DEPARTMENT OF MATHEMATICS AND STATISTICS

Project Report

Paper : MTH416A – Regression Analysis

Semester II

Study of Presence of Diabetes Based on Possible Risk Factors with the Help of Multiple Logistic Regression

Submitted by -

Srijani Adhikary :	**211397**
Pallavi Chakravarty :	**211441**
Sreeja Deb Ray :	**211466**
Sovon Gayen :	**211465**
Anis Pakrashi :	**211264**

-under the supervision of

Prof. Dr. Sharmishtha Mitra

ABSTRACT

Diabetes is a prevalent health issue among millions of people around the world, more so in the recent years. It's always good to be aware of the possible risk factors and detecting the disease at the earliest is crucial for a healthy living. In this project, we look into the survey data of 2015 collected by CDC which consists of a response variable, *Diabetes_binary* having two classes: 0 for no diabetes and 1 for prediabetes or diabetes; and 21 feature variables, some of which are high BP, high cholesterol, BMI and age. There are 70692 observations in all. The objective of this project is to fit a multiple logistic regression model using 70% of the data and test its performance on the rest 30% of the data using performance diagnostics from confusion matrix, like percentage of misclassification, specificity rate and sensitivity rate. In the process of building the model, we shall perform initial exploratory data analysis, choose significant subset of regressors using stepwise selection, check model adequacy measures on the 70% data. After fitting the model, we interpret the importance of each regressor towards predicting the nature of diabetes, using the remaining 30% of data. Finally, as an alternative, we try to predict the presence of diabetes using other ML techniques, and we compare our results with that of the fitted multiple logistic regression model.

Keywords: Diabetes, Multiple Logistic Regression, Deviance Statistic, Random Forest classifier

ACKNOWLEDGEMENT

We present our project report on “presence of diabetes based on analysis of its possible risk factors.” However, as mere students of statistics, any achievement, in terms of a substantial project, would have been an uphill task, considering just our own efforts. It is apparent that there has always been a constant encouragement of noble minds for our efforts to bring this project to a successful completion. The satisfaction that accompanies the effectiveness of any task would be incomplete without mentioning those who made it possible for us to complete our desired project.

First and foremost, we want to express our sincere gratitude to our instructor Dr. Sharmishtha Mitra for her constant help, support and advice throughout the project preparation, which acted as a catalyst for effective completion of our work. Without her valuable guidance and motivation, it would have been nearly impossible to work as a team and imbibe the practical aspects of the course “MTH416A: Regression Analysis”.

Besides, we are thankful to all faculty members of our department and our seniors, because without their support at various stages, this project would not have materialized.

We are also thankful to our friends for their critical appreciation and their questions and counter-questions solved a huge spectrum of doubts related to our work.

Last but not the least, we are grateful to our parents for their constant motivation on the way of materializing the project.

Srijani Adhikary
Pallavi Chakravarty
Sreeja Deb Ray
Sovon Gayen
Anis Pakrashi

CONTENTS

1. INTRODUCTION

1.1 Context

Diabetes mellitus, commonly called diabetes, is a serious chronic metabolic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood. The human pancreas produces the hormone insulin, which moves sugar from the blood into our cells to be stored or used for energy. With the onset of the disorder, the human body either doesn't make enough insulin or can't effectively use the insulin it generates. Untreated high blood glucose levels due to diabetes can damage our nerves, eyes, kidneys and other parts of the body.

Diabetes is among the most prevalent chronic diseases around the globe, wreaking havoc on the health of numerous people each year, leading to reduced quality of life and life expectancy. It also exerts a significant burden on the finance and economy. Diabetic patients are susceptible to complications like heart disease, vision loss, lower-limb amputation, and kidney disease, due to chronically high levels of glucose remaining in the bloodstream. What makes diabetes a serious concern is that there is no cure for it. However, medical practitioners prescribe strategies like losing weight, eating balanced diet, involving in activity, besides providing optimum medical treatments, with the aim of mitigating the evils of this disease in many patients.

Early diagnosis can lead to making predictive models for the risk of diabetes. Once we are well versed with the important causes of diabetes, we may develop ways to minimize risks of occurrence of the disease.

In the context of the United States, the Centers for Disease Control and Prevention (CDC) has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk. Diabetes may appear in a variety of forms; nevertheless, it is the type II diabetes which is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health.

Various modelling techniques have been developed to address the lack of standardized processes that incorporate the perspectives of all healthcare investors. Such models can back the decision-making process aimed at achieving specific clinical outcomes, and at the same time, guide the allocation of healthcare resources and reduce costs. Prognostic and predictive models are predominantly

relevant to healthcare, mainly in the clinical sector, with implications for payers, patients, and providers. Significant improvisation in the clinical decision-making process leads to a boom in the usage of such models to achieve better outcomes, while reducing overall healthcare costs. Some models are aimed at predicting a clinical outcome, whereas others focus on identifying patients who may be at risk for the development of a particular condition.

Logistic Regression is gaining importance everyday due to its extensive use in real life problems and its ability to explain the complexities of statistical decision-making in simple terms to a commoner. Its use in predictive analysis, mainly in the medical and financial sectors, is a gemstone in the current context where predictions have gained significant importance, to save time and money and also to boost accuracy.

1.2 Objectives

In our project we try to consider a large spectrum of possible risk factors leading to the disorder and determine the most effective ones in inducing diabetes in people. Using a multiple logistic regression setup, we try to estimate the importance of various factors in the model used to predict the presence or absence of diabetes. We also check for model adequacy and judge how well our model fits into the prediction purposes. Later on, we employ Machine Learning techniques to compare the accuracy of the previously obtained model with the one obtained using ML.

2. DATASET DESCRIPTION

This is a dataset of Diabetes health indicators. This is a secondary and survey data. This feature variables are collected either by directly asked questions to individuals or calculated variables based on individual participant responses. This survey was done by CDC's BRFSS (The behavioural risk factor surveillance system). For this project we are using the dataset available on Kaggle. This is a clean dataset containing 70692 survey responses with an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable diabetes has two classes, 0 is for no diabetes and 1 is for prediabetes or

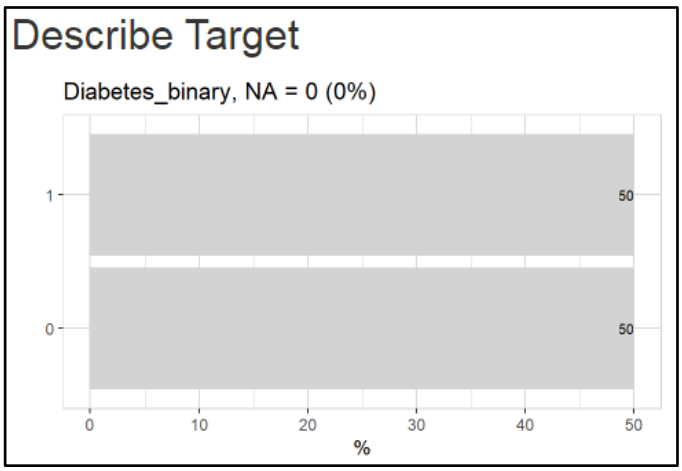
diabetes. Dataset has 21 feature variables and is balanced. The variables are described as questionnaire, they are listed as follows:

- **Diabetes_binary:** 0 for no diabetes and 1 for prediabetes or diabetes.
- **HighBP:** 1 if an individual has high blood pressure and 0 for no high blood pressure.
- **HighChol:** 1 if an individual has high cholesterol, 0 for no high cholesterol,
- **Cholcheck:** 0 if an individual has not done cholesterol check in last 5 years and 1 for cholesterol check in last 5 years
- **BMI:** Body Mass Index
- **Smoker:** 1, if the person has smoked at least 100 cigarettes in his/her entire life, 0 otherwise
- **Stroke:** 0 indicates the individual never had stroke, 1 otherwise.
- **HeartDiseaseorAttack:** 1 if the person has coronary heart disease (CHD) or myocardial infarction (MI). 0 otherwise.
- **PhysActivity:** 1 if the individual has done physical activity in past 30 days (not including job) , 0 otherwise.
- **Fruits:** 1 if the person has consumed fruit 1 or more times per day,0 otherwise.
- **Veggies:** If the person has consumed vegetables 1 or more times per day, 0 otherwise.
- **HvyAlcoholConsump:** 1 for individuals having more than 14 drinks per week in case of adult men and more than 7 drinks per week in case of adult women, 0 otherwise.
- **AnyHealthcare:** 1 if the person has any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. , 0 otherwise.
- **NoDocbcCost:** 1 if the person needed to see a doctor in the past 12 months but could not because of cost ,0 otherwise.
- **GenHlth:** What would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
- **MentHlth:** Number of days an individual faced problems related mental health which includes stress, depression, emotions for last 30 days.
- **PhysHlth:** Number of days an individual faced problems related physical illness and injury for last 30 days.
- **DiffWalk:** 1 if the person has serious difficulty walking or climbing stairs ,0 otherwise.
- **Sex:** 0 for Female, 1 for Male
- **Education:** Education level of the respondents. Scale 1-6: 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High

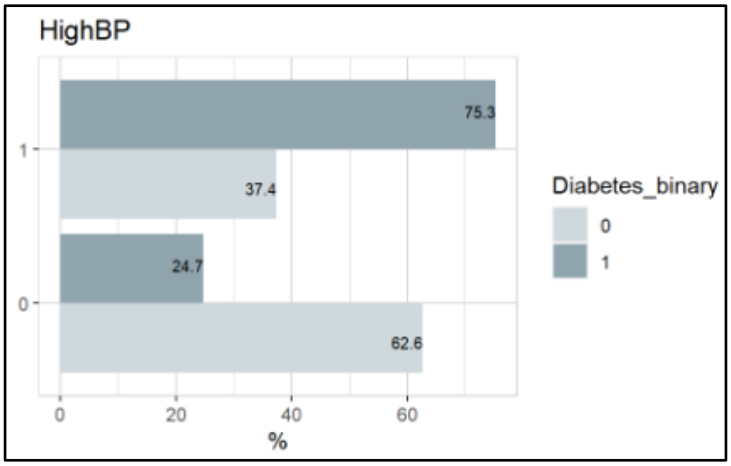
school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate).(Education Category)

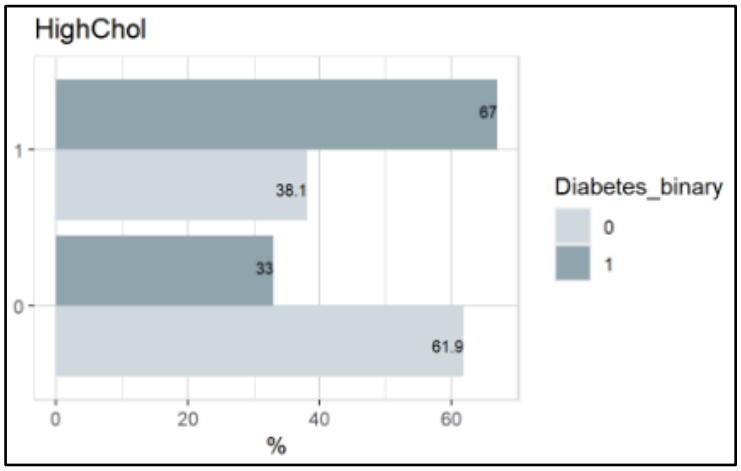
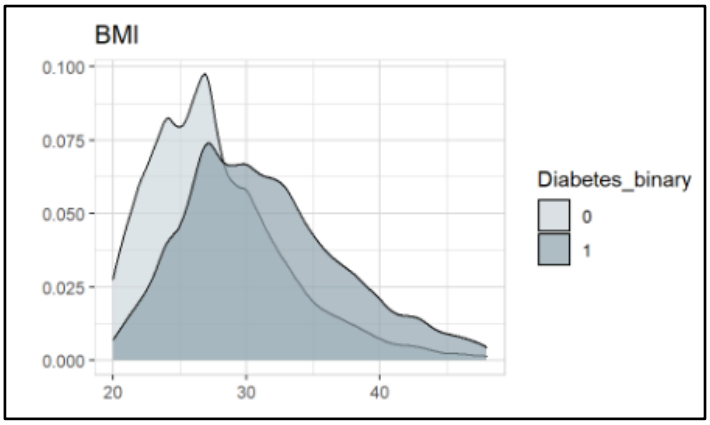
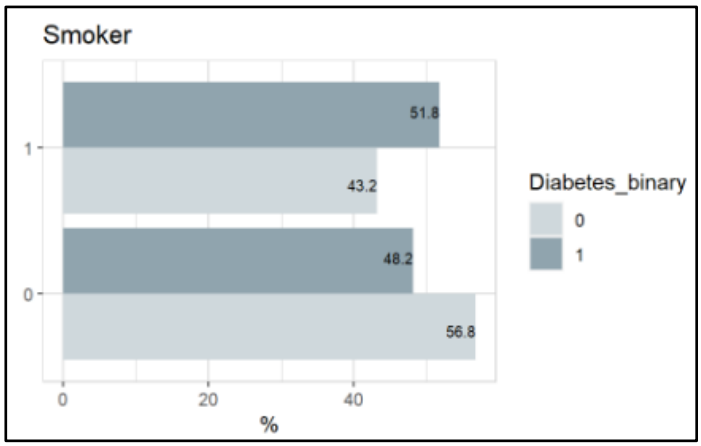
- **Income:** Income scale. Scale 1-8: 1 = < $10,000, 5 = < $35,000 8 = $75,000 or more.
- **Age:** Age of the respondents. In this survey responses are categorized in 13-level age category (13-level AGECategory). Where 1= age 18-24, 2=25-29, 8=age 55-59, 13=age 80 or more.

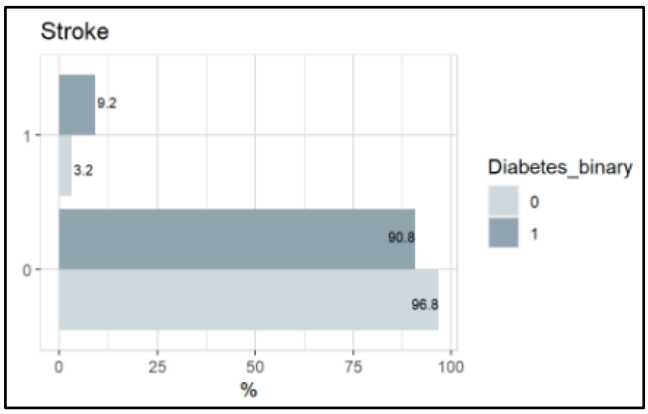
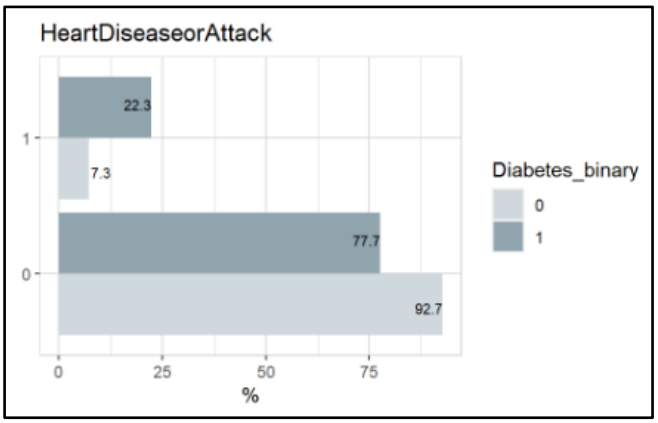
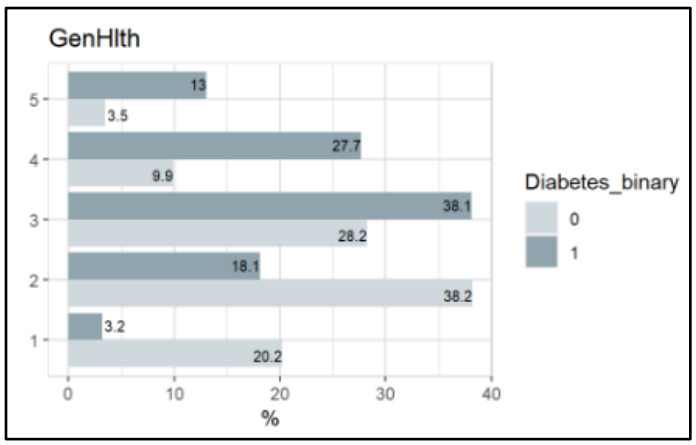
3. DATA EXPLORATION

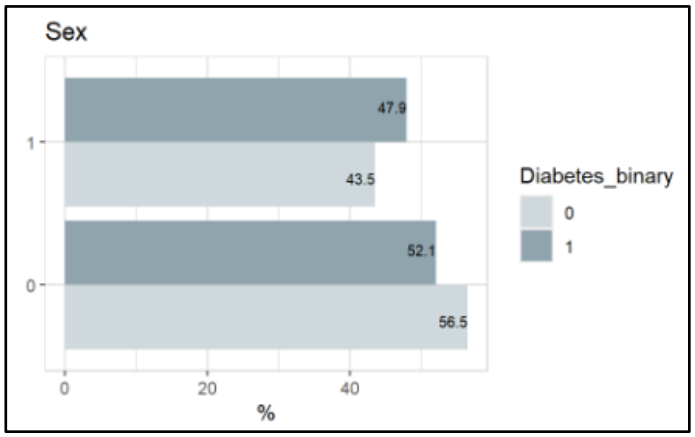
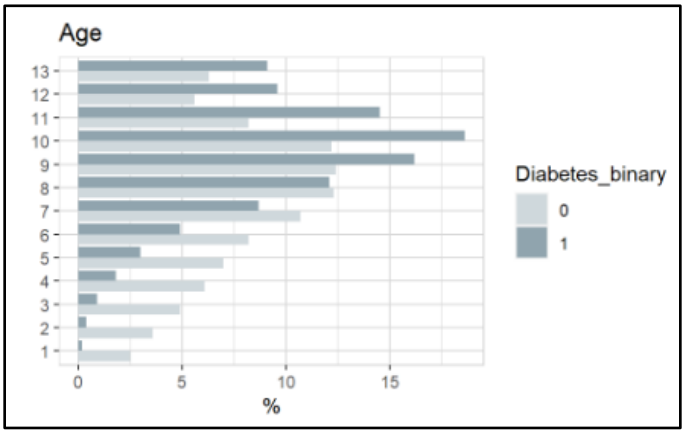
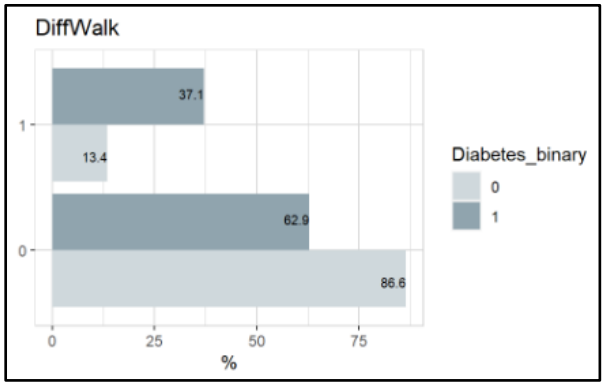


For our target variable diabetes (If an individual has diabetes or not) we see an equal split (50-50) in responses. No missing values have been found.

Plots of Feature variables vs Target variable	Interpretation
	Around 75.3% of the respondents having diabetes are suffering from high blood pressure whereas 37.4% of non-diabetic individuals have high blood pressure. A positive association can be seen between diabetes and highBP.

Plots of Feature variables vs Target variable	Interpretation
	67% diabetic individuals have high cholesterol and 38.1% of the non-diabetic individuals have high cholesterol. Here also we can see a positive association between diabetes and HighChol.
	From the graph we see that people having diabetes are likely to have more BMI.
	51.8% of the diabetic individuals have responded as smokers which means they have smoked at least 100 cigarettes, 56.8% of the nondiabetic respondents are non-smoker.

Plots of Feature variables vs Target variable	Interpretation
	9.2% of the diabetic respondents reported as they have had a stroke and 3.2% of the non-diabetic respondents had a stroke
	22.3% of the diabetic individuals are facing heart problems or had a heart attack where as 7.3% of the non-diabetic respondents are also facing similar problem
	Our general health condition was scaled from 1 to 5. The graph shows us that on an average non diabetic individuals are in better health condition which is quite obvious to some extent. 38.2% and 20.2% of non-diabetic respondents' health condition is very good and excellent respectively. On the other hand, 27.7% and 38.1% of diabetic people are in fair and good health condition respectively.

Plots of Feature variables vs Target variable	Interpretation
	52.1% of diabetic respondents were female and 56.5% of non-diabetic respondents were female.
	In this survey age responses were categorized in 13-level category, where 1=age 18-24, 8= age 55-59, 13=age 80 or more. Majority of the diabetic respondents are falling above 7 level categories which is people of above age 50. Proportions of non-diabetic individuals are higher in below level 7 categories.
	37.1% of the diabetic respondents are facing serious difficulties in walking or climbing stairs and 13.4% of the non-diabetic respondents are facing the same issue.

4. MODEL FITTING

4.1 Multiple Logistic Regression

Let Y be the binary response variable with 2 categories quantified as an indicator variable 0 and 1.

Let, P(Y=1)= π

and P(Y=0)= 1- π , 0< π <1

Let, there are p regressors $x_1, x_2, \dots, x_p$

Let us consider the generalised linear model (GLM) :

$$Y_i = f\left(\underline{\beta}|\underline{x_i}\right) + \varepsilon_i \text{ , i=1(1)n}$$

$$where\ \underline{x_i} = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)'$$

Or equivalently,

$$E\left(Y|\underline{x_i}\right) = f\left(\underline{\beta}|\underline{x_i}\right) = \pi_i$$

Let, $\eta_i = \underline{x_i}'\underline{\beta}$

Logistic Regression Model is a particular GLM where

$$f\left(\underline{\beta}|\underline{x_i}\right) = \frac{e^{\underline{x_i}'\underline{\beta}}}{1+e^{\underline{x_i}'\underline{\beta}}} = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \pi_i$$

Or, $$\eta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

Here, η is called the logit link function.

Let, there be k groups of observations with the i^{th} group having n_i observations with $\sum_{i=1}^{k} n_i = n$

Let, $y_i \sim Binomial(n_i, \pi_i)$, independently

The Likelihood function is given by :

$$L(\pi|\underline{y}) = \text{(constant)}. \prod_{i=1}^{k} {\pi_i}^{y_i} (1-\pi_i)^{n_i - y_i}$$

Solving $\frac{\delta(\ln L)}{\delta \underline{\beta}} = 0$, we get the set of (p+1) score equations :

$$\sum_{i=1}^{k} (y_i - \eta_i \pi_i)\, \underline{x_i} = 0$$

Score equations are solved numerically to get the estimates of the parameters.

4.2 Defining Dummy Variables

Let there be k categorical regressors in the model $(v_1, v_2, \dots, v_k)$ where v_j has c_j categories , $j = 1(1)k$

Then we have to introduce $\sum_{j=1}^{k}(c_j - 1)$ dummy variables in total.

The $c_j - 1$ dummy variables for v_j are defined as :

$$z_{ik} = \begin{cases} 1, & \text{if the } i^{th} \text{observation belongs to the } k^{th} \text{ category} \\ 0, & \text{otherwise} \end{cases}$$

k= 1(1)($c_j - 1$)

4.3 Train – Test Split

We randomly divide the whole dataset into two parts. We take 70% of the total observations as train dataset on which the model is initially fit. Remaining 30% of the total observations is used as test dataset. The number of observations in 2 parts are 49485 and 21207 respectively.

4.4 Variable Selection

To follow the principle of parsimony we have to choose a subset of best regressors from a pool of regressors. We use stepwise selection i.e. a combination of forward selection and backward elimination for the variable selection purpose.

The selection procedure is based on Deviance Statistic which is defined as :

$$D = -2\ln\left(\frac{L(\underline{x} \mid y_1, y_2, \ldots, y_n, fitted\ model)}{L(\underline{x} \mid y_1, y_2, \ldots, y_n, saturated\ model)}\right)$$

$$= -2\sum_{i=1}^{n}\left\{y_i \ln\frac{\hat{\pi}_i}{y_i} + (1-y_i)\ln\frac{1-\hat{\pi}_i}{1-y_i}\right\}$$

[*The model for which the data fits perfectly is called saturated model.*]

$$D \sim \chi^2_{n-p-1}$$

Small value of D indicates that the fitted model explains the data as efficiently as the saturated model.

Testing significance of parameters with deviance statistic :

$$H_0 : \beta_j = 0 \qquad against \qquad H_1 : \beta_j \neq 0$$

Define $G = D(model\ without\ X_j) - D(model\ with\ X_j)$

$$G \sim \chi^2_1$$

The steps for variable selection are as follows :

- We calculate G for all 1 regressor model (with intercept) and choose that X_j for which G is maximum among the significant G values (G > $\chi_{1-\alpha\,;1}$ at level α)
- We add each of the remaining (p-1) variables one at a time to the existing model and compute G for each case. We again choose the regressor with the maximum significant G value.
- We proceed in this way and at each step check for possible deletion of all previously included regressors by backward elemination i.e. calculate G after removing the previously included variables one at a time and drop the variable with minimum insignificant G value.
- We continue till there is no variable left with significant G value.

In case of our data the stepwise selection procedure is shown in the table below :

Table : Steps Showing Included Variable and Corresponding Subset Model Deviance in the Stepwise Selection Procedure

Step	Included Variable	Deviance
1	GenHlth	59404.50
2	HighBP	55415.34
3	Age	53686.52
4	BMI	51635.35
5	HighChol	50992.77
6	CholCheck	50782.99
7	Income	50588.55
8	HvyAlcoholConsump	50419.06
9	Sex	50252.15
10	HeartDiseaseorAttack	50150.13
11	Education	50120.47
12	DiffWalk	50092.43
13	Stroke	50077.74
14	PhysHlth	50071.79

No variable was deleted in the stepwise selection method as no insignificant G value was found.

So, we fit the model with 14 regressors listed in the 2nd column of the above table.

Now, selecting the cut-off probability as 0.5, the predicted probabilities of Y taking value 1, i.e. $\hat{\pi}_1$ are transformed such that, if $\hat{\pi}_1 < 0.5\ then\ \widehat{Y} = 0, \text{otherwise}\ \widehat{Y} = 1.$

The confusion matrix, thus obtained is as follows:

Observed (Y)	**Predicted ($\widehat{Y}$)**	
	0	1
0	17995	6816
1	5540	19134

Table 2: Summary of the fitted logistic model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.24891	0.381311	-19.0105	1.40E-80
GenHlth2	0.738117	0.046038	16.03261	7.56E-58
GenHlth3	1.443559	0.045545	31.69516	1.81E-220
GenHlth4	1.868223	0.052073	35.87667	7.06E-282
GenHlth5	1.989909	0.068355	29.11121	2.59E-186
HighBP1	0.731581	0.023738	30.81895	1.46E-208
Age2	0.022975	0.192977	0.119054	0.905233
Age3	0.359252	0.174278	2.061371	0.039268
Age4	0.684222	0.165098	4.144352	3.41E-05
Age5	0.993807	0.161056	6.170583	6.80E-10
Age6	1.184491	0.158539	7.471292	7.94E-14
Age7	1.44325	0.1566	9.216148	3.08E-20
Age8	1.50662	0.155792	9.670739	4.01E-22
Age9	1.745566	0.155396	11.23302	2.81E-29
Age10	1.894015	0.155311	12.19495	3.31E-34
Age11	2.068561	0.156223	13.24107	5.08E-40
Age12	1.948536	0.157678	12.3577	4.43E-35
Age13	1.886448	0.157706	11.96178	5.63E-33
BMI	0.074455	0.001892	39.3429	0
HighChol1	0.543851	0.022753	23.90201	2.92E-126
CholCheck1	1.333717	0.096485	13.82307	1.85E-43
Income2	0.063538	0.067261	0.944643	0.344841
Income3	-0.01526	0.064305	-0.23737	0.812373
Income4	-0.07352	0.062149	-1.18289	0.236853
Income5	-0.12505	0.060969	-2.05104	0.040263
Income6	-0.21652	0.059672	-3.62841	0.000285
Income7	-0.20177	0.059789	-3.3747	0.000739
Income8	-0.39659	0.058597	-6.76812	1.30E-11
HvyAlcoholConsump1	-0.74453	0.058286	-12.7738	2.30E-37
Sex1	0.270966	0.022698	11.93793	7.51E-33
HeartDiseaseorAttack1	0.304857	0.033994	8.968081	3.02E-19
Education2	0.389154	0.336863	1.15523	0.247996
Education3	0.307855	0.332138	0.92689	0.353984
Education4	0.227399	0.329149	0.69087	0.489647
Education5	0.272471	0.329276	0.827487	0.407961
Education6	0.140381	0.329406	0.426165	0.669988
DiffWalk1	0.167135	0.030769	5.431873	5.58E-08
Stroke1	0.188847	0.048998	3.854178	0.000116
PhysHlth	-0.00352	0.001444	-2.43967	0.014701

5. MODEL DIAGNOSTICS

After fitting the model with 14 regressors to the training data, we proceed to check the performance of the model by the measures: Pearson's Chi-Squared test, φ Coefficient and Contingency Coefficient.

Pearson's Chi-Squared test:

This test is carried out to check for the statistical independence of the observed responses (Y) and the fitted responses ($\hat{Y}$).

In order to carry out this test, the joint probability distribution of Y and $\hat{Y}$ is estimated by the known proportions from the confusion matrix.

Consider the testing problem-

$\mathbf{H_0}$: $Y \text{ and } \hat{Y} \text{ are independent}$ against $\mathbf{H1}$: $Y \text{ and } \hat{Y} \text{ are dependent}.$

The Pearson test statistic is given by:

$$\chi_p^2 = \sum_{j=1}^{J} \frac{\sum_{k=1}^{K} (f_{jk} - \frac{f_{j.} f_{.k}}{n})^2}{\frac{f_{j.} f_{.k}}{n}}$$

where, f_{jk}: the (j,k)th cell frequency of the confusion matrix

$f_{j.} = \sum_{k=1}^{K} f_{jk}$

$f_{.k} = \sum_{j=1}^{J} f_{jk}$

J, K: number of categories. Here, J=K=2.

Under the null hypothesis, the test statistic χ_p^2 follows a $\chi^2_{(1-\alpha);(J-1)(K-1)}$ distribution.

Taking α=0.05, $\chi^2_{(1-\alpha);(J-1)(K-1)} = \chi^2_{0.95,1} = 3.84146$

Decision: The test procedure is such that, if the observed value of the Pearson test statistic is greater than $\chi^2_{(1-\alpha);(J-1)(K-1)}$ then, H_0 is rejected, i.e. $Y \text{ and } \hat{Y} \text{ are not independent}$

Observation: From the confusion matrix, the value of the test statistic is obtained as 12436.18 which is much greater than $\chi^2_{0.95,1} = 3.84146$. Hence, Y and $\hat{Y}$ are not independent. This means that the predicted response is highly dependent on the observed response and the model is successful in the task of prediction using the information in the data.

Since, the hypothesis that $Y \; and \; \hat{Y} \; are \; independent$ is rejected, we can further check the following measures of association:

a. Φ Coefficient: It is a measure of association between two binary variables given by,

$$\varphi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1.}f_{0.}f_{.0}f_{.1}}}$$

Higher the value of φ coefficient, stronger is the association between the two variables.

From our confusion matrix, φ=0.5013508.

b. Contingency Coefficient: It is given by,

$$P = \sqrt{\frac{\chi_p^2}{\chi_p^2+n}}$$

where, χ_p^2: Pearson's Chi-square statistic

From our data it is calculated as, P = 0.4481.

Now, observing the values of the measures, we can be assured of the adequate performance of the fitted model on the training data. So, we proceed to interpret few of the model parameters.

6. INTERPRETATION OF THE PARAMETERS

One of the advantages of logistic regression is the ease and clarity with which the model coefficients can be interpreted in order to understand the effect of each regressor on the response.

In logistic regression, the j^{th} regression coefficient β_j denotes the change in the logit function with respect to a unit change in X_j, keeping all other covariates fixed.

For better understanding, let us consider a single categorical explanatory variable X with two levels 0 and 1. Then,

Odds Ratio(OR)=$\frac{Odds\ in\ favour\ of\ Y=1\ when\ X=1}{Odds\ in\ favour\ of\ Y=1\ when\ X=0} = \frac{Odds\ for\ X=1}{Odds\ for\ X=0} = e^{\beta_1}$. Here, β_1 is the coefficient of X.

In case of k(>2) levels of a categorical explanatory variable, we create (k-1) dummy variables, considering one level to be the reference level. Then the interpretation of the β estimate of the j^{th} dummy variable is done by looking into the odds ratio between the j^{th} level and the reference level.

In this way, we have interpreted the coefficient estimates in our model to get an idea of how each factor affects the chance of a person being diabetic.

- According to the estimates of the various levels of the variable **Age**, people of age 55 years and above are around 4 to 6 times more likely to have diabetes, compared to people falling in age category 18-24, keeping other possible factors to be fixed. Specifically, people of age group 70-74 are almost 8 times more likely to have diabetes.
- According to the estimates of the various levels of the variable **GenHlth**, people having poor health are 7 times more likely to have diabetes, compared to people claiming to have excellent health.
- People having high **blood pressure** are around 2 times more likely to have diabetes than the people with low blood pressure.
- People with high **cholesterol** are 1.7 times more likely to have diabetes, compared to people having low cholesterol.
- People who have not **checked cholesterol** in last 5 years are around 4 times more likely to have diabetes compared to those who have not.
- People having coronary **heart disease** (CHD) or myocardial infarction (MI) are 1.35 times more likely to have diabetes than the people without heart disease.
- Males are 1.3 times more likely to have diabetes than female.
- 10 units increase in **BMI** increases possibility of having diabetes by 2 times.

7. MODEL PERFORMANCE ON TEST DATA

The confusion matrix obtained from the test data is as follows:

Observed (Y)	**Predicted ($\widehat{Y}$)**	
	0	1
0	7562	2973
1	2392	8280

Some of the diagnostic metrics from the confusion matrix are:

1. **Percentage of misclassification:** It is the percentage of the total number of predictions that are incorrect.

   $$P.M.C. = \frac{f_{01} + f_{10}}{n} \times 100\%$$

   The lower the percentage of misclassification, the better.

   Here, P.M.C. = 25.298%

2. **Accuracy**: It is the number of correct predictions divided by the total number of predictions made by the model.

   $$Accuracy = \frac{f_{11} + f_{00}}{n}$$

   The higher the accuracy, the better.

   Here, Accuracy= 0.747

3. **Sensitivity Rate/True Positive Rate**: It is the proportion of positives correctly identified as positive.

   It is defined as: $\frac{f_{11}}{f_{01}+f_{11}}$

   Here, Sensitivity rate= 0.7759.
   This shows that the proportion of the patients predicted to be having diabetes by the model out of those actually having diabetes is 0.7759.

4. **Specificity Rate/ True Negative Rate:** It is the proportion of negatives correctly identified as negatives.

   It is defined as: $\frac{f_{00}}{f_{00}+f_{10}}$

   Here, Specificity Rate= 0.7178
   This shows that the proportion of the patients predicted to be non-diabetic out of those who are actually non-diabetic is 0.7178.

8. CLASSIFICATION USING ML

We now try to predict the presence of diabetes using the same training and testing dataset with the help of Random Forest Classifier.

Random Forest Classifier

Decision Tree is a supervised machine learning algorithm that uses a set of rules to make decisions where the set of rules are based on the feature vector of the learning sample and the output is the assigned class. Random Forest consists of large numbers of individual decision trees that operate as an ensemble. For classification task, the output of the random forest is the class selected by most trees.

Results Obtained from The Test Dataset

The following confusion matrix and the subsequent diagnostic measures give a clear picture of the model accuracy.

Confusion Matrix

Observed (Y)	**Predicted ($\widehat{Y}$)**	
	0	1
0	7365	3170
1	2475	8197

Diagnostic measures based on the confusion matrix

1. Percentage of misclassification=26.61%
2. Accuracy=73.39%
3. Sensitivity Rate=0.768085
4. Specificity Rate=0.699098

We observe that the Random Forest Classifier gives lesser accuracy, sensitivity and specificity rates than that of the fitted logistic regression model.

9. CONCLUSION

In this paper, we aim to predict presence of diabetes in a person by using logistic regression model. Upon splitting the entire dataset into training and testing sets, we use stepwise selection to fit a parsimonious model on the training set. We found that people, of age more than 55 years, are around 4 to 6 times more susceptible to diabetes as compared to people in the age group 18 to 24. People having high cholesterol and high blood pressure are more susceptible to have diabetes. Also, it is noted that people having higher BMI have higher probability of having diabetes. We use this fitted model on the test dataset and found an accuracy of 74.7%. Alternatively, we use Random Forest Classification to cater to the same problem. We found the accuracy of the classifier to be 73.39%. We may conclude that logistic regression acts as a better classifier than Random Forest in our dataset.

10. REFERENCES

Below is a list of references we used for the completion of this project:

- Lecture Notes of Regression Analysis (MTH416A), instructed by Dr. Sharmishtha Mitra
- *Applied Logistic Regression* – David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant
- *An Introduction to Statistical Learning with Applications in R* – Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani