St. Joseph's University

Bangalore, Karnataka

PROJECT REPORT

ON

Advance Statistical Methods

Submitted by:

Anwesh Roy - 222BDA45

Irene Sarah Biju - 222BDA23

Joel Bastin X - 222BDA32

Submitted to:

JAYATI KAUSHIK

Assistant Professor

Department of Advanced Computing

St. Joseph's University

Logistic Regression for Diabetes Analysis

Statistically, the association between a binary dependent variable (such as having diabetes or not) and one or more independent variables (such as age, BMI, etc.) can be examined using logistic regression. Finding the most accurate model that estimates the likelihood that the dependent variable (such as having diabetes) will occur once given the independent factors is the aim of logistic regression.

We have a dataset of patients with diabetes and their respective independent variables such as age, BMI, blood pressure, etc. We can use logistic regression to analyze the relationship between these independent variables and whether a patient has diabetes or not. We also did some EDA to understand the relationship between different features in order to get better insights about the dataset.

For hypothesis purpose we defined null and alternative hypothesis.The alternative hypothesis (H1) would be that there is a significant association, while the null hypothesis (H0) would be that there is no significant relationship between the independent factors and diabetes.

Next, we used a significance test (chi-square test, Shapiro-Wilk Test,MannWhitneyU Test) to determine whether our findings are statistically significant. If the p-value is less than our predetermined significance level (usually 0.05), we can reject the null hypothesis and conclude that there is a significant relationship between the independent variables and diabetes.

The degree and direction of the association between each independent variable and diabetes can be determined by interpreting the logistic regression coefficients, assuming we reject the null hypothesis and discover a significant relationship. For instance, a positive coefficient for BMI would mean that the risk of having diabetes rises as BMI rises.

In general, logistic regression can be an effective method for examining the association between independent variables and a binary dependent variable, like diabetes. We can learn more about the causes of diabetes and determine which patients are most likely to develop it by conducting hypothesis testing and analysing the coefficients.

Domain- Medical Field

Problem Statement- Logistic regression on medical data

Introduction

Diabetes is a very common disease prevailing among all the individuals irrespective of their age group. There are almost 450 million people being diagnosed with diabetes all over the earth. In our study we have used logistic Regression to predict diabetes in a person. Along with that many hypothesis analysis have been done for the better understanding of the data and to come across a accurate result. Our model and all the tests done will help people to know the factors that is making them at risk to diabetes.

We have acquired the dataset of women which includes the key features contributing to diabetes, this will eventually help us to study more in depth about the diabetes. Also it performs as a predictive model which helps us to predict the disease. Prediction model will eventually screen the pre-diabetes or people more prone to diabetes. We have combined our knowledge in Logistic Regression and machine learning to do the above mentioned.

Literature Review

Diabetes is a severe, chronic condition that develops when the pancreas either produces insufficient insulin, a hormone that controls blood sugar or glucose, or when the body struggles to utilize the insulin that is produced. One of the four noncommunicable diseases (NCDs) that top the list of priorities for action by world leaders is diabetes, which is a significant public health issue. Over the past few decades, there has been a consistent rise in both the incidence and prevalence of diabetes.
Studies on the role of bariatric (metabolic) surgery in relation to sustained weight loss and the effects of weight loss on improvements in diabetes mellitus have been conducted.

The management of diabetes mellitus has been demonstrated to be impacted by metabolic surgery through a number of methods, both weight-dependent and weight-independent. Although the precise mechanisms are unknown, we are getting closer to comprehending how metabolic surgery contributes to improved glucose homeostasis and, consequently, better management of diabetes mellitus. Improved cardiovascular health, one of the main causes of morbidity and mortality in the United States, is one of the positive effects of improved diabetes control after metabolic surgery.

The clinical evidence regarding the link between weight variability and CVD and microvascular complications of diabetes has been the subject of another significant study. Researchers have compiled the potential methods for reducing weight fluctuations, such as particular dietary plans and readily available glucose-lowering medications.

There are few studies on depressive and antipsychotic drug prescribing in relation to diabetes outcomes, and those that do exist have limitations and inconsistent results. People with diabetes who are taken antidepressants and antipsychotics should be monitored, have risk factors treated appropriately, and have complications screened until further information is known, per the recommendations in general diabetes guidelines.

Aim of the work:

The aim of the project is to predict whether a person is affected by diabetes using logistic regression and using some hypothesis-testing techniques to validate the project.

Methods and Materials:

- We have used Pandas and NumPy library for basic operations with dataset.
- We have used Matplotlib and Seaborn libraries for plotting graphs for Exploratory Data Analysis.
- We have used the math library to perform some mathematical operations.
- We have used the sklearn library to create a predictive model.
- From the sklearn library we imported Logistic Regression and created the model for predicting the outcome of our model.
- From the sklearn library we import the accuracy score and confusion matrix to know how accurate the model is predicting and to validate the model.
- From the sklearn library we used train_test_split which is used to split the data into training data and testing data, where training data is used to train the model and testing data is compared with the predicted output and it is checked whether the model is working correctly or not.
- We installed Scipy and stats models to do Anova test, Shapiro test and Chi-square test

Study Design

Problem Statement:

Create a model for predicting whether a patient is affected by disease or not using Logistic Regression. And do hypothesis testing with the given data.

Data Collection:

Data was collected from Kaggle.

Data Cleaning:

In data cleaning we removed the NaN values, replaced it with median and removed the unnecessary outliers.

Exploratory Data Analysis:

We used violin plots, heatmaps and pie charts for doing our Exploratory data analysis with the help of Matplotlib and Seaborn libraries

Model Building:

With the given data, we divided the dataset into independent and dependent variables as x and y. Then we are splitting the whole data into testing data and training data. Then we fit the training data into the Logistic Regression and with the predict function we are finding the predicted output. And with predicted output and test data we can get the accuracy of the model.

Exploratory Data Analysis (EDA)

(a) For better understanding of the data, we did an exploratory data analysis. The idea is to find how each feature are interrelated with each other. We used violin plots for visualizing the distribution of each variable in relation to the outcome (diabetes or non-diabetes) and for identifying any patterns or trends in the data.

However, if the minimum value for each of the variables (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI) is zero, it is likely that these values are missing or incomplete rather than actual values. It is not possible for these variables to have a value of zero in a living human being.

We carefully reviewed the data and identified any missing or incomplete values, and implemented an appropriate imputation method to tackle missing values or excluded the incomplete cases from the analysis.

From the graphs we got to know that:

- Glucose levels tend to be higher in individuals with diabetes compared to those without diabetes.
- Blood Pressure is similar for both groups
- Skin Thickness tends to be higher in individuals with diabetes
- Insulin levels are higher in individuals with diabetes
- BMI tends to be higher in individuals with diabetes
- Diabetes pedigree function doesn't appear to be significantly different between two groups.

(b) We then moved on to plot a heatmap to find correlation between various numerical parameters in dataset. The Heatmap shows the Pearson's correlation between various parameters. Most of the variables do not show a very strong correlation

BMI and Skin Thickness are the only 2 variables to show a correlation > 0.5. Although it is still not a very strong correlation number.

(c) We were willing to see how much of positive and negative diabetic patients were there in the dataset. So we plotted a pie chart with labels "Positive" and "Negative". The percentage value inside each slice indicates the proportion of the total dataset that is positive or negative.

34.9% of women have diabetes.

65.1% of women do not have diabetes.

Results

After EDA, as described in the story we will move forward to the hypothesis and testing phase.

We used Shapiro test, MannWhitneyU test (if the dataset are not normally distributed), T-test, etc.

The following hypothesis we did are:-

(i) *Blood Pressure vs Outcome*

The hypothesis is like comparing the average blood pressure. The null hypothesis (H0) states that the difference between the mean blood pressure of diabetics and non-diabetics is less than or equal to zero and alternate hypothesis(H1) states that the difference between the mean blood pressure of diabetics and non-diabetics is greater than zero. We used Shapiro test and concluded that Blood Pressure is higher for diabetic women.

(ii) Blood Pressure vs Pregnancies

Here our null hypothesis(H0) is like whether the blood pressure of all the pregnancies are same and alternative hypothesis(H1) is average blood pressure of at least one pregnancy is different. We tried Kruskal test and we fail to reject H0. We got that, blood pressure doesn't change with pregnancies.

(iii) BMI vs Outcome

We compare the mean BMI of diabetic and non-diabetic individuals. The null hypothesis (H0) states that the difference between the mean BMI of diabetics and non-diabetics is less than or equal to zero and (H1) states that the difference between the mean BMI of diabetics and non-diabetics is greater than zero.We tried Shapiro test but since the datasets were not normal we went for MannWhitneyU test and conclude that the BMI is higher for Diabetic women.

(iv) BMI vs Pregnancies

Null hypothesis(H0) is the average BMI of all the pregnancies are same and alternate hypothesis(H1) is that average BMI of at least one pregnancy is different. We went with Kruskal test and found that $p<0.05$, test statistics falls in the rejection region. We conclude that BMI is dependant on number of pregnancies a women has had.

(v) Diabetes Pedigree Function vs Outcome

We compare the mean diabetic pedigree function of diabetic and non-diabetic individuals. The null hypothesis (H0) states that difference between the mean DPF of diabetics and non-diabetics is less than or equal to zero. And alternate hypothesis (H1) states that the difference between mean DPF of diabetics and non-diabetics is greater then zero. We tested with MannwhitneyU test and concluded that Test statistics falls in the rejection region($p<0.05$).We therefore reject H0.We conclude that Diabetes Pedigree Function for diabetic women is greater than for non-diabetic women.

(vi) Glucose vs Outcome

The null hypothesis (H0) is that the difference between the mean glucose levels of diabetics and non-diabetics is less than or equal to zero and (H1) states that the difference between the mean glucose levels of diabetics and non-diabetics is greater than zero. After conducting the MannWhitneyU test we conclude that glucose level for diabetic women is higher than for non-diabetic women.

We also conducted hypothesis testing on Glucose vs Pregnancies, Skin Thickness vs Outcome, Insulin Vs Outcome, Insulin vs Pregnancies and Age vs Outcome and verified all the key factors that were important for the analysis.

Model

The model that was suitable for our problem statement was logistic regression. With this we predicted weather a person is diabetic or non diabetic.

Codes

Here is the link to the GitHub repository:

https://github.com/Anwesh154/ADS_Project

Conclusion

1. **Glucose**: Rise in glucose levels could lead to increasing risk of diabetes. However, pregnancy does not impact the glucose levels.
2. **Blood Pressure:** High blood pressure indicates an increased risk of Diabetes. However, number of pregnancies that a woman has had does not have an impact on her blood pressure.
3. **BMI**: Increasing BMI could increase the risk of diabetes. Pregnancies could cause a change in BMI leading to change in risk of getting diabetes. Surprisingly, women who have not been pregnant even once tend to have a higher BMI.
4. **Skin Thickness:** Increase in thickness of the triceps skin fold can also cause increase in risk of getting diabetes. This is a means of measuring the body fat percentage. In the same case as BMI, increase in body fat percentage could lead to higher risk of getting diabetes. Number of pregnancies does not impact the skin thickness of a woman
5. **Diabetes Pedigree Function:** This is a genetic function that gives us an idea of our risk of getting diabetes. Naturally, higher the value higher the risk.
6. **Insulin**: Decrease in insulin level could again lead to increasing risk of contracting diabetes. Insulin level changes with the number of pregnancies a woman has had.
7. **Age:** Naturally number of pregnancies increase with age. Also, older women are at a higher risk of getting diabetes.

Future Works

The way we conduct research on diabetes epidemiology is drastically altering as a result of the use of digital health data, such as that gathered from social media or linked devices, as well as the methods of artificial intelligence used to analyse it. Due to a better integration of the impact of diabetes on people's daily lives, the concepts of deep digital phenotyping and digital twins, which are based on the development of large cohort studies of extremely well characterised people with diabetes, will make it possible to further personalise the monitoring and management of the disease. Similar to how voice biomarkers would assist diabetes patients' remote monitoring, the creation of digital biomarkers will pave the way for the telemonitoring and telemedicine of the future.

We are increasingly focusing on the analysis of the experience and perception of the disease given by the individuals with diabetes themselves in order to better manage diabetes and go beyond the standard risk factors already identified. It is crucial to keep in mind that a person with diabetes sees a medical practitioner for 6 hours on average every year. However, it takes more than 600 hours a year to manage the illness on your own, including daily decisions about the disease's course of treatment, food, physical activity, and relationships with family and friends as well as the workplace. And it does not include the nearly constant mental strain caused by diabetes. Before very recently, everything that occurred outside of consultations or biological tests—yet serving as a rich source of data to better understand the various varieties of diabetes—was either disregarded in clinically-focused medical research, or was either inaccessible or not gathered. With the use of digital technologies, "real life data" will now play a larger role in patient-centered research and medical care going forward.

Digital biomarkers for diabetes research and remote monitoring

Digital data has the potential to enhance remote diabetes monitoring, which will be beneficial for both ongoing clinical practise and future diabetes research. We may now recognise digital biomarkers linked to diabetes or its complications, whether via the use of closed-loop devices or through digital remote monitoring. Avoiding the inter-consultation black hole, during which the T1D patient has little to no touch with medical experts, is one of the key challenges of telemedicine and tele-support. Therefore, digital biomarkers

enable remote monitoring of crucial indicators for telemonitoring diabetes, as well as the potential to activate an alert system and set an earlier appointment date for those who require it.

One of the most active areas of research nowadays is the use of novel data sources, including speech, to identify digital vocal biomarkers in addition to monitoring traditional glycaemic control parameters via medical devices. A vocal biomarker is an audio signal, or a set of audio signals, that is linked to a health condition or occurrence that one wants to monitor, identify, or forecast. It must possess the same qualities as a conventional biomarker, such as analytical validation, clinical benefit demonstration, and practical use. Even though the Food and Drug Administration and the European Medicines Agency have not yet approved any vocal biomarkers, more technological advancements including vocal biomarkers for medical applications are anticipated in the upcoming years.

The most socially connected group of chronically ill persons worldwide is the diabetic community. In the future, patients with diabetes may be identified by hundreds of values for clinical, biochemical, genetic, emotional, or environmental characteristics rather than just one or two recent assessments of their HbA1c. These data, which were generated by digital technology, and the artificial intelligence techniques used to analyse them have changed how we view the many types of diabetes, their therapies, and their management. Although these new opportunities are encouraging, they also present a number of difficulties that should not be disregarded. Above all, these advancements must systematically work in favour of diabetics who, with the help of digital data, must be liberated from the burden of their condition and the accompanying mental strain. We need to make sure that no one with diabetes is left out of these major changes we are currently going through. The main challenges of today and tomorrow include reducing social disparities caused by digital technology, encouraging co-construction of innovations, ensuring an ethical analysis of these vast amounts of data, improving training for health professionals and researchers in artificial intelligence, and facilitating the adoption of these advancements by people with diabetes. The possibility of systemic bias against particular minority will therefore be reduced by diversifying the demographics of diabetes patients in cohorts and trials. The growth of open data and open source in medical research would speed up scientific advancements that will significantly improve the lives of people with diabetes. This will be accomplished by halting research "in silos," funding large-scale projects, and putting an end to research "in silos".

Key Learnings

Diabetes is a complex disease with many different aspects to study, including risk factors, prevention, diagnosis, and treatment. Here are some key findings in diabetes research:

Prevalence -Diabetes is a global health problem, with an estimated 463 million adults living with diabetes worldwide in 2019. The prevalence of diabetes is increasing, particularly in low- and middle-income countries.

Risk factors - There are several risk factors associated with the development of diabetes, including obesity, physical inactivity, unhealthy diet, smoking, and genetics. Identifying and addressing these risk factors can help in preventing or delaying the onset of diabetes.

Diagnosis - Early diagnosis of diabetes is important for effective management and prevention of complications. Several tests are available for diagnosing diabetes, including fasting plasma glucose, oral glucose tolerance test, and glycated haemoglobin (A1C) test.

Treatment - Treatment for diabetes involves managing blood glucose levels, blood pressure, and cholesterol levels, as well as adopting a healthy lifestyle. Treatment options include medications, insulin therapy, and lifestyle changes such as regular exercise and a healthy diet.

Since we used the logistic Regression model to study our data we were able to identify many significant factors such as Age,BMI,Blood Pressure.These were some of the factors to help us predict the risk of the disease. For example, it can be used to determine how much the risk of diabetes increases for every unit increase in BMI.We used the Logistic regression model as a predictive model to try and understand the probability of an individual developing diabetes. The model has been tested for its accuracy and other analysis has also been done.Individuals who are identified at high risk for diabetes are given the chance prelook the damage and take necessary precautive measures.

References

1. https://diabetesaction.org/current-research
   National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
   CURE FOR TYPE 1 DIABETES
2. https://www.sciencedaily.com/releases/2021/01/210127122356.htm
   Scientists discover a new promising target for diabetes treatment
   January 27, 2021,Helmholtz Zentrum München - German Research Center for Environmental Health
3. https://www.nature.com/subjects/type-2-diabetes-mellitus
   Type 2 diabetes articles from across Nature Portfolio
4. https://www.kaggle.com/datasets/kandij/diabetes-dataset
   Diabetics prediction using logistic regression