Diabetes Analysis and Prediction

Asa Bejraputra, Alireza Palizibasmanj, Alfiya Ahmed, Akshay Malik, Natthaphong Pinthong
Data Science (M.Sc.) Student, winter semester 2024/25
University of Europe for Applied Sciences, 14469 Potsdam, Germany.

***Abstract*—This report is our analysis on the annual healthcare survey of US citizens. We will figure out the risk factors which associated to diabetes disease, then use statistical procedure to find out the difference in characteristics between the diabetes patient compared to others. Then build-up the regression model to predict the risk to diabetes.**

***Index Terms*—Diabetes, Body Mass Index ("BMI"), Cholesterol, Hypertension**

I. Introduction

Diabetes is a chronic disease that occurs either when the pancreas does not produce insulin or when the body cannot effectively use the insulin it makes. Insulin is a hormone that regulates blood glucose. Lack of effective blood glucose regulation may lead to Hyperglycemia, which affects many of the body's systems, especially nerves and blood vessels.

Nowadays, diabetes become one of the four main non-communicable diseases ("NCD") – besides Cardiovascular diseases, Cancer, and Chronic respiratory. In 2022, around 14% of adults aged 18 years old and over suffer from diabetes. Moreover, many of them were not covered by medical treatment, especially in low- and middle-income countries.

Our analysis is to use correlation analysis to find out which input attributes have a valid association with diabetes. Then, find out the difference between those who have diabetes and those who don't.

II. Data And Methodology

A. Data Source

In this analysis, we used the datasets provided from the Behavioral Risk Factor Surveillance System (hereinafter called **"BRFSS"**, or **"The Survey"**), which the Survey is conducted annually by The United States Center of Disease Control (hereinafter called **"US CDC"**). In our analysis, we used the Survey of 2023, which is the latest published survey as of the date we analyze the data.

B. General Information of Dataset

As already mentioned, BRFSS is the Survey conducted annually by US CDC. The Survey itself contains numbers of questions and topocis related to survey respondents' daily life behavior across the USA: their demographic information, their risk behaviors, chronic health information, healthcare access, their mental and physical health, and so on.

C. Methodologies

The Survey contains hundreds of attributes. However, since our analysis is focused on 'Diabetes' disease, we selected only some questions considered as 'relevant' to the diabetes diseases. Then, we will split the observations into 2 main groups: Those who have told before the survey that they have diabetes, and those who don't (hereinafter called **"the Groups"**).

Among the selected columns, we used statistical tools to find out 5 attributes with the highest correlation with diabetes diseases. Then, we performed statistical testing to conclude if the value of each attribute has the significant difference between the Groups.

Then, we built up the logistic regression model to build up the model that predicts the risk to diabetes illness. Due to the large difference of number of observation between the Groups, the sample balancing technique was applied with those datasets without sample balancing techniques. Then, we compared the prediction accuracy to find out which model has more accurate prediction.

In terms of tools used, we used python with other libraries related to dataset management, statistical testing, and regression model e.g. pandas, scikit-learn, etc. as main tools in our analysis.

III. Data Preprocessing

In this section, we will explain how we process, clean, and preliminary analyze the data.

A. Data Preprocessing

1) Preliminary Attribute Selection: Initially, the Survey provided the number of attributes in our datasets (350 attributes). We selected some of it as we considered it should be relevant to diabetes disease as listed as shown in Table I:

2) Data Transformation: Then, for our analysis purpose, we need to do some data transformation as follows:

- drop the null values using *pandas.dropna()* method.
- Reduce their classification in 'DIABETE4' column from various kinds of output to binary values. Classify the person who in prediabetes or in the boundary stage of diabetes into the group of those without diabetes disease. Then, remove the uninterpretable values e.g. Those who don't know or refuse to answer the question.
- Divide the values in '_BMI5' by 100 (as they assume the 2 additional decimal digits)

Variable name	Variable description
DIABETE4	Respondent's known diabetes status
_BMI5	Computed Body Mass Index (BMI)
_RFBMI5	Overweight of obese
_RFHYPE6	High Blood Pressure
TOLDHI3	Respondent's known Cholesterol level
_CHOLCH3	Cholesterol checked
CHCKDNY2	Ever told you have kidney diseases
SMOKE100‘	Respondent smoke at least 100 cigarettes
_RFDRHV8	Heavy Alcohol consumption
CVDSTRK3	Ever diagnosed with strokes
_MICHD	Ever had CHD or MI
_TOTINDA	Leisure time physical activity
GENHLTH	General health
PHYSHLTH	Number of days which physical health not good
MENTHLTH	Number of days which mental health not good
DIFFWALK	Difficulty on walking or climbing stairs
_HLTHPL1	Have any health insurance
MEDCOST1	Could not afford to see doctor
CHECKUP1	Length of time since last routine checkup
_SEX	Survey respondent's sex
_AGEG5YR	Age categories
EDUCA	Respondent's education
INCOME	Respondent's income

TABLE I
SELECTED COLUMN NAME, AND THEIR DESCRIPTION

- For categorial attributes like '_RFBMI5', '_RFHYPE6', '_TOLDHI3', '_CHOLCH3', '_CHCKDNY2', 'SMOKE100', '_RFDRHV7', 'CVDSTRK3', '_MICHD', '_TOTINDA', 'DIFFWALK', '_HLTHPLN', 'MEDCOST1', , we dropped the observations with uninterpretable value (values shows that Respondent either doesn't know the value, or refuse to answer.). Then, change the values to binary of 1 and 0.
- For float and ordina; attributes like 'GNHLTH', 'PHYSHLTH', 'MENTHLTH', 'CHECKUP1', '_AGEG5YR', 'EDUCA', 'INCOME3', we dropped the observations with uninterpretable values. Then, leave the scalable values as it is.

At this point, we renamed the columns' name in accordance with Table I for readability purpose, then export this dataframe into new DataFrame (hereinafter called **"Binary Dataset"**, or **"Imbalanced Dataset"**). We will use these dataset in further statistical analysis and logistic regression model development.

B. Exploratory Data Analysis

In this part, we performed Exploratory Data Analysis (EDA) and found the main characteristics of the Survey's population as follows.

1) Portion of people who have diabetes disease: In accordance with the Survey, the portions of people between the Groups are as shown in Fig. 1.

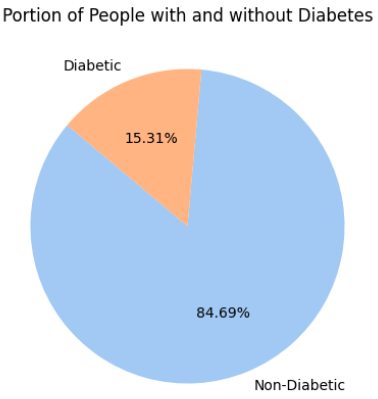

Fig. 1. Pie chart of who has diabetes, and who don't.

2) Age group of survey Respondent: We explore the age group of survey respondent and find the age group and sex as shown in Fig. 2.

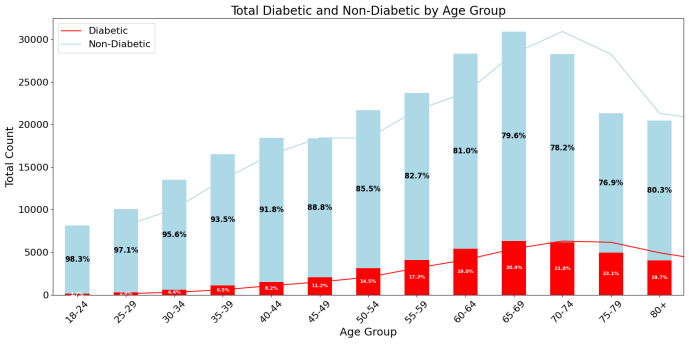

Fig. 2. Total Diabetic and Non-Diabetic by Age Group.

3) Body Mass Index: Body Mass Index is calculated values, which comes from:

$$BMI = \frac{Weight(kg.)}{Height(m.)^2} \quad (1)$$

We explored the data and build up the distribution of respondent's BMI between the diabetes groups (Fig. 3). Then, we found that the distribution is almost in the bell-shaped curve form, but there are some outliers identified in the right-ended side of BMI range. So, we built the box plot to confirm our assumption (Fig. 4)

In accordance with boxplot, we found that there are 6,613 BMI observations (2.5% of our observations) considered as outliers in our datasets. We considered to retain these outliers for our further analysis from the following reasons.

- In accordance with source data documentation (as known as 'codebook'), BMI score was calculated from the weight and height, which is also converted into metric units. So, we believe that there is low risk exposure to mistake due to human errors.

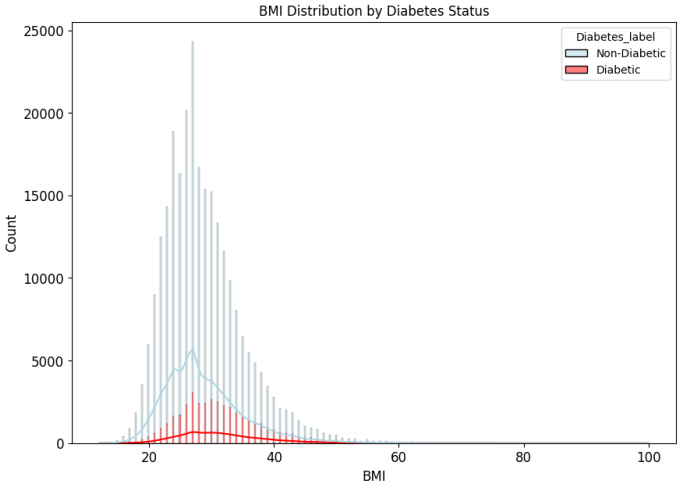

Fig. 3. BMI Distribution by Diabetes Status.

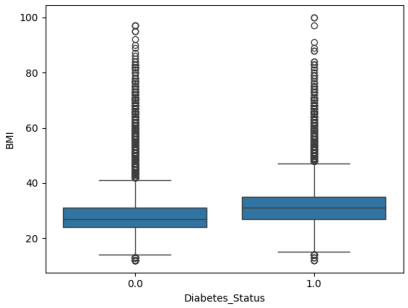

Fig. 4. BMI Boxplot by Diabetes Status.

- We believe that there is potent association between high BMI (Obesity) and diabetes disease.

IV. Statistical Analysis

A. Statistical Testing Assumption

According to Exploratory Data Analysis procedure performed, we need to declare some assumptions before proceed with further statistical analysis.

- The Survey data performed by third parties are accurate to the actual sample.
- The Survey conductor used a non biased sampling method to make a sample.
- All numeric statistical values observed from this survey are normally distributed.

B. Correlation Analysis

We calculated the Pearson correlation value for those selected attributes to find out what are 5 attributes with highest correlation to diabetes disease. As a result, we find that these 5 attributes have the highest correlation coefficient with diabetes.

- General Health
- Hypertension
- High Cholesterol
- Walking/climbing stairs difficulties
- Body Mass Index

Then, we performed statistical testing onto these attributes to conclude if there is significant difference between the Groups. However, since each attributes has their own characteristics in value, so it needs to be performed with different statistical testing method.

C. Chi-Square Analysis

In this section, we performed Chi-Square test onto the attributes with categorical value. We assumed the Null Hypothesis (H_0) that each attribute was not statistically different between the Groups of samples. The statements above could be written in the statistical manner as follows.

$$H_0 : \mu_1 = \mu_2$$
$$\mathrm{H}_1 : \mu_1 \neq \mu_2$$

Assume that we have determined the confidence interval level as 95%. As a consequence, we have a significant value (alpha) by 0.05

1) Hypertension: We performed Chi-Square test to find out if there is difference between the Groups in terms of Hypertension and found the results as shown in Table II.

Statistic	Value
Chi-Square Statistic	15473.21
Degrees of Freedom	1
P-Value	0.000

TABLE II
Chi-Square Test Results: Association between Diabetes Status and Hypertension

We found that p-value is lower than our defined significant value. Therefore, we reject the Null Hypothesis in terms of Hypertension between each group of samples, which means there is difference in Hypertension between the Groups.

2) High Cholesterol: We performed Chi-Square test to find out if there is difference between the Groups in terms of High Cholesterol and found the results as shown in Table III.

Statistic	Value
Chi-Square Statistic	8750.46
Degrees of Freedom	1
P-Value	0.000

TABLE III
Chi-Square Test Results: Association between Diabetes Status and Age Group

We found that p-value is lower than our defined significant value. Therefore, we reject the Null Hypothesis in terms of High Cholesterol between each group of samples, which means there is difference in High Cholesterol level between the groups.

3) Walking/Climbing Stairs Difficulty: We perform Chi-Square test to find out if there is difference between the Groups in terms of walking or climbing upstairs difficulty and found the results as shown in Table IV.

Statistic	Value
Chi-Square Statistic	10371.31
Degrees of Freedom	1
P-Value	0.000

TABLE IV
CHI-SQUARE TEST RESULTS: ASSOCIATION BETWEEN DIABETES STATUS AND DIFFICULTY WALKING

We found that p-value is lower than our defined significant value. Therefore, we reject the Null Hypothesis in terms of walking or climbing upstairs difficulty between each group of samples, which means there is difference in walking or climbing upstairs difficulty between the groups.

D. Independent t-test analysis

Since our sample between the Groups is independent to each other, so we performed independent t-test onto our numerical attributes. We assumed the Null Hypothesis (H_0) as test attributes was not statistically different between each group of samples. Those statements above could be written in the statistical manner as follows.

$$H_0 : \mu_1 = \mu_2$$
$$\mathrm{H}_1 : \mu_1 \neq \mu_2$$

Assume that we determined the confidence interval level as 95%. As a consequence, we have a significant value (alpha) by 0.05

1) General Health: We performed independent t-test between the Groups to find out if there is difference in the general health condition score between the Groups.

Statistic	Value
T-Statistic	137.73
Degrees of Freedom	1
P-Value	0.000

TABLE V
T-TEST RESULTS: GENERAL HEALTH COMPARISON BETWEEN DIABETES AND NON-DIABETES GROUPS

From Table V, we found that p-value is lower than our defined significant value. Therefore, we reject the Null Hypothesis in terms of General Health conditions, which means there is a difference in General Health score between the Groups.

2) Body Mass Index: We performed independent t-test between the Groups to find out if there is difference in the Body Mass Index (BMI) level between the Groups.

Statistic	Value
T-Statistic	99.45
Degrees of Freedom	1
P-Value	0.000

TABLE VI
T-TEST RESULTS: BMI COMPARISON BETWEEN DIABETES AND NON-DIABETES GROUPS

From Table VI, We found that p-value is lower than our defined significant value. Therefore, we reject the Null Hypothesis in terms of Body Mass Index, which means there is a difference in Body Mass Index level between the Groups.

E. Logistic Regression Analysis

1) Imbalanced Data Handling: In accordance with Exploratory Data Analysis procedure performed, we found that there is imbalance in numbers between the Groups (referred to 1. This imbalance could cause model inaccuracy so we need to handling those imbalance as follows.

- We found that 39,787 (16%) respondents have diabetes, while 220,064 (84%) respondent don't have diabetes.
- At this point, we could do either Under-sampling or Over-sampling approach to handle this imbalance. Since we found that our minority (those who have diabetes) still have a large number of observations, we chose to use **Under-sampling** approach to handling this imbalance
- Then, we need to do the sampling of those who don't have diabetes by the same number with who has diabetes disease (39,787 observations). Then build up the new Datasets from those Groups, Hereinafter, we will referred to this DataFrame as **"Balanced Dataset"**.
- The Balanced dataset contains 79,574 observations due to the procedures performed as shown above. Then, we were able to proceed with logistic regression modeling.

However, we developed the logistic regression model using both "Balanced" and "Imbalanced" datasets, then compared the prediction accuracy between each other.

2) Logistic Regression: After preparation of Datasets, we developed the logistic regression model, then evaluated the accuracy to measure the prediction efficiency.

As our dependent variable is the classified attributes, we built up the logistic regression model using scikit-learn python library. We set the regression parameters as follows;

- x (features) = Those 5 attributes with highest correlation to diabetes in accordance with correlation analysis procedure performed.
- y (values) = Diabetes status (Have you ever told that you have diabetes?, 1 = yes, 0 = no).
- Testing data portion = 30% of all observation in each Dataset.

Models built from	Accuracy (%)
Balanced Dataset	70.47%
Imbalanced Dataset	84.97%

TABLE VII
ACCURACY COMPARISON BETWEEN LOGISTIC REGRESSION MODEL

As mentioned earlier, we built up 2 Logistic Regression models from both Balanced and Imbalanced dataset. Then we will compare the accuracy between each other.

3) Results: After developing the model, then we compared the accuracy of each regression model. The accuracy results is shown in Table VII.

Then, we could imply that the Imbalanced Datasets provides more prediction accuracy. We assumed that those accuracy comes from the larger size of observations used in model training.

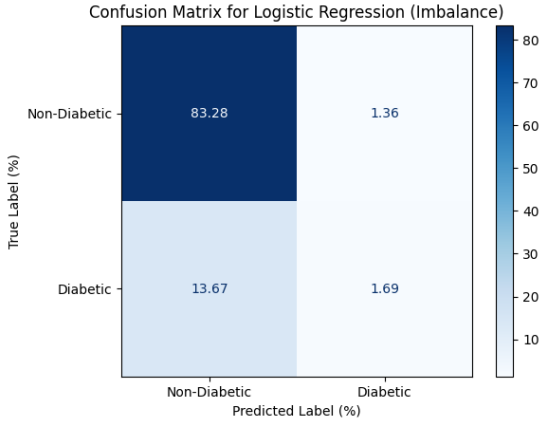

Fig. 5. Confusion Matrix for Logistic Regression (Imbalance).

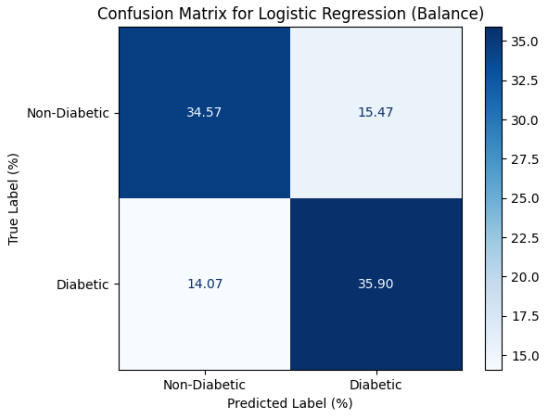

Fig. 6. Confusion Matrix for Logistic Regression (Balance).

ACKNOWLEDGMENT

From our studies, we found that **Body Mass Index ("BMI"), Walking difficulties, Hypertension, Cholesterol, and General Health Conditions** are 5 factors with most correlation with diabetes. We performed statistical testing and could confirm that there are significant differences in our selected factors between the Groups. Then, we built up the logistic regression model to predict the risk of diabetes based on those attributes. The developed logistic regression model has satisfying prediction accuracy.

REFERENCES

[1] Diabetes Data Analysis. https://www.kaggle.com/code/rahul713/diabetes-data-analysis

[2] Diabetes Mellitus Detection. https://www.kaggle.com/code/muskansah/diabetes-mellitus-detection

[3] Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

[4] Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. [https://www.cdc.gov/pcd/issues/2019/19_0109.htm]