**CS608 – Advance Techniques in Data Science**

**Final Project Submission**

Submitted to: **Professor Dr. Muhammad Awais Hassan**

Submitted by: **Rizwan Riaz Mir**

Roll No: **2023–Phd–CS-8**



**Department of Computer Science,**

**University of Engineering and Technology,**

**Lahore.**

Table of Contents

[1 Problem Identification and Definition 4](#_Toc186204113)

[**1.1** **Problem Title** 4](#_Toc186204114)

[**1.2** **Problem Background** 4](#_Toc186204115)

[**1.3** **Problem Statement** 4](#_Toc186204116)

[**1.4** **Problem Significance** 4](#_Toc186204117)

[**1.5** **Objectives** 4](#_Toc186204118)

[**1.6** **Potential Impacts of Studies** 5](#_Toc186204119)

[2 Drafting the Right Questions and their Relevance 5](#_Toc186204120)

[**2.1** **Is there a significant relationship between BMI, cholesterol, blood pressure and the likelihood of diabetes?** 5](#_Toc186204121)

[**2.2** **Is there a pattern between age, gender, education, income levels, and the prevalence of diabetes?** 5](#_Toc186204122)

[**2.3** **Does the intake of fruits and vegetables along with being physically active reduces the risk of diabetes?** 5](#_Toc186204123)

[**2.4** **Is there any significant correlation between diabetes, cardiovascular disease, alcohol and smoking?** 5](#_Toc186204124)

[**2.5** **How do mental and physical health directly correlate with worsening of diabetes?** 6](#_Toc186204125)

[3 Data Source and Data Description 6](#_Toc186204126)

[**3.1** **Data Source** 6](#_Toc186204127)

[**3.2** **Description of Dataset Columns** 6](#_Toc186204128)

[4 Data Wrangling and Cleaning 8](#_Toc186204129)

[**4.1** **Handling Missing Values** 8](#_Toc186204130)

[**4.2** **Convert Datatypes in Dataset** 9](#_Toc186204131)

[**4.3** **Feature Engineering for Numerical Features** 10](#_Toc186204132)

[**4.3.1** **High Blood Pressure Column** 10](#_Toc186204133)

[**4.3.2** **High Cholesterol Column** 11](#_Toc186204134)

[**4.3.3** **Age Column** 11](#_Toc186204135)

[**4.3.4** **Income Column** 11](#_Toc186204136)

[**4.4** **Feature Engineering for Categorical Features** 13](#_Toc186204137)

[**4.4.1** **Diabetes Column** 13](#_Toc186204138)

[**4.4.2** **General Health Column** 13](#_Toc186204139)

[**4.4.3** **Difficult in walking Column** 13](#_Toc186204140)

[**4.4.4** **Gender Column** 14](#_Toc186204141)

[**4.4.5** **Education Column** 14](#_Toc186204142)

[**4.5** **Detecting and Addressing Outliers** 15](#_Toc186204143)

[5 Exploratory Data Analysis (EDA) 17](#_Toc186204144)

[**5.1** **Correlations between features / columns** 17](#_Toc186204145)

[**5.2** **Finding Answers to the Questions** 18](#_Toc186204146)

[**5.2.1** **Is there any correlation between Diabetes and Blood Pressure?** 18](#_Toc186204147)

[**5.2.2** **Is there any correlation between Diabetes and Cholesterol?** 19](#_Toc186204148)

[**5.2.3** **If a patient has high blood pressure and high cholesterol as well then what is the correlation of them with diabetes?** 20](#_Toc186204149)

[**5.2.4** **What is a chance of being diabetic if a person has an abnormal BMI range?** 20](#_Toc186204150)

[**5.2.5** **What is the correlation between diabetes, smoking and alcohol consumption?** 22](#_Toc186204151)

[**5.2.6** **What is the correlation between diabetes, stroke and heart diseases?** 22](#_Toc186204152)

[**5.2.7** **What are effects of life style activities on Diabetes?** 24](#_Toc186204153)

[**5.2.8** **What are the correlation of diabetes with general, mental and physical health conditions?** 26](#_Toc186204154)

[**5.2.9** **What are the correlation between Diabetes and Gender, Age, Income and Education?** 26](#_Toc186204155)

[**5.2.10** **Separation of Major and Least Effective Features on the basis of EDA?** 29](#_Toc186204156)

[6 Predictive Analysis 30](#_Toc186204157)

[**6.1** **Predictive Analysis using Random Forest Classifier** 30](#_Toc186204158)

[**6.2** **Predictive Analysis using Logistic Regression Classifier** 30](#_Toc186204159)

[**6.3** **Predictive Analysis using Decision Tree Classifier** 32](#_Toc186204160)

[**6.4** **Comparative Analysis for Used Classifiers to detect Diabetes** 33](#_Toc186204161)

# Problem Identification and Definition

## **Problem Title**

Data-Driven Insights for Various Factors affecting Diabetes

## **Problem Background**

Diabetes is known as one of the oldest chronic diseases first reported 3000 years back in an Egyptian manuscript (81-133AD). Diabetes Mellitus (DM) is classified into four types that are Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM), Pre-Diabetes, and Gestational Diabetes. Most common among the four types and approximately 90 - 95% of the total diabetes mellitus cases are of T2DM. When a human body is not able to use insulin well within a body due to impaired insulin secretion, insulin resistance, or both, then the condition is referred to as T2DM. Pre-Diabetes and T2DM refer to the malfunctioning of Beta (β) cells of the pancreas. In general, Diabetes is referred as Pre-Diabetes (Early Diabetes Symptoms) and Type 2 Diabetes (T2DM).

## **Problem Statement**

Diabetes is a rapidly growing global health concern that affects millions of people worldwide. If it left unattended and unmanaged, vital organs and senses (Kidney, Heart, Vision, Neuropathy) failure is done by the chronic conditions that put extra burden on national healthcare systems and financial conditions of patient and his family. In 2017, the average annual cost for an individual in the USA with diabetes was $16,750. The increasing prevalence of diabetes and its associated complications and cost require a comprehensive understanding of factors causing and affecting diabetes.

## **Problem Significance**

In 2022, 537 Million of the world's population approximately was living with diabetes from which 360 Million people were urban This is supposed to increase to 783.2 Million by 2045 having 600 Million people urbanized. In 2021, the mortality of 6.7 Million adults (Age: 20 - 79) was estimated as a result of diabetes or its complications. Pakistan became third in the list for diabetes from age 20 to 79 was 33 Million followed by China and India in numbers that will grow to 62.2 Million population diagnosed with diabetes till 2045.

## **Objectives**

The primary objective of this study is to uncover meaningful patterns and trends in diabetes related dataset. The influence of the study is to focus on demographic, lifestyles and clinical factors available in dataset to understand the linkages between these factors and their impact on diabetes risk calculations and predictions.

Additionally, a predictions modeling techniques are used to forecast the disease outcomes to foreseen diabetes management and intervention proactively.

## **Potential Impacts of Studies**

The study will improve understanding of diabetes risk factors by analyzing demographic and clinical trends. It will help to predict the disease severity in early stages that will prevent the leading complications of the diabetes. That will ultimately help to reduce the burden on national healthcare systems and economic pressures on governments, patients and their families as well.

# Drafting the Right Questions and their Relevance

In the "Asking the Right Questions" section, it's essential to identify the critical factors influencing diabetes and related health conditions. This study aims to uncover significant patterns, correlations, and risk factors by exploring how variables like high blood pressure, cholesterol levels, BMI, lifestyle habits, and demographics interact. These insights will guide data-driven recommendations for improved disease prediction, management and prevention.

## **Is there a significant relationship between BMI, cholesterol, blood pressure and the likelihood of diabetes?**

Examining the link between BMI, high cholesterol, and diabetes helps identify key health metrics that contribute to the disease. Recognizing these relationships can improve screening and preventive measures, leading to better management of diabetes risk factors.

## **Is there a pattern between age, gender, education, income levels, and the prevalence of diabetes?**

Understanding how demographic factors like age, gender, education, and income levels correlate with the prevalence of diabetes can help identify vulnerable populations. This insight can guide targeted interventions and policies to address and prevent diabetes more effectively.

## **Does the intake of fruits and vegetables along with being physically active reduce the risk of diabetes?**

Determining the impact of a healthy diet and physical activity on diabetes risk can highlight the importance of lifestyle modifications. This can encourage public health initiatives aimed at promoting healthier behaviors to prevent diabetes.

## **Is there any significant correlation between diabetes, cardiovascular disease, alcohol and smoking?**

Investigating the associations between diabetes, cardiovascular disease, alcohol consumption, and smoking can reveal interconnected risk factors. Understanding these correlations can support comprehensive health strategies that address multiple chronic conditions simultaneously.

## **How do mental and physical health directly correlate with worsening of diabetes?**

Exploring the direct relationship between mental and physical health and diabetes progression underscores the holistic nature of disease management. It can lead to more integrated healthcare approaches that address both psychological and physical aspects of diabetes care.

# Data Source and Data Description

## **Data Source**

We have selected the CDC Diabetes Health Indicators dataset from the UCI Machine Learning Repository for its comprehensive coverage of health-related factors associated with diabetes. This dataset includes a rich variety of attributes such as BMI, physical activity, smoking habits, and more, across a broad demographic spectrum. The extensive nature of the data allows for in-depth analysis and the identification of key patterns and correlations that are crucial for understanding and managing diabetes effectively. This robust dataset serves as a valuable resource for developing data-driven strategies to improve public health outcomes related to diabetes.

Table - Description of Diabetes Dataset

|  |  |
| --- | --- |
| **Category** | **Description** |
| Data Source | CDC Diabetes Health Indicators |
| Availability | September 25, 2023 |
| Data Link | <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators> |
| Dataset Population | United States of America |
| Dataset Characteristics | Tabular |
| Area of Subject | Healthcare and Medicine |
| Instances | 253680 |
| Features | 21 |
| Missing Values | No |

## **Description of Dataset Columns**

Table - Descriptions of Columns in Dataset

| **Variable Name** | **Role** | **Type** | **Demographic** | **Description** | **Missing Values** |
| --- | --- | --- | --- | --- | --- |
| Diabetes\_binary | Target | Binary |  | 0 = no diabetes 1 = prediabetes or diabetes | no |
| HighBP | Feature | Binary |  | 0 = no high BP 1 = high BP | no |
| HighChol | Feature | Binary |  | 0 = no high cholesterol 1 = high cholesterol | no |
| CholCheck | Feature | Binary |  | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years | no |
| BMI | Feature | Integer |  | Body Mass Index | no |
| Smoker | Feature | Binary |  | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes | no |
| Stroke | Feature | Binary |  | (Ever told) you had a stroke. 0 = no 1 = yes | no |
| HeartDiseaseorAttack | Feature | Binary |  | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes | no |
| PhysActivity | Feature | Binary |  | physical activity in past 30 days - not including job 0 = no 1 = yes | no |
| Fruits | Feature | Binary |  | Consume Fruit 1 or more times per day 0 = no 1 = yes | no |
| Veggies | Feature | Binary |  | Consume Vegetables 1 or more times per day 0 = no 1 = yes | no |
| HvyAlcoholConsump | Feature | Binary |  | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes | no |
| AnyHealthcare | Feature | Binary |  | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes | no |
| NoDocbcCost | Feature | Binary |  | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes | no |
| GenHlth | Feature | Integer |  | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor | no |
| MentHlth | Feature | Integer |  | 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? scale 1-30 days | no |
| PhysHlth | Feature | Integer |  | 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? scale 1-30 days | no |
| DiffWalk | Feature | Binary |  | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes | no |
| Sex | Feature | Binary | Sex | 0 = female 1 = male | no |
| Age | Feature | Integer | Age | 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older | no |
| Education | Feature | Integer | Education Level | Education level (EDUCA see codebook) 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) | no |
| Income | Feature | Integer | Income | Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more | no |

# Data Wrangling and Cleaning

## **Handling Missing Values**

There are no missing values in the data as claimed by the data source. It is cross verified by isnull() function and results are provided below:

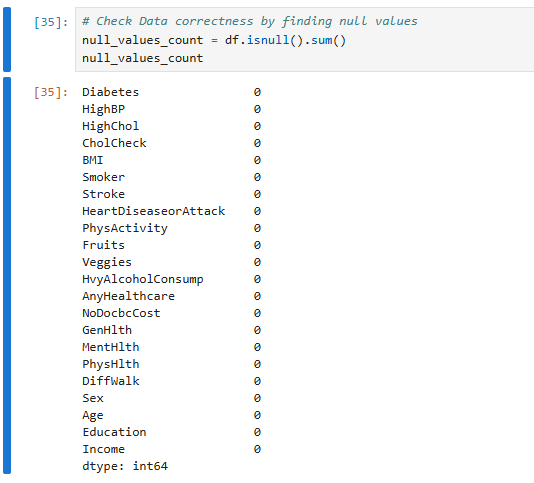


Figure - Counting Null values in the Dataset

## **Convert Datatypes in Dataset**

To perform exploratory data analysis (EDA), we convert the related columns from the encoded format to the normal values as described in section 3.2 according to the guidelines and description of dataset available on source webpage. A column that describes the binary values to describe the need of doctor and not visit due to money is dropped from the data frame and python command is shown in figure 2. While, new columns and their data types are shown in figure 3.

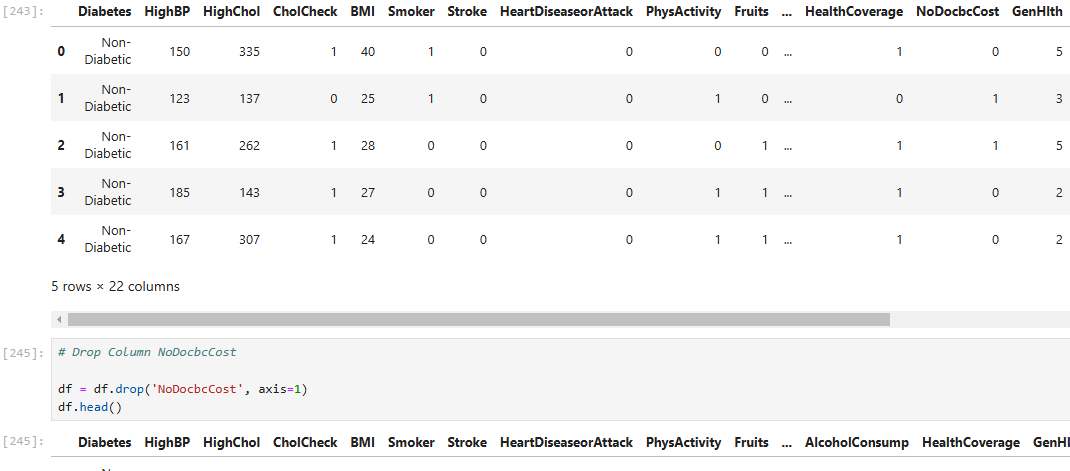


Figure - Drop the Column NoDocbcCost

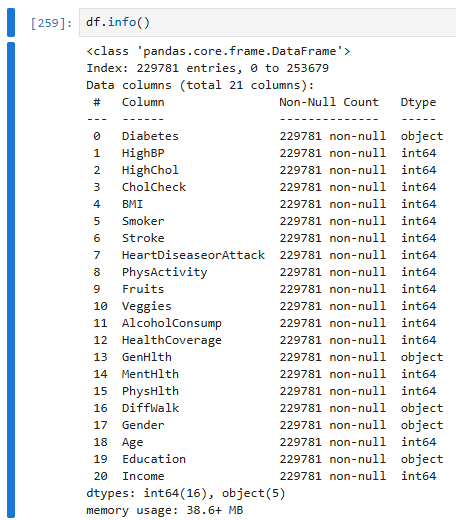


Figure - Information of Column Datatypes after Conversion

## **Feature Engineering for Numerical Features**

Feature Engineering for numerical features are performed one four features / columns. Details for each feature conversion is described below.

### **High Blood Pressure Column**

In actual dataset downloaded from source, the reference values of high blood pressure were encoded into zero and one. Zero means normal blood pressure while one means high blood pressure.

By using random function, the zero values in blood pressure column is replaced by normal blood pressure values ranges from 110 to 130 mmHg. And ones in the column are replaced by high blood pressure values ranges from 131 to 220 mmHg. Screenshots of python commands are displayed below to see the HighBP column conversion to realistic numerical values.

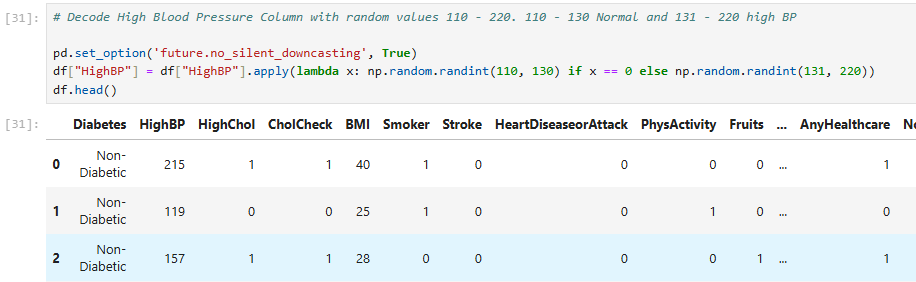


Figure - Conversion of HighBP column to real numerical numbers

### **High Cholesterol Column**

In actual dataset downloaded from source, the reference values of high Cholesterol were encoded into zero and one. Zero means normal cholesterol while one means high values of cholesterol.

By using random function, the zero values in cholesterol column is replaced by normal cholesterol values recommended by WHO ranges from 120 to 200 mg/dL. And ones in the column are replaced by high real world values ranges from 201 to 400 mg/dL. Screenshots of python commands are displayed below to see the HighChol column conversion to realistic numerical values.

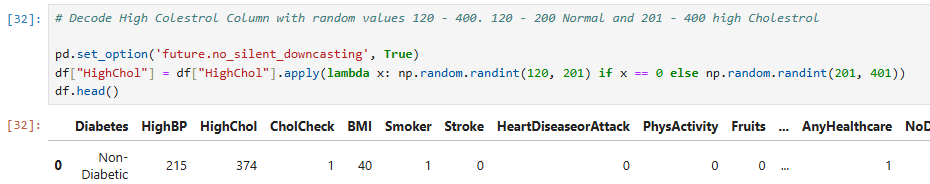


Figure - Conversion of HighChol column to real numerical numbers

### **Age Column**

According to the description of dataset in Age column, the age is converted from numeric category number to the age resides within the category. Age number is provided in the column with the help of random function covering the age. A function named as generate\_random\_age and its criteria is mentioned in the screenshot below in figure 6.

### **Income Column**

As in age column, the income column is again decoded into real salary numbers as per description provided in Table 2 in section 3.2. The salary number is randomly generated with the help of category description to generate a real number for salary against the category from 1 – 8. Figure 7 is the screenshot of code snippet of Python to generate salary in numbers.



Figure - Function to generate a numerical age value against age category

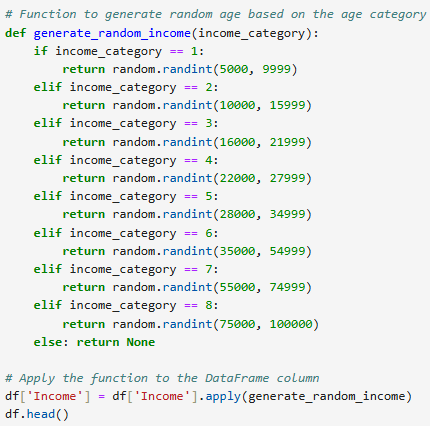


Figure - Function to generate a numerical salary value in USD against income category

## **Feature Engineering for Categorical Features**

### **Diabetes Column**

In diabetes column, values in original dataset was in range from zero to two (0 – 2). Zero code is assigned to patient having no diabetes, while one for pre-diabetic patient and two is for diabetic or T2DM patient. So, except code, a description is added for each value to increase understand of values before EDA. Figure 8 shows the python code to show the conversion of codes to description.

### **General Health Column**

General health is described in encoded form from 1 to 5 values that are also converted to description of the column values to increase understandability while performing EDA. Figure 9 shows the conversion from encode values to descriptive values.

### **Difficult in walking Column**

This column describes the patient condition that faced difficulty to walk or climb stairs. Zero value is assigned if no difficulty was faced by patient and one for patients facing difficulty while climbing stairs or walking. Python code to convert the numerical encoding to categorical value is shown in figure 10.

### **Gender Column**

Column name is changed from Sex to Gender while numerical values of zero and ones are converted to female (for zeros) and male (against ones). Figure 11 is describing the conversion for the Sex Column.

### **Education Column**

Education in the original dataset is also assigned by the category number that are translated to the actual education of the patient from numerical to values according to the description of column in Table 2. In figure 12, a python code is shown to describe the conversion according to description in Table 2 present in section 3.2.

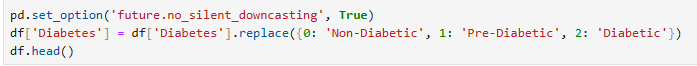


Figure - Convert Diabetes numeric values to equivalent Categorical values

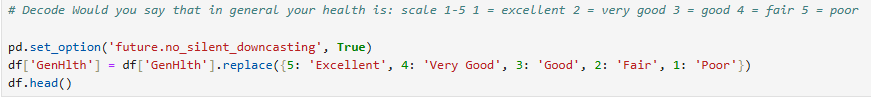


Figure - Convert GenHlth numeric values to equivalent Categorical values

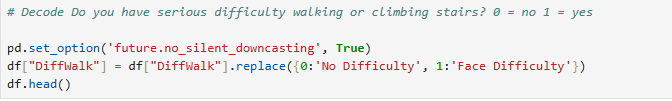


Figure - Convert DiffWalk numeric values to equivalent Categorical values

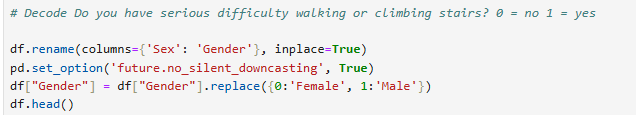


Figure – Change Column name and convert Gender numeric values to equivalent Categorical values

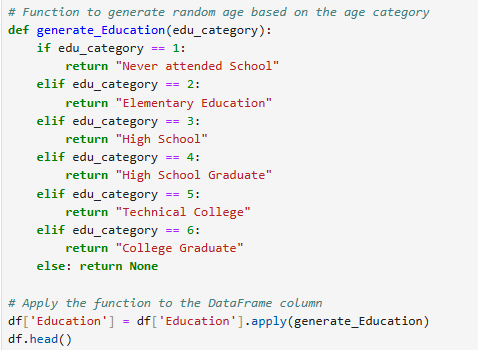


Figure - Convert Education numeric values to equivalent values

## **Detecting and Addressing Outliers**

To detect outliers, we are choosing the columns with continuous values in them. To differentiate between discrete and continuous columns, we use histograms as shown in figure 13.

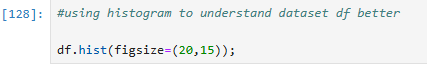


Figure - Checking for Continuous and Discrete Values in Dataset Columns

The output of the histogram command is provided below in figure 14 and 15.

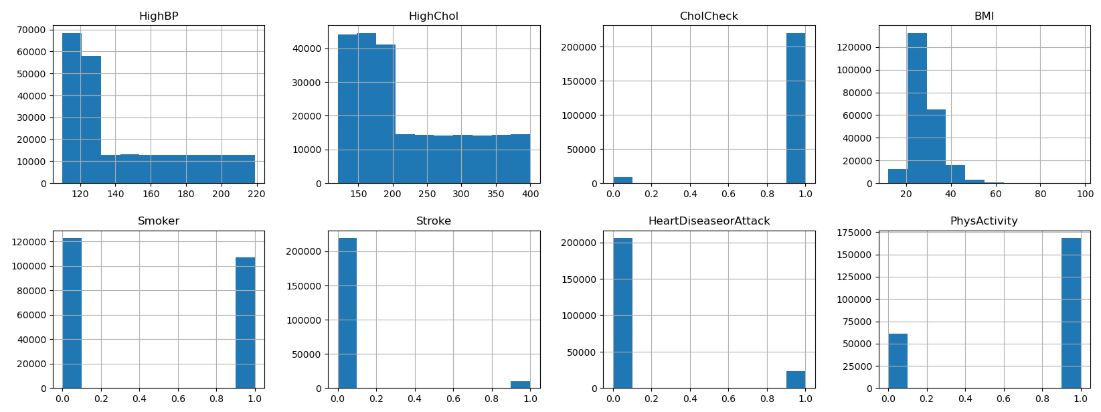


Figure - Result 1 of histograms

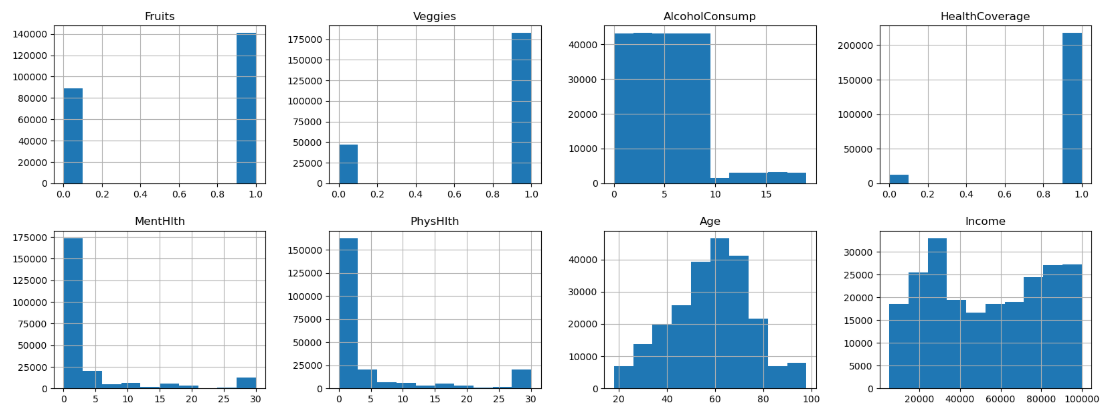


Figure - Result 2 of histograms

So, as we visualize the output of histogram and we depict that columns HighBP, HighChol, BMI, AlcoholConsump, MentHlth, PhysHlth, Age and Income are continuous columns. And for continuous columns, we have to find the outliers within these columns having continuous values in them.

For detecting outliers within continuous columns mentioned above, is performed with the following python commands to draw boxplots as shown in figure 16. And results of boxplots are shown in figure 17, 18 and 19.



Figure - Command to draw boxplots

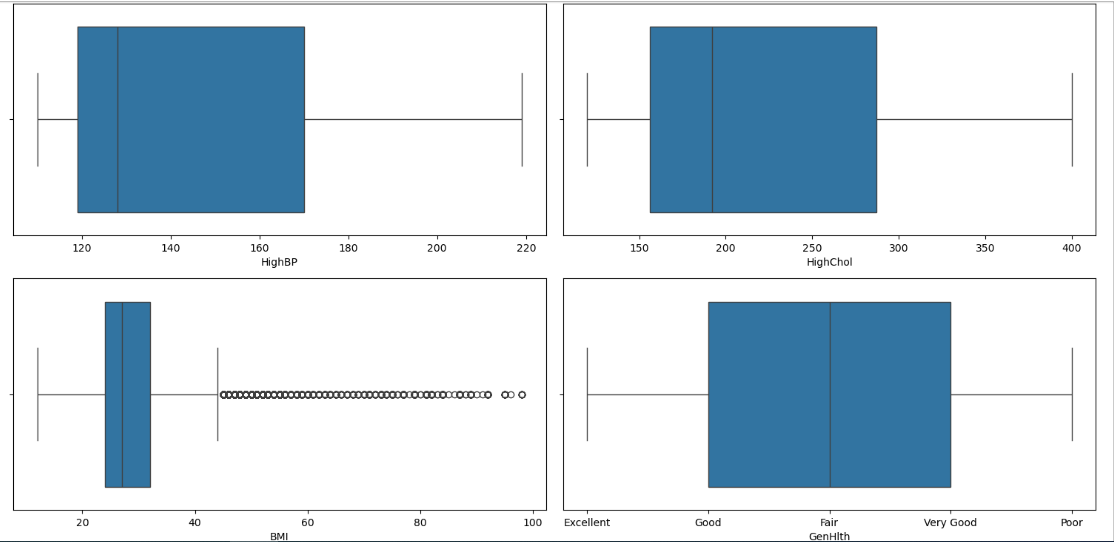


Figure - Boxplots to find outliers - a

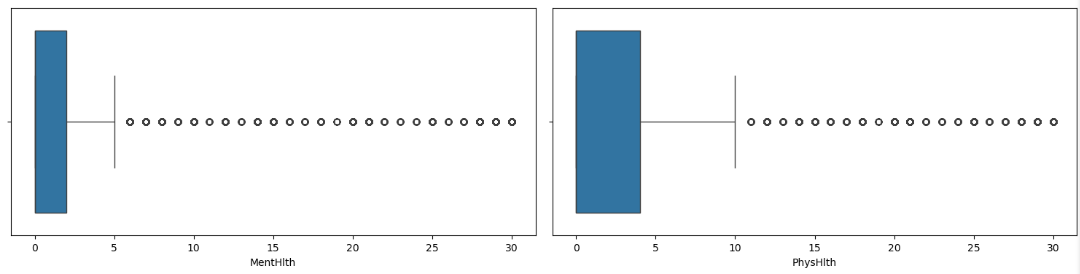


Figure - Boxplots to find outliers - b

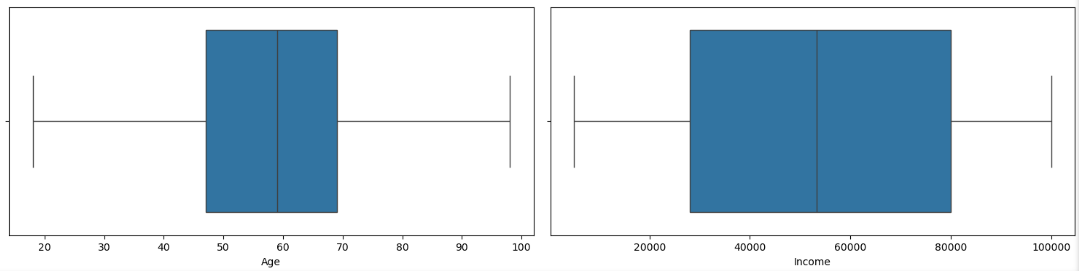


Figure - Boxplots to find outliers – c

At the end, we don’t have outliers present in our continuous columns as it is evident from the visualization we used. Columns MentHlth and PhysHlth have values for the number of days a patient feels problems in their mental and physical health that needs to be addressed so we need that values and does not consider the segregated data as outliers.

# Exploratory Data Analysis (EDA)

## **Correlations between features / columns**

Correlations between different columns are find by the help of heat map and shown in figure 20 that explains relations that are positive and negative in nature.

Correlation heat map show relation between columns:

(GenHlth, PhysHlth), (PhysHlth, DiffWalk), (GenHlth, DiffWalk) are highly correlated with each other that results in a positive relation between features. While, (GenHlth, Income), (DiffWalk, Income) are highly correlated with each other but they are not logically related to each other so these two affiliations or relations are considered to be as Negative relation between features and these negative relationships are ignored while performing EDA.

## 

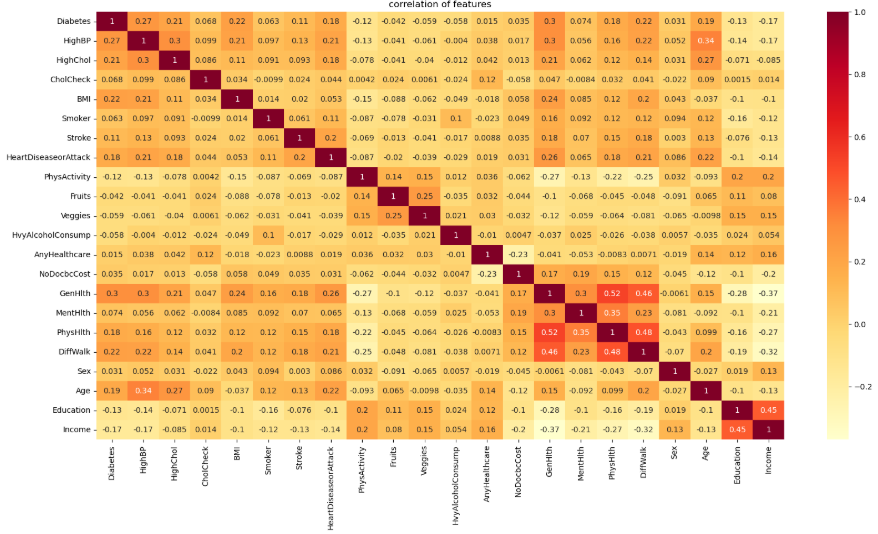


Figure - Correlations between features

## **Finding Answers to the Questions**

### **Is there any correlation between Diabetes and Blood Pressure?**

**Result:** According to the data and its analysis, there’s a strong correlation between Diabetes and Blood pressure. When a person has high blood pressure ranges from (higher than 130 mmHg) then the occurrence of diabetes in a patient is also increased. The result is calculated and analyzed by the data as shown in figure 21 and 22.

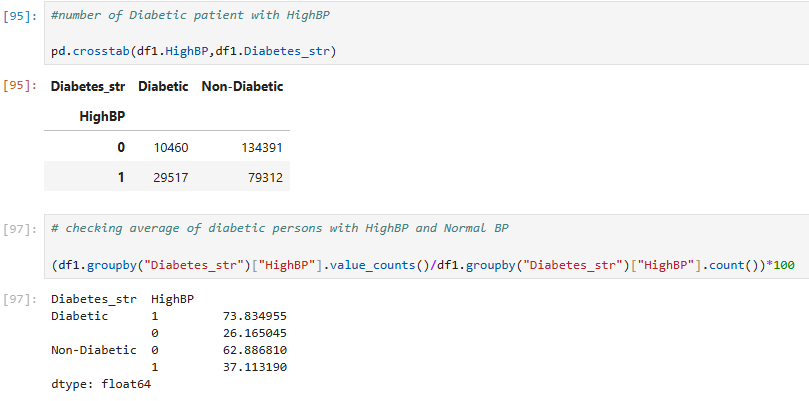


Figure - Corelation between Diabetes and High BP

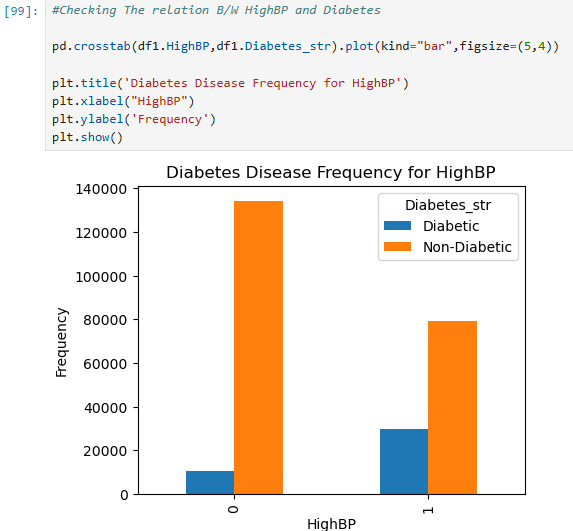


Figure - Correlation between Diabetes and High BP

### **Is there any correlation between Diabetes and Cholesterol?**

**Result:** According to the data and its analysis, there’s a strong correlation between Diabetes and Cholesterol. When a person has high cholesterol in blood ranges from (higher than 200 mmd/L) then the occurrence of diabetes in a patient is also increased. The result is calculated and analyzed by the data as shown in figure 23.

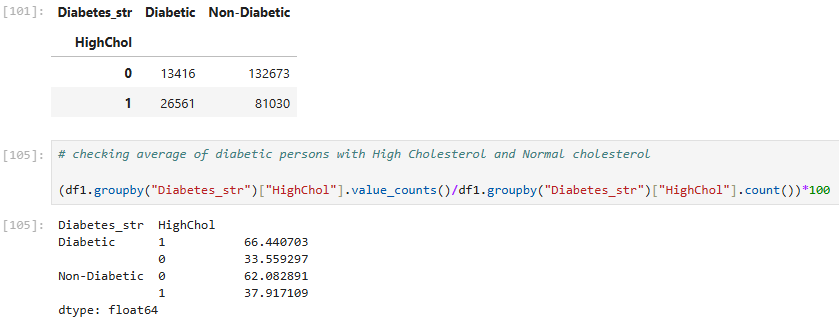


Figure - Correlation between Diabetes and High Cholesterol

### **If a patient has high blood pressure and high cholesterol as well then what is the correlation of them with diabetes?**

**Result:** This analysis suggests that having both high blood pressure (HighBP) and high cholesterol (HighChol) simultaneously can raise the likelihood of developing diabetes, indicating a combined effect of these conditions on diabetes risk. The analysis done is shown in figure 24 and 25.

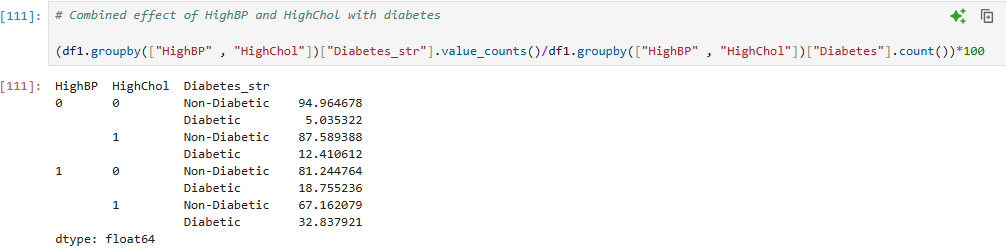


Figure - Occurrence of HighBP and HighChol along with Diabetes

### **What is a chance of being diabetic if a person has an abnormal BMI range?**

**Result:** The data indicates a clear correlation between BMI categories and the prevalence of diabetes. Among individuals with a normal BMI, only 6.68% are diabetic, suggesting that a healthy BMI is associated with a lower likelihood of diabetes. However, the percentage of diabetic patients increases significantly with higher BMI categories: 20.65% of overweight individuals and 28.14% of those classified as obese are diabetic. This trend highlights the strong relationship between higher BMI and an increased risk of diabetes, underscoring the importance of maintaining a healthy weight for diabetes prevention. The analysis results are shown in figure 26.

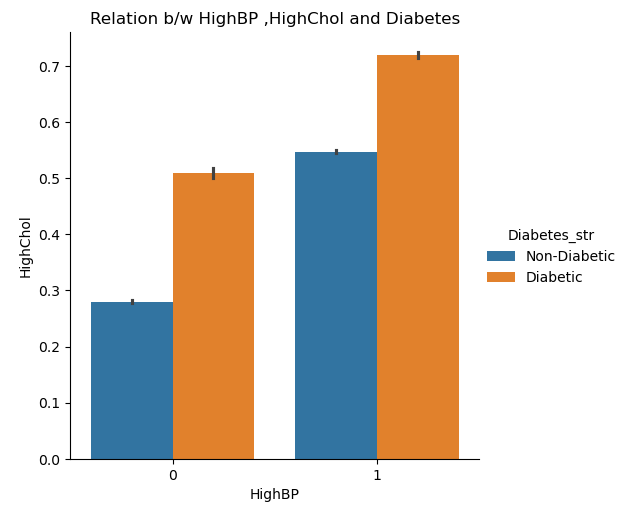


Figure -Correlation of HighBP and HighChol with Diabetes

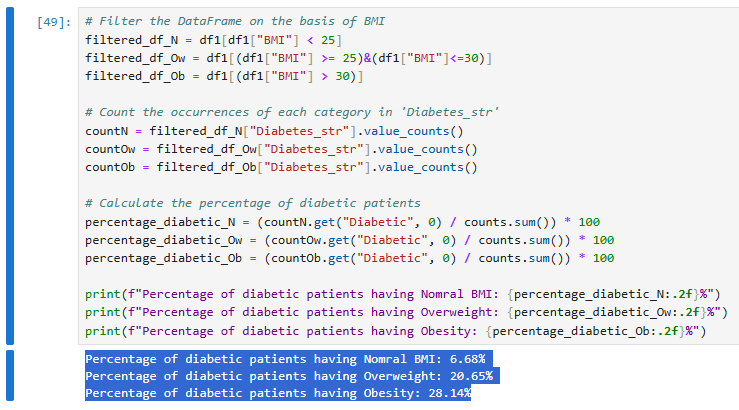


Figure - Correlation between BMI and Diabetes

### **What is the correlation between diabetes, smoking and alcohol consumption?**

**Result:** The results highlight the relationships between smoking, heavy alcohol consumption, and diabetes prevalence. Among non-smokers, those who do not consume alcohol heavily have a lower prevalence of diabetes (14.03%) compared to their non-diabetic counterparts (85.97%). Non-smokers who engage in heavy alcohol consumption have an even lower diabetes prevalence of 5.33%, with 94.67% being non-diabetic. Among smokers, the prevalence of diabetes increases. Smokers who do not consume alcohol heavily have a diabetes rate of 19.22%, while 80.78% are non-diabetic. Interestingly, smokers who engage in heavy alcohol consumption show a lower diabetes prevalence (8.35%) compared to non-smokers but still have a majority (91.65%) who are non-diabetic. These trends suggest complex interactions between smoking, alcohol consumption, and diabetes risk, potentially influenced by lifestyle or health factors. The analysis results are shown in figure 27.

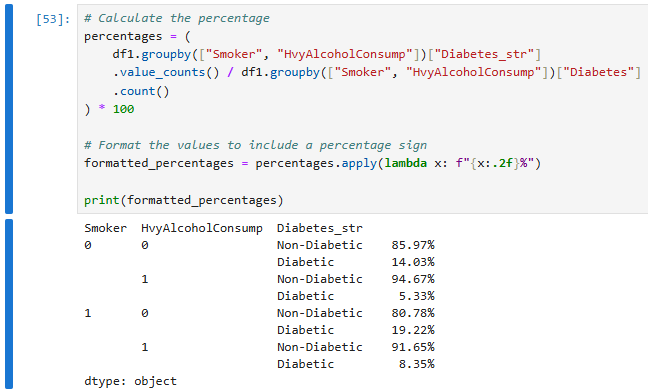


Figure - Correlation between Diabetes, Smoking and Alcohol Consumption

### **What is the correlation between diabetes, stroke and heart diseases?**

**Result:** The data reveals significant differences in diabetes prevalence based on the presence of stroke and heart disease or attack. Among individuals without a history of stroke or heart disease, 86.73% are non-diabetic, while 13.27% are diabetic. However, for those without a stroke but with a history of heart disease, the prevalence of diabetes increases notably to 34.10%. For individuals with a history of stroke but no heart disease, 28.26% are diabetic, and the proportion rises even further to 44.12% for those with both stroke and heart disease. These findings indicate a strong association between the presence of stroke, heart disease, or both and an increased likelihood of diabetes, emphasizing the interconnected nature of these health conditions. The analysis results are shown in figure 28 and figure 29.

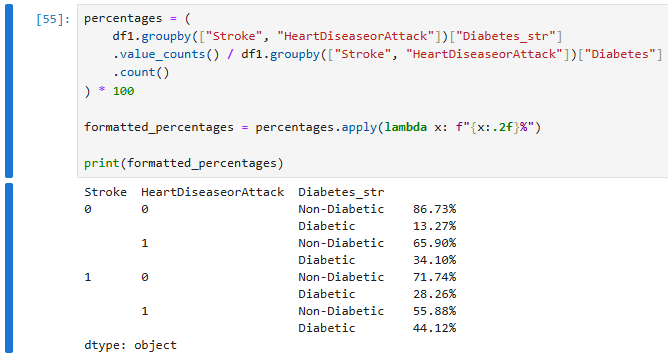


Figure - Correlation between Stroke, Heart Diseases and Diabetes

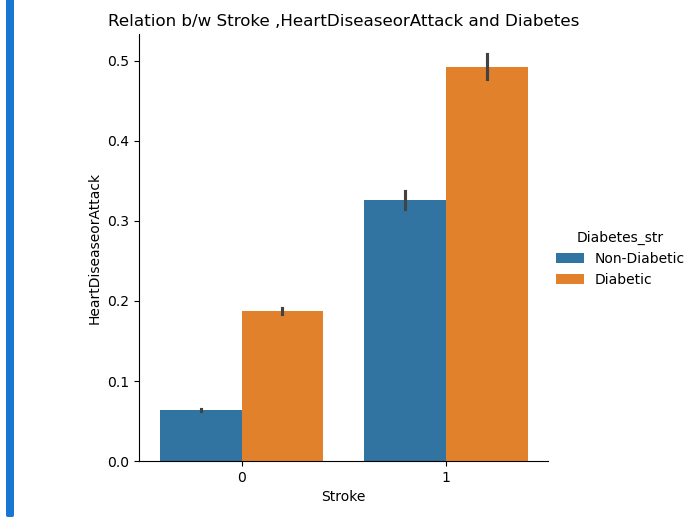


Figure - Correlation of Stroke, Heart Disorders and Diabetes

### **What are effects of life style activities on Diabetes?**

**Result:** The data demonstrates the relationship between fruit and vegetable consumption and diabetes prevalence. Individuals who consume neither fruits nor vegetables have a higher proportion of diabetes, with 20.05% being diabetic. Among those who consume vegetables but not fruits, the prevalence of diabetes decreases slightly to 16.71%. Similarly, individuals who consume fruits but not vegetables have a diabetes prevalence of 20.53%. The lowest prevalence, 13.84%, is observed among those who consume both fruits and vegetables, highlighting the potential benefits of a balanced diet rich in both.

In addition to diet, physical activity plays a crucial role in diabetes management. Regular exercise improves blood sugar control, enhances insulin sensitivity, and supports weight management, all of which are vital for mitigating diabetes-related complications. A combination of a nutrient-rich diet and an active lifestyle can significantly reduce diabetes risk and improve outcomes for those living with the condition. The analysis results are shown in figure 30, figure 31 and figure 32.

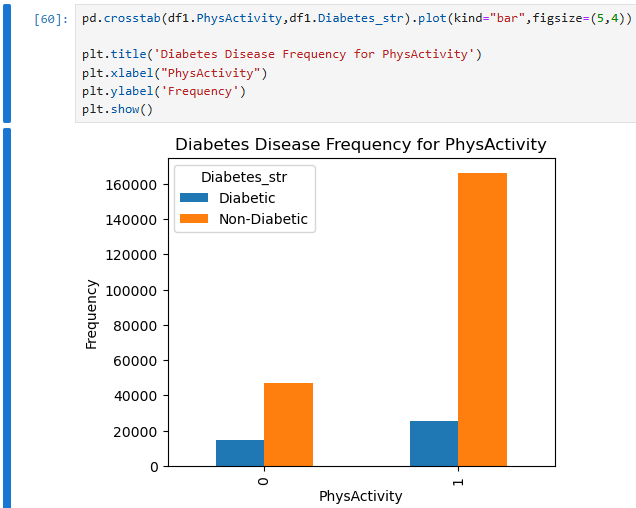


Figure - Diabetes and Physical Activity Correlation

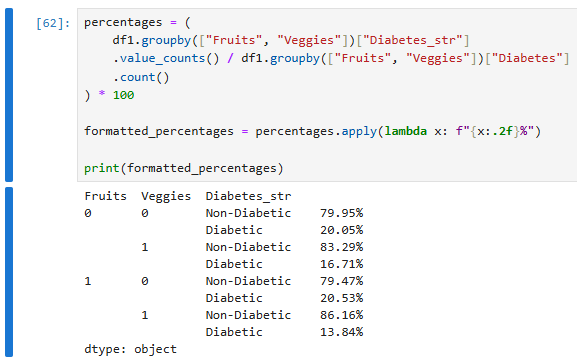


Figure - Correlation of Fruits and Veggies intake on Diabetes



Figure - Combined effect of Physical activity, fruits and vegetables intake on diabetes

### **What is the correlation of diabetes with general, mental and physical health conditions?**

**Result:** The results underscore the significant impact of various health factors on diabetes prevalence. General health (GenHlth) plays a critical role, as poor general health is strongly associated with an increased risk of diabetes. When individuals report their overall health as not good, the likelihood of developing diabetes rises sharply. This highlights the importance of maintaining overall well-being to reduce the risk of diabetes.

Mental health (MentHlth) is another major contributing factor. Persistent instability in mental health significantly increases the risk of diabetes, likely due to the interplay between stress, hormonal changes, and lifestyle factors such as poor diet and lack of physical activity. This underscores the need for addressing mental health as part of diabetes prevention and management strategies.

Similarly, physical health (PhysHlth) also has a profound effect. Long-term issues with physical health are linked to a higher risk of diabetes, further emphasizing the interconnectedness of physical ailments and metabolic disorders. Maintaining good physical health through regular exercise, balanced nutrition, and timely medical care is crucial in mitigating this risk.

Lastly, mobility challenges, such as difficulty walking or climbing, are strongly associated with diabetes, with nearly half of those experiencing such difficulties being diabetic. This may be attributed to limited physical activity, which exacerbates diabetes risk and complicates its management. Addressing mobility issues and encouraging adaptive physical activity can play a vital role in improving outcomes for these individuals.

### **What is the correlation between Diabetes and Gender, Age, Income and Education?**

**Result:** The analysis reveals that diabetes affects males and females equally, indicating no significant gender-based disparity in vulnerability. However, age is a critical factor, as individuals over 45 are significantly more prone to diabetes than younger people. The prevalence of diabetes increases steadily with age, underscoring the need for targeted preventive measures and early interventions among older populations to manage and reduce risk.

Education and income also influence diabetes prevalence. As education levels increase, the number of individuals with diabetes decreases, suggesting that higher education may correlate with better health awareness and access to resources for maintaining a healthy lifestyle. Conversely, people with lower incomes are at greater risk of developing diabetes compared to those with higher incomes, likely due to limited access to nutritious food, healthcare, and lifestyle options. Notably, the variables age, education, and income show no outliers, ensuring the reliability and consistency of these findings in representing the broader population trends.

The results are shown in figure 33 to figure 36.

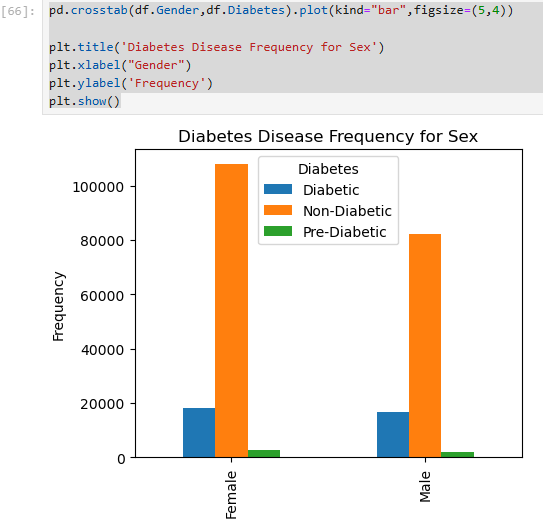


Figure - Correlation between Diabetes and Gender

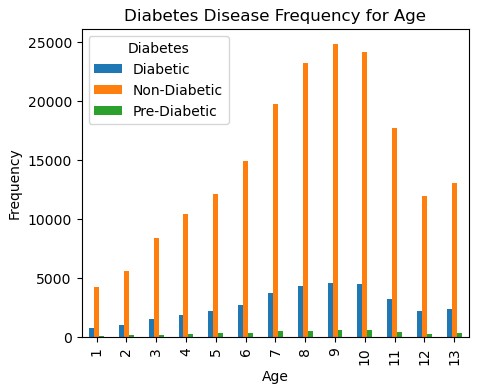


Figure - Correlation between Diabetes and Age Groups

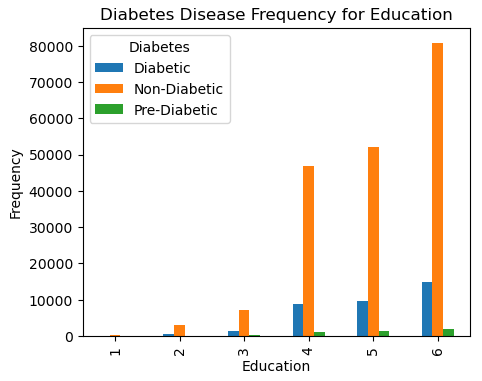


Figure - Correlation between Diabetes and Education

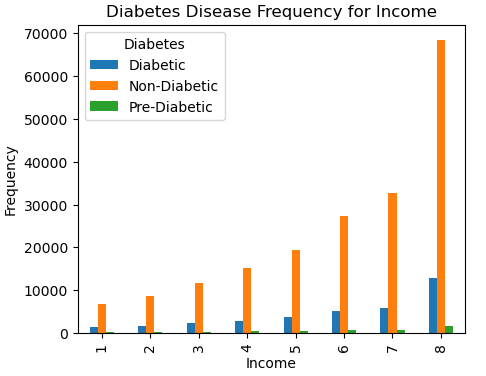


Figure - Correlation between Income and Diabetes

### **Separation of Major and Least Effective Features on the basis of EDA?**

**Result:** Result of major and minor or least effective features are detected on the basis of EDA performed and shown in table 3 provided below.

Table - Major and Minor Features detected in Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. #** | **Attributes** | **Major Features** | **Least Effective Features** |
| 1 | Diabetes | - | - |
| 2 | HighBP | Yes | - |
| 3 | HighChol | Yes | - |
| 4 | CholCheck | - | Yes |
| 5 | BMI | Yes | - |
| 6 | Smoker | Yes | - |
| 7 | Stroke | Yes | - |
| 8 | HeartDiseaseorAttack | Yes | - |
| 9 | PhysActivity | Yes | - |
| 10 | Fruits | Yes | - |
| 11 | Veggies | Yes | - |
| 12 | AlcoholConsump | - | Yes |
| 13 | HealthCoverage | - | Yes |
| 14 | GenHlth | Yes | - |
| 15 | MentHlth | Yes | - |
| 16 | PhysHlth | Yes | - |
| 17 | DiffWalk | - | Yes |
| 18 | Gender | - | Yes |
| 19 | Age | Yes | - |
| 20 | Education | Yes | - |
| 21 | Income | Yes | - |

# Predictive Analysis

## **Predictive Analysis using Random Forest Classifier**

**Results**: The classification report for the Random Forest Classifier predicting diabetes reveals significant insights and challenges. Class 0 (no diabetes) dominates the dataset, showing excellent performance with high precision (0.87), recall (0.95), and F1-score (0.91), resulting in an overall accuracy of 83%. However, this high accuracy is primarily driven by the overrepresentation of Class 0, leading to biased predictions. On the other hand, the classifier struggles significantly with Class 1 (diabetes), with extremely poor precision (0.03) and recall (0.00), reflecting its inability to detect diabetes cases. For Class 2 (pre-diabetes), the model performs moderately, with a precision of 0.41 and a recall of 0.23, but still fails to consistently identify instances from this category.

The imbalance in the dataset is a major challenge, with Class 0 vastly outnumbering Classes 1 and 2. This imbalance skews the performance metrics, as evidenced by the macro averages (precision 0.44, recall 0.39, F1-score 0.40) being much lower than the weighted averages (precision 0.79, recall 0.83, F1-score 0.80). These metrics indicate that while the model performs well overall, it fails to generalize across all classes, particularly the underrepresented Class 1. The inability to accurately identify diabetes cases (Class 1) is a critical shortfall, as it undermines the model’s utility in a real-world scenario where identifying such cases is essential. Results are shown in figure 37.

## **Predictive Analysis using Logistic Regression Classifier**

**Results**: The classification report for the Logistic Regression Classifier predicting diabetes highlights significant performance disparities across the three classes. Class 0 (no diabetes) dominates the predictions with high precision (0.86), recall (0.98), and an F1-score of 0.92, contributing to an overall accuracy of 85%. However, this high accuracy is largely due to the class imbalance, as Class 0 accounts for the majority of the dataset. The classifier struggles significantly with Class 1 (diabetes), achieving a precision, recall, and F1-score of 0.00, which means it fails to correctly identify any diabetes cases. For Class 2 (pre-diabetes), the model demonstrates moderate precision (0.52) but a very low recall (0.16), resulting in an F1-score of 0.24, indicating inconsistent performance for this class.

The macro averages for precision (0.46), recall (0.38), and F1-score (0.38) reflect the model’s poor ability to generalize across all classes, particularly for Classes 1 and 2. The weighted averages (precision 0.80, recall 0.85, F1-score 0.81) appear better but are biased due to the dominance of Class 0 in the dataset. This imbalance skews the model's focus toward the majority class, leaving minority classes poorly represented and predicted. The failure to identify any instances of Class 1 (diabetes) is especially concerning, as these cases are critical for practical medical applications. Results are shown in figure 38.

A screenshot of a computer

Description automatically generated

Figure - Predictive Analysis using Random Forest Classifier

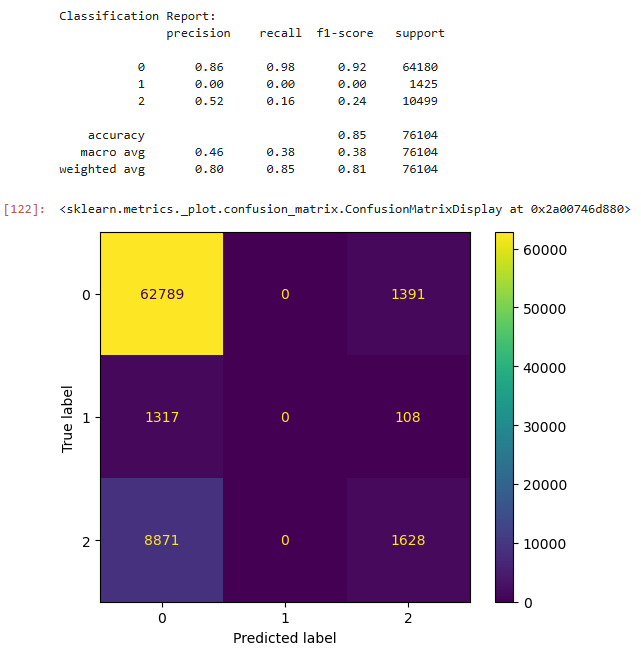


Figure - Predictive Analysis using Logistic Regression

## **Predictive Analysis using Decision Tree Classifier**

**Results**: The classification report for the Decision Tree Classifier predicting diabetes reveals a mixed performance across the three classes, with significant disparities in prediction accuracy. Class 0 (no diabetes) achieves high metrics, including a precision, recall, and F1-score of approximately 0.88, indicating strong performance for this majority class. This contributes to the overall accuracy of 79%. However, the classifier shows poor performance for Class 1 (diabetes), with all metrics—precision, recall, and F1-score—at just 0.03, signifying an inability to effectively identify or correctly classify instances of diabetes. For Class 2 (pre-diabetes), the results are slightly better, with a precision of 0.31, recall of 0.29, and an F1-score of 0.30, but these values still highlight inconsistencies in identifying this category.

The macro averages for precision (0.41), recall (0.40), and F1-score (0.40) reflect poor overall generalization across all classes, driven by the underperformance in Classes 1 and 2. The weighted averages (precision 0.78, recall 0.79, F1-score 0.78) are skewed by the large support for Class 0, masking the deficiencies in minority class predictions. The failure to distinguish between the minority classes, particularly Class 1, underscores the limitations of the Decision Tree model in handling imbalanced datasets. Results are shown in figure 39.

A screenshot of a computer

Description automatically generated

Figure - Predictive Analysis using Decision Tree Classifier

## **Comparative Analysis for Used Classifiers to detect Diabetes**

Across all three classifiers, Class 0 (no diabetes) is well-predicted, but the performance for Classes 1 (diabetes) and 2 (pre-diabetes) is largely insufficient. The **Random Forest** and **Logistic Regression** classifiers perform similarly in that they both fail to effectively identify diabetes cases (Class 1), with precision and recall at near-zero for Class 1. The **Decision Tree** model also faces similar issues, although its performance for Class 2 (pre-diabetes) is slightly better than Logistic Regression. The underlying issue across all classifiers is the **class disparity**, with the majority class (Class 0) dominating performance metrics, making the classifiers biased and underperforming for minority classes. Techniques such as resampling, class-weight adjustment, or using more sophisticated models like ensemble methods or neural networks could improve overall performance, particularly for Class 1 (diabetes) and Class 2 (pre-diabetes).