.					
b. The cut value is .500					

The output of Block 0 for a model consists only the intercept. Given the base rates of the two decision options, 67.4% reported RSV negative, and 32.6% reported RSV positive. Using this strategy, the correctly classified percentage is 67.4%.

Table 5.5: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-0.727	0.096	57.639	1	0.000	0.484

Table 5.5 indicates that the logistic regression model with the constant term can be written as:
$\ln\left(\frac{P}{1-P}\right) = -0.727$. By exponentiation both sides of this expression, the predicted odds could be found as $Exp(B) = 0.484$. That is, the predicted odd of RSV positive is 0.484. Since 162 children are diagnosed as positive RSV and 335 children reported negative RSV, observed odds are 162 out of 335.

All 37 categorical variables and three non-categorical variables are included in the model. The Omnibus test of model coefficients is used to test whether this new model gives an improvement or not over the initial model (constant only model). The test hypothesis is given below:

H_0: Adding new variables to the model has not significantly increased the ability to predict the status of RSV

H_1: Adding new variables to the model has significantly increased the ability to predict the status of RSV

Table 5.6: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 8	Step	5.364	1	0.021
	Block	91.816	8	0.000
	Model	91.816	8	0.000

Table 5.6 explains that the chi-square value of 91.816 is resulting in a significant p-value (0.000). Therefore, the null hypothesis is rejected at 5% significant level. Hence, adding new variables to the model has significantly increased the ability to predict the status of the RSV of a patient.

Table 5.7: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
8	535.674	0.169	0.235

Table 5.7 exhibits the model summary. -2 Log-likelihood statistic measures how poorly the model predicts the decision, which is 535.674. Furthermore, between 16.9% and 23.5% of the total variation is explained by the model according to Cox & Snell and Nahelkerke R-Squared values respectively.

Table 5.8: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
8	11.956	8	0.153

H_0: Observed and expected proportions are equal

H_1: Observed and expected proportions are not equal

According to Table 5.8, the Hosmer and Lemeshow test statistic value is 11.956, and the corresponding p-value is 0.153. Since the p-value is greater than 5% significant level, the null hypothesis is not rejected. Therefore, it can be concluded that the predictions of RSV in this model fit well with the observed status of the RSV of children. The estimated coefficients of the fitted binary logistic model are given in the Table 5.9:

Table 5.9: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[a]	Age in months	-0.038	0.011	11.212	1	0.001	0.962
	Fever Days	-0.386	0.085	20.542	1	0.000	0.680
	Cough(1)	2.062	0.765	7.263	1	0.007	7.864
	Conjunctivitis(1)	-1.445	0.699	4.275	1	0.039	0.236
	Tachypnoea(1)	0.512	0.250	4.202	1	0.040	1.669
	Fatigue(1)	-1.391	0.578	5.795	1	0.016	0.249
	Possibility of Inhaling Toxic Fumes(1)	-0.823	0.253	10.627	1	0.001	0.439
	Pets at Home(1)	0.632	0.239	7.000	1	0.008	1.880
	Constant	-0.631	0.818	0.595	1	0.441	0.532
a. Variable(s) entered on step 8: Conjunctivitis.							

H_0: The explanatory variable is not significant

H_1: The explanatory variable is significant

According to the Table 5.9, all eight explanatory variables are significant as p-values are less than 0.05. Furthermore, the estimated constant -0.631 is not significant at 5% significance level (p-value = 0.441).

It can be summarized that two out of three non-categorical variables are significant (Age in months and Fever Days) at 5% significant level. The other non-categorical variable “Number of People at Home” is not reduced into the final level.

Data is gathered under seven risk factors among children. Among them, since their p-values are (≤ 0.05) lesser than desired 5% significant level, the presence of RSV only depends on the Possibility of Inhaling Toxic Fumes and Pets at Home. Moreover, according to the Table 5.9, twenty-five symptom variables were used for the analysis, and only four of them are statistically significant at 5% significant level: Cough, Conjunctivitis, Tachypnoea, and Fatigue.

Table 5.10: Classification Table

	Observed		Predicted		
			RSV		Percentage Correct
			Negative	Positive	
Step 8	RSV	Negative	292	43	87.2
		Positive	101	61	37.7
	Overall Percentage				71.0
a. The cut value is .500					

According to the Table 5.10, 71.0% of patients were correctly classified irrespective of the status of RSV.

The false-positive error is predicted status of RSV is positive; given that the observed status is negative. The decision rule predicted the status of RSV as positive 104 times. That prediction was wrong 43 times, with a false positive rate of 43/104 = 41.35%.

The false-negative error is predicted status of RSV is negative; given that the observed status is positive. The decision rule predicted 393 times that the status of RSV negative. On the other hand, that prediction came wrong 101 times, with a false-negative rate of 101/393 = 25.70%. Then the statistics, sensitivity (see Table 5.11) and specificity are calculated according to the equations 3.9 and 3.10.

Table 5.11: Statistics on the Confusion Matrix

Sensitivity	37.65%
Specificity	87.16%
False Positive Rate	41.35%
False Negative Rate	25.70%
Overall % Correct	71.0%

The plotted ROC curve for the binary logistic model is given in Figure 5.1 below:

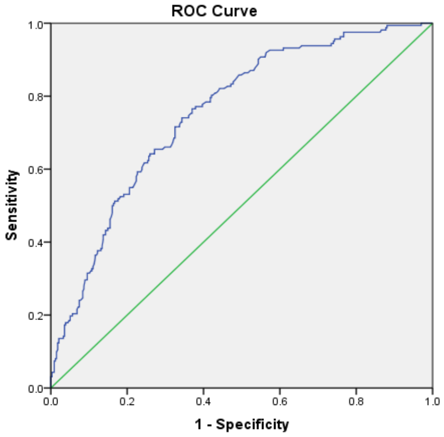


Figure 5.1:ROC Curve for the Binary Logistic Model

ROC is a probability curve, and it indicates how capable the model is in classifying the status of RSV. Higher AUC indicates better the model at predicting negatives as negatives and positives as positives. The null hypothesis, where the model discriminates two outcomes, is tested and given in the Table 5.12:

Table 5.12: Area Under the Curve

Test Result Variable(s): Predicted Probability				
Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.754	0.022	0.000	0.711	0.798
Null Hypothesis: true area = 0.5				

H_0: AUC = 0 .5 (No discrimination between status of RSV)

H_0: AUC ≠ 0 .5 (No discrimination between status of RSV)

Table 5.12 indicates that p-value to check the null hypothesis is 0.000, which is less than the desired 1% significant level. Therefore, the null hypothesis is rejected at 1% significant level. It can be concluded that the fitted model can be used to discriminate two outputs. Furthermore, the area under the curve is 0.754.

5.3 Approach of the Genetic Algorithm

Two different GAs are developed (Section 3.6) by taking two fitness functions called AUC and Likelihood to reduce the number of explanatory variables to predict the status of RSV. The number of iterations is predefined as the stopping criteria of maximizing fitness functions separately. Developed R codes were attached in Appendix D and E. Figure 3.2 is modified for this study (see Figure 5.2):

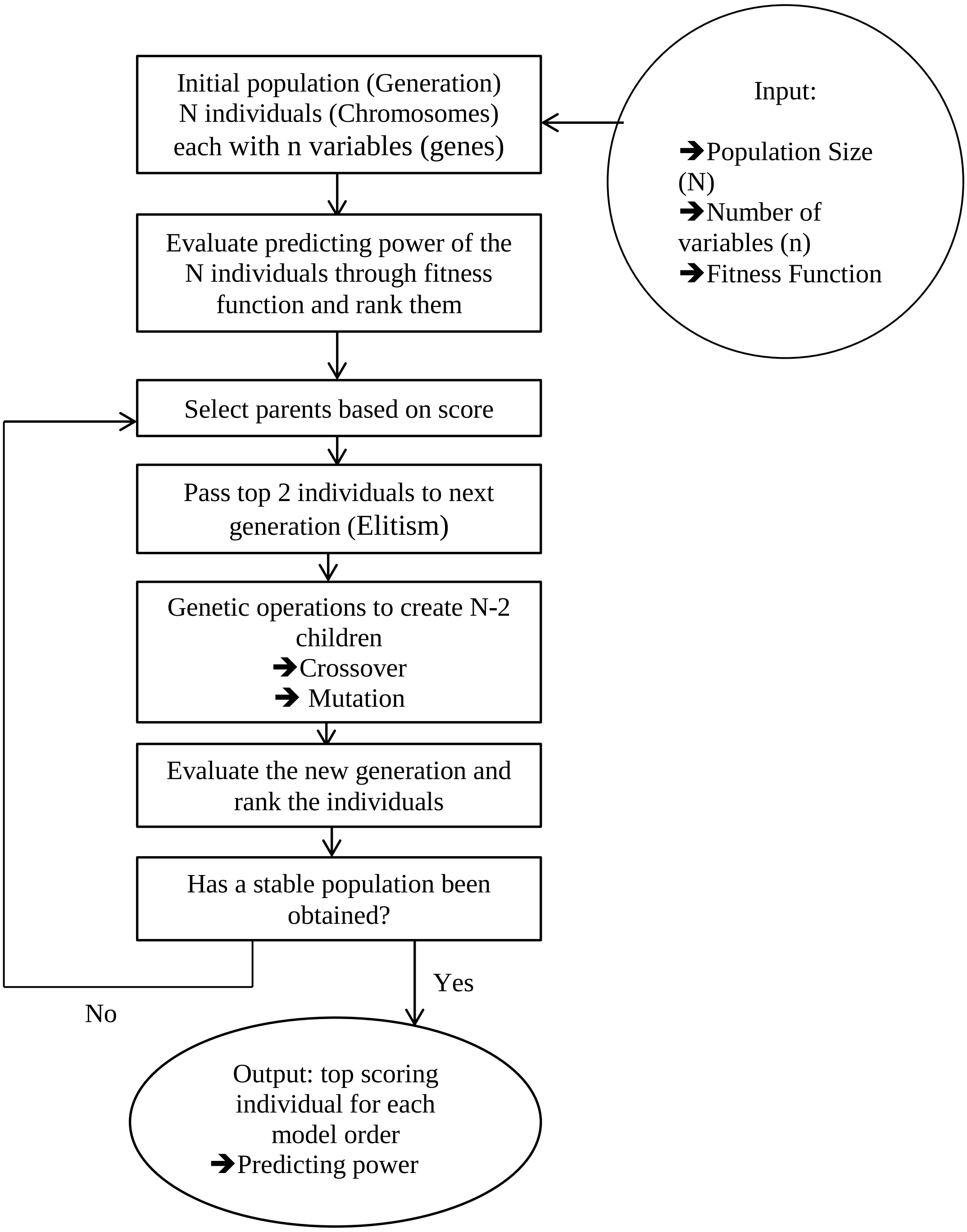

Initial population (Generation)
N individuals (Chromosomes)
each with n variables (genes)
Input:
➔Population Size (N)
➔Number of variables (n)
➔Fitness Function
Evaluate predicting power of the N individuals through fitness function and rank them
Select parents based on score
Pass top 2 individuals to next generation (Elitism)
Genetic operations to create N-2 children
➔Crossover
➔ Mutation
Evaluate the new generation and rank the individuals
Has a stable population been obtained?
Yes
No
Output: top scoring individual for each model order
➔Predicting power

Figure 5.2: Adapted GA

5.3.1 AUC Genetic Algorithm (AUC-GA)

The fitness function of the GA is developed according to Section 3.5.3. The objective is to maximize the AUC. The GA is repeated until the best individual in the population converges to an optimal solution, i.e. until one of these stopping criteria is met:

1. Maximum number of generations is reached - 100 iterations
2. The goal has not been changed over 50 generations – Goal is set to 0.9

The size of the chromosome is 20. That implies the models consist of 20 variables. If one of the above two stopping criteria is met, then the process of evolution will stop, and the process will proceed to another iteration of the search for a better solution (Zhang et al., 2018). The final iteration of AUC-GA is in the Table 5.13:

Table 5.13: Final Step of the AUC-GA Method

[e] Starting: Fitness Goal=0.9, Generations=(10 : 200)						
[e]	Elapsed Time	Generation	Fitness	%Fit	[Next Generations]	
[e]	0h 0m 0s	0	0.6192	68.8%	+-+--+--+++++-++-+--	
[e]	0h 3m 58s	20	0.66564	73.96%	+-+-++--+-+--++-++-+	
[e]	0h 7m 60s	40	0.69344	77.05%	--+-+-++--+-+-+-+--+	
[e]	0h 12m 2s	60	0.70145	77.94%	+---+--++-+-++--++--	
[e]	0h 16m 7s	80	0.65635	72.93%	-+-++-+-+--+--+--++-	
[e]	0h 20m 11s		100	0.70375	78.19%	+-+++-++-++-+-++-+-+
[e]	0h 24m 51s		120	0.68732	76.37%	---+-++-+--+-+-+----
[e]	0h 29m 25s		140	0.68347	75.94%	++++-+-+--+--++-+-+-
[e]	0h 33m 58s		160	0.66836	74.26%	-+--+-+-+--+-+--++--
[e]	0h 38m 2s	180	0.69183	76.87%	+--+-++-+-+++-+--+-+	
[e]	0h 42m 43s	***	200	0.75075	83.42%	NO SOLUTION, best : 24 7 14 2 30 11 26 37 26 24 13 4 7 34 33 24 24 25 3 29
[Bb]	100 0	NO Sol	0.75075	83.42%	200	2563.22s 251602s

According to the Table 5.13, the first row shows the fitness goal and the second column shows the elapsed time. The current number of generations is represented by the third column, which refreshed by 20 generations. Existing best fitness is given by the “Fitness” column and in”%Fit” shows the relative percentage with respect to the defined fitness goal. The behaviour of the next generation is illustrated by the “[NextGenerations]”. Following notations are used to represent the new generations:

"."	Maximum fitness of the current population has not been changed.
"+"	Maximum fitness of the current population has been increased
"-"	Maximum fitness of the current population has been decreased
"G"	The fitness goal has been reached

The numbers following "FINISH" is the indices of variables account for the best-fit chromosome. It just mentions "NO SOLUTION" if the fitness function does not achieve the defined goal. After 100 iterations, the top 5 variables are in black colour and named (Figure 5.3). Level of the contribution of other explanatory variables is positioned from left to right.

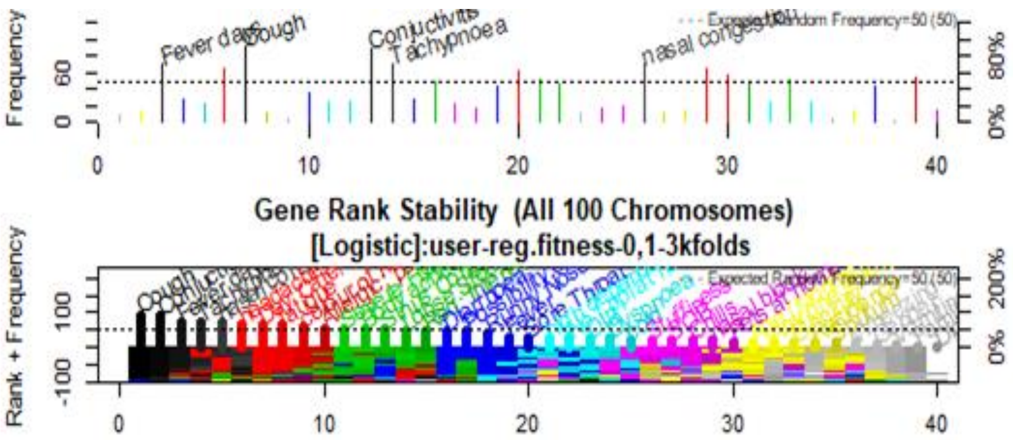


Figure 5.3: Level of Contribution to Predict the Status of RSV

The fitness values (Figure 5.4) of all AUC-GA evolutionary processes are traced across generations (thin grey lines). The blue curve shows the average fitness value across generations for all chromosomes.

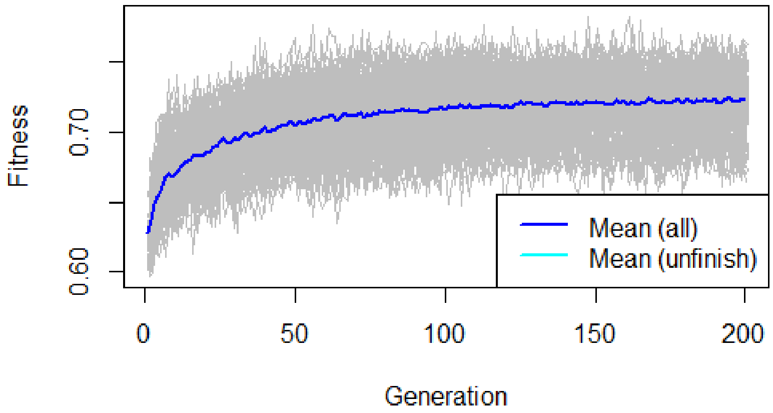


Figure 5.4: Convergence of Fitness Value on AUC-GA

The dependency of each explanatory variable can be illustrated using a network plot (Figure 5.5)

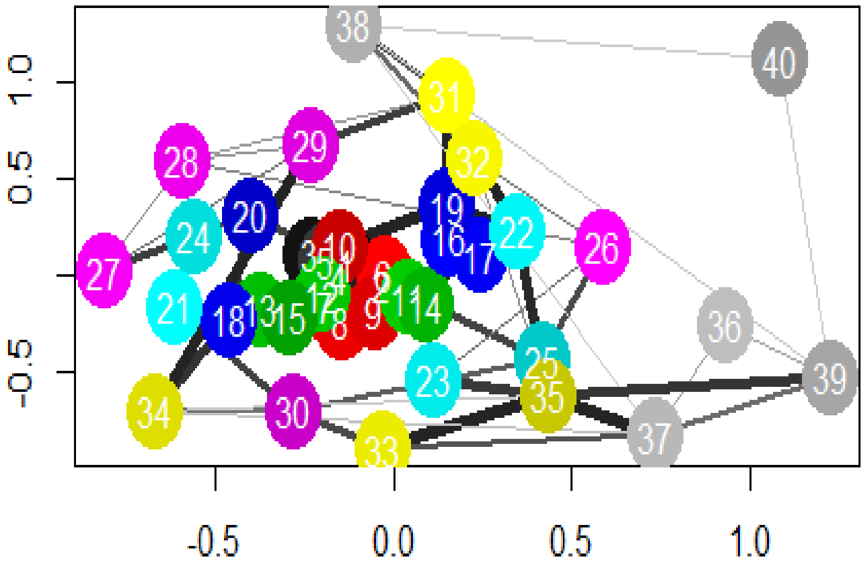


Figure 5.5: Dependency of each Explanatory Variable

The priority of variables on status of RSV is represented by numbers in Figure 5.5. The thickness of a line between two variables illustrates the strength of dependency between these two variables.

Table 5.14: Fitted Function Value for Several Models

AUC-GA	AUC
First 5 variables	0.6910
First 10 variables	0.7068
First 14 variables	0.7285
First 17 variables	0.7419

Top 5 variables, 10 variables, 14 variables and 17 explanatory variables were selected separately and fitted binary logistic regression models to measure AUC. Table 5.14 explains that value of AUC is increased when the number of explanatory variables is increased. The model fitted using the first 17 variables including one non-categorical variable, three risk factors, twelve symptom variables and one general variable give the maximum fitness function value (Table 5.14).

Non Categorical Variables	Fever Days
Risk Factors:	Does Patient Attend Daycare / Preschool
	Allergic to Pollen or Dust
	Possibility of Inhaling Toxic or Fumes
Symptom Variables	Conjunctivitis
	Cough
	Tachynoea
	Nasal Congestion
	Headache
	Fatigue
	Colour of Nasal Secretion
	Body Aches
	Severe Hydration
	Sinus Congestion
	Loose Stool
	Diarrhoea

and one general variable “Required Intensive Care” is the featured explanatory variables found by AUC-GA method to predict the status of RSV.

5.3.2 Log-likelihood Genetic Algorithm (LL-GA)

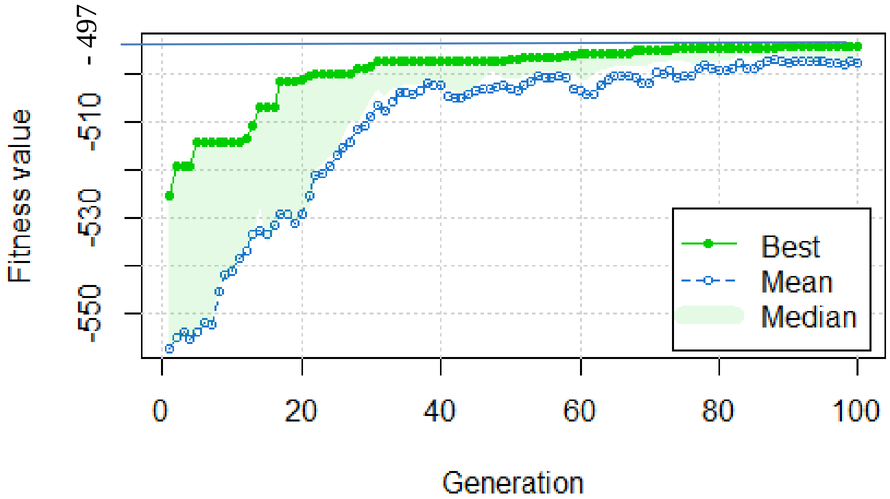


Figure 5.6: Convergence of Fitness Value on LL-GA Method

The fitness function of the GA is developed according to the Section 3.6.1. Therefore, the objective is to maximize $-(-2\, log\, likelihood)$ value. Stopping criteria is defined as 100 iterations. The summary of the LL-GA is given in the Table 5.15. Developed R codes were attached in Appendix C.

Table 5.15: Summary of LL-GA

-- Genetic Algorithm -------------------
GA settings:
Type = binary
Population size = 50
Number of generations = 100
Elitism = 2
Crossover probability = 0.9
Mutation probability = 0.1
GA results:
Iterations = 60
Fitness function value = -494.1027
Solution =
Gender `Age in months` `Fever days` Fever cold headache Cough Sputum Vomiting `Runny Nose`
[1,] 1 1 1 1 1 1 1 1 1 1
... `Allergic: pollen/dust` `Pets at home`

The population size is taken to be 50, and the number of future generations is limited to 100 to implement LL-GA. Defining elitism helps to keep the best chromosome, and automatically it goes to the next generation. Here, elitism is defined to be 2. Crossover and mutation probabilities are defined as 0.9 and 0.1, respectively (Johnson et al., 2014). After iterating 100 generations, the fitted function value took -494.1027. Figure 5.6 illustrates the convergence of the fitted function value.

In the Figure 5.6, the best-fitted function value represented by green colour and mean of it indicated in blue colour. It can be observed that the best-fitted value approaches to -497 after the last 60 iteration and best 17 explanatory variables including all three non-categorical variables, five risk factors, eight symptom variables and one general variable were identified to predict the status of RSV.

Non categorical Variables	Age in months
	Fever Days
	Number of People at Home
Five Risk Factors	Family History of Asthma Fumes
	Possibility of Inhaling Toxic or Fumes
	Did Patient Travel Recently
	Allergic to Pollen or Dust
	Pets at Home
Eight Symptom Variables	Runny Nose
	Sinus Congestion
	Body Aches
	Loose Stool
	Nasal Block
	Nasal Congestion
	Colour of Nasal Secretion
	Severe Hydration

and one general variable "Required Intensive Care" is the featured explanatory variable found by LLGA method to predict the status of RSV.

CHAPTER 6

COMPARISON OF THREE METHODS

In medical researches, one of the objectives is to identify a disease in the highest degree of accuracy. A set of risk factors and the types of symptoms were used to identify the disease at first and later biological tests will be carried out for further investigation. Forward and backward variable selection methods in binary logistic regression analysis is one of the most frequently used statistical methodologies to evaluate multiple explanatory variables over a binary dependent variable in clinical trials. Even though GA and data mining methods such as supportive vector methods and neural network methods are commonly used methods in other fields, they have not been used widely in the field of medicine. In this study, a set of clinical data is analysed using variable selection methods in binary logistic regression model and two different GAs. The aim of this study is to identify a combination of factors which affects the status of RSV. A comparison of the goodness of fit measurements between the Three Methods is made to identify the best predictive method to reduce variables.

The best set of explanatory variables to predict the status of RSV is analyzed by three methods explained in sections 5.2, 5.3.1 and 5.3.2. Table 6.1 summarised the goodness of fit statistics for three methods:

Table 6.1: Summary Statistics

Method	Number of variables	-2loglikelihood	AUC	Cox & Snell R-squared	Nagelkerke R Squared	Correctly Classified Percentage	Specificity	Sensitivity
LR	8	535.674	0.754	0.169	0.235	71.0%	87.16%	37.65%
AUC-GA	17	539.160	0.742	0.167	0.233	71.40%	87.8%	35.8%
LL-GA	17	495.204	0.790	0.212	0.289	75.10%	70.8%	75.6%

LL-GA emerged as the best method to predict the status of RSV according to Table 6.1. For LL-GA *-2 log-likelihood* (495.204), Cox & Snell (0.212), Nagelkerke R Squared (0.289) and the sensitivity (75.6%) are higher than that of logistic regression and AUC-GA methods. The AUC was also found to be higher in the other two models. However, the LR approach stated 71.0% while LL-GA has recorded the 75.10% as the correctly classified percentage, which is higher than that of AUC-GA. Besides, LL-GA method has 70.8% specificity while both LR and AUC-GA have a percentage of 87.16% and 87.8% of specificity respectively. In general, higher sensitivity leads to lower specificity and vice versa. When the sensitivity increases, the ability to correctly identify people who have the disease also increases. Among the three methods used in this study indicate that the highest sensitivity percentage is scored by LL-GA. Moreover, LL-GA gives better predictions when considering both sensitivity and specificity since they are higher than 70%.

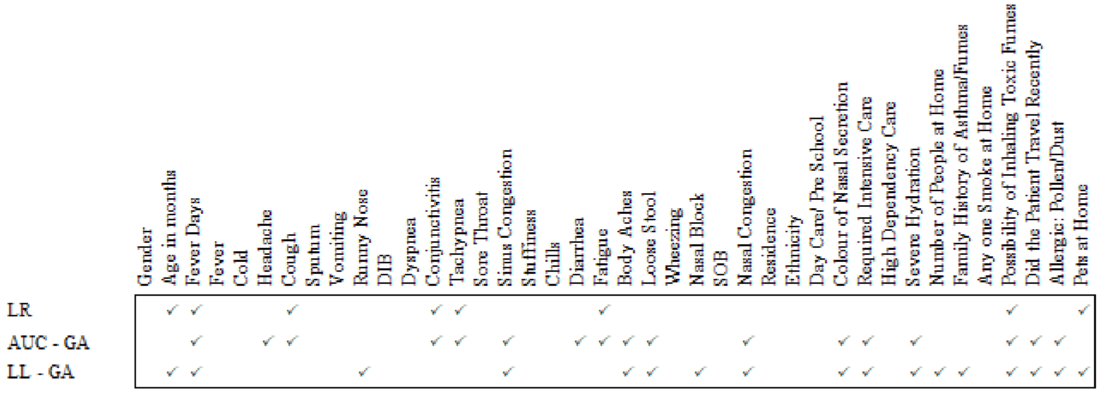
On the other hand, when considered about the number of reduced variables in the final model, LR gives the lesser variables, where it gives all the above explained goodness of statistics only for 8 variables while LL-GA scored better goodness of fit measurements by reducing the number of variables into 17. Contrastively, LL-GA method gives 19 variables when it has around 511 for the *-2 loglikelihood* measurement. Even though the LR method has relatively lower accuracy of predicting the status of RSV according to the goodness of fit measurements, it accumulates only for 8 explanatory variables. Change of the goodness of fit measurements could be insignificant compared to the dimensions of optimum solutions obtained by LR and LL-GA. Therefore, it can be identified that LR has a higher power of prediction capacity of the status of RSV although LL-GA gives slightly better goodness of measurements.

Selected variables from the three methods were summarized in Table 6.2.

Table 6.2: Reduced Variables by Three Methods

LL-GA	**LL-AUC**	**LR**	
17 Variables	17 Variables	8 Variables	p-value
Age in months	Fever Days	Age in months	0.001
Fever Days	Does Patient Attend Daycare / Preschool	Fever Days	0.000
Number of People at Home	Allergic to Pollen or Dust	Cough(1)	0.007
Family History of Asthma Fumes	Possibility of Inhaling Toxic/Fumes	Conjunctivitis(1)	0.039
Possibility of Inhaling Toxic/Fumes	Conjunctivitis	Tachypnoea(1)	0.040
Did Patient Travel Recently	Cough	Fatigue(1)	0.016
Allergic to Pollen or Dust	Tachypnea	Possibility of Inhaling Toxic/Fumes(1)	0.001
Pets at Home	Nasal Congestion	Pets at Home(1)	0.008
Runny Nose	Headache		
Sinus Congestion	Fatigue		
Body Aches	Colour of Nasal Secretion		
Loose Stool	Body Aches		
Nasal Block	Severe Hydration		
Nasal Congestion	Sinus Congestion		
Colour of Nasal Secretion	Loose Stool		
Severe Hydration	Diarrhea		
Required Intensive Care	Required Intensive Care		

Table 6.2 indicates that “Fever Days” and “Possibility of Inhaling Toxic/Fumes” have been identified as significant factors out of all three methods to detect the status of RSV. Moreover, these two variables are significant at 5% significance level in the logistic regression model. The variables “Age in months” and “Pets at Home” are found to be common in logistic regression and LL-GA methods. These two variables are also significant at 5% significance level in the logistic regression model.



	Gender	Age in months	Fever Days	Fever	Cold	Headache	Cough	Sputum	Vomiting	Runny Nose	DIB	Dyspnea	Conjunctivitis	Tachypnea	Sore Throat	Sinus Congestion	Stuffiness	Chills	Diarrhea	Fatigue
LR		✓	✓				✓						✓	✓						✓
AUC - GA			✓			✓	✓						✓	✓		✓			✓	✓
LL - GA		✓	✓							✓						✓				

	Body Aches	Loose Stool	Wheezing	Nasal Block	SOB	Nasal Congestion	Residence	Ethnicity	Day Care/ Pre School	Colour of Nasal Secretion	Required Intensive Care	High Dependency Care	Severe Hydration	Number of People at Home	Family History of Asthma/Fumes	Any one Smoke at Home	Possibility of Inhaling Toxic Fumes	Did the Patient Travel Recently	Allergic: Pollen/Dust	Pets at Home
LR																	✓			✓
AUC - GA	✓	✓				✓				✓	✓		✓				✓	✓	✓	
LL - GA	✓	✓		✓		✓				✓	✓		✓	✓	✓		✓	✓	✓	✓

Figure 6.1: Reduced Variables to Predict the Status of RSV

In summary (see Figure 6.1), “Age in months”, “Fever days”, “Cough”, “Conjunctivitis” , “Tachypnoea”, “Fatigue”, “Possibility of Inhaling Toxic/Fumes” and “Pets at Home” are the identified combination of variables for logistic regression method and “Age in months”, “Fever days”, “Number of People at Home”, “Family History of Asthma Fumes”, “Possibility of Inhaling Toxic or Fumes”, “Did Patient Travel Recently”, “Allergic to Pollen or Dust”, “Pets at Home”, “Runny Nose”, “Sinus Congestion”, “Body Aches”, “Loose Stool”, “Nasal Block”, “Nasal Congestion”, “Colour of Nasal Secretion”, “Severe Hydration” and “Required Intensive Care” are the identified combinations of variables for prediction of progression of the status of RSV for a child in LL-GA method.

CHAPTER 7

CONCLUSION, RECOMMENDATIONS AND FUTURE WORK

7.1 Conclusion

We have presented two GA based methods and binary logistic regression method to select variables and observe the predictive ability of a combination of variables that can estimate the progression of a disease. A comparison of variable selection methods in GA over logistic regression technique is carried out by testing the same set of clinical data. Results indicate that the efficiency of the resulting classifier increased when using GA by implementing log-likelihood as the fitness function to reduce variables within the framework of GA. It is highlighted that contributions of variable sets are more important than the contributions of individual variables for the accuracy of the model.

On the other hand, GA method to reduce variables provides the decision-maker with a broader opportunity for setting parameters, such as in assigning the parameters of GA, mutation method, selecting method and choosing necessary results of the pool of solutions. The advantage of using GA to select variables is that the data does not need to satisfy any predefined assumptions in the background of the study. Moreover, variables selection procedure over GA is more suitable when solving high dimensional classification problems in a large, complex, or vaguely understood searching space.

However, the logistic regression method has the ability to reduce the number of variables into a lower number compared to GA. Hence, accurate predictions can be made using the logistic regression method. Furthermore, logistic regression method is easy to apply and it consumes less time to build a model.

7.2 Recommendations

Important recommendations can be stated when selecting the method of variable selection as follows.

1. Choose logistic regression when:
    a. it needs to filter less number of predictor variables which are most important.
    b. time limitation for the solution is given.

2. Choose GA when:
    a. predefined assumptions are not satisfied.
    b. the problem involves high dimensional predictors with a complex searching space.
    c. need to design the objective (fitness function) to target specific research question directly.

7.3 Future Work

Some basic aspects of classification concepts have been explored in this study. There is a considerable potential future work can be done using the concept of GA for statistical approaches by:

a. designing different types of classifiers (such as SVM and neural networks) to evaluate the prediction.
b. considering situations where the misclassification has not the same weight from the application point of view.
c. developing GAs using different fitness functions and test in order to maximize the classification performance.
d. different stopping criteria can be defined when the algorithm does not show a significant improvement for a certain number of generations.

References

Abdulqader, S. A., & Saeed, B. A. (2017). Characteristics of patients attending the child and adolescent psychiatric outpatient clinic in Erbil city. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0209418

Abedin, T., Chowdhury, M. Z. I., Afzal, A., Turin, T., & Turin, T. C. (2016). Application of Binary Logistic Regression in Clinical Research Corresponding Author. *Journal of National Heart Foundation of Bangladesh*, *5*(1), 8–11. https://www.researchgate.net/publication/320432727

Agresti, A. (2007). Building and Applying Logistic Regression Models. In *An Introduction to Categorical Data Analysis*. https://doi.org/10.1002/9780470114759.ch5

Al-Douri, Y. K., Hamodi, H., & Lundberg, J. (2018). Time series forecasting using a two-level multi-objective genetic algorithm: A case study of maintenance cost data for tunnel fans. *Algorithms*, *11*(8), 4–9. https://doi.org/10.3390/a11080123

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. In *Accident Analysis and Prevention* (Vol. 34, Issue 6, pp. 729–741). https://doi.org/10.1016/S0001-4575(01)00073-2

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, *54*(10), 979–985. https://doi.org/10.1016/S0895-4356(01)00372-9

Bergtold, J. S., Yeager, E. A., & Featherstone, A. M. (2018). Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *Journal of Applied Statistics*, *45*(3), 528–546. https://doi.org/10.1080/02664763.2017.1282441

Bhattacharyya, S., & Pendharkar, P. C. (1998). Inductive, evolutionary, and neural computing techniques for discrimination: A comparative study. *Decision Sciences*, *29*(4), 871–899. https://doi.org/10.1111/j.1540-5915.1998.tb00880.x

Bujang, M. A., Sa'At, N., Tg Abu Bakar Sidik, T. M. I., & Lim, C. J. (2018). Sample size guidelines for logistic regression from observational studies with large population: Emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malaysian Journal of Medical Sciences*, *25*(4), 122–130. https://doi.org/10.21315/mjms2018.25.4.12

Cateni, S., Colla, V., & Vannucci, M. (2010). Variable selection through genetic algorithms for classification purposes. *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications, AIA 2010*, *January 2020*, 6–11. https://doi.org/10.2316/p.2010.674-080

Chiang, L. H., & Pell, R. J. (2004). Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control*, *14*(2), 143–155. https://doi.org/10.1016/S0959-1524(03)00029-5

Chauhan, V., Suman, H. K., & Bolia, N. B. (2016). Binary logistic model for estimation of mode shift into Delhi Metro. The Open Transportation Journal, 10(1).

Cockburn, W. C., & Assaad, F. (1973). Some observations on the communicable diseases as public health problems. *Bulletin of the World Health Organization*, *49*(1), 1–12.

Cooper, M. (2010). Advanced Bash-Scripting Guide An in-depth exploration of the art of shell scripting Table of Contents. *Okt 2005 Abrufbar Uber Httpwww Tldp OrgLDPabsabsguide Pdf Zugriff 1112 2005*, *2274*(November 2008), 2267–2274. https://doi.org/10.1002/hyp

F.R., & S., (2009). Economic instability and its impact on decision making in health care. *P and T*.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information and Management*. https://doi.org/10.1016/S0378-7206(99)00051-8

G.M., B., R.F., B., D.R., F., A.O., A., B., A., L., C., M.K., N., B.S., F., V., O., H., N., P.M., O., D.O., M., G.O., A., B., O., M.A., K., J.M., M., & D.C., B. (2013).

Epidemiology of respiratory syncytial virus infection in rural and urban Kenya. *Journal of Infectious Diseases*.

Gayou, O., Das, S. K., Zhou, S. M., Marks, L. B., Parda, D. S., & Miften, M. (2008). A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes. *Medical Physics*, *35*(12), 5426–5433. https://doi.org/10.1118/1.3005974

Hoffman, J. I. E. (2019). Basic biostatistics for medical and biomedical practitioners. In *Biostatistics for Medical and Biomedical Practitioners*. https://doi.org/10.1016/C2018-0-02190-8

Hosmer, D. W., Lemeshow, S., & Klar, J. (1988). Goodness-of-Fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small. *Biometrical Journal*. https://doi.org/10.1002/bimj.4710300805

Iquebal, M. A., Prajneshu, & Ghosh, H. (2012). Genetic algorithm optimization technique for linear regression models with heteroscedastic errors. *Indian Journal of Agricultural Sciences*, *82*(5), 422–425.

Jayaweera, J. A. A. S., Noordeen, F., Morel, A., Pitchai, N., Kothalawala, S., Abeykoon, A. M. S. B., & Peiris, J. S. M. (2016). Viral burden in acute respiratory tract infections in hospitalized children in the wet and dry zones of Sri Lanka. *International Journal of Infectious Diseases*, *45*, 463. https://doi.org/10.1016/j.ijid.2016.02.980

Johnson, P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P., Macaulay, L. S., Ellis, K. A., Szoeke, C., Martins, R. N., Rowe, C. C., Masters, C. L., Ames, D., & Zhang, P. (2014). Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics*, *15*(Suppl 16), S11. https://doi.org/10.1186/1471-2105-15-S16-S11

Kassem, E., Na, W., Bdair-amsha, A., Zahalkah, H., & Muhsen, K. (2019). *Comparisons between ethnic groups in hospitalizations for respiratory syncytial virus bronchiolitis in Israel*.

Mahikul, W., White, L. J., Poovorawan, K., Soonthornworasiri, N., Sukontamarn, P., Chanthavilay, P., Medley, G. F., & Pan-Ngum, W. (2019). Modeling household dynamics on Respiratory Syncytial Virus (RSV). *PLoS ONE*, *14*(7), 1–13. https://doi.org/10.1371/journal.pone.0219323

Manouchehrian, A., Sharifzadeh, M., Hamidzadeh Moghadam, R., & Nouri, T. (2013). Selection of regression models for predicting strength and deformability properties of rocks using GA. *International Journal of Mining Science and Technology*, *23*(4), 495–501. https://doi.org/10.1016/j.ijmst.2013.07.006

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253–267. https://doi.org/10.1080/09720502.2010.10700699

Minerva, T., & Paterlini, S. (2002). Evolutionary approaches for statistical modelling. *Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002*, *2*(January), 2023–2028. https://doi.org/10.1109/CEC.2002.1004554

Minghua, S., Qingxian, X., Benda, Z., & Feng, Y. (2017). Regresijsko modeliranje zasnovano na poboljšanom genetičkom algoritmu. *Tehnicki Vjesnik*, *24*(1), 63–70. https://doi.org/10.17559/TV-20160525104127

Muthulingama, A., Noordeena, F., & Morelb, A. J. (2014). Viral etiology in hospitalized children with acute respiratory tract infection in the Kegalle area of Sri Lanka. *Journal of Pediatric Infectious Diseases*, *9*(4), 167–170. https://doi.org/10.3233/JPI-140432

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v053.i04

Shi, T., McAllister, D. A., O’Brien, K. L., Simoes, E. A. F., Madhi, S. A., Gessner, B. D., Polack, F. P., Balsells, E., Acacio, S., Aguayo, C., Alassani, I., Ali, A., Antonio, M., Awasthi, S., Awori, J. O., Azziz-Baumgartner, E., Baggett, H. C., Baillie, V. L., Balmaseda, A., … Nair, H. (2017). Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *The Lancet*, *390*(10098), 946–958. https://doi.org/10.1016/S0140-

6736(17)30938-8

Sukono, Sholahuddin, A., Mamat, M., & Prafidya, K. (2014). Credit scoring for Cooperative of financial services using logistic regression estimated by genetic algorithm. *Applied Mathematical Sciences*, *8*(1–4), 45–57. https://doi.org/10.12988/ams.2014.310600

Sze, N. N., & Wong, S. C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. *Accident Analysis and Prevention*. https://doi.org/10.1016/j.aap.2007.03.017

Trevino, V., & Falciani, F. (2006). GALGO: An R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, *22*(9), 1154–1156. https://doi.org/10.1093/bioinformatics/btl074

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in medicine*, *19*(24), 3417-3432.

Vinterbo, S., & Ohno-Machado, L. (1999). A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*.

Weber, M. W., Mulholland, E. K., & Greenwood, B. M. (1998). Respiratory syncytial virus infection in tropical and developing countries. *Tropical Medicine and International Health*, *3*(4), 268–280. https://doi.org/10.1046/j.1365-3156.1998.00213.x

Wu, A. S., & Norat, R. (2017). *A Genetic Algorithm Approach to Predictive Modeling of Medicare Payments to Physical Therapists*. 311–316.

Zhang, Z., Trevino, V., Hoseini, S. S., Belciug, S., Boopathi, A. M., Zhang, P., Gorunescu, F., Subha, V., & Dai, S. (2018). Variable selection in Logistic regression model with genetic algorithm. *Annals of Translational Medicine*, *6*(3), 45–45. https://doi.org/10.21037/atm.2018.01.15

APPENDIX A: Description of Variables

Table 1: Description of Variables

Type	No	Name	Description	Levels
Dependent Variable	1	RSV	The child is RSV infected or not	RSV positive =1 RSV negative =0
Independent Variables	1	Age in months	Age of the child in months	Non categorical variable

	2	Gender	Gender of the child	Female =1
				Male =0

	3	Residence	Residence of the child	Urban =1
				Rural =2
				Estate =3

	4	Ethnicity	Ethnicity of the child	Sinhala =1
				Tamil =2
				Muslim =3

	5	Fever Days	Number of days that the child have fever	Non categorical variable

	6	Fever	Does the child have fever	Yes = 1
				No =0

	7	Cold	Does the child feel cold	Yes =1
				No =0

	8	Headache	Does the child have a headache	Yes =1
				No = 0

	9	Cough	Does the child have a cough	Yes =1
				No= 0

	10	Sputum	Does the child have thick mucus that is coughed up from the lungs	Yes = 1
				No = 0

	11	Vomiting	Does the child vomit	Yes = 1
				No =0

	12	Runny Nose	Does the child have excess nasal drainage	Yes =1
				No = 0

	13	DIB	Does the child have Difficulty In Breathing	Yes =1
				No = 0

Type	No	Name	Description	Levels

	14	Dyspnoea	Does the child have not being able to breathe well enough	Yes =1
				No = 0

	15	Conjunctivitis	Does the child have an inflammation of the membrane covering the surface of the eyeball	Yes =1
				No = 0

	16	Tachypnea	Does the child have abnormal rapid breathing	Yes =1
				No = 0

	17	Sore Throat	Does the child have a sore throat	Yes =1
				No = 0

	18	Sinus Congestion	Does the child have a result of inflammation of his nasal passages and the sinuses	Yes =1
				No = 0

	19	Stuffiness	Does the child feel of lacking good airflow or ventilation	Yes =1
				No = 0

	20	Chills	Does the child feel of being cold without an apparent cause	No =0
				Yes =1

	21	Diarrhea	Does the child's body's solid waste is more liquid than usual	Yes =1
				No = 0

	22	Fatigue	Does the child feel being extremely tired	Yes =1
				No = 0

	23	Body Aches	Does the child have anybody aches	Yes =1
				No = 0

	24	Loose Stool	Does the child have loose stool	Yes =1
				No = 0

	25	Wheezing	Does the child have wheezing	Yes =1
				No = 0

	26	Nasal Block	Does the child have a nasal block	Yes =1
				No = 0

	27	SOB	Does the child have shortness of breath	Yes =1
				No = 0

Type	No	Name	Description	Levels

	28	Nasal Congestion	Does the child feel nasal congestion	Yes =1
				No = 0

	29	Day Care/ Pre School	Does the child go to daycare of preschool	Yes =1
				No = 0

	30	Colour of Nasal Secretion	What does the child's nasal secretion colour	White = 1 Off white = 2

	31	Required Intensive Care	Does the child need intensive care (One nurse to care for the patient)	Yes =1
				No = 0

	32	High Dependency Care	Does the child need high dependency care (Many madrigal professionals should attend to the patient)	Yes =1
				No = 0

	33	Severe Hydration	Does the child have severe hydration	Yes =1
				No = 0

	34	Number of People at Home	Number of people live with the child at home	Non categorical variable

	35	Family History of Asthma Fumes	Is there a family history of respiratory disease/asthma	Yes =1
				No = 0

	36	Anyone Smokes at Home	Does anyone smoke at home	Yes =1
				No = 0

	37	Possibility of Inhaling Toxic / Fumes	Is there a possibility of inhaling toxic chemicals/fumes at home or surrounding	Yes =1
				No = 0

	38	Did Patient Travel Recently	Does the child travel recently (past 14 days)	Yes =1
				No = 0

	39	Allergic Pollen/Dust	Does the child have an allergic reaction to pollen/dust	Yes =1
				No = 0

	40	Pets at Home	Does the child have pets at home	Yes =1
				No = 0

APPENDIX B: Relationship between Independent Variables

Table 2: Relationship between independent variables

		Age in months	Fever Days	Headache	Cough	Dyspnea	Conjunctivitis	Tachypnea	Stuffiness	Diarrhea	Fatigue	High Dependency Care	Severe Hydration	Number of People at Home	Possibility of Inhaling Toxic Fumes	Pets at Home
Age (in months)	Strength															
	p-value	1.00														
Fever Days	Strength	0.07														
	p-value	0.10	1.00													
Headache	Strength	5.69	2.65													
	p-value	0.02	0.10	1.00												
Cough	Strength	4.17	0.10	91.18												
	p-value	0.04	0.76	0.00	1.00											
Dyspnea	Strength	4.96	4.18	2.64	0.17											
	p-value	0.03	0.04	0.10	0.68	1.00										
Conjunctivitis	Strength	0.18	0.34	76.24	52.90	0.46										
	p-value	0.68	0.56	0.00	0.00	0.50	1.00									
Tachypnea	Strength	0.63	0.54	0.87	1.12	174.94	1.89									
	p-value	0.43	0.46	0.35	0.29	0.00	0.17	1.00								
Stuffiness	Strength	6.88	0.08	1.31	0.75	6.73	3.12	4.39								
	p-value	0.01	0.78	0.25	0.39	0.01	0.77	0.04	1.00							
Diarrhea	Strength	2.45	0.23	2.95	1.68	28.57	2.04	39.74	0.45							
	p-value	0.12	0.63	0.09	0.20	0.00	0.15	0.00	0.51	1.00						
Fatigue	Strength	2.90	3.68	28.16	0.67	0.86	0.89	1.12	0.75	0.19						
	p-value	0.09	0.06	0.00	0.41	0.36	0.35	0.29	0.39	0.66	1.00					
High Dependency Care	Strength	2.81	0.95	0.65	0.00	22.68	0.01	43.35	0.77	1.54	0.00					
	p-value	0.09	0.33	0.42	1.00	0.00	0.92	0.00	0.38	0.22	1.00	1.00				
Severe Hydration	Strength	0.55	1.32	0.35	0.20	10.50	0.19	4.00	0.07	5.34	0.20	18.17				
	p-value	0.47	0.25	0.55	0.65	0.001	0.67	0.05	0.79	0.02	0.65	0.00	1.00			
Number of People at Home	Strength	0.03	0.06	1.56	6.52	10.49	5.66	9.39	0.05	6.58	2.08	1.74	1.38			
	p-value	0.47	0.15	0.21	0.01	0.00	0.02	0.00	0.83	0.01	0.15	0.19	0.24	1.00		
Possibility of Inhaling Toxic Fumes	Strength	1.47	0.63	4.67	4.80	0.00	14.73	0.65	0.13	0.05	4.99	0.01	0.89	1.42		
	p-value	0.23	0.43	0.03	0.03	0.95	0.00	0.42	0.72	0.82	0.03	0.90	0.35	0.23	1.00	
Pets at Home	Strength	0.04	12.24	3.60	3.53	9.30	6.78	8.07	0.00	1.30	7.35	6.55	0.05	6.64	5.67	
	p-value	0.85	0.00	0.06	0.06	0.00	0.01	0.00	1.00	0.25	0.01	0.01	0.82	0.01	0.02	1.00

APPENDIX C: Six Variable Reduction Methods in SPSS

Table.3: Forward Conditional Method

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[a]	Age in months	-.038	.011	11.212	1	.001	.962
	Fever Days	-.386	.085	20.542	1	.000	.680
	Cough(1)	2.062	.765	7.263	1	.007	7.864
	Conjunctivitis(1)	-1.445	.699	4.275	1	.039	.236
	Tachypnoea(1)	.512	.250	4.202	1	.040	1.669
	Fatigue(1)	-1.391	.578	5.795	1	.016	.249
	Possibility of Inhaling Toxic Fumes(1)	-.823	.253	10.627	1	.001	.439
	Pets at Home(1)	.632	.239	7.000	1	.008	1.880
	Constant	-.631	.818	.595	1	.441	.532

a. Variable(s) entered on step 8: Conjunctivitis.

Table.4: Forward Likelihood Ratio

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[a]	Age in months	-.038	.011	11.212	1	.001	.962
	Fever Days	-.386	.085	20.542	1	.000	.680
	Cough(1)	2.062	.765	7.263	1	.007	7.864
	Conjunctivitis(1)	-1.445	.699	4.275	1	.039	.236
	Tachypnoea(1)	.512	.250	4.202	1	.040	1.669
	Fatigue(1)	-1.391	.578	5.795	1	.016	.249
	Possibility of Inhaling Toxic Fumes(1)	-.823	.253	10.627	1	.001	.439
	Pets at Home(1)	.632	.239	7.000	1	.008	1.880
	Constant	-.631	.818	.595	1	.441	.532

a. Variable(s) entered on step 8: Conjunctivitis.

Table.5: Forward Wald

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[a]	Age in months	-.038	.011	11.212	1	.001	.962
	Fever Ddays	-.386	.085	20.542	1	.000	.680
	Cough(1)	2.062	.765	7.263	1	.007	7.864
	Conjunctivitis(1)	-1.445	.699	4.275	1	.039	.236
	Tachypnoea(1)	.512	.250	4.202	1	.040	1.669
	Fatigue(1)	-1.391	.578	5.795	1	.016	.249
	Possibility of Inhaling Toxic Fumes(1)	-.823	.253	10.627	1	.001	.439
	Pets at Home(1)	.632	.239	7.000	1	.008	1.880
	Constant	-.631	.818	.595	1	.441	.532
a. Variable(s) entered on step 8: Conjunctivitis.							

Table.6: Backward Conditional Method

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 27[a]	Age in months	-.035	.012	8.982	1	.003	.966
	Fever Days	-.388	.088	19.435	1	.000	.678
	Cough(1)	1.973	.774	6.500	1	.011	7.193
	Conjunctivitis(1)	-1.322	.673	3.859	1	.049	.267
	Sinus Congestion(1)	-23.043	17247.087	.000	1	.999	.000
	Stuffiness(1)	1.755	.766	5.247	1	.022	5.782
	Diarrhoea(1)	1.252	.486	6.639	1	.010	3.497
	Fatigue(1)	-1.446	.585	6.105	1	.013	.236
	Severe Hydration(1)	21.319	23119.284	.000	1	.999	1813747269.875
	No of People at Home	-.148	.082	3.238	1	.072	.862
	Possibility of Inhaling Toxic Fumes(1)	-.928	.257	13.009	1	.000	.395
	Allergic Pollen Dust(1)	-22.294	40192.970	.000	1	1.000	.000
	Pets at Home(1)	.823	.246	11.224	1	.001	2.276
	Constant	.125	.901	.019	1	.890	1.133

Table.7: Forward Likelihood Ratio

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 27[a]	Age in months	-.035	.012	8.982	1	.003	.966
	Fever Days	-.388	.088	19.435	1	.000	.678
	Cough(1)	1.973	.774	6.500	1	.011	7.193
	Conjunctivitis(1)	-1.322	.673	3.859	1	.049	.267
	Sinus Congestion(1)	-23.043	17247.087	.000	1	.999	.000
	Stuffiness(1)	1.755	.766	5.247	1	.022	5.782
	Diarrhoea(1)	1.252	.486	6.639	1	.010	3.497
	Fatigue(1)	-1.446	.585	6.105	1	.013	.236
	Severe Hydration(1)	21.319	23119.284	.000	1	.999	1813747269.875
	No of People at Home	-.148	.082	3.238	1	.072	.862
	Possibility of Inhaling Toxic Fumes(1)	-.928	.257	13.009	1	.000	.395
	Allergic Pollen Dust(1)	-22.294	40192.970	.000	1	1.000	.000
	Pets at Home(1)	.823	.246	11.224	1	.001	2.276
	Constant	.125	.901	.019	1	.890	1.133

Table.8: Backward Wald

Variables in the Equation[b]							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 32[c]	Age in months	-.035	.012	9.069	1	.003	.966
	Fever Days	-.398	.087	21.022	1	.000	.672
	Cough(1)	2.037	.771	6.979	1	.008	7.670
	Conjunctivitis(1)	-1.321	.670	3.890	1	.049	.267
	Sinus Congestion(1)	-22.920	17234.144	.000	1	.999	.000
	Stuffiness(1)	1.751	.772	5.152	1	.023	5.763
	Diarrhoea(1)	1.241	.478	6.722	1	.010	3.457
	Fatigue(1)	-1.474	.585	6.361	1	.012	.229
	Possibility of Inhaling Toxic Fumes(1)	-.917	.255	12.956	1	.000	.400
	Pets at Home(1)	.742	.240	9.553	1	.002	2.100
	Constant	-.550	.819	.450	1	.502	.577

APPENDIX C: Categorical Variables Coding

Table 3: Categorical Variables Coding

No	Variable		Frequency	Parameter	
				(1)	(2)
1	Residence	Urban	219	.000	.000
		Rural	269	1.000	.000
		Estate	9	.000	1.000
2	Ethnicity	Sinhala	477	.000	.000
		Tamil	13	1.000	.000
		Muslim	7	.000	1.000
3	Pets at home	No	361	.000	
		Yes	136	1.000	
4	Chills	No	435	.000	
		Yes	62	1.000	
5	Stuffiness	No	486	.000	
		Yes	11	1.000	
6	Sinus Congestion	No	493	.000	
		Yes	4	1.000	
7	Sore Throat	No	467	.000	
		Yes	30	1.000	
8	Tachypnoea	No	396	.000	
		Yes	101	1.000	
9	Conjunctivitis	No	468	.000	
		Yes	29	1.000	
10	Dyspnoea	No	386	.000	
		Yes	111	1.000	
11	Difficulty In Breathing	No	257	.000	
		Yes	240	1.000	
12	Runny Nose	No	53	.000	
		Yes	444	1.000	
13	Vomiting	No	409	.000	
		Yes	88	1.000	
14	Sputum	No	325	.000	
		Yes	172	1.000	
15	Cough	No	31	.000	
		Yes	466	1.000	
16	Headache	No	445	.000	
		Yes	52	1.000	
17	Cold	No	133	.000	
		Yes	364	1.000	

No	Variable		Frequency	Parameter coding	
18	Fever	No	101	.000	
		Yes	396	1.000	
19	Diarrhoea	.0	473	.000	
		1.0	24	1.000	
20	Fatigue	No	466	.000	
		Yes	31	1.000	
21	Allergic: pollen/dust	No	496	.000	
		Yes	1	1.000	
22	Did the patient travel recently	No	480	.000	
		Yes	17	1.000	
23	Possibility of inhaling toxic fumes	No	113	.000	
		Yes	384	1.000	
24	Anyone smoke at home	No	416	.000	
		Yes	81	1.000	
25	Family history of asthma/fumes	No	416	.000	
		Yes	81	1.000	
26	Severe Hydration	No	494	.000	
		Yes	3	1.000	
27	High dependency care	No	465	.000	
		Yes	32	1.000	
28	Colour of nasal secretion	Yes	470	.000	
		2.0	27	1.000	
29	Day Care/ Pre School	No	491	.000	
		Yes	6	1.000	
30	Body aches	No	467	.000	
		Yes	30	1.000	
31	Loose Stool	No	475	.000	
		Yes	22	1.000	
32	Wheezing	No	469	.000	
		Yes	28	1.000	
33	Nasal block	No	469	.000	
		Yes	28	1.000	
34	Nasal congestion	No	490	.000	
		Yes	7	1.000	
35	SOB	No	426	.000	
		Yes	71	1.000	
36	Gender	Male	302	.000	
		Female	195	1.000	

APPENDIX D: R Code for AUC-GA

```
library(pROC)
library(galgo)
library(rtkore)
library(Rcpp)
library(aod)
library(GA)
library(MASS)

reg.fitness = function(chr, parent,tr,te,res) {
  try=as.factor(parent$data$classes[tr])
  trd = data.frame(parent$data$data[tr,as.numeric(chr)])
  trm = nnet::nnet(try ~ ., data = cbind(trd,try=try),trace=F,size = 20)
  tey = as.factor(parent$data$classes[te])
  ted = data.frame(parent$data$data[te,as.numeric(chr)])
  pred=predict(trm,newdata = cbind(ted,tey=tey),type = "raw")
  if(res){
    roc(tey,pred,levels=levels(tey), direction = "<")$auc
  } else{
    predict(trm,newdata=cbind(ted,tey=tey),type="class")
  }
}


reg.bb = configBB.VarSel(data=t(subset(RSV_Data_full, select= -c(RSV))),
                         classes=RSV ,
                         classification.method="user",
                         classification.userFitnessFunc=reg.fitness,
                         chromosomeSize=20 ,niches=1, maxSolutions=100,
                         goalFitness = 0.9, saveVariable="reg.bb",
                         saveFrequency=50, saveFile="reg.bb.Rdata",
                         main="Logistic")


blast(reg.bb)

############################ Summary ######################
y=as.factor(RSV)
x1=`Fever days`
x2=as.factor(Conjuctivitis)
x3=as.factor(Cough)
x4=as.factor(Tachypnoea )
x5=as.factor(`nasal congestion`)
x6=as.factor(headache)
x7=as.factor(`Day Care/ Pre School`)
x8=as.factor(fatigue)
```

```
x9=as.factor(`colour of nasal secreation`)
x10=as.factor(`Allergic: pollen/dust`)
x11=as.factor(`Body aches`)
x12=as.factor(`Severe Hydration`)
x13=as.factor(`Sinus Congestion`)
x14=as.factor(`Loose Stool`)
x15=as.factor(`Req: Intensive care`)
x16=as.factor(Diarrhoea)
x17=as.factor(`Possibility of inhaling toxic fumes`)

model=(glm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x16+x1
7,family = binomial(link="logit")))
probRSV=predict(model,type=c("response"))
roccurve<- roc(y ~ probRSV)
auc(roccurve)

###################To Take plots#######################
Plot(blast(reg.bb))

predicted <- round(probRSV)
tab <- table(Predicted = predicted, Reference = RSV_Data_full$RSV)
plot(reg.bb, type="fitness")
plot(reg.bb,type="geneoverlap",cex=0.6)
plot(reg.bb, type="genenetwork")
plot(reg.bb, type="genecoverage",   coverage=c(0.5, 0.75,1))
plot(reg.bb, type="confusion")
rchr <- lapply(reg.bb$bestChromosomes[1:50], robustGeneBackwardElimination,
reg.bb, result="shortest")
barplot(table(unlist(lapply(rchr,length))), main="Length of Shortened
Chromosomes")
fsm <- forwardSelectionModels(reg.bb)
```

APPENDIX E: R Code for LL-GA

```
library(pROC)
library(galgo)
library(rtkore)
library(Rcpp)
library(aod)
library(GA)
library(MASS)

## Logistic model
model=glm(RSV~.,family = binomial(link="logit"),data=RSV_Data_full)

############################## GA starts here ##########################
x <- model.matrix(model)[, -1]
is.na(x)
fitness=function(string){
  inc=which(string==1)
  X=cbind(1,x[,inc])
  mod=glm.fit(X,RSV,family = binomial(link="logit"))
  class(mod)="glm"
  -(-2*logLik(mod))
}

GA <-
ga("binary",lower=c(0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,
0,0,0,0,0,0), pmutation = 0.1,pcrossover = 0.9,crossover = gareal_spCrossover,
selection = gabin_rwSelection,popSize=50, maxiter=100, upper
=c(1,66,8,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,3,1,2,0,1,1,11,1,1,1,1,1,1),
fitness = fitness, nBits = ncol(x),names = colnames(x))

plot(GA)
summary(GA)

## to take logistic regression model with reduced number of variables by GA
variables=data.frame(x[,GA@solution == 1])
modelGA <- glm(RSV~ ., data = data.frame(x[,GA@solution == 1]))
summary(modelGA)
```