

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 Electronic Theses & Dissertations
www.lib.mrt.ac.lk 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



CHAPTER 1

INTRODUCTION

1.1 Introduction to Diabetes

Diabetes mellitus is a group of metabolic diseases in which a person has high blood glucose either because insulin production is inadequate or because cells do not respond to the insulin that is produced or both. According to the current World Health Organization [WHO], classification of disorders of diabetes comprise four types; Type 1 Diabetes Mellitus [T1DM], Type 2 Diabetes Mellitus [T2DM], Gestational Diabetes and other Specific Types of Diabetes (World Health Organization 1999).

1.2 Diabetes Epidemic in the World

Diabetes mellitus is one of the growing vitally fatal diseases which is growing at an alarming rate. The T2DM and associated cardiovascular complications pose a major health-care burden worldwide and present a significant challenge to patients, health-care systems, and national economies (Ramachandran et al. 2010). According to an estimate of International Diabetes Federation (2006) comparative prevalence of Diabetes during 2007 is 8.0 % and likely to increase to 7.3% by 2025. WHO estimates that 347 million people world-wide have diabetes (World Health Organization 1999). Furthermore, they report that this number is likely to get more than double by 2030 without intervention. Almost 80% of diabetes deaths occur in low- and middle-income countries. WHO projects that diabetes will be the 7th leading cause of death in 2030. In the United States, nearly 16 million people have been diagnosed with diabetes, which means that about 1 of every 20 people have this disease. Recent data indicates that South Asia is one of the major sites of epidemic of T2DM with a projected 72% increase in

the number of subjects with T2DM in the next 20 years (Ramachandran et al. 2010). One of the first nation-wide studies conducted in Sri Lanka (Katulanda et al. 2008) reveals that one in five adults have either diabetes or pre-diabetes and one-third of those with diabetes are undiagnosed. Sri Lanka is seeing a rapid epidemiological transition where non-communicable diseases are becoming major causes of mortality and morbidity. A good assessment of the current status of diabetes in the country and associated risk factors is essential to meet this challenge. Facing the challenge of controlling the rapidly emerging non-communicable diseases, one of which is diabetes, need to be regarded as a prime need of the present day.

1.3 Types of Diabetes

The four types of Diabetes are Type 1 Diabetes, Type 2 Diabetes, Gestational Diabetes and other Specific Types of Diabetes.

A body with T1DM does not produce insulin and people usually develop T1DM at their teenage years, often before their 40th year. According to an estimate of International Diabetes Federation (2013) 490,100 children below the age of 15 years are living with type 1 diabetes. The T2DM is the most common form of diabetes and occurs due to inadequate body production of insulin for proper function, or due to insulin resistance, which means cells in the body not reacting to insulin. The World Health Organization (WHO) estimates that 90 percent of people around the world who suffer from diabetes suffer from type 2 diabetes. The gestational diabetes cause when pregnant women without a previous diagnosis of diabetes develop high blood glucose levels during pregnancy. According to an estimate of International Diabetes Federation (2013). 21 million cases of high blood glucose in pregnancy are estimated to contribute to the

global burden of diabetes. Other forms of diabetes mellitus include congenital diabetes, which is due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids, and several forms of monogenic diabetes. The majority of patients with T2DM initially have pre-diabetes. Their blood glucose levels are higher than normal, but not high enough to signal as diabetes at diagnosis. In that case, cells in the body are becoming resistant to insulin.

1.4 The Diagnosis of Type 2 Diabetes

There are several ways to diagnose diabetes. The HbA1c test measures the average blood glucose for the past 2 to 3 months. The fasting plasma glucose test checks fasting blood glucose levels. [Fasting means after not having anything to eat or drink [except water] for at least 8 hours before the test]. The oral glucose tolerance test [OGTT] is a two-hour test that checks blood glucose levels before and 2 hours after drinking a special sweet drink. It tells the doctor how the body processes glucose. American Diabetes Association criteria (American Diabetes Association 2014) defines blood sugar levels in diabetes and this criterion has been used for the diagnosis of diabetes in the majority of epidemiological studies of diabetes. In that the diabetes is defined as when fasting plasma glucose ≥ 126 mg/dL [≥ 7.0 mmol/l], oral glucose tolerance ≥ 200 mg/dL [≥ 11.1 mmol/l] and HbA1c $\geq 6.5\%$.

1.4.1 The Risk Factors for Type 2 Diabetes

Katulanda P. et al. (2006) have explained the gravity of the diabetes epidemic by figuring out diabetes prevalence of 14.2% among males and 13.5% among females in Sri Lanka. Furthermore, they have revealed that physical inactivity, raised body mass

index [BMI] and central obesity along with urban living as strongly associated with the increased risk of T2DM. Jamal Zafar et al. (2011) have established the existence of strong link between family history and the development of T2DM. Exercise is also reported as another important factor for preventing or delaying the onset of diabetes due to its impact on the body's use of glucose. Another possible cause that is not yet well established is the stress level of the person. Changes in dietary styles is another factor that needs to be focused. A comprehensive study of the effects of these factors in the Sri Lankan context is deemed to be important.

The chances of developing T2DM is known to increase with age (Jamal Zafar et al. 2011). However, recent studies show that there is a considerable percentage who face inflated blood sugar levels at young age (Diabetes.co.uk 2014). Without further investigation people use sample mean as a measure of this average age in the population. However, this measure may not work well in all situations, especially when the distribution of age at diagnosis is skewed.



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

1.4.2 Complications of Type 2 Diabetes

Type 2 diabetes leads to serious and life-threatening complications. Over time, diabetes can damage the heart, blood vessels, eyes, kidneys, nerves etc. Diabetes increases the risk of heart disease and stroke. In a multinational study, 50% of people with diabetes die of cardiovascular disease (primarily heart disease and stroke) (Morrish et al. 2001). Combined with reduced blood flow, neuropathy (nerve damage) in the feet increases the chance of foot ulcers, infection and eventual need for limb amputation. Diabetic retinopathy is an important cause of blindness, and occurs as a result of long-term accumulated damage to the small blood vessels in the retina. One percent of global

blindness can be attributed to diabetes and also diabetes is among the leading causes of kidney failure (World Health Organization 2012). The overall risk of dying among people with diabetes is at least double the risk of their peers without diabetes (Roglic et al. 2005).

1.4.3 Symptoms and Treatment for Type 2 Diabetes

Type 2 diabetes symptoms develop slowly in a body. Some of the symptoms include increased thirst and frequent urination, increased hunger, weight loss, fatigue, blurred vision, slow-healing sores and areas of darkened skin etc. (Mayo Foundation for Medical Education and Research 2016).

Although there is no exact cure for T2DM, with proper care and attention it can be kept under control. Making a proper diet, monitoring blood glucose, and routine exercises, and making other life-style commitments can help keep diabetes under control (National Diabetes Education Program 2014).

1.5 Objectives of the Study

On view of the above description the objectives of this study are to:

- Identify the significant risk factors for the type 2 diabetes.
- Develop a prediction model using binary logistic regression model to identify a patient having T2DM or not.
- Validate the model.

1.6 Significance of the Study

Type 2 diabetes is one of the fatal diseases which is growing at an alarming rate. Further it leads to serious and life-threatening complications in a body. Therefore it is important to know and control the risk factors for type 2 diabetes. In this study it reveals the significant risk factors for type 2 diabetes which will help to keep diabetes under control. In addition, the knowledge of the relationship of the risk factors with the presence of type 2 diabetes is useful in predicting the occurrence of the disease. But the predictive ability of the developed models are problematic due to unavailability of validation techniques. In the study relationship of the risk factors with the diabetes was modeled through binary logistic regression approach and predictive ability was validated using the cross validation techniques in order to produce an accurate predictive model.



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

1.7 Outline of the Dissertation

The organization of the dissertation is as follows:

Chapter 2 provides literature review related to the research topic. The previous researches discussed on two sub topics; **Studies on prevalence of Type 2 Diabetes and associated risk factors** and **Studies on type 2 diabetes through model building**. At the end of the chapter 2 a summary is provided.

The third chapter of the dissertation enclose with materials and methods used in the study. The data source with the description of the variables and theories behind the study given in detail within the chapter.

The Chapter 4 comprises of the results of the study with interpretations. The results of statistical analyses are presented, in order to determine reliability, validity, and the statistical significance of the findings.

The fifth and the final chapter of the dissertation is Conclusions and Recommendations. This chapter sums up the entire findings with respect to objectives. Further, it provides the recommendations on future research.



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

CHAPTER 2

LITERATURE REVIEW

2.1 Studies on Prevalence of Type 2 Diabetes and Associated Risk Factors

Mahen et al. (2011) have evaluated in Sri Lanka, intensive (3-monthly) lifestyle modification advice to a less-intensive (12 monthly; control group) lifestyle modification advice and found that intensive program was superior than less-intensive program; They supposed to introduce this low cost program as a prevention tool in Sri Lanka. Further they evaluated the prevalence of diabetes, pre-diabetes and cardio-metabolic risk factors among young urban Sri Lankans.

Katulanda et al. (2006) have explained the gravity of the diabetes epidemic by figuring out diabetes prevalence of 14.2% among males and 13.5% among females in Sri Lanka. Further they believed that rapid changes in lifestyle are important in underlying causes for type 2 diabetes. Katulanda et al. (2006) have encouraged to study the problem of type 2 diabetes among young adults.

Katulanda et al. (2008) have determined the prevalence of diabetes mellitus and pre-diabetes [impaired fasting glucose and impaired glucose tolerance] in adults in Sri Lanka. Projections for the year 2030 and factors associated with diabetes and pre-diabetes were also presented. Katulanda et al. (2011) have conducted a cross sectional study to assess knowledge, awareness and practices relating to management of Diabetes Mellitus among Sri Lankan general practitioners. On average, the knowledge on the management of type 2 diabetes in pregnancy was very poor. The concept of strict glycaemia control in preference to symptom control was appreciated only by 68%. Appropriate use of HbA1c and urine micro albumin was known by 15.2% and 39.2% respectively.

The Framingham offspring study in the U.S. collected information on 5124 subjects revealed that age, physical activity, alcohol consumption, and cigarettes smoked were important predictors of the risk factors for type 2 diabetes. Moreover, weight and height were found to differentially affect the probabilities of having type 2 diabetes (Alok 2003). Zahid & Muhammad (2006) have found the effect of different risk factors on diabetes in a cross-sectional hospital based study in Lahore, Pakistan. In the overall analysis the risk factors obesity, exercise and hypertension were significant. The obesity and hypertension are positively associated with Diabetes Mellitus whereas; exercise is negatively associated with Diabetes Mellitus. In the male Person's analysis, family history of diabetes is only the main significant risk factor. In the females; obesity, exercise and hypertension are the main significant risk factors.

Syed (2007) reported that in each year 7 million people develop Diabetes and the most dramatic increases in type 2 Diabetes have occurred in populations where there have been rapid and major changes in lifestyle, demonstrating the important role played by lifestyle factors and the potential for reversing the global epidemic. According to this article a person with type 2 diabetes is 2 – 4 times more likely to get cardiovascular disease, and 80% of people with Diabetes will die from it. Premature mortality caused by diabetes results in an estimated 12 to 14 years of life lost. A person with Diabetes incurs medical costs that are two to five times higher than those of a person without diabetes. Diabetes is the fourth leading cause of death in most developed countries. Complications from Diabetes, such as coronary artery and peripheral vascular disease, stroke, diabetic neuropathy, amputations, renal failure and blindness are resulting in increasing disability, reduced life expectancy and enormous health costs. Finally this article conclude that the diabetes as the one of the most challenging health problems in the 21st century.

According to the American Diabetes Association (2015) in 2012, 29.1 million Americans, or 9.3% of the population, had diabetes. Diabetes remains the 7th leading cause of death in the United States in 2010, with 69,071 death certificates listing it as the underlying cause of death, and a total of 234,051 death certificates listing diabetes as an underlying or contributing cause of death. In 2010, after adjusting for population age differences, hospitalization rates for heart attack were 1.8 times higher among adults aged 20 years or older with diagnosed diabetes than among adults without diagnosed diabetes. In 2005–2008, of adults with diabetes aged 40 years or older, 4.2 million (28.5%) people had diabetic retinopathy, damage to the small blood vessels in the retina that may result in loss of vision.

2.2 Studies on Type 2 Diabetes through Model Building



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

Madhavi et al. (2012) have designed a classifier for the detection of Diabetes. They proposed a method for designing a classifier using a neural network implementation of the fuzzy k-nearest neighbor algorithm. They have used neural networks due to their dynamic nature of learning and future application of knowledge. Further, Fuzzy logic allows partial membership and rule base that allows direct mapping between human thinking and machine results. Finally they have shown the results of the proposed system are expected to perform better than those in the current literature survey in detecting diabetes.

Murali et al. (1999) have discussed about the classification problems which is often encountered in medical diagnosis. They presents an introduction to the classification theory and shows how artificial neural networks can be used for classification. They

also map out a bootstrapped procedure for interval estimation of posterior probabilities. The entire procedure is illustrated using the diabetes mellitus data in Pima Indians.

The performance of recently developed neural network structure, General Regression Neural Network (GRNN), is examined by Kamer & Tulay (2003). Pima Indian Diabetes data set is chosen to study and had been examined by more complex neural network structures in the past. The results of early studies and of the GRNN structure is compared. Close classification accuracy to the reference work using ARTMAP-IC structured model, which is the best result obtained since now, is achieved by using GRNN, which has a simpler structure. The performance of the standard Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) feed forward neural networks are also examined for the comparison as they are the most general and commonly used neural network structures. The performance of the MLP was tested for different types of back-propagation training algorithms.



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

A multilayer neural network structure which was trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used to classify diabetes by Hasan et al. (2009). The results of the study were compared with the results of the previous studies on diabetes classification. Data mining techniques has being applied to find useful patterns in medical diagnosis and treatment. Rajesh & Sangeetha (2012) have applied several data mining techniques to classify the diabetes. Further they have tested the performances of those techniques and proved that the C4.5 algorithm is performed better than the other algorithms by giving 95% classification rate for diabetes.

Karegowda et al. (2012) present a hybrid model for classifying Pima Indian diabetic database. The model consists of two stages. In the first stage, the K-means clustering is

used to identify and eliminate incorrectly classified instances. The continuous data is converted to categorical form by approximate width of the desired intervals, based on the opinion of medical expert. In the second stage a fine tuned classification is done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify the cascaded K-means clustering and Decision tree C4.5 has enhanced classification accuracy of C4.5. Further rules generated using cascaded C4.5 tree with categorical data are less in numbers and easy to interpret compared to rules generated with C4.5 alone with continuous data. The proposed cascaded model with categorical data obtained the classification accuracy of 93.33 %.

Chun-Liang et al. (2007) have developed a logistic regression model and neural networks in order to identify and validate predictive factors for Glycemic control. For the cross validity purpose, 512 middle-aged patients, enrolled in Diabetes Healthcare Quality Improvement Program, were divided into training data and holdout data in a teaching hospital in Taiwan. The findings revealed that neural networks is more accurate than logistic regression. The important factors influence glycemic control are Years of diabetes onset, Education status, Body mass index, Months of enrolled in Diabetes Healthcare Quality Improvement Program, and Patient-Physician relationship.

Sadowski (2010) in his thesis, time series models are developed to explore the correlates of blood glucose fluctuation of diabetic patients. In particular, it is investigated whether certain human activities and lifestyle events (e.g. food and medication consumption, physical activity, travel and social interaction) influence blood glucose, and if so, how. A unique dataset is utilized consisting of 40 diabetic patients who participated in a 3-day study involving continuous monitoring of blood glucose at five minute intervals, combined with measures for sugar; carbohydrate; calorie and insulin intake; physical activity; distance from home; time spent traveling via public transit and private

automobile; and time spent with other people, dining and shopping. Using a dynamic regression model fitted with Auto-Regressive Integrated Moving Average (ARIMA) components, the influence of independent predictive variables on blood glucose levels is quantified, while at the same time the impact of unknown factors is defined by an error term. Models were developed for individuals with overall findings demonstrating the potential for continuous monitoring of diabetic patients who are trying to control their blood glucose. Model results produced significant blood glucose predicting variables that include food consumption, exogenous insulin administration and physical activity.

Zahid & Muhammad (2006) have found the effect of different risk factors on diabetes in a cross-sectional hospital based study in Lahore, Pakistan. In the analysis they have fitted a binary logistic regression model in order to predict the presence of diabetes. From the total of 580 persons, 398 (68.6%) are correctly predicted by the model. The value of model chi-square is 75.190 (P-value = 0.000) with d.f = 5. This is highly significant therefore they were 95% confident that the fitted model is appropriate.

2.3 Summary

The criteria used to diagnose diabetes mellitus were different between studies. Fasting blood glucose was used in some studies, and the oral glucose tolerance test in others. The earlier studies have used the WHO criteria (World Health Organization 1999), and the newer studies seem to be using the American Diabetic Association [ADA] criteria (American Diabetes Association 2014) for classification of diabetes.

Many researchers have estimated the prevalence and risk factors for type 2 diabetes mellitus in Sri Lanka and other countries over the last years (Zahid & Muhammad 2006,

Mahen et al. 2011, Katulanda et al. 2006, Katulanda et al. 2008, Katulanda et al. 2011, Alok 2003). They have identified different risk factors for type 2 diabetes. A comprehensive study of the effects of many risk factors is deemed to be important.

Considerable number of studies on type 2 diabetes through model building can be found in the last decade. Further it will be the one of the most popular area in the 21st century. The application of Data Mining techniques on diabetes and comparison of these applications with neural networks, time series models and multivariate techniques can be found in the literature survey. These studies will not be sufficient to control the world wide epidemic of type 2 diabetes.



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

CHAPTER 3

MATERIALS AND METHODS

3.1 Data Collection

Data have collected from the UCI machine learning repository which consists of 768 observations of female patients with and without diabetes and their records on several risk factors; Number of times pregnant, Plasma glucose concentration after 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (μ U/ml), Body mass index, Diabetes pedigree function and Age (years).

The data set is available at <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

3.2 Description of the Variables used for the Study

The following variables were used in the study:

Number of Times Pregnant: This variable denotes the pregnant frequency of each of the patient.

Plasma Glucose Concentration after 2 hours in an Oral Glucose Tolerance Test:

The glucose tolerance test is a medical test in which glucose is given and blood samples taken afterward to determine how quickly it is cleared from the blood. The test is usually used to test for diabetes, insulin resistance, impaired beta cell function and sometimes reactive hypoglycemia and acromegaly, or rarer disorders of carbohydrate metabolism. In the most commonly performed version of the test, an oral glucose tolerance test (OGTT), a standard dose of glucose is ingested by mouth and blood levels are checked two hours later.

Diastolic Blood Pressure: This number indicates the pressure in the arteries when the heart rests between beats. A normal diastolic blood pressure is 80 mmHg. A diastolic blood pressure between 80 mmHg and 89 mmHg indicates prehypertension. A diastolic blood pressure number of 90 mmHg or higher is considered to be hypertension or high blood pressure.

Triceps Skin Fold Thickness: The anthropometry of the upper arm is a set of measurements of the shape of the upper arms. The principal anthropometry measures are the upper arm length, the triceps skin fold (TSF), and the (mid-) upper arm circumference (MUAC). Although they are not directly convertible into measures of overall body fat, weight and density, researchers have been used these measures as rough indicators of body fat.

2-Hour Serum Insulin: The 2-hour insulin level is an effective indicator of IR (Insulin resistance) and can aid in diagnosing IR.



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

Body Mass Index: The Body Mass Index (BMI) value is calculated as follows:
[Body mass index = (weight in kg) / (height in meters)²]

Diabetes Pedigree Function: It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gave an idea of the hereditary risk one might have with the onset of diabetes mellitus (Jason 2014).

Age: This provides the age of the each of the patient.

Diabetes: This is a class variable indicates whether the patient has diabetes or not.

3.3 The Point-Biserial Correlation

The Point-Biserial Correlation has applied in the context in order to find the associations between the type 2 diabetes and each of the remaining variables.

The following five requirements should be met in order to apply the Point-Biserial correlation (Gregory & Dale 2009).

Requirement 1: One of the two variables should be measured on a continuous scale.

Requirement 2: The other variable should be dichotomous.

Requirement 3: There should be no outliers for the continuous variable for each category of the dichotomous variable.

Requirement 4: The continuous variable should be approximately normally distributed for

each category of the dichotomous variable.

Requirement 5: The continuous variable should have equal variances for each category of the dichotomous variable.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.mo.mt.ac.lk

The hypotheses test in the Point-biserial correlation as follows:

H₀: No correlation between the two variables

H₁: There is a correlation between the two variables

3.4 Binary Logistic Regression

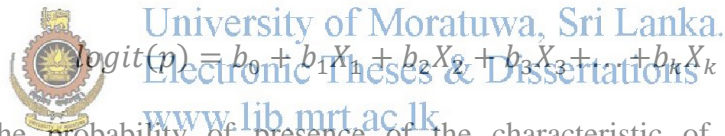
In order to model the relationship between the type 2 diabetes and the remaining variables the binary logistic regression technique has applied.

The binary logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In binary logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (True, success, etc.) or 0 (False, failure, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable, response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

Logistic regression equation can be written as follows:



$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the log odds as:

$$\text{odds} = \frac{P}{1 - P} = \frac{\text{probability of presence of the characteristic}}{\text{probability of absence of the characteristic}}$$

and

$$\text{Logit}(p) = \ln\left(\frac{p}{1 - p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

Binary logistic regressions, by design, overcome many of the restrictive assumptions of linear regressions. For example, linearity, normality and equal variances are not assumed, nor is it assumed that the error variance is normally distributed.

The major assumptions are:

- Dependent variable should be measured on a dichotomous scale.
- One or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable).
- Independence of observations and the dependent variable should have mutually exclusive and exhaustive categories.
- No (or little) multicollinearity among independent variables.
- There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.



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

3.5 Model Diagnostics

3.5.1 Likelihood Ratio Test

The likelihood ratio test is performed by estimating two models and comparing the fit of one model to the fit of the other. Removing predictor variables from a model will almost always make the model fit less well (i.e., a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant. The likelihood ratio test does this by comparing the log likelihoods of the two models, if this difference is statistically significant, then the less restrictive model (the one with more variables) is said to fit the data significantly better than the more restrictive model. If one

has the log likelihoods from the models, the likelihood ratio(lr) test is fairly easy to calculate.

The formula for the likelihood ratio test statistic is:

$$lr = -2 \ln \left(\frac{L(m1)}{L(m2)} \right) = 2[l1(m2) - l1(m1)]$$

Where $L(m^*)$ denotes the likelihood of the respective model (either model 1 or model 2), and $l1(m^*)$ the natural log of the model's final likelihood (i.e., the log likelihood), $m1$ is the more restrictive model, and $m2$ is the less restrictive model.

The resulting test statistic is distributed chi-squared, with degrees of freedom equal to the number of parameters that are constrained.

3.5.2 Hosmer – Lemeshow Goodness of Fit Test

The Hosmer-Lemeshow test is a statistical test for goodness of fit for the logistic regression model. The data are divided into approximately ten groups defined by increasing order of estimated risk. The observed and expected number of cases in each group is calculated and a Chi-square statistic is calculated as follows:



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

Hosmer & Lemeshow equation:

$$\chi_{HL}^2 = \sum_{g=1}^n \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)}$$

where O_g , E_g and n_g are the observed events, expected events and number of observations for the g^{th} risk decile group, and n is the number of groups. The test statistic follows a Chi-squared distribution with $n-2$ degrees of freedom.

A large value of Chi-squared (with small p -value < 0.05) indicates poor fit and small Chi-squared values (with larger p -value closer to 1) indicate a good logistic regression model fit.

The Contingency Table for Hosmer and Lemeshow Test table shows the details of the test with observed and expected number of cases in each group.

3.5.3 Pseudo R² for Logistic Regression

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, therefore the OLS (Ordinary Least squares) approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squared values have been developed. These are "pseudo" R-squared values because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squared values never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squared values can arrive at very different values.

Commonly Encountered Pseudo R-Squared Values are:



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations

www.lib.mrt.ac.lk

- McFadden's pseudo R²

$$R^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$

Where M_{Full} = Model with predictors

M_{Intercept} = Model without predictors

\hat{L} = Estimated likelihood

- Cox & Snell pseudo R²

$$R^2 = 1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}$$

- Nagelkerke / Cragg & Uhler's pseudo R²

$$R^2 = \frac{1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$

3.5.4 Classification Tables

The classification table is another method to evaluate the predictive accuracy of the logistic regression model. In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off value, for example $p = 0.50$) are cross-classified.

3.5.5 Wald Test

Wald χ^2 statistics are used to test the significance of individual coefficients in the model and are calculated as follows:

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Where, $\hat{\beta}_i$ is the estimate of the coefficient of the independent variable x_i and $SE(\hat{\beta}_i)$ is the standard error of $\hat{\beta}_i$. The squared value of the Wald statistics as indicated below is chi-square distributed with one degree of freedom.

The Wald Statistics tests the following hypotheses:

$$H_0 : \beta_i = 0, \text{ for } i = 1, 2, \dots, p \text{ and,}$$

$$H_1 : \beta_i \neq 0, \text{ for } i = 1, 2, \dots, p$$

Each Wald statistic is compared with a χ^2 distribution with 1 degree of freedom. Wald statistics are easy to calculate but their reliability is questionable, particularly for small samples. For data that produce large estimates of the coefficient, the standard error is often inflated, resulting in a lower Wald statistic, and therefore the explanatory variable may be incorrectly assumed to be unimportant in the model. Likelihood ratio tests are generally considered to be superior.

3.5.6 ROC Curve

A Receiver Operating Characteristic Curve (ROC) is a standard technique for summarizing classifier performance over a range of trade-offs between true positive (TP) and false positive (FP) error rates. ROC curve is a plot of sensitivity (the ability of the model to predict an event correctly) versus 1-specificity for the possible cut-off classification probability values.

A model with high discrimination ability will have high sensitivity and specificity simultaneously, leading to an ROC curve which goes close to the top left corner of the plot.

A model with no discrimination ability will have an ROC curve which is the 45 degree diagonal line.

For logistic regression it can create a 2×2 classification table of predicted values from your model for your response if $\hat{y} = 0$ or 1 versus the true value of $y = 0$ or 1 . The prediction if $\hat{y} = 1$ depends on some cut-off probability π_0 . For example, $\hat{y} = 1$ if $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ if $\hat{\pi}_i \leq \pi_0$. The most common value for $\pi_0 = 0.5$. Then sensitivity = $P(\hat{y}=1/y=1)$ and specificity = $P(\hat{y}=0/y=0)$.

The ROC curve is more informative than the classification table since it summarizes the predictive power for all possible π_0 .

The position of the ROC on the graph reflects the accuracy of the diagnostic test. It covers all possible thresholds (cut-off points). The ROC of random guessing lies on the diagonal line. The ROC of a perfect diagnostic technique is a point at the upper left corner of the graph, where the TP proportion is 1.0 and the FP proportion is 0.

The Area Under the Curve (AUC), also referred to as index of accuracy, or concordance index, c is an accepted traditional performance metric for a ROC curve. The higher the area under the curve the better prediction power the model has. The $c = 0.8$ can be interpreted to

mean that a randomly selected individual from the positive group has a test value larger than that for a randomly chosen individual from the negative group 80 percent of the time.

3.5.7 Odds Ratios

An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Odds ratios are most commonly used in case-control studies, however they can also be used in cross-sectional and cohort study designs as well (with some modifications and/or assumptions).

When a logistic regression is calculated, the regression coefficient (b_1) is the estimated increase in the log odds of the outcome per unit increase in the value of the exposure. In other words, the exponential function of the regression coefficient (e^{b_1}) is the odds ratio associated with a one-unit increase in the exposure.

Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

OR = 1 Exposure does not affect odds of outcome

OR > 1 Exposure associated with higher odds of outcome

OR < 1 Exposure associated with lower odds of outcome

$$OR = \frac{a/c}{b/d}$$

Where,

a = Number of exposed cases

b = Number of exposed non-cases

c = Number of unexposed cases

d = Number of unexposed non-cases

3.5.8 Variance Inflation Factor

As the name suggests, a variance inflation factor (VIF) quantifies how much the variance is inflated. The variances of the estimated coefficients are inflated when multicollinearity exists. Therefore, the variance inflation factor for the estimated coefficient b_k denoted VIF_k is just a factor by which the variance is inflated.

Consider a model with correlated predictors:



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

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i$$

If some of the predictors are correlated with the predictor x_k , then the variance of b_k is inflated. It can be shown that the variance of b_k is:

$$Var(b_k) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}$$

The reciprocal of the variance inflation factor is called as tolerance and it can be used as an indication for multicollinearity.

3.6 Cross Validation

Cross validation (rotation estimation) is one of the commonly used internal validation methods. It is used for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants

to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like over fitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc. (Cross validation (Statistics) 2016)

One round of cross validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds (Cross validation (Statistics) 2016).

Furthermore, one of the main reasons for using cross validation instead of using the conventional validation (e.g. partitioning the data set into two sets of 70% for training and 30% for test) is that the error (e.g. Root Mean Square Error) on the training set in the conventional validation is not a useful estimator of model performance and thus the error on the test data set does not properly represent the assessment of model performance. This may be because there is not enough data available or there is not a good distribution and spread of data to partition it into separate training and test sets in the conventional validation method (Cross validation (Statistics) 2016).

In summary, cross validation combines (averages) measures of fit (prediction error) to correct for the optimistic nature of training error and derive a more accurate estimate of model prediction performance (Cross validation (Statistics) 2016).

The cross validation has applied in this study in order to validate the fitted binary logistic regression model.

3.6.1 *k*-fold Cross Validation

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k* – 1 subsamples are used as training data. The cross-validation process is then repeated *k* times (the *folds*), with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general *k* remains an unfixed parameter (Cross validation (Statistics) 2016).



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

The 5-fold cross validation technique has applied in the study in order to validate the fitted model.

3.6.2 Apparent Performance

Austin and Steyerberg (2014) reported apparent performance as follows: A simple approach is to assess model performance directly in the sample in which it was developed. Using the fitted regression model, the predicted probability of the outcome is determined for each subject in the analytic sample. A summary measure of model performance, such as the c-statistic, is then reported. A limitation of this approach is that the model is optimized for performance in the sample in which it was developed. Subsequent predictions in subjects who were not used in model development are likely to have poorer accuracy than that which was reported in the analytic sample. Apparent performance estimates are hence optimistic.

The magnitude of optimism is expected to decrease as the effective sample size of the model derivation sample increases.

3.6.3 Model Optimism

The optimism is defined as the difference between the bootstrap performance (apparent performance of the bootstrap model) and its out-of-sample performance. This bootstrap process is then repeated. The Model optimism is the averaged value across all the bootstrap iterations. As a final step, subtracts this optimism estimate from the apparent performance to obtain the optimism-corrected performance estimate.



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

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data Cleaning

Data cleaning is a mandatory process that should be considered before going to apply any data analysis technique. The total number of cases available in the data set is 768. Some patients have glucose value as 0 which is impossible in the nature. With the same reasoning body mass index values of 0, diastolic blood pressure values of 0, skin fold thickness readings of 0 and serum insulin levels of 0 were deleted. After deleting the abnormal cases 392 cases were available for the analysis.

4.2 Descriptive Statistics

Table 4.1: Descriptive Statistics of the variables under study

	No of times Pregnant	Glucose Tolerance	Diastolic Blood Pressure	Skin Fold Thickness	2-hour Insulin	Body Mass Index	Diabetes Pedigree Function	Age	Diabetes
Positive Count									130(33.2%)
Negative Count									262(66.8%)
Mean	3.30	122.63	70.66	29.15	156.06	33.086	.52305	30.86	
Std. Error of Mean	.162	1.559	.631	.531	6.002	.3550	.017450	.515	
Median	2.00	119.00	70.00	29.00	125.50	33.200	.44950	27.00	
Mode	1	100	70	32	105	32.0	.692	22	
Std. Deviation	3.211	30.861	12.496	10.516	118.842	7.0277	.345488	10.201	
Skewness	1.336	.518	-.088	.209	2.165	.663	1.959	1.404	
Kurtosis	1.486	-.483	.795	-.458	6.357	1.557	6.367	1.738	
Range	17	142	86	56	832	48.9	2.335	60	
Minimum	0	56	24	7	14	18.2	.085	21	
Maximum	17	198	110	63	846	67.1	2.420	81	

The useful summary statistics of the variables are shown in the Table 4.1. The table indicates the measures of dispersion, central tendency and distribution for the data values. 130 (33.2%) of the cases positive for the diabetes. The highest right-skewed variable is 2-hour

insulin. Further the table indicates that how each of the variable varies between its minimum and maximum values with mean and standard error of the mean.

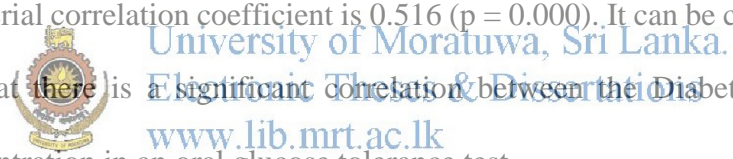
4.3 Identification of the Significant Risk Factors for the Type 2 Diabetes

4.3.1 The Association Between the Diabetes and the Plasma Glucose Concentration in an Oral Glucose Tolerance Test

Table 4.2: The Point-biserial correlation between the Plasma glucose concentration in an oral glucose tolerance test and the Diabetes

		Diabetes
	Pearson Correlation	.516
Glucose Tolerance	Sig. (2-tailed)	.000

The point-biserial correlation coefficient is 0.516 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Plasma glucose concentration in an oral glucose tolerance test.



4.3.2 The Association Between the Diabetes and the Number of Times Pregnant

Table 4.3: The Point-biserial correlation between the Number of times pregnant and the Diabetes

		Diabetes
No of times Pregnant	Pearson Correlation	.257
	Sig. (2-tailed)	.000

The point-biserial correlation coefficient is 0.257 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Number of times pregnant.

4.3.3 The Association Between the Diabetes and the Diastolic Blood Pressure

Table 4.4: The Point – Biserial Correlation between the Diabetes and the Diastolic blood pressure

		Diastolic Blood Pressure
Diabetes	Pearson Correlation	.193
	Sig. (2-tailed)	.000

According to the Table 4.4 the point-biserial correlation coefficient is 0.193 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Diastolic blood pressure.

4.3.4 The Association Between the Diabetes and the Triceps Skin Fold Thickness

Table 4.5: The Point – Biserial Correlation between the Diabetes and the Triceps skin fold thickness

		Skin Fold Thickness
Diabetes	Pearson Correlation	.256
	Sig. (2-tailed)	.000

According to the Table 4.5 the point-biserial correlation coefficient is 0.256 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Triceps skin fold thickness.

4.3.5 The Association Between the Diabetes and the 2-Hour Serum Insulin

Table 4.6: The Point – Biserial Correlation between the Diabetes and the 2-Hour serum insulin

		2-hour Insulin
Diabetes	Pearson Correlation	.301
	Sig. (2-tailed)	.000

According to the Table 4.6 the point-biserial correlation coefficient is 0.301($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the 2-Hour serum insulin.

4.3.6 The Association Between the Diabetes and the Body Mass Index

Table 4.7: The Point – Biserial Correlation between the Diabetes and the Body mass index

		Body Mass Index
Diabetes	Pearson Correlation	.270
	Sig. (2-tailed)	.000

According to the Table 4.7 the point-biserial correlation coefficient is 0.270 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Body mass index.

4.3.7 The Association Between the Diabetes and the Diabetes Pedigree Function

Table 4.8: The Point – Biserial Correlation between the Diabetes and the Diabetes pedigree function

		Diabetes Pedigree Function
Diabetes	Pearson Correlation	.209
	Sig. (2-tailed)	.000

According to the table 4.8 the point-biserial correlation coefficient is 0.209 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Diabetes pedigree function.

4.3.8 The Association Between the Diabetes and the Age

Table 4.9: The Point – Biserial Correlation between the Diabetes and the Age

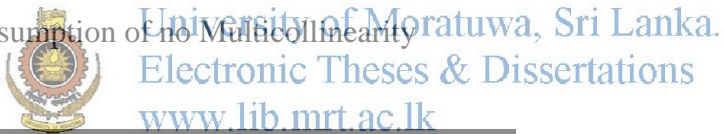
		Age
Diabetes	Pearson Correlation	.351
	Sig. (2-tailed)	.000

According to the Table 4.9 the point-biserial correlation coefficient is 0.351 ($p = 0.000$). It can be concluded with 95% confidence that there is a significant correlation between the Diabetes and the Age.

4.4 Model Building through Binary Logistic Regression

4.4.1 Assumptions

Table 4.10: Assumption of no Multicollinearity



Variable	Collinearity Statistics	
	Tolerance	VIF
No of times Pregnant	.526	1.901
Glucose Tolerance	.599	1.670
Diastolic Blood Pressure	.812	1.232
Skin Fold Thickness	.540	1.853
2-hour Insulin	.643	1.556
Body Mass Index	.505	1.980
Diabetes Pedigree Function	.944	1.059
Age	.470	2.129

According to Habshah et al. (2010) a tolerance of 0.1 or less as well as VIF of 2.5 or above is an indication of multicollinearity. Therefore the Table 4.10 suggests that that there is no multicollinearity among the explanatory variables.

4.4.2 The Binary Logistic Regression Model

Table 4.11: The coefficients of the full binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)
Preg	.082	.055	2.197	.138	1.086
Glucose	.038	.006	44.025	.000	1.039
BP	-.001	.012	.014	.904	.999
Thickness	.011	.017	.431	.511	1.011
Insulin	-.001	.001	.399	.528	.999
BMI	.071	.027	6.655	.010	1.073
Pedigree	1.141	.427	7.125	.008	3.130
Age	.034	.018	3.412	.065	1.035
Constant	-10.041	1.218	67.994	.000	.000

According to the Table 4.11 the binary logistic regression model can be written as:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = -10.041 + 0.038 \text{ GlucoseTest} + 0.071 \text{ BMI} + 1.141 \text{ Pedigree}$$

Where p is the probability of categorizing as a diabetes patient.

There is a significant contribution from the variables Glucose level, BMI and Pedigree for the type 2 diabetes at 5% significance level. Therefore, only these three variables were selected for the logistic regression equation.

For every one-unit increase in Glucose there is a 0.038 increase in the log-odds of Diabetes holding all other independent variables constant. For every one-unit increase in BMI there is a 0.071 increase in the log-odds of Diabetes holding all other independent variables constant. Similarly, for every one-unit increase in Pedigree there is a 1.141 increase in the log-odds of Diabetes holding all other independent variables constant.

4.4.3 Model Diagnostics

Table 4.12: Hosmer and Lemeshow Test of the full binary logistic model

Chi-square	df	Sig.
3.905	8	.866

Table 4.13: Some diagnostic measures of the full binary logistic model

	Cox & Snell R Square	Nagelkerke R Square
-2 Log likelihood		
344.021	.325	.452

Table 4.14: Omnibus Tests of Model Coefficients of the full binary logistic model

	Chi-square	df	Sig.
Step 1 Step	154.077	8	.000
Block	154.077	8	.000
Model	154.077	8	.000

Table 4.15: Classification Table of the full binary logistic model

Observed		Predicted		Percentage Correct
		Diabetes		
		negative for diabetes	positive for diabetes	
Diabetes	negative for diabetes	233	29	88.9
	positive for diabetes	56	74	56.9
Overall Percentage				78.3

The cut value is .500

Table 4.16: -2 log likelihood value of the null model

-2 Log likelihood
498.098

The $-2 \log$ likelihood value of the null model (only constant in the model) is 498.098 and it reduces to 344.021 in the full model. The difference between these two measures is the model chi-square value of the Omnibus test ($154.077 = 498.098 - 344.021$) and is tested for statistical significance. The corresponding p value of 0.000 reveals that the improvement in the model associated with the additional variables is statistically significant. The Cox and Snell R^2 and Nagelkerke R^2 values reveals that independent variables able to explain between 32.5% and 45.2% of the variance in type 2 diabetes. The Hosmer and Lemshow Test is not statistically significant ($p=0.866$) indicates predicted group memberships correspond closely to the actual group memberships which provides a good model fit. The overall classification rate of 78.3% indicates that the model fit the data at an acceptable level.

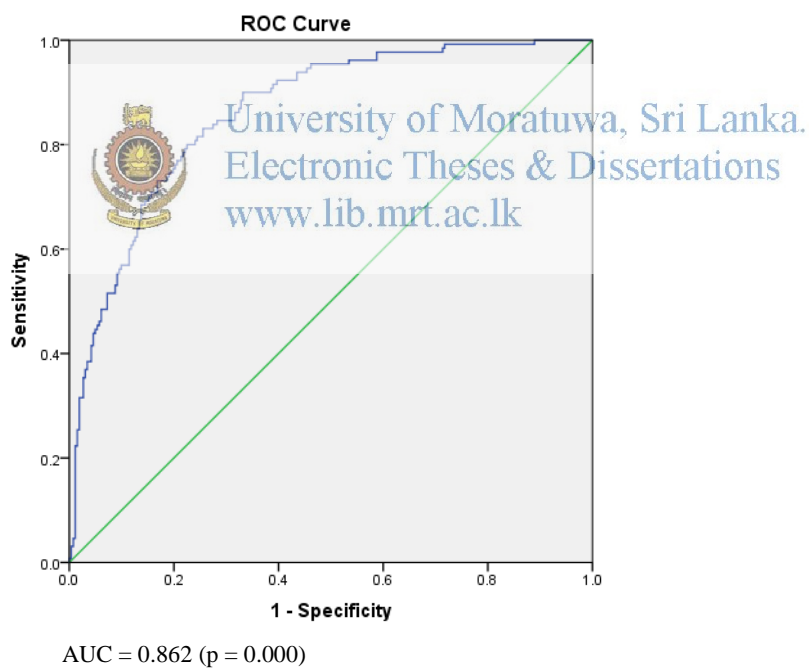


Figure 4.1: ROC and AUC of the full binary logistic regression model

The area under the ROC curve quantifies the overall ability of the model to discriminate between individuals with and without the diabetes. The higher the value of Area under the curve ($AUC = 0.862$) indicates good discrimination power. In Figure 4.1 the significance of the AUC reveals that the correct overall discrimination is not due to chance.

4.5 Application of 5- fold Cross Validation

The data were randomly divided into 5 parts with approximately equal sizes. Then a part is used for the validation of the binary logistic regression model and the remaining 4 parts used for the development of the binary logistic regression model (Training set). Then the process repeated 5 times which each of the 5 parts used exactly once as the validation set.

4.5.1 The Five Binary Logistic Regression Models

Table 4.17: The coefficients of the first binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)
Preg	.073	.060	1.472	.225	1.076
Glucose	.038	.006	36.638	.000	1.039
BP	-.005	.013	.128	.720	.995
Thickness	.011	.019	.351	.554	1.011
Insulin	.001	.001	.552	.457	1.999
BMI	.081	.030	7.207	.007	1.084
Pedigree	1.139	.464	6.042	.014	3.125
Age	.033	.019	2.918	.088	1.034
Constant	-10.078	1.315	58.748	.000	.000

According to the Table 4.17 the binary logistic regression model can be written as:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = -10.078 + 0.038 \text{ GlucoseTest} + 0.081 \text{ BMI} + 1.139 \text{ Pedigree}$$

Where p is the probability of categorizing as a diabetes patient.

There is a significant contribution from the variables Glucose level, BMI and Pedigree for the type 2 diabetes at 5% significance level. Therefore, only these three variables were selected for the logistic regression equation.

Table 4.18: The coefficients of the second binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
						Lower	Upper
Preg	.087	.062	1.976	.160	1.091	.966	1.231
Glucose	.039	.007	32.984	.000	1.040	1.026	1.054
BP	-.001	.013	.004	.951	.999	.973	1.026
Thickness	.020	.020	.958	.328	1.020	.980	1.062
Insulin	.000	.002	.046	.831	1.000	.997	1.003
BMI	.067	.031	4.677	.031	1.069	1.006	1.135
Pedigree	1.154	.509	5.143	.023	3.172	1.170	8.603
Age	.021	.020	1.028	.311	1.021	.981	1.062
Constant	-10.135	1.418	51.082	.000	.000		

According to the Table 4.18 the binary logistic regression model can be written as:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = -10.135 + 0.039 \text{ GlucoseTest} + 0.067 \text{ BMI} + 1.154 \text{ Pedigree}$$

Where p is the probability of categorizing as a diabetes patient.

There is a significant contribution from the variables Glucose level, BMI and Pedigree for the type 2 diabetes at 5% significance level. Therefore, only these three variables were selected for the logistic regression equation.

Table 4.19: The coefficients of the third binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
						Lower	Upper
Preg	.113	.067	2.843	.092	1.119	.982	1.276
Glucose	.039	.007	33.904	.000	1.039	1.026	1.053
BP	-.006	.014	.205	.651	.994	.968	1.021
Thickness	.007	.019	.135	.714	1.007	.970	1.046
Insulin	.000	.002	.007	.934	1.000	.997	1.003
BMI	.089	.032	7.878	.005	1.093	1.027	1.164
Pedigree	1.419	.516	7.566	.006	4.131	1.503	11.351
Age	.034	.023	2.317	.128	1.035	.990	1.082
Constant	-10.624	1.448	53.817	.000	.000		

According to the Table 4.19 the binary logistic regression model can be written as:

$$\begin{aligned} \text{Logit}(p) &= \log\left(\frac{p}{1-p}\right) \\ &= -10.624 + 0.039 \text{ GlucoseTest} + 0.089 \text{ BMI} + 1.419 \text{ Pedigree} \end{aligned}$$

Where p is the probability of categorizing as a diabetes patient.

There is a significant contribution from the variables Glucose level, BMI and Pedigree for the type 2 diabetes at 5% significance level. Therefore, only these three variables were selected for the logistic regression equation.

Table 4.20: The coefficients of the fourth binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
						Lower	Upper
Preg	.057	.061	.877	.349	1.058	.940	1.192
Glucose	.037	.006	32.229	.000	1.037	1.024	1.051
BP	.004	.014	.095	.757	1.004	.978	1.031
Thickness	.012	.020	.401	.526	1.012	.974	1.052
Insulin	-.002	.001	1.653	.199	.998	.996	1.001
BMI	.057	.031	3.498	.061	1.059	.997	1.125
Pedigree	.839	.472	3.167	.075	2.315	.918	5.833
Age	.057	.022	6.735	.009	1.059	1.014	1.106
Constant	-10.095	1.390	52.757	.000	.000		

According to the Table 4.20 the binary logistic regression model can be written as:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = -10.095 + 0.037 \text{ GlucoseTest} + 0.057 \text{ Age}$$

Where p is the probability of categorizing as a diabetes patient. There is a significant contribution from the variables Glucose level and Age for the type 2 diabetes at 5% significance level. Therefore, only these two variables were selected for the logistic regression equation.

Table 4.21: The coefficients of the fifth binary logistic regression model

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
						Lower	Upper
Preg	.074	.064	1.339	.247	1.076	.950	1.220
Glucose	.039	.006	37.822	.000	1.040	1.027	1.052
BP	-.001	.013	.011	.918	.999	.973	1.025
Thickness	.012	.019	.411	.521	1.012	.976	1.049
Insulin	-.001	.001	.388	.533	.999	.996	1.002
BMI	.056	.030	3.320	.068	1.057	.996	1.122
Pedigree	1.118	.443	6.379	.012	3.058	1.285	7.282
Age	.031	.020	2.466	.116	1.031	.992	1.072
Constant	-9.473	1.307	52.491	.000	.000		

According to the Table 4.21 the binary logistic regression model can be written as:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = -9.473 + 0.039 \text{Glucose} + 1.118 \text{Pedigree}$$

Where p is the probability of categorizing as a diabetes patient. There is a significant contribution from the variables Glucose level and Pedigree for the type 2 diabetes at 5% significance level. Therefore, only these two variables were selected for the logistic regression equation.

4.5.2 Model Diagnostics of the Five Binary Logistic Regression Models

Table 4.22: Diagnostic measures of the five binary logistic regression models - Summary

Model	Overall Classification %		-2 log likelihood	Cox and Snell R ²	Nagelkerke R ²	Chi-Square (Omnibus test)	Chi-square (Hosmer & Lemeshow test)	Significant variables	AUC of ROC		Optimism
	Training	Validation							Training	Validation	
1	79.1	78.8	297.107	0.317	0.443	Sig.	NSig.	BMI,Glucose ,pedigree	0.861(p=0.000)	0.854(p=0.000)	0.007
2	79.9	75.3	259.775	0.310	0.438	Sig.	Nsig.	BMI,Glucose ,pedigree	0.861(p=0.000)	0.861(p=0.000)	0.000
3	80.1	76.9	252.375	0.360	0.498	Sig.	NSig.	BMI,Glucose ,pedigree	0.875(p=0.000)	0.817(p=0.000)	0.058
4	79.4	81.3	266.818	0.329	0.452	Sig.	Nsig.	Glucose,Age	0.854(p=0.000)	0.865(p=0.000)	-0.011
5	79.0	81.3	292.090	0.323	0.447	Sig.	NSig.	Glucose,Pedigree	0.860(p=0.000)	0.862(p=0.000)	0.000
Average	79.5	78.72	273.633	0.3278	0.4556	Significant	Not Significant		0.8622	0.8518	0.0108

According to the internal validation, Cox and Snell R^2 and Nagelkerke R^2 values reveals that independent variables able to explain between 32.8% and 45.6% of the variance in type 2 diabetes. The Hosmer and Lemshow Test is not statistically significant indicates predicted group memberships correspond closely to the actual group memberships which provides a good model fit. The overall classification rate of training set is 79.5% which indicates that the model fit the data at an acceptable level. Further the overall average classification rate of the validation set is 78.7% implies that the stability of the model performances. The higher the value of AUC in both training and validation (0.862 and 0.852) indicates good discrimination power. The significances of the AUCs reveal that the correct overall discriminations are not due to chance. The model optimism value of 0.0108 indicates low over-fitting in predicting the presence of diabetes.



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

4.6 Model Performance through Cross Validation

The apparent c-statistic of 0.862 (Figure 4.1) indicates a reasonable discriminative ability in predicting the presence of diabetes. The optimism-corrected performance estimate is equal to the 0.8512(0.862 – 0.0108). The low value of optimism and high value of c-statistic (0.8512) indicate an acceptable discrimination power of the fitted model.

4.7 Summary

According to the analysis it can be concluded with 95% confidence that all the variables under study are significantly associated with the type 2 diabetes. The full binary logistic regression model was fitted using all the variables entered at once. The diagnostic measures were discussed. 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. The low value of optimism and high value of c-statistic indicate an acceptable discrimination power of the fitted model.



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

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The size of the original data set is equal to 768. Although the size of the data set was reduced to 392 after the data cleaning it is recommend to apply the cleaning process using the logical arguments in order to obtain an accurate results and findings.

The majority of the patients of the data set (66.8%) are positive for the type 2 diabetes. There is a need of a prediction model to classify the patient as diabetes or not based on the observed characteristics.

The results of the point-biserial correlation have identified the variables; Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-Hour serum insulin, Body mass index, Diabetes pedigree function and Age as the significant factors for the type 2 diabetes.

With the aim of modeling the relationship between the diabetes and the significant risk factors the binary logistic regression approach was applied. The fitted model concludes that there is a significant contribution from the variables Glucose, BMI and Pedigree for the Diabetes at 5% significance level. Further, the positive coefficients of the variables; Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Triceps skin fold thickness, Body mass index, Diabetes pedigree function and Age in the model equation shows the increase in the log-odds of Diabetes in unit increase of each of the variable. The -2 log likelihood value of the null model (only constant in the model) is 498.098 and it reduces to 344.021 in the full model. The

difference between these two measures is the model chi-square value of the Omnibus test ($154.077 = 498.098 - 344.021$) and is tested for statistical significance. The corresponding p value of 0.000 reveals that the improvement in the model associated with the additional variables is statistically significant. The Cox and Snell R^2 and Nagelkerke R^2 values reveals that independent variables able to explain between 32.5% and 45.2% of the variance in type 2 diabetes. The Hosmer and Lemshow Test is not statistically significant ($p=0.866$) indicates predicted group memberships correspond closely to the actual group memberships which provides a good model fit. The overall classification rate of 78.3% indicates that the model fit the data at an acceptable level.

The performance of a predictive model is overestimated when simply determined on the sample of subjects that was used to construct the model. Several internal validation methods are available that aim to provide a more accurate estimate of model performance in new subjects. In order to validate the performance of the fitted model the 5 –fold cross validation was applied. The easiest and fast method of internal validation of split-sample method is not tested in the analysis because of its low accuracy than cross validation. The results of the cross validations ended up with low model optimism value of 0.0108. Hence the optimism-corrected performance estimate is equal to the 0.8512. The low value of model optimism and high value of optimism-corrected performance estimate indicate an acceptable discrimination power of the fitted model.

5.2 Recommendations

All of the variables under study are significantly associated with type 2 diabetes. This reveals that it will be important to keep the controllable significant factors under control to reduce the risk for type 2 diabetes.

The fitted model can be used as a good prediction model in order to predict whether a patient has type 2 diabetes based on observed characteristics of the patient (Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2-Hour serum insulin, Body mass index, Diabetes pedigree function and Age).

The application of cross validation technique is a novel one and it is efficient than Jack-knife validation in the presence of large data sets.

5.3 Future Research

This study aims to develop an accurate predictive model in order to predict the presence of diabetes through cross validation. The more advanced internal validation techniques such as bootstrap methods can be applied to develop a prediction model with stable estimates of model performances. Additionally, the influence of the events per variable (EPV) can be investigated by changing the number of observations in each of the variable.

Further, the application of external validation techniques to confirm the generalizability of the findings are encouraged.



REFERENCES

Alok, B 2003, 'A longitudinal analysis of the risk factors for diabetes and coronary heart disease in the Framingham Offspring Study', *Population Health Metrics*, vol.1,no. 3, retrieved on 02 November 2014, <http://www.pophealthmetrics.com/content/1/1/3>.

American Diabetes Association 2014, *Diagnosing Diabetes and Learning About Pre-diabetes*, retrieved on 02 November 2014, <http://www.diabetes.org/diabetes-basics/diagnosis/>.

American Diabetes Association 2015, *Statistics about diabetes*, retrieved on 7 December 2015, <http://www.diabetes.org/diabetes-basics/statistics/>.

Austin, PC & Steyerberg, EW 2014, 'Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models', *Statistical Methods in Medical Research*, vol.0, no.0, pp. 1–13.

Chun-Liang, L., Chung-Liang, L., Show-Wei, C, Kwoting, F 2007, 'Identification and Validation of Predictive Factors for Glycemic Control: Neural Networks vs. Logistic Regression', International Conference on Computer Engineering and Applications, Australia, 17-19 January, pp. 300-305.

Cross-validation (statistics) 2016, retrieved on 05 January 2016, [https://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics)).

Diabetes.co.uk 2014, *Children and Diabetes*, retrieved on 05 November 2014, <http://www.diabetes.co.uk/children-and-diabetes.html>.

Gregory, WC & Dale, IF 2009, *Non-parametric statistics for non-statisticians*, 2nd edn, John Wiley & Sons, New Jersey.

Habshah, M, Sarkar, SK, Sohel, R 2010, 'Collinearity diagnostics of binary logistic regression model', *Journal of Interdisciplinary Mathematics*, vol.13,no.3,pp. 253-267.

Hasan Temurtasa, Nejat Yumusakb, Feyzullah Temurtasc 2009, 'A comparative study on diabetes disease diagnosis using neural networks', *Expert Systems with Applications*, vol.36, no. 4, pp.8610-8615.

International Diabetes Federation 2006, *IDF Diabetes Atlas*, 3rd edn, Brussels, Belgium.

International Diabetes Federation 2013, *IDF Diabetes Atlas*, 6th edn, Basel, Switzerland.

Jamal, Z, Fiaz, B, Nasim, A, Uzma, R, Rizwan, B, Saima, H, Ayesha, W, Fardah, Y, Madeesha, N, Umaima 2011, 'Prevalence and risk factors for diabetes mellitus in a selected urban population of a city in Punjab', *J Pak Med Assoc*, vol. 61, no. 1, pp.40-47.

Jason Brownlee 2014, *Case Study: Predicting the Onset of Diabetes within Five Years (part 1 of 3)*, retrieved on 28 December 2015, <http://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/>.



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

Karegowda, AG, Punya, V, Jayaram, MA, Manjunath, AS 2012,' Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5', *International Journal of Computer Applications*, vol.45, no.12, pp. 45-50.

Katulanda, P, Constantine, GR, Mahesh, JG, Sheriff, R, Seneviratne, RDA, Wijeratne, S, Wijesuriya, M, McCarthy, MI, Adler, AI, Matthews, DR 2008, 'Prevalence and projections of diabetes and pre-diabetes in adults in Sri Lanka--Sri Lanka Diabetes', *Diabetic Medicine*, vol. 25, pp.1062-1069.

Katulanda, P, Constantin, GR, Weerakkody, MI, Perera, YS, Jayawardena MG, Wijegoonawardena P, Matthews DR, Sheriff MH 2011, 'Can we bridge the gap? Knowledge and practices related to Diabetes Mellitus among general practitioners in a developing country: A cross sectional study', *Asia Pac Fam Med*, vol. 5, pp. 10-15.

Katulanda P, Sheriff MH, Matthews DR 2006, 'The diabetes epidemic in Sri Lanka - a growing problem', *Ceylon Med*, vol. 51, pp. 26-28.

Kayaer, K, Yıldırım, T 2003, 'Medical diagnosis on Pima Indian diabetes using general regression neural networks', *Proceedings of the international conference on artificial neural networks and neural information processing ICANN/ICONIP*, Turkey, pp. 181-184.

Madhavi P, Ketki K, Parag N, Ajinkya P, Eknath P 2012, 'Design of Classifier for Detection of Diabetes using Neural Network and Fuzzy k-Nearest Neighbor Algorithm', *International Journal Of Computational Engineering Research*, vol. 2, no. 5, pp. 1384-1387.

Mahen, W, Martin, G, Laksha, V, Giancarlo, V, Luigi, G, Janaka, K 2011, 'DIABRISK - SL Prevention of cardio-metabolic disease with life style modification in young urban Sri Lankan's - study protocol for a randomized controlled trial', *BioMed*, vol. 12, retrieved on 01 November 2014, <http://www.trialsjournal.com/content/12/1/209>.

Mayo Foundation for Medical Education and Research 2016, *Diseases and conditions - Type 2 diabetes*, retrieved on 25 December 2015, <http://www.mayoclinic.org/diseases-conditions/type-2-diabetes/basics/symptoms/con-20031902>.



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

Morrish NJ, Wang SL, Stevens LK, Fuller JH, Keen H. Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes. *Diabetologia*2001, 44 Suppl 2:S14–S21.

Murali Shanker, MYHu, Hung MS 1999, *Estimating Probabilities of Diabetes Mellitus Using Neural Networks*, retrieved on 27 December 2015, http://www.personal.kent.edu/~mshanker/personal/Zip_files/sar_2000.pdf.

National Diabetes Education Program 2014, *Manage Your Diabetes*, retrieved on 03 November 2014, <http://www.ndep.nih.gov/i-have-diabetes/ManageYourDiabetes.aspx>.


Rajesh, K, Sangeetha, V 2012, 'Application of Data Mining Methods and Techniques for Diabetes Diagnosis', *International Journal of Engineering and Innovative Technology*, vol.2, no.3, pp. 224-229.

Ramachandran, A, Ma, RCW, Snehalatha, C 2010, 'Diabetes in Asia', *Lancet*, vol.30, no. 375, pp. 408-18.

Roglic, G, Unwin, N, Bennett, PH, Mathers, C, Tuomilehto, J, Nag S et al. The burden of mortality attributable to diabetes: realistic estimates for the year 2000. *Diabetes Care*, 2005, 28(9):2130–2135.

Sadowski, EA 2010, 'A Time Series Analysis: Exploring the Link between Human Activity and Blood Glucose Fluctuation', MA thesis, Wilfrid Laurier University, Canada.

Syed AT 2007, 'Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century?', *International Journal of Health Sciences*, vol. 1, no. 2, pp.5-8

World Health Organization: Definition 1999, *Diagnosis and Classification of Diabetes Mellitus and its Complications. Report of a WHO Consultation. Part 1: Diagnosis and Classification of Diabetes Mellitus*, Geneva, retrieved on 01 November 2014, <http://www.who.int>.
 University of Moratuwa, Sri Lanka
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

World Health Organization 2012. *Global data on visual impairments 2010*, Geneva, retrieved on 16 March 2015, [http://www.who.int/blindness/GLOBALDATAFINALforweb .pdf](http://www.who.int/blindness/GLOBALDATAFINALforweb.pdf).

Zahid A, Muhammad KP 2006, 'Risk Factors and Diabetes Mellitus (Statistical Study of Adults in Lahore, Pakistan)', *Journal of Statistics*, vol. 13, no.1, pp. 1684 – 8403.