MODELLING THE RISK FOR TYPE 2 DIABETES USING LOGISTIC REGRESSION APPROACH

A. M. C. H. ATTANAYAKE

(138851 B)

Degree of Master of Science

Department of Mathematics

University of Moratuwa

Sri Lanka

May 2016

MODELLING THE RISK FOR TYPE 2 DIABETES USING LOGISTIC REGRESSION APPROACH

A. M. C. H. ATTANAYAKE

(138851 B)

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

Department of Mathematics

University of Moratuwa

Sri Lanka

May 2016

Declaration of the Candidate

"I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any University or other 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"

Signature:

…… ……

A.M.C.H. Attanayake Date

138851B

Declaration of the Supervisors

"I have supervised and accepted the thesis titled 'Modelling the Risk for Type 2 Diabetes using Logistic Regression Approach' for the submission of the degree."

Signature of the supervisors:

…… ……

Dr. (Mrs.) D.D.M. Jayasundara Date
Senior Lecturer,
Head of the Department,
Department of Statistics & Computer Science,
Faculty of Science,
University of Kelaniya.

…… ……

Prof. T.S.G. Peiris Date
Professor in Applied Statistics,
Head of the Department,
Department of Mathematics,
Faculty of Engineering,
University of Moratuwa.

ABSTRACT

Type 2 diabetes is one of the growing vitally fatal diseases all over the world. The knowledge of the significant risk factors for type 2 diabetes will be useful to keep the diabetes under control. This study has identified eight significant risk factors for type 2 diabetes in the data set of UCI machine learning repository by using point-biserial correlation. With the aim of developing an accurate predictive model to predict the presence of diabetes based on identified significant risk factors a binary logistic regression approach was applied. The performance of a predictive model is overestimated when simply determined on the sample of subjects that was used to construct the model. Therefore five-fold cross validation technique has applied in order to validate the predictive ability of the developed model. Results reveal that low value of optimism (0.0108) and high value of c-statistic (0.8512) in the fitted model indicating an acceptable discrimination power of type 2 diabetes. There is a significant influence by Glucose level, BMI and Pedigree for the diabetes on the classification of the patient as type 2 diabetes.

Key Words: Binary logistic regression, BMI, C-statistic, Five-fold cross validation, Glucose level, Optimism, Pedigree, Point-biserial correlation, Risk factors, Type 2 diabetes

ACKNOWLEDGEMENT

The work on this study would not have been possible without encouragement and support given by many people. First and foremost, I would like to express my deepest gratitude to my supervisors Dr. (Mrs.) D.D.M. Jayasundara, Senior Lecturer in the Department of Statistics & Computer Science, University of Kelaniya and Prof. T.S.G. Peiris, Department of Mathematics, University of Moratuwa for their guidance and providing useful insight towards making this report a success.

Additionally, I would like to thank Prof. Ewout W. Steyerberg and Dr. Daan Nieboer from the University Medical Center at Netherland for providing me useful suggestions on cross validation through emails.

Furthermore, my special gratitude goes to my family members and relatives for their encouragement and support given to complete the degree.

At last but not least, I would like to express my thanks to my friends and colleagues who were with me whenever I need a help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ADA	American Diabetic Association
ARIMA	Auto-Regressive Integrated Moving Average
AUC	Area Under the Curve
BMI	Body Mass Index
EPV	Events Per Variable
FP	False Positive
GRNN	General Regression Neural Network
HbA1c	Glycated Haemoglobin A1c
IR	Insulin Resistance
MLP	Multilayer Perceptron
MUAC	(Mid-)Upper Arm Circumference
OGTT	Oral Glucose Tolerance Test
OLS	Ordinary Least Squares
OR	Odds Ratio
RBF	Radial Basis Function
LM	Levenberg–Marquardt
ROC	Receiver Operating Characteristic
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
TP	True Positive
TSF	Triceps Skin Fold
VIF	Variance Inflation Factor
WHO	World Health Organization