**Identification of the Associated Risk Factors of Diabetes in Sylhet Region.**



Advance Data Analysis Lab

07.07.2021

Submitted to

**Dr. Mohammed Taj Uddin**

Professor, Department of Statistics

Shahjalal University of Science and Technology, Sylhet-3114

Submitted By

**Arunava Chandan Roy**

Registration No: 2016134016

Department of Statistics

Shahjalal University of Science and Technology, Sylhet-3114

Contents

[ACKNOWLEDGEMENT 3](#_heading=h.30j0zll)

[ABSTRACT 3](#_heading=h.1fob9te)

[INTRODUCTION 4](#_heading=h.3znysh7)

[1.1 Background of the Study and Statement of the Problem 4](#_heading=h.2et92p0)

[1.2 Objective of the Study 5](#_heading=h.3dy6vkm)

[LITERATURE REVIEW 6](#_heading=h.44sinio)

[METHODOLOGY 7](#_heading=h.1t3h5sf)

[3.1 Data Description and the Variables 7](#_heading=h.4d34og8)

[3.2 Data Source 8](#_heading=h.2s8eyo1)

[3.3 DESCRIPTIVE STATISTICS 8](#_heading=h.17dp8vu)

[3.4 Logistic Regression Model for shop preference. 9](#_heading=h.3rdcrjn)

[Results and Discussion 10](#_heading=h.26in1rg)

[4.1 Descriptive Statistics 10](#_heading=h.lnxbz9)

[4.4 Logistic Regression Model 18](#_heading=h.2jxsxqh)

[CONCLUSION 22](#_heading=h.35nkun2)

[REFERENCES 22](#_heading=h.1ksv4uv)

# ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest sense of faith and gratitude to Almighty Allah, the merciful to all and all good things that happened by his wish, for giving me the strength, endurance and ability to carry out my research work . It is a great opportunity for me to convey my deepest indebtedness and sincere gratefulness to my reverend teacher and supervisor Dr. Mohammed Taj Uddin, Professor, Department of Statistics, Shahjalal University of Science & Technology, Sylhet-3114, Bangladesh for his constant encouragement, valuable suggestion and excellent personal guidance for carrying out my thesis. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly and gave me the opportunity to grow in this field of research.

# ABSTRACT

Diabetes is one of the fastest growing chronic life threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. The early diagnosis of diabetes is only possible by proper assessment of both common and less common sign symptoms, which could be found in different phases from disease initiation up to diagnosis. Data mining classification techniques have been well accepted by researchers for risk prediction model of the disease. To predict the likelihood of having diabetes requires a dataset, which contains the data of newly diabetic or would be diabetic patient.

In this study, secondary data has been used. I have used a dataset of 520 instances, which has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. Binary logistic regression model has been performed to identify the associated risk factors.

Age, gender, genital thrush, polyuria, polydipsia, itching, irritability, and partial paresis have been identified as the main risk factors for receiving a diabetes diagnosis.

# Keywords: Diabetes, Risk Factors, Logistic Regression.

# INTRODUCTION

## 1.1 Background of the Study and Statement of the Problem

Diabetes Mellitus, a chronic metabolic disorder, is one of the fastest growing health crises of this era regardless of geographic, racial, or ethnic context. Commonly, we know about two types of diabetes called type 1 and type-2 diabetes. Type-1 diabetes occurs when the immune system mistakenly attacks the pancreatic beta cells and very little insulin is released to the body or sometimes even no insulin is released to the body. On the other hand, type 2 diabetes occurs when our body doesn't produce proper insulin or the body becomes insulin resistant. Some researchers divided diabetes into Type 1, Type 2, and gestational diabetes [1]. Gestational diabetes is a type of diabetes which occurs only in pregnancy due to hormonal changes. The common symptoms of diabetes are polyuria, polydipsia, polyphagia, sudden weight loss (usually Type 1), weakness, obesity (usually Type 2), delayed healing, visual blurring, itching, irritability, genital thrush, partial paresis, muscle stiffness, alopecia, etc. [1, 2]. This could be a clear evidence that, according to WHO, the number of the diabetic patient had been sharply increased from 108 million in 1980 to 422 million in 2014 (3). The most alarming fact is that more than 80% of diabetic people were from low- and middle-income countries in 2013 and the prevalence is surging up in these countries. Recently, Diabetes Australia has published that, Diabetes even may exist up to 7 years before clinical diagnosis [4], which was even up to 12 years previously noted by Harris et al. [5]. Within this time frame, people may gradually suffer from fatal complications like heart attacks, strokes, eye damage resulting in blindness, foot ulcer, amputation of the affected limb, kidney damage, and other forms of multi-organ damage (5). Most of the cases, these complications would be easily controlled or even prevented in some cases with early detection and treatment initiation that could possibly save around 1415 AUD [4). The degree of diabetic com- plication is more when the period between onset of disease and treatment initiation is longer (5). According to Diabetes Australia, failure in early detection of TYPE 2 could cost the Australian healthcare system more than 700 million dollars each year (4). In 2017, the total expenditure of diagnosed diabetes in the United States alone was 327 billion USD (2). In [6], in the year 2011, China had experienced 90 million (9% of the population), India had 61.3 million (8% of the population) and Bangladesh had 8.4 million (10% of the population). Comparing to developed countries like Australia and The USA, low and middle-income countries cannot afford the burden of managing such a costly disease like diabetes, the prevalence of which is increasing at an alarming rate. Therefore, early diagnosis and initiation of appropriate.

## 1.2 Objective of the Study

Explore the prevalence of being diagnostic with diabetes in Sylhet region as well as identifying the significant risk factors.

Chapter II

# METHODOLOGY

## 2.1 Data Description and the Variables

Secondary data has been used for this study. Each observation have 17 variables containing age, gender, diabetes etc.

This data is source from UCI machine learning repository

<https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>.

## 2.2 Data Source

A research concerned with a careful study of a problem in order to discover new facts or information from it and reaching in some decisions. The quality of a research depends on the findings based on quality of data and methodology that are implemented. If data are not reliable, the corresponding findings will be unacceptable. So it is of supreme importance to use reliable data in order to perform a worthy study and to make it creditable. An authentic research, therefore arrives at proper decisions with the evidence of reliable data. This chapter is constructed to indicate the brief description of the study area, the source of data.

## 3.3 DESCRIPTIVE STATISTICS

**Frequency Table:**

Frequency distribution is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

**Cross Tabulation:**

In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They provide a basic picture of the interrelation between two variables and can help find interactions between them.  The term contingency table was first used by [Karl Pearson](https://en.wikipedia.org/wiki/Karl_Pearson) in "On the Theory of Contingency and Its Relation to Association and Normal Correlation", part of the [Drapers' Company](https://en.wikipedia.org/wiki/Drapers%27_Company) Research Memoirs Biometric Series I published in 1904.The degree of association between the two variables can be assessed by a number of coefficients.

## 3.4 Logistic Regression Model for shop preference.

Regression analysis is a statistical technique used to measure the relationships between variables for the purpose of predicting future values. When the outcome variable is continuous, we usually use simple linear regression. But in many cases the outcome variable is categorical in nature. To deal with categorical outcome variables, logistic regression was introduced as an alternative to simple linear regression. Logistic regression was developed in late 1960s and early 1970s and became popular among researches in various fields.

Logistic regression can be of different types, such as binomial (binary), multinomial, or ordinal depending on the nature of outcome variable. When the outcome variable can have only two categories, then binomial or binary logistic regression is used, if the outcome variable have more than two categories which are not ordered then multinomial logistic regression is used and if the outcome variable is ordered then ordinal logistic regression is used.

The mathematical concept of logistic regression is to express the relationship between outcome variable and predictor variables (independent variables) in terms of logit: the natural logarithm of odds. Let’s consider as simple case where Y is a dichotomous outcome variable categorized as “1” and “0” and X is a continuous predictor variable. Now if we draw a scatter plot we will have two parallel lines corresponding to each outcome variable category. The relationship does not follow a linear trend and hence not possible to describe through a simple linear regression. Logistic regression facilitates this situation by logit transformation on the outcome variable Y. The simplest form of logistic regression model can be written as:

Logit(Y) = = β0 + β1X … … … … … (1)

Here π is the probability of occurring the outcome Y and π/(1−π) is the odds of success; the ratio of the probability of occurring the outcome Y and the probability of not occurring the outcome Y. β0 and β1 are called intercept and slope (regression coefficient) respectively. By taking antilog on both sides of equation (1) we can estimate the probability of the occurrence of outcome Y for a given value of predictor X:

π = P(Y|X=x) = … … … … (2)

The predictor variable X can be either continuous or categorical. We can extend the logistic model for more than one predictor as well,

Logit(Y) = = β0 + β1X1 + … + βp Xp … … … … … (3)

Equation (3) is the general form of logistic regression model for p number of predictors. Regression parameter βs can be estimated by either maximum likelihood (ML) method or weighted least square method. The value of regression coefficients β1...βp indicate the relationship between X’s and logit of Y. Coefficient value bigger than 0 indicates an increase in logit of Y with an increase in X and coefficient value smaller than 0 indicates a decrease in logit of Y with an increase in X. When the coefficient value is 0, it indicates there is no linear relationship among logit of Y and predictors X (Fig. 2). For the ease of interpretation we usually report the odds ratio along with regression coefficient. Odds ratio can be calculated by the following formula,

OR =

Statistical significance of the regression coefficient can generally be tested using Wald’s test and overall model significance can be tested by likelihood ratio test or pseudo R2 test. The null hypothesis underlying the overall model states that all β’s equal zero. A rejection of this null hypothesis implies that at least one βdoes not equal zero in the population, which means that the logistic regression equation predicts the probability of the outcome better than the mean of the dependent variable Y. The interpretation of results is rendered using the odds ratio for both categorical and continuous predictors.

## 3.5SOFTWARE AND TECHNICAL SUPPORT.

To find an efficient result of any study it is important to use a proper software and complete the analysis with a good technical support. In this study, the entire analysis is done in personal computer. The following tools were used for the total analysis:

## 3.5.1 SPSS:

A well-known statistical package named SPSS (Statistical packages for Social Science) software, Windows version 20 was used for data processing, creating descriptive and frequency tables & for Cross-tabulation. This software is built around the SPSS programming language. Menus and dialog boxes are useful because they give the visual reminders of the options with each step of analysis. SPSS can handle a large amount of data and can perform a wide range of analysis.

## 3.5.2 MICROSOFT OFFICE 2016:

Report writing and different types of high quality charts were produced using Microsoft office 2016.

CHAPTER IV

# Results

## 4.1 Descriptive Statistics

**Frequency Tables:**

| Gender | | Frequency | Percent |
| --- | --- | --- | --- |
|  | Female | 192 | 36.9 |
| Male | 328 | 63.1 |
| Total | 520 | 100.0 |

63.1% data has been collected from male respondents.

| **Diabetes** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
|  | Negative | 200 | 38.5 | 38.5 | 38.5 |
| Positive | 320 | 61.5 | 61.5 | 100.0 |
| Total | 520 | 100.0 | 100.0 |  |

In this data 61.5% respondents have diabetes.

| **Partial paresis** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
|  | No | 296 | 56.9 | 56.9 | 56.9 |
| Yes | 224 | 43.1 | 43.1 | 100.0 |
| Total | 520 | 100.0 | 100.0 |  |

In this data 43.1% respondents have been suffering from partial paresis.

**Cross Tabulation:**

| **Gender \* Diabetes Crosstabulation** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | | Diabetes | | Total |
| Negative | Positive |
| Gender | Female | Count | 19 | 173 | 192 |
| % within Gender | 9.9% | 90.1% | 100.0% |
| % within Diabetes | 9.5% | 54.1% | 36.9% |
| % of Total | 3.7% | 33.3% | 36.9% |
| Residual | -54.8 | 54.8 |  |
| Male | Count | 181 | 147 | 328 |
| % within Gender | 55.2% | 44.8% | 100.0% |
| % within Diabetes | 90.5% | 45.9% | 63.1% |
| % of Total | 34.8% | 28.3% | 63.1% |
| Residual | 54.8 | -54.8 |  |
| Total | | Count | 200 | 320 | 520 |
| % within Gender | 38.5% | 61.5% | 100.0% |
| % within Diabetes | 100.0% | 100.0% | 100.0% |
| % of Total | 38.5% | 61.5% | 100.0% |

90.1% of the females are suffered from diabetes and only 44.8% of the males are suffered from diabetes. so it is clear that females are more suffered from diabetes than males.

| **Polydipsia \* Diabetes Crosstabulation** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | | Diabetes | | Total |
| Negative | Positive |
| Polydipsia | No | Count | 192 | 95 | 287 |
| % within Polydipsia | 66.9% | 33.1% | 100.0% |
| % within Diabetes | 96.0% | 29.7% | 55.2% |
| % of Total | 36.9% | 18.3% | 55.2% |
| Residual | 81.6 | -81.6 |  |
| Yes | Count | 8 | 225 | 233 |
| % within Polydipsia | 3.4% | 96.6% | 100.0% |
| % within Diabetes | 4.0% | 70.3% | 44.8% |
| % of Total | 1.5% | 43.3% | 44.8% |
| Residual | -81.6 | 81.6 |  |
| Total | | Count | 200 | 320 | 520 |
| % within Polydipsia | 38.5% | 61.5% | 100.0% |
| % within Diabetes | 100.0% | 100.0% | 100.0% |
| % of Total | 38.5% | 61.5% | 100.0% |

96.6% patients who has diabetes are also suffering from polydipsia.

| **Polyphagia \* Diabetes Crosstabulation** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | | Diabetes | | Total |
| Negative | Positive |
| Polyphagia | No | Count | 152 | 131 | 283 |
| % within Polyphagia | 53.7% | 46.3% | 100.0% |
| % within Diabetes | 76.0% | 40.9% | 54.4% |
| % of Total | 29.2% | 25.2% | 54.4% |
| Residual | 43.2 | -43.2 |  |
| Yes | Count | 48 | 189 | 237 |
| % within Polyphagia | 20.3% | 79.7% | 100.0% |
| % within Diabetes | 24.0% | 59.1% | 45.6% |
| % of Total | 9.2% | 36.3% | 45.6% |
| Residual | -43.2 | 43.2 |  |
| Total | | Count | 200 | 320 | 520 |
| % within Polyphagia | 38.5% | 61.5% | 100.0% |
| % within Diabetes | 100.0% | 100.0% | 100.0% |
| % of Total | 38.5% | 61.5% | 100.0% |

79.7% patients who has diabetes are also suffering from polyphagia.

**Logistic Regression Model Output:**

| **Variables in the Equation** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I .for EXP(B) | |
| Lower | Upper |
| Step 1a | Age | -.051 | .025 | 4.070 | 1 | .044 | .950 | .904 | .999 |
| Gender(1) | 4.351 | .598 | 52.910 | 1 | .000 | 77.570 | 24.017 | 250.538 |
| Polyuria(1) | -4.440 | .705 | 39.625 | 1 | .000 | .012 | .003 | .047 |
| Polydipsia(1) | -5.070 | .829 | 37.420 | 1 | .000 | .006 | .001 | .032 |
| sudden weight loss(1) | -.190 | .548 | .121 | 1 | .728 | .827 | .283 | 2.418 |
| weakness(1) | -.817 | .537 | 2.317 | 1 | .128 | .442 | .154 | 1.265 |
| Polyphagia(1) | -1.194 | .534 | 5.007 | 1 | .025 | .303 | .107 | .862 |
| Genitalthrush(1) | -1.864 | .553 | 11.345 | 1 | .001 | .155 | .052 | .459 |
| visual blurring(1) | -.916 | .651 | 1.978 | 1 | .160 | .400 | .112 | 1.434 |
| Itching(1) | 2.803 | .673 | 17.362 | 1 | .000 | 16.493 | 4.413 | 61.643 |
| Irritability(1) | -2.341 | .591 | 15.712 | 1 | .000 | .096 | .030 | .306 |
| delayed healing(1) | .392 | .550 | .507 | 1 | .476 | 1.479 | .503 | 4.348 |
| partial paresis(1) | -1.159 | .525 | 4.880 | 1 | .027 | .314 | .112 | .877 |
| muscle stiffness(1) | .729 | .580 | 1.578 | 1 | .209 | 2.073 | .665 | 6.462 |
| Alopecia(1) | -.150 | .620 | .059 | 1 | .808 | .860 | .255 | 2.901 |
| Obesity(1) | .289 | .544 | .282 | 1 | .595 | 1.335 | .459 | 3.880 |
| Constant | 12.324 | 2.322 | 28.168 | 1 | .000 | 225056.967 |  |  |
| a. Variable(s) entered on step 1: Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscles tiff ness, Alopecia, Obesity. | | | | | | | | | |

CHAPTER V

# CONCLUSION

It has been found that age, gender, polyuria, polydipsia, genital thrush, itching, irritability and partial paresis are the significant risk factors associated with being diagnostic with diabetes.

CHAPTER VI

# REFERENCES

1. Angus O. UNEGBU1 and David IREFIN2 (1Accounting Faculty, American University of Nigeria, Nigeria. 2Department of Economics, University of Maiduguri, Borno State, Nigeria.); Impact of VAT on economic development of emerging nations; Journal of Economics and International Finance (Vol. 3(8), pp. 492-503, August 2011).
2. Rob Pike, Mark Lewis, Daniel Turner; Impact of VAT reduction on the consumer price indices; Economic & Labor Market Review(August 2009, Volume 3, Issue 8, pp 17–21)
3. Scott Menard; Applied Logistic Regression Analysis, Volume 106; Volume 2002
4. [Chao-Ying Joanne Peng](https://www.tandfonline.com/author/Peng%2C+Chao-Ying+Joanne),[KukLida Lee](https://www.tandfonline.com/author/Lee%2C+Kuk+Lida) &[Gary M. Ingersoll](https://www.tandfonline.com/author/Ingersoll%2C+Gary+M);An Introduction to Logistic Regression Analysis and Reporting;[The Journal of Educational Research](https://www.tandfonline.com/toc/vjer20/current); Volume 96, (2002 - [Issue 1](https://www.tandfonline.com/toc/vjer20/96/1))
5. Edwards, A. L. (1985). *A series of books in psychology. Multiple regression and the analysis of variance and covariance.*New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co.