

**MODELLING THE RISK FOR TYPE 2 DIABETES USING
LOGISTIC REGRESSION APPROACH**

A. M. C. H. ATTANAYAKE

(138851 B)



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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)



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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  University of Moratuwa, Sri Lanka.
138851B www.lib.mrt.ac.lk Electronic Theses & Dissertations
Date

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.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

.....

.....

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 (0.008) 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.  www.lib.mrt.ac.lk
University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations

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

Declaration of the Candidate	i
Declaration of the Supervisors	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
LIST OF TABLES	vii
LIST OF ABBREVIATIONS.....	ix
CHAPTER 1 INTRODUCTION	1
1.1 Introduction to Diabetes.....	1
1.2 Diabetes Epidemic in the World	1
1.3 Types of Diabetes.....	2
1.4 The Diagnosis of Type 2 Diabetes	3
1.4.1 The Risk Factors for Type 2 Diabetes	3
1.4.2 Complications of Type 2 Diabetes.....	4
1.4.3 Symptoms and Treatment for Type 2 Diabetes.....	5
1.5 Objectives of the Study.....	5
1.6 Significance of the Study.....	6
1.7 Outline of the Dissertation.....	6
CHAPTER 2 LITERATURE REVIEW	8
2.1 Studies on Prevalence of Type 2 Diabetes and Associated Risk Factors.....	8
2.2 Studies on Type 2 Diabetes through Model Building.....	10
2.3 Summary	13
CHAPTER 3 MATERIALS AND METHODS	15
3.1 Data Collection	15
3.2 Description of the Variables used for the Study	15
3.3 The Point-Biserial Correlation	17
3.4 Binary Logistic Regression	17
3.5 Model Diagnostics.....	19
3.5.1 Likelihood Ratio Test.....	19
3.5.2 Hosmer – Lemshow Goodness of Fit Test.....	20
3.5.3 Pseudo R ² for Logistic Regression	21
3.5.4 Classification Tables	22
3.5.5 Wald Test	22
3.5.6 ROC Curve	23
3.5.7 Odds Ratios	24

3.5.8 Variance Inflation Factor	25
3.6 Cross Validation.....	25
3.6.1 <i>k</i> -fold Cross Validation	27
3.6.2 Apparent Performance.....	27
3.6.3 Model Optimism	28
CHAPTER 4 RESULTS AND DISCUSSION	29
4.1 Data Cleaning	29
4.2 Descriptive Statistics	29
4.3 Identification of the Significant Risk Factors for the Type 2 Diabetes	30
4.3.1 The Association Between the Diabetes and the Plasma Glucose Concentration in an Oral Glucose Tolerance Test.....	30
4.3.2 The Association Between the Diabetes and the Number of Times Pregnant.....	30
4.3.3 The Association Between the Diabetes and the Diastolic Blood Pressure	31
4.3.4 The Association Between the Diabetes and the Triceps Skin Fold Thickness.....	31
4.3.5 The Association Between the Diabetes and the 2-Hour Serum Insulin	31
4.3.6 The Association Between the Diabetes and the Body Mass Index	32
4.3.7 The Association Between the Diabetes and the Diabetes Pedigree Function.....	32
4.3.8 The Association Between the Diabetes and the Age.....	33
4.4 Model Building through Binary Logistic Regression.....	33
4.4.1 Assumptions.....	33
4.4.2 The Binary Logistic Regression Model.....	34
4.4.3 Model Diagnostics.....	35
4.5 Application of 5- fold Cross Validation.....	37
4.5.1 The Five Binary Logistic Regression Models.....	37
4.5.2 Model Diagnostics of the Five Binary Logistic Regression Models	41
4.6 Model Performance through Cross Validation	42
4.7 Summary	42
CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS	44
5.1 Conclusions.....	44
5.2 Recommendations	45
5.3 Future Research	46
REFERENCES	47



LIST OF TABLES


Table No	Title	Page No
Table 4.1	Descriptive Statistics of the variables under study	29
Table 4.2	The Point-biserial correlation between the Plasma glucose concentration in an oral glucose tolerance test and the Diabetes	30
Table 4.3	The Point-biserial correlation between the Number of times pregnant and the Diabetes	30
Table 4.4	The Point – Biserial Correlation between the Diabetes and the Diastolic blood pressure	31
Table 4.5	The Point – Biserial Correlation between the Diabetes and the Triceps skin fold thickness	31
Table 4.6	The Point – Biserial Correlation between the Diabetes and the 2-Hour serum insulin	31
Table 4.7	 The Point – Biserial Correlation between the Diabetes and the Body mass index <i>University of Moratuwa, Sri Lanka. Electronic Theses & Dissertations www.lib.mrt.ac.lk</i>	32
Table 4.8	The Point – Biserial Correlation between the Diabetes and the Diabetes pedigree function	32
Table 4.9	The Point – Biserial Correlation between the Diabetes and the Age	33
Table 4.10	Assumption of no multicollinearity	33
Table 4.11	The coefficients of the full binary logistic regression model	34
Table 4.12	Hosmer and Lemeshow Test of the full binary logistic model	35
Table 4.13	Some diagnostic measures of the full binary logistic model	35
Table 4.14	Omnibus Tests of Model Coefficients of the full binary logistic model	35

Table 4.15	Classification Table of the full binary logistic model	35
Table 4.16	-2 log likelihood value of the null model	35
Table 4.17	The coefficients of the first binary logistic regression model	37
Table 4.18	The coefficients of the second binary logistic regression model	38
Table 4.19	The coefficients of the third binary logistic regression model	38
Table 4.20	The coefficients of the fourth binary logistic regression model	39
Table 4.21	The coefficients of the fifth binary logistic regression model	40
Table 4.22	Diagnostic measures of the five binary logistic regression models - Summary.	41



University of Moratuwa, Sri Lanka.
 Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk