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

LIST OF FIGURES

LIST OF TABLES

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