BUCHAREST UNIVERSITY OF ECONOMIC STUDIES
FACULTY OF BUSINESS ADMINISTRATION

What risk factors are most predictive of diabetes risk?

- a Scientific Paper -

Bucharest, 2024

Authors:
Daniela Manole
Yasmin Khader
Carmen Grigoras

1. INTRODUCTION

1.1 Global Diabetes Epidemic

The escalating global occurrence of diabetes demands an in-depth exploration of its predictive risk factors. As modern lifestyles continue to evolve, understanding the interplay between demographic, lifestyle, and health-related variables becomes crucial. This study delves into an extensive analysis of a dataset comprising 70,692 patients, utilizing machine-learning techniques to uncover the most influential factors contributing to diabetes risk.

Diabetes, a chronic metabolic disorder characterized by elevated blood sugar levels, has reached epidemic proportions globally. According to recent estimates by the International Diabetes Federation (IDF), approximately 463 million adults aged 20-79 were living with diabetes in 2019. Alarmingly, this number is projected to rise to 700 million by 2045, underscoring the urgent need for comprehensive research to understand and mitigate the impact of this pervasive condition.

1.2 The Burden of Diabetes

Beyond its sheer prevalence, diabetes imposes a staggering burden on individuals, healthcare systems, and economies worldwide. In 2019 alone, diabetes was responsible for an estimated 4.2 million deaths, making it a leading cause of mortality. The economic toll is equally profound, with direct healthcare expenditures on diabetes surpassing USD 760 billion globally.

Despite ongoing efforts to address the diabetes crisis, the burden of this condition continues to escalate, posing a profound threat to global health. Notably, diabetes-related deaths have been on a steady rise, with a 5% increase reported in the past decade. The economic impact is equally alarming, as diabetes-related healthcare expenditures are expected to soar to unprecedented levels, surpassing USD 845 billion by 2045. This financial strain extends beyond direct medical costs, encompassing indirect expenses related to reduced workforce productivity and increased disability.

Moreover, the societal and humanistic toll of diabetes cannot be overstated, with individuals grappling not only with physical health complications but also with psychological challenges. The urgency to comprehensively understand and address the underlying risk factors contributing to the diabetes epidemic has never been more critical. This study, situated at the intersection of public health, epidemiology, and data science, seeks to unravel the complexities of diabetes risk factors, providing insights that can catalyze effective preventive strategies and alleviate the burgeoning global burden of this pervasive condition.

1.3 The Need for Targeted Research

In light of these circumstances, there is a critical need for targeted research to identify and comprehend the risk factors contributing to the onset of diabetes. Understanding these factors is paramount for developing effective preventive strategies, personalized interventions, and informed healthcare policies. This study seeks to address this imperative by unraveling the complex interplay of demographic, lifestyle, and health-related variables influencing diabetes risk.

Beyond the statistics, diabetes profoundly impacts individuals' quality of life. Complications associated with diabetes, including cardiovascular diseases, kidney failure, and blindness, exact a heavy toll on affected individuals. The societal consequences are equally significant, with productivity losses, increased healthcare utilization, and the strain on healthcare infrastructures.

1.4 Research Objectives

In light of this, our research aims to answer a pivotal question: “What risk factors are most predictive of diabetes risk?” By delving into this question, we aim to contribute not only to the academic understanding of diabetes epidemiology but also to provide actionable insights that can inform healthcare practices, public health initiatives, and policy decisions.

Our primary objective is to unravel the complex factors influencing diabetes risk. By combining data analysis, visualization, and machine learning, we aim to provide a comprehensive understanding of the diverse elements contributing to the onset of diabetes. This paper details the methodologies employed and the profound insights achieved from an exploration of various patient attributes.

As the global burden of diabetes continues to escalate, there exists a critical knowledge gap in our understanding of the specific factors contributing to the onset of diabetes in diverse populations. This research strives to bridge this gap by employing advanced data analysis techniques, machine learning, and a nuanced exploration of a diverse set of variables. By uncovering the most predictive risk factors, we aim to empower healthcare professionals, policymakers, and individuals to take proactive measures in preventing and managing diabetes, ultimately contributing to a healthier and more resilient global population.

2. METHODS

2.1 Data Collection & Processing

The dataset utilized in this research was sourced from Kaggle, a prominent platform for data science and machine learning enthusiasts. Specifically, the dataset

titled "Diabetes Health Indicators Dataset" provided by Alex Teboul was instrumental in our investigation. This comprehensive dataset, available at **this Kaggle link**, encompasses a wealth of health indicators derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey. The dataset, encompassing 21 features and a binary diabetes indicator for each patient, forms the foundation of our analysis. We meticulously scrutinized the data for completeness, ensuring a clean dataset for robust analysis. Demographic information, lifestyle choices, and health metrics such as BMI, cholesterol levels, and reported physical and mental health were included.

2.2 Overview of Methodology

- Basic understanding of the data: Dataset columns, shape, types of data
- Histograms for each feature (distribution of values in the data frame)
- Violin plots (Correlation between the distribution of a feature and if the patient has diabetes)
- Correlation Matrix (Heatmap)
- Train CatBoost model (Gradient Boosted Decision Trees) and determine feature importance (SHAP values)

2.3 Features

Independent Variables (Target Feature):

- Diabetes_binary: No=0, Diabetes=1

Other Features:

- HighBP: Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional --> Yes=1, No=0
- HighChol: Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> Yes=1, No=0
- CholCheck: Cholesterol check within past five years --> Yes=1, No=0
- BMI: Body Mass Index (BMI) --> BMI value
- Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] --> Yes=1, No=0
- Stroke: (Ever told) you had a stroke. --> Yes=1, No=0
- HeartDiseaseorAttack: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) --> Yes=1, No=0
- PhyActiviity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job --> Yes=1, No=0
- Fruits: Consume Fruit 1 or more times per day --> Yes=1, No=0
- Veggies: Consume Vegetables 1 or more times per day --> Yes=1, No=0
- HvyAlcoholConsump: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) --> Yes=1, No=0

- AnyHealthcare: 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? --> Yes=1, No=0
- NoDocbcCost: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? --> Yes=1, No=0
- GenHlth: Would you say that in general, your health is: --> Ordinal feature: 1=Excellent - 5=Poor
- MentHlth: 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? --> Ordinal=number of days
- PhysHlth: 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? --> Ordinal=number of days
- DiffWalk: Do you have serious difficulty walking or climbing stairs? --> Yes=1, No=0
- Sex: Indicate sex of respondent. --> Female=0, Male=1
- Age: Age category --> Ordinal: age 18-24=1, all the way up to 13, which is 80 and older. 5 year increments.
- Education: What is the highest grade or year of school you completed? --> Ordinal: 1 being never attended school or kindergarten only up to 6 being college 4 years or more
- Income: Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") --> Ordinal: 1 being less than 10,000 all the way up to 8 being 75,000 or more

In [21]:
```
# Preparing packages
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
```

In [23]:
```
# Check for data quality
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70692 entries, 0 to 70691
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Diabetes_binary       70692 non-null  float64
 1   HighBP                70692 non-null  float64
 2   HighChol              70692 non-null  float64
 3   CholCheck             70692 non-null  float64
 4   BMI                   70692 non-null  float64
 5   Smoker                70692 non-null  float64
 6   Stroke                70692 non-null  float64
 7   HeartDiseaseorAttack  70692 non-null  float64
 8   PhysActivity          70692 non-null  float64
 9   Fruits                70692 non-null  float64
 10  Veggies               70692 non-null  float64
 11  HvyAlcoholConsump     70692 non-null  float64
 12  AnyHealthcare         70692 non-null  float64
 13  NoDocbcCost           70692 non-null  float64
 14  GenHlth               70692 non-null  float64
 15  MentHlth              70692 non-null  float64
 16  PhysHlth              70692 non-null  float64
 17  DiffWalk              70692 non-null  float64
 18  Sex                   70692 non-null  float64
 19  Age                   70692 non-null  float64
 20  Education             70692 non-null  float64
 21  Income                70692 non-null  float64
dtypes: float64(22)
memory usage: 11.9 MB
```

In [24]:
```
data['Diabetes_binary'].value_counts()
```

```
Diabetes_binary
```

2.4 Basic understanding of the data

We have a Dataset of 22 columns (21 features + target feature) and 70692 rows (patients).

The dataset is clean with 0 missing values, and all the features are numeric and data types are float.

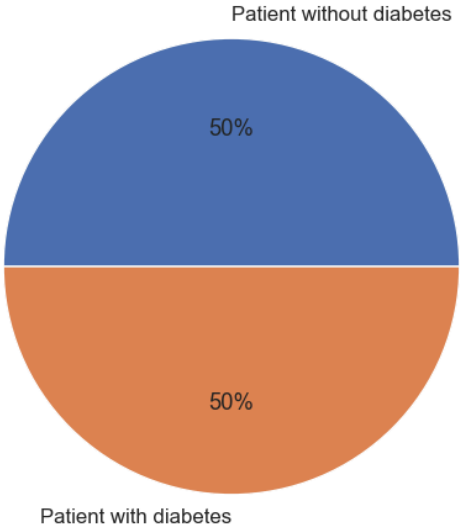
In [22]:
```
# Importing and visualising the contents of the Dataset
data = pd.read_csv("./diabetes_5050.csv")
print(data.shape)
data.head()
```

```
(70692, 22)
```

Out[22]:

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
0	0.0	1.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	0.0
1	0.0	1.0	1.0	1.0	26.0	1.0	1.0	0.0	0.0	1.0
2	0.0	0.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	1.0
3	0.0	1.0	1.0	1.0	28.0	1.0	0.0	0.0	1.0	1.0
4	0.0	0.0	0.0	1.0	29.0	1.0	0.0	0.0	1.0	1.0

5 rows × 22 columns



In [25]:
```
# Check the dataset of how many patients have diabetes
patients_diabetes = {
    0: sum(data["Diabetes_binary"].apply(lambda x: 1 if x == 0 else 0)),
    1: sum(data["Diabetes_binary"].apply(lambda x: 1 if x == 1 else 0)),
}
plt.figure(figsize=(8, 6))
plt.pie(x=patients_diabetes.values(), labels=['Patient without diabetes', 'Patient with diabetes'],
plt.show()
```

The dataset has an equal percentage of patients with diabetes and patients without diabetes.

2.5 Histograms and Violin Plots

In order to better visualize the data we have at our disposal, we decided to use two types of plots:

- a histogram, which will show us the distribution of values on the respective column in the data frame
- a violin plot, which will give us important insights into whether the data influences the outcome in a visible way.

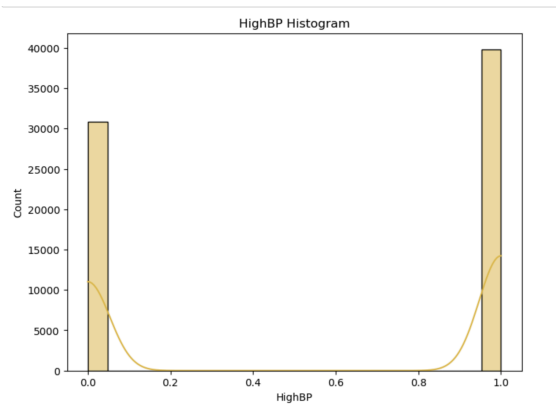
Defining a plotting function to use it for histograms and violin plots

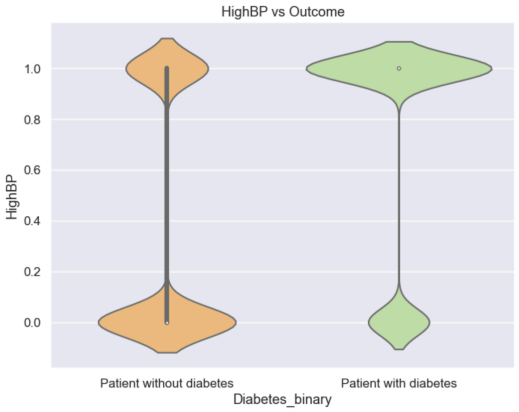
In [26]:
```
def plot_function(dataframe, feature_to_plot, type_of_plot):
    plt.figure(figsize=(8, 6))
    if type_of_plot=='violinplot':
        sns.violinplot(data=dataframe,
                       x='Diabetes_binary',
                       y=feature_to_plot,
                       palette="Spectral").set_xticklabels(['Patient without diabetes', 'Patient wi
        plt.title(f'{feature_to_plot} vs Outcome')
        plt.show()
    elif type_of_plot=='histogram':
        sns.histplot(data=dataframe,
                     x=feature_to_plot,
                     color='#e4b634',
                     kde=True)
        plt.title(f'{feature_to_plot} Histogram')
        plt.show()
    else:
        print("Wrong type of plot")
    return None
```

Feature Histogram and Violinplot

```
In [27]: # Plotting each column (feature) in histogram and violin plot
         %matplotlib inline

         for column in data.columns:
             if column == "Diabetes_binary":
                 continue
             plot_function(data, column, "histogram")
             plot_function(data, column, "violinplot")
```

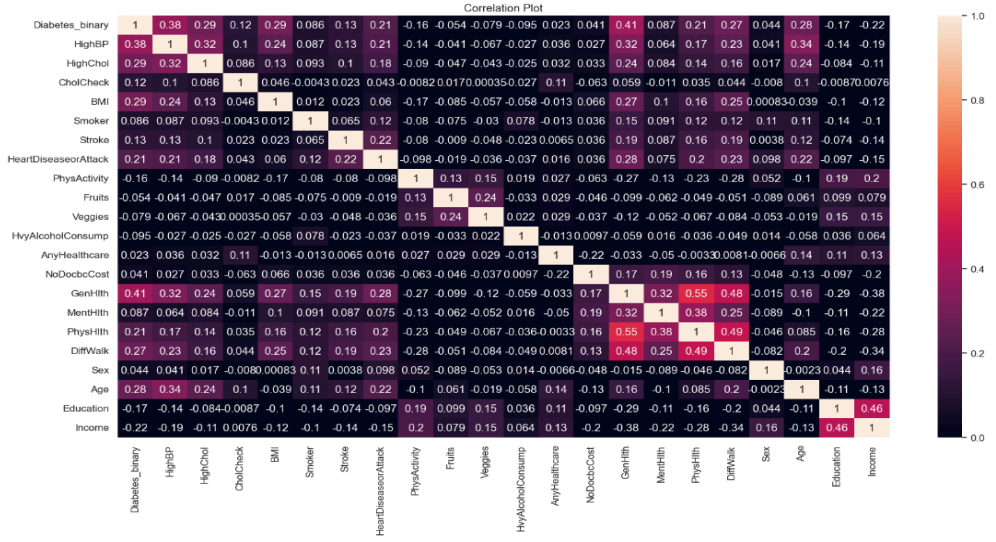




2.6 Correlation Heatmap

A correlation matrix heatmap visually represents the correlation between variables using colors. It helps identify relationships, detect multicollinearity, select variables, recognize patterns, and spot outliers in a dataset. The color scheme indicates the strength and direction of correlations, simplifying insights into complex relationships.

```
In [28]: # Plotting the heatmap
         plt.figure(figsize=(20, 10))
         sns.set(font_scale=1.1)
         heatmap = sns.heatmap(data.corr(), vmin=0, vmax=1, annot=True)
         heatmap.set_title('Correlation Plot')
         plt.show()
```



2.7 Train CatBoost Classifier

CatBoost is a powerful gradient-boosting algorithm specifically designed to handle categorical features seamlessly. It stands out by eliminating the need for extensive preprocessing of categorical data, making it a robust choice for machine learning tasks. CatBoost incorporates strategies to efficiently handle categorical variables, optimizing training speed and model performance. This algorithm is particularly useful in scenarios where feature importance interpretation and categorical data management are critical aspects of the machine learning process.

In [29]:
```
# splitting the train and the test data
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

std_scaler = StandardScaler()
X = data.drop(columns='Diabetes_binary', axis=1)
y = data['Diabetes_binary']

X = std_scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, shuffle=True)
```

In [30]:
```
# training the CatBoost model
from catboost import CatBoostClassifier
cbc = CatBoostClassifier()
cbc.fit(X_train, y_train)
```

2.8 Feature Importance

Feature importance in machine learning refers to the assessment of each input variable's contribution to the predictive performance of a model. It helps us identify which features or variables have a more significant impact on the model's outcomes. Feature Importance is extracted from the CatBoost Model; it is determined by permutation importance applied within this particular model; Permutation importance means the following: let’s say we have some data, with the above columns. If we were to randomly shuffle the values on one column without changing the others, how would the outcome be impacted?

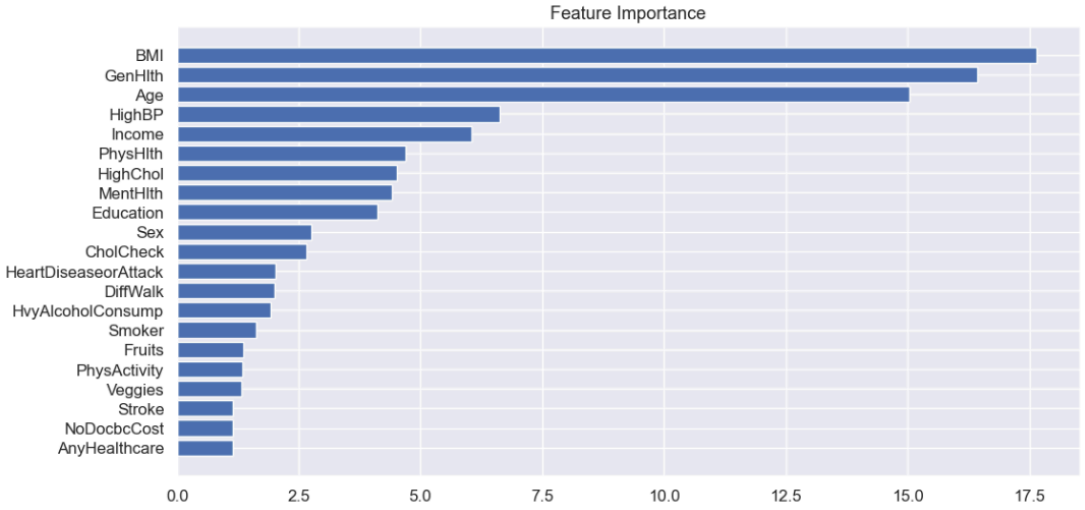
In [31]:
```
# selecting the features without the target feature
features = data.drop(columns='Diabetes_binary', axis=1).columns
features = [str(x) for x in features]
features
```

Out[31]:
```
['HighBP',
 'HighChol',
 'CholCheck',
 'BMI',
 'Smoker',
 'Stroke',
 'HeartDiseaseorAttack',
 'PhysActivity',
 'Fruits',
 'Veggies',
 'HvyAlcoholConsump',
 'AnyHealthcare',
 'NoDocbcCost',
 'GenHlth',
 'MentHlth',
 'PhysHlth',
 'DiffWalk',
 'Sex',
 'Age',
 'Education',
 'Income']
```

Here we are using the permutation feature importance from our trained model, and then we are mapping the values in the figure below.

```
In [32]: feature_importance = cbc.feature_importances_
         sorted_idx = np.argsort(feature_importance)
         fig = plt.figure(figsize=(12, 6))
         plt.barh(range(len(sorted_idx)), feature_importance[sorted_idx], align='center')
         plt.yticks(range(len(sorted_idx)), np.array(features)[sorted_idx])
         plt.title('Feature Importance')
```

```
Out[32]: Text(0.5, 1.0, 'Feature Importance')
```



2.9 SHAP Values

SHAP values as a tool that helps us understand and interpret the impact of different variables on our model's predictions. The impact comes from a vector v with contributions of each feature to the prediction for every input object and the expected value of the model prediction for the object (average prediction given no knowledge about the object).

```
In [33]: # plotting SHAP values based on our test values
         import shap

         shap_explainer = shap.TreeExplainer(cbc)
         shap_values = shap_explainer.shap_values(X_test)

         shap.summary_plot(shap_values, X_test, max_display=X_test.shape[1], show=False, feature_names=featu
```

3. RESULTS & DISCUSSION

The outcomes derived from **histograms and violin plots** provide valuable insights into the factors determining diabetes. The histogram analysis reveals that almost all patients had undergone cholesterol checks within the past five years, indicating a proactive approach to monitoring cardiovascular health. Furthermore, almost all patients in the study did not have a history of stroke, suggesting a potential correlation between stroke occurrence and diabetes.

The violin plots offer a more detailed perspective on the relationships between diabetes and

various factors. It is evident that patients with diabetes are more likely to have **High Blood Pressure and High Cholesterol**, emphasizing the interconnected nature of these health conditions. Additionally, the prevalence of **heart disease or attacks** is found to be most likely correlated with diabetes. Interestingly, strokes and smoking do not exhibit a direct correlation with diabetes, implying that these factors may not be primary determinants of diabetes in this population.

Contrary to common assumptions, the study reveals that consumption of fruits, vegetables, and physical activity are not directly correlated with diabetes. This challenges conventional beliefs about the impact of lifestyle choices on diabetes risk. However, the results indicate a significant correlation between **poor general health, poor physical health**, and an increased likelihood of having diabetes. The **difficulty of walking** is also positively associated with diabetes, suggesting a potential link between mobility issues and the disease.

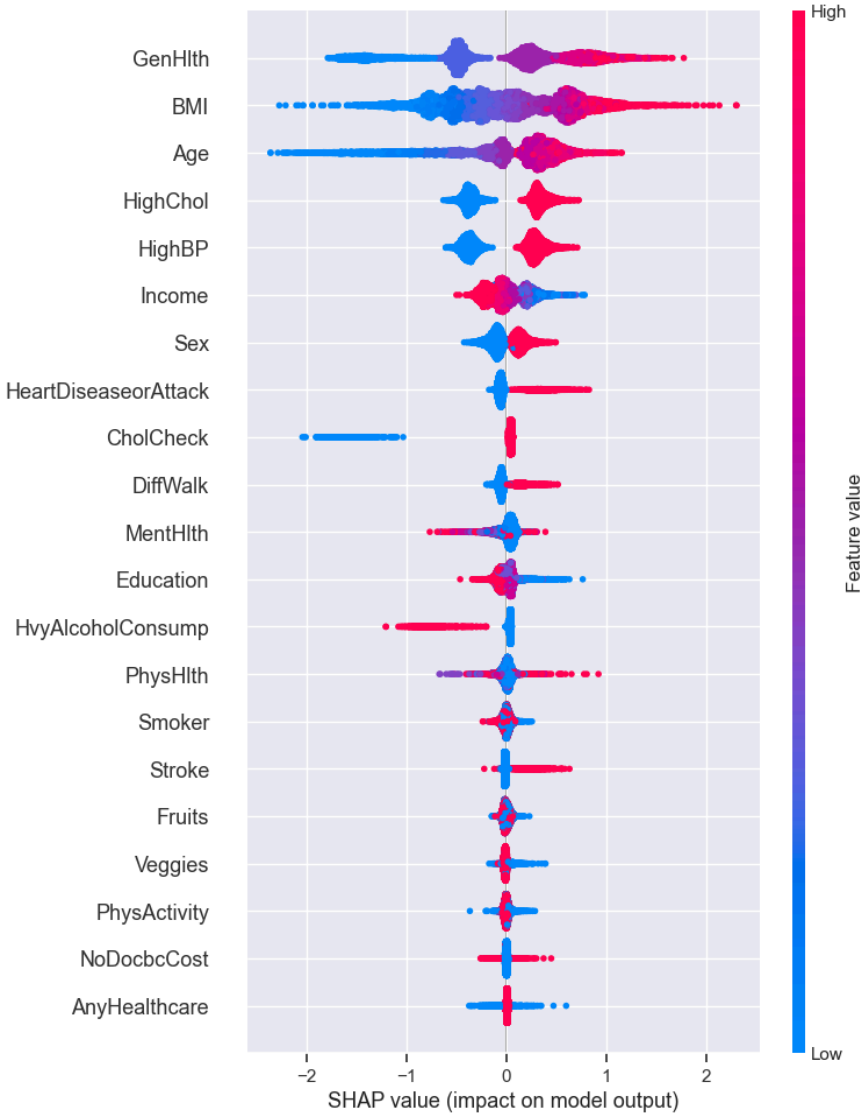
Demographic factors further describe the diabetes landscape. The data indicates a higher prevalence of diabetes in men compared to women, particularly in the age range of **65-70 years**. Additionally, **income levels** play a role, as higher income is associated with a lower likelihood of having diabetes. This socio economic aspect adds a nuanced layer to understanding the diverse factors influencing diabetes prevalence.

The Correlation Matrix heatmap has provided valuable insights into the relationships between various variables in the context of diabetes research. The correlation coefficients, ranging from -1 to 1, indicate the strength and direction of these relationships. Here are the key outcomes:

- **Diabetes and General Health** (Correlation = 0.41): A moderate positive correlation suggests that individuals with diabetes are more likely to report poorer general health.
- **Diabetes and High Blood Pressure** (Correlation = 0.38): A positive correlation implies that there is a connection between diabetes and high blood pressure.
- **Diabetes and High Cholesterol** (Correlation = 0.29): A positive correlation indicates that individuals with diabetes are more likely to have higher cholesterol levels.
- **Diabetes and BMI** (Correlation = 0.29): A positive correlation between diabetes and Body Mass Index (BMI) suggests that higher BMI may be associated with an increased likelihood of having diabetes.
- **Diabetes and Age** (Correlation = 0.28): A positive correlation suggests that age is a factor in diabetes, with older individuals being more likely to have diabetes.
- **Diabetes and DiffWalk** (Correlation = 0.27): A positive correlation indicates that individuals with diabetes are more likely to experience difficulties with walking.
- **General Health and Physical Health** (Correlation = 0.55): A strong positive correlation highlights that individuals with poorer general health are also likely to have poorer physical health.
- **General Health and DiffWalk** (Correlation = 0.48): A positive correlation suggests that individuals with poorer general health are more likely to face difficulties with walking.
- **Income and Education** (Correlation = 0.46): A positive correlation indicates that higher levels of education are associated with higher income.

In our Feature Importance assesment, we identified the **top 5 most important features** in predicting diabetes risk:

- **General Health:** The analysis indicates that general health is the most influential factor in predicting diabetes risk. This suggests that individuals with poorer general health are at a higher risk of developing diabetes.
- **BMI (Body Mass Index**): BMI emerges as the second most important feature in predicting diabetes risk. This aligns with existing knowledge that obesity and higher BMI are significant risk factors for diabetes.
- **Age**: Age is identified as a key predictor of diabetes risk, ranking third in importance. This result is consistent with the understanding that diabetes prevalence tends to increase with age.
- **High Blood Pressure:** The presence of high blood pressure is highlighted as a crucial factor in predicting diabetes risk, ranking fourth in importance. This reinforces the well-established association between diabetes and cardiovascular health.
- **Income**: Income is identified as the fifth most important feature in predicting diabetes risk. This socioeconomic factor suggests that individuals with higher income levels may be at a lower risk of developing diabetes.



When generating the SHAP values plot, certain patterns become evident. Apart from High Cholesterol, the initial five values align directly with the identified features correlated to the risk of diabetes, as observed in the feature importance analysis. Additionally, it is noteworthy that Income significantly influences the risk, but with an inverse relationship, contrasting with the direct proportionality observed in the other identified factors.

In Conclusion, this study aimed to identify important health factors that contribute to the prediction of diabetes in patients. Through histograms, violin plots, correlation matrix, feature importance, and SHAP values, a nuanced understanding of the variables influencing diabetes has been achieved. The findings collectively contribute to a more comprehensive understanding of diabetes determinants, encompassing health, demographic, and socioeconomic dimensions. These insights can guide targeted interventions, inform healthcare strategies, and enhance our overall approach to diabetes prevention and care.