**Documentation**

**Diabetes Health Indicators**

## **About Dataset**

### **Context**

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes.

### **Content**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.

## **Purpose**

The purpose of this code notebook is to clean BRFSS data into a useable format for machine learning alogrithms. The dataset originally has 330 features (columns), but based on diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions, only select features are included in this analysis.

#### **Important Risk Factors**

Research in the field has identified the following as **important risk factors** for diabetes and other chronic illnesses like heart disease (not in strict order of importance):

* blood pressure (high)
* cholesterol (high)
* smoking
* diabetes
* obesity
* age
* sex
* race
* diet
* exercise
* alcohol consumption
* BMI
* Household Income
* Marital Status
* Sleep
* Time since last checkup
* Education
* Health care coverage
* Mental Health

**Independent Variables:**

**High Blood Pressure**

* Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional --> \_RFHYPE5

**High Cholesterol**

* Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> TOLDHI2
* Cholesterol check within past five years --> \_CHOLCHK

**BMI**

* Body Mass Index (BMI) --> \_BMI5

**Smoking**

* Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] --> SMOKE100

**Other Chronic Health Conditions**

* (Ever told) you had a stroke. --> CVDSTRK3
* Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) --> \_MICHD

**Physical Activity**

* Adults who reported doing physical activity or exercise during the past 30 days other than their regular job --> \_TOTINDA

**Diet**

* Consume Fruit 1 or more times per day --> \_FRTLT1
* Consume Vegetables 1 or more times per day --> \_VEGLT1

**Alcohol Consumption**

* Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) --> \_RFDRHV5

**Health Care**

* Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? --> HLTHPLN1
* Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? --> MEDCOST

**Health General and Mental Health**

* Would you say that in general your health is: --> GENHLTH
* Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? --> MENTHLTH
* Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? --> PHYSHLTH
* Do you have serious difficulty walking or climbing stairs? --> DIFFWALK

**DATA ANALYSIS REPORT**

**Problem statement: Predict whether the patient is having diabetes or not .**

1. **Imputing missing values**

* Categorical = Mode
* Real Values = Mean, Interpolation, BFill, FFill etc.

1. **Detecting outliers**

* Z-Score method
* Boxplot
* Scatter plot

1. **Treatment for outliers**

* IQR Method
* Z-Score method
* K-NN Method

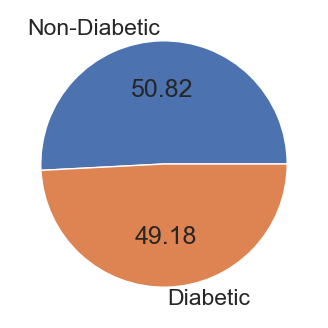
1. **Visualization**
2. **Univariate Analysis**

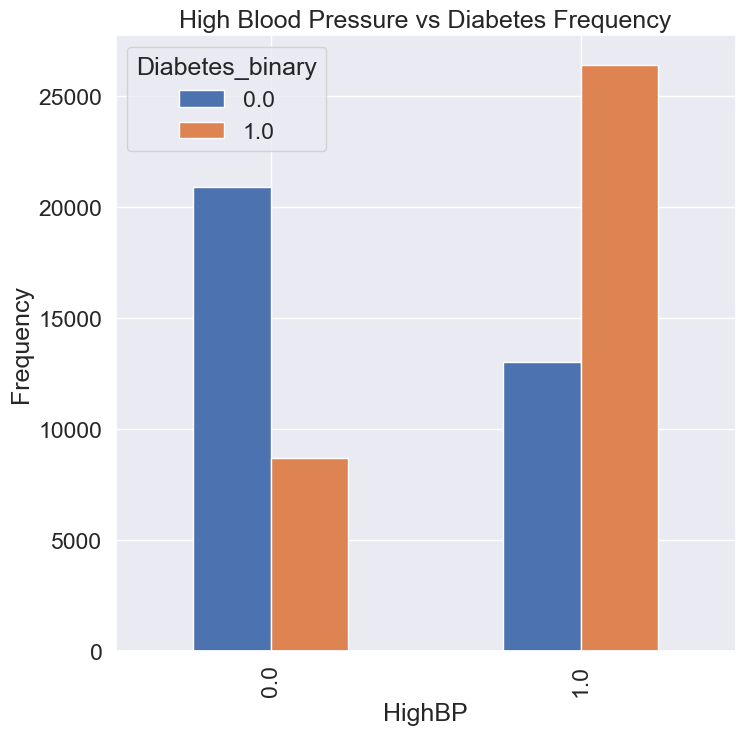
* **Numerical**
* Histogram plot
* Density plot
* **Categorical**
* Count plot

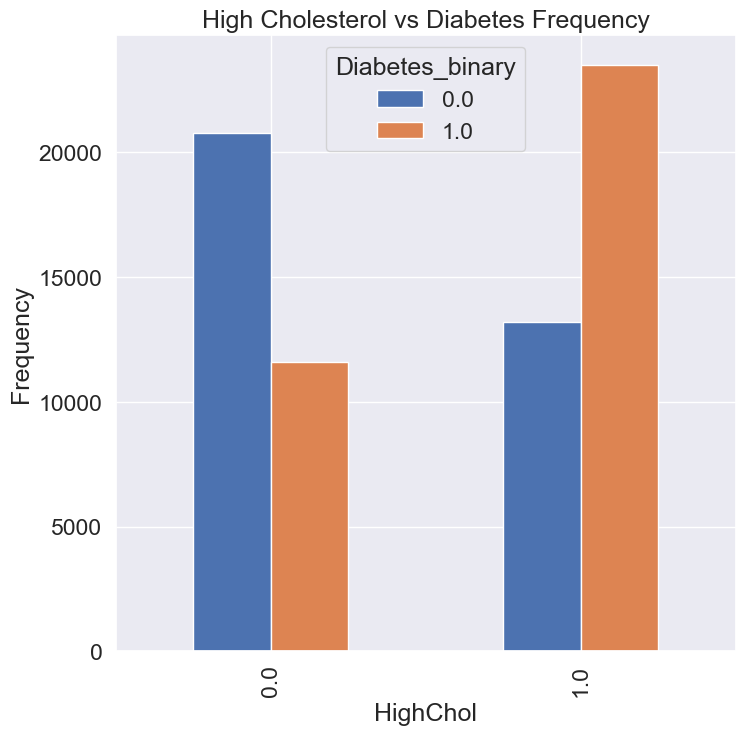
1. **Bivariate Analysis**

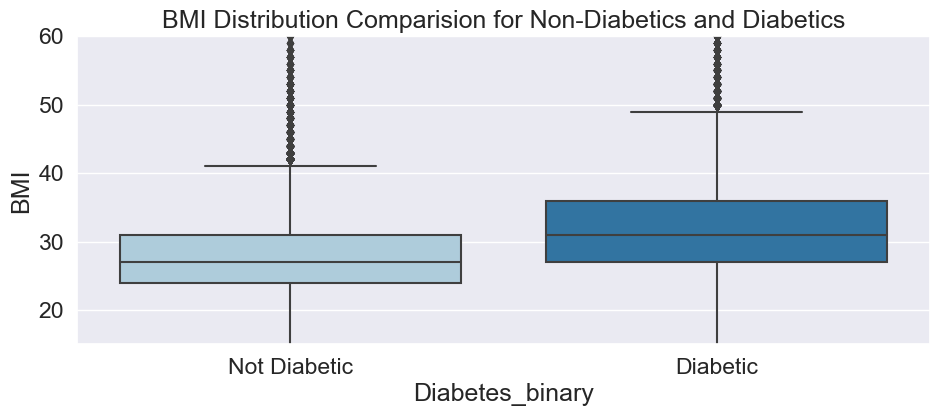
* **Numerical vs numerical**
* Scatter plot
* Histogram plot
* **Numerical vs Categorical**
* Boxplot
* **Categorical vs Categorical**
* Stacked bar

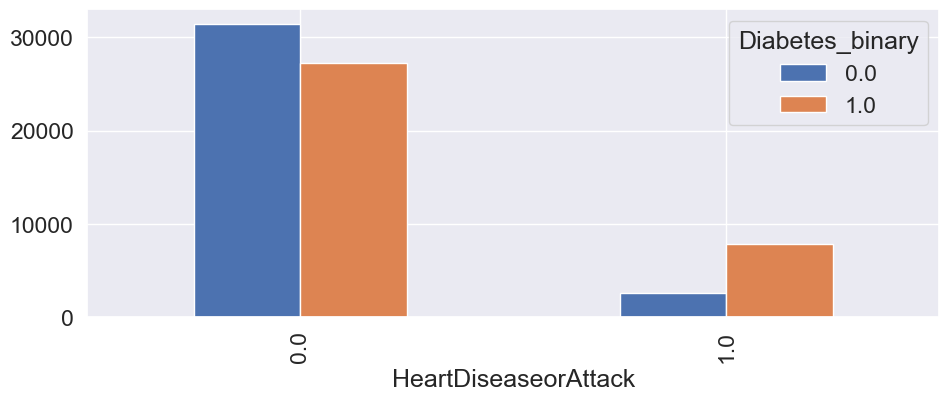
1. **Multi-Variate Analysis**

* **Heat Map**
* **Pair Plot**
* According to data, Pie\_chart tells that 50% people are Non-Diabetic and 50% people are Diabetic.



* According to Bar\_Plot, we can say that people who are having highBP are divided into two categories, people who are having Diabetes in highBP are more than people who doesn’t have Diabetes in highBP.
* By seeing this Bar\_plot we can say that people who are High\_Cholesterol is divided into two categories who are having diabetes and non\_diabetes, so people who are suffuring from Diabetes are more in who have High\_Cholestrol.



* According to Box\_Plot ,those who have high BMI they are highly Diabetic.
*  By seeing this plot we can say that people who are having HeartDisease are also suffering from Diabetes but less in count.

**Model Building:**

**Data Cleaning Report:**

After the completion of data cleaning, Now we are ready to send our data to Neural Network.

But, before sending our data to the neural network, we must observe our data, whether it is properly scaled or not, because our numerical data should be in the same scale then only our neural network properly works and training could be speed.

So, we must Rescale our data by using **Normalisation techniques or Standardisation techniques**, so the values will lie between 0 to 1.

This technique helps us to train our model very fast and our optimizers will run very fast.

**Model Report:**

To create the ANN Model we used the Keras Tuner. Where keras tuner can help us to return the best model corresponding to No.of.hidden layers and no.of.Neurons and which activation function to be used, which optimizer and weight initializers etc.

Our target variable is binary classification, so we used the loss function as “binary\_crossentropy”.

**Model Accuracy:**

Finally, my model accuracy is 74%.

The total hidden layers are 7, activation is Relu, optimizer is RMSPROP.