Exploratory Data Analysis (EDA) Report: Diabetes Health Indicator Dataset

**Report prepared by:** Nyakana Gilbert

**RegNo:** 2024/HD05/21952U

**Date:** 11th September 2024

# Introduction

This report outlines the findings of an exploratory Data Analysis performed on the Diabetes Health Indicator dataset. The dataset used has three csv files with the largest file having over 250,000 records and 22 features. The goal of this exercise is to understand the relationship between the features, data distribution, identify missing values and generate insights for the proceeding processes.

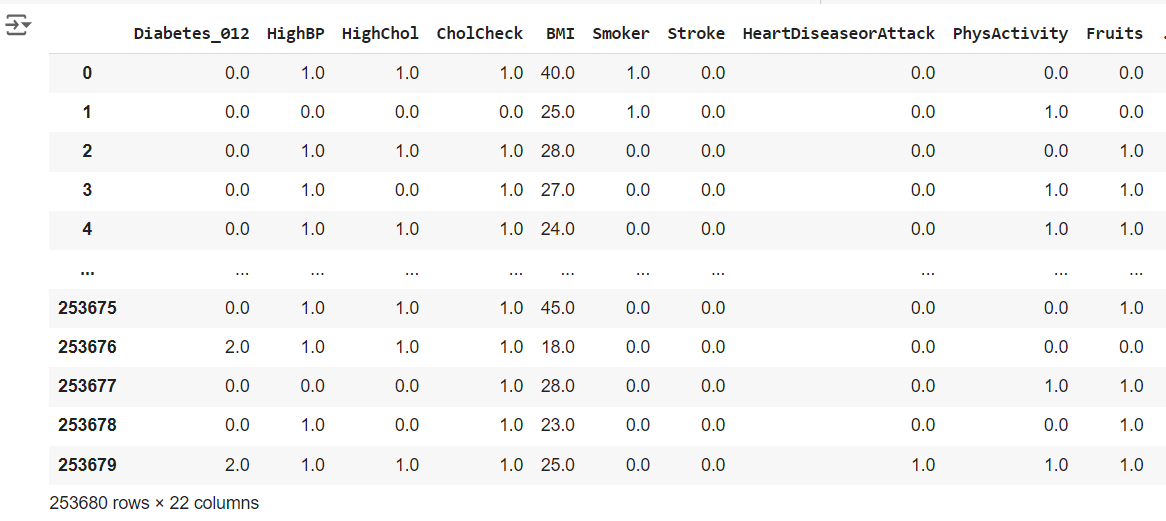
# Dataset Overview

The purpose of the dataset is to predict diabetes basing on lifestyle. It falls in the classification category and is tabular in nature. Illustrated here are a summary of the dataset.

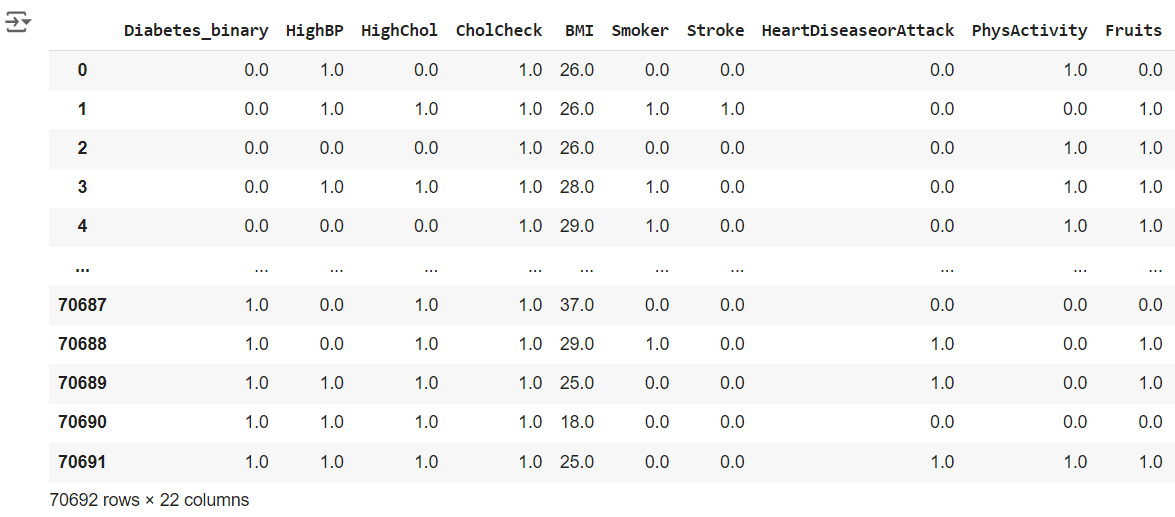
The table below shows the summary of the initial three datasets used

|  |  |  |
| --- | --- | --- |
| Dataset | Number of records | Number of features |
| Dataset1 | 253680 | 22 |
| Dataset2 | 70692 | 22 |
| Dataset3 | 253680 | 22 |

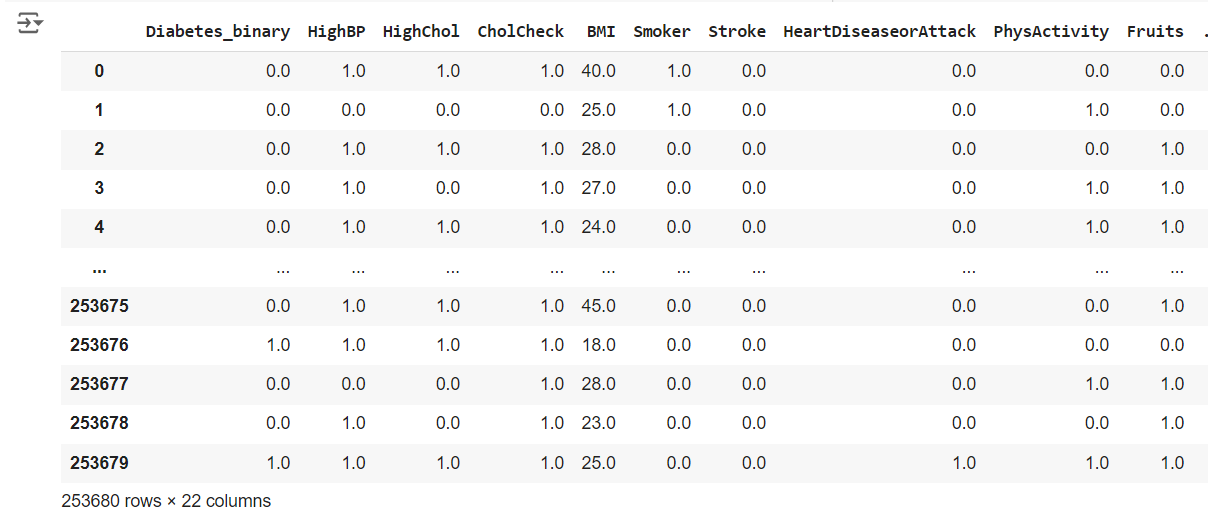
Overview of the first dataset



Overview of the second dataset



Overview of the third dataset



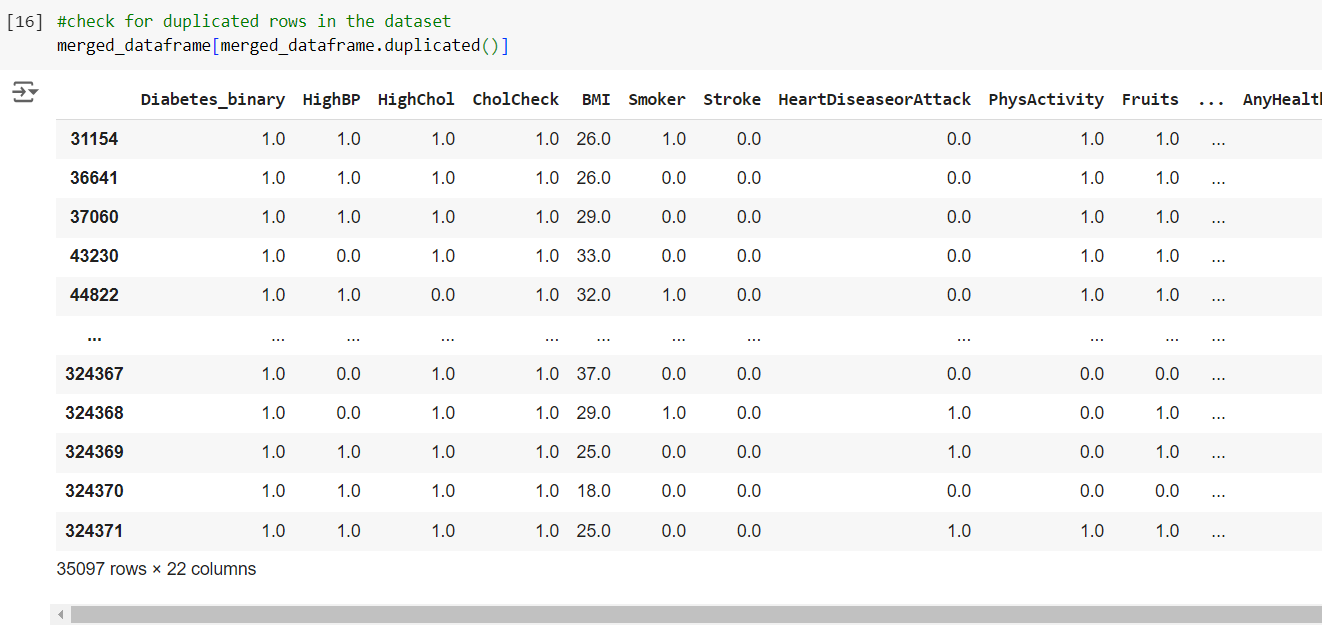
The table below shows the datasets and the corresponding features

|  |  |  |  |
| --- | --- | --- | --- |
| SN | Dataset1 | Dataset2 | Dataset3 |
|  | Diabetes\_binary | Diabetes\_binary | Diabetes\_012 |
|  | HighbP | HighbP | HighbP |
|  | HighChol | HighChol | HighChol |
|  | CholCheck | CholCheck | CholCheck |
|  | BMI | BMI | BMI |
|  | Smoker | Smoker | Smoker |
|  | Stroke | Stroke | Stroke |
|  | HeartDiseaseorAttack | HeartDiseaseorAttack | HeartDiseaseorAttack |
|  | PhysActivity | PhysActivity | PhysActivity |
|  | Fruits | Fruits | Fruits |
|  | Veggies | Veggies | Veggies |
|  | HvyAlcoholConsump | HvyAlcoholConsump | HvyAlcoholConsump |
|  | AnyHealthcare | AnyHealthcare | AnyHealthcare |
|  | NoDocbcCost | NoDocbcCost | NoDocbcCost |
|  | GenHlth | GenHlth | GenHlth |
|  | MentHlth | MentHlth | MentHlth |
|  | PhysHlth | PhysHlth | PhysHlth |
|  | DiffWalk | DiffWalk | DiffWalk |
|  | Sex | Sex | Sex |
|  | Age | Age | Age |
|  | Education | Education | Education |
|  | Income | Income | Income |

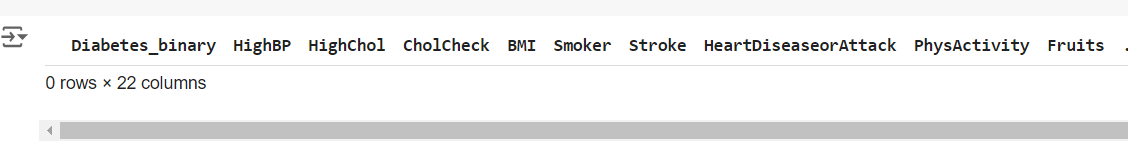
# Analysis performed

The following actions were taken during the EDA process on the selected dataset.

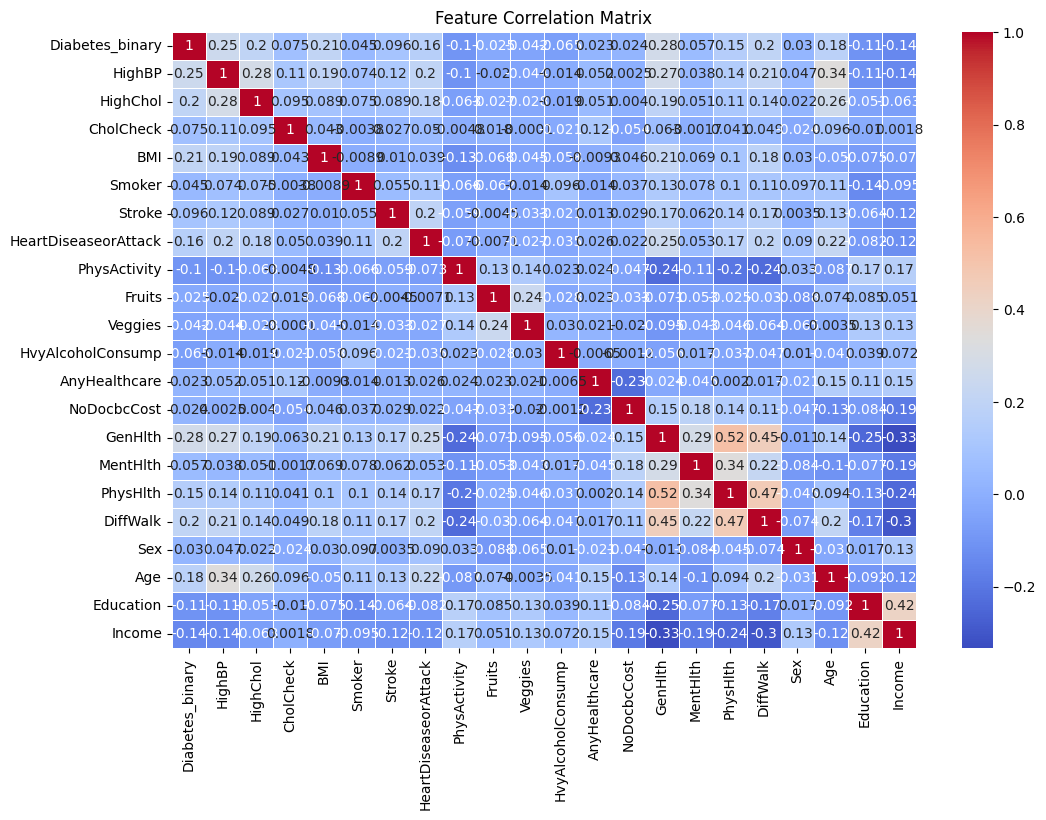
* Performed some normalization on one of the datasets that had a different column name.
* For easy analysis, the three dataset files were merged into one by concatenation.
* After merging, the target feature was identified to be diabetes\_binary
* Performed a check on the missing values after merging the files. The check results showed that there are no missing values in the dataset
* Performed a check on the unique values of each feature in the dataset.
* The unique values for the target feature were three [0,2,1] and I only need two values, 1 for the ones with diabetes and 0 for those that do not have diabetes. For this reason, I replaced 2 with 1
* The previous action introduced 35097 duplicates



* Dropped the duplicates and on the doublechecking, no duplicates were found.



* Performed a correlation check to see how the different features relate with each other.

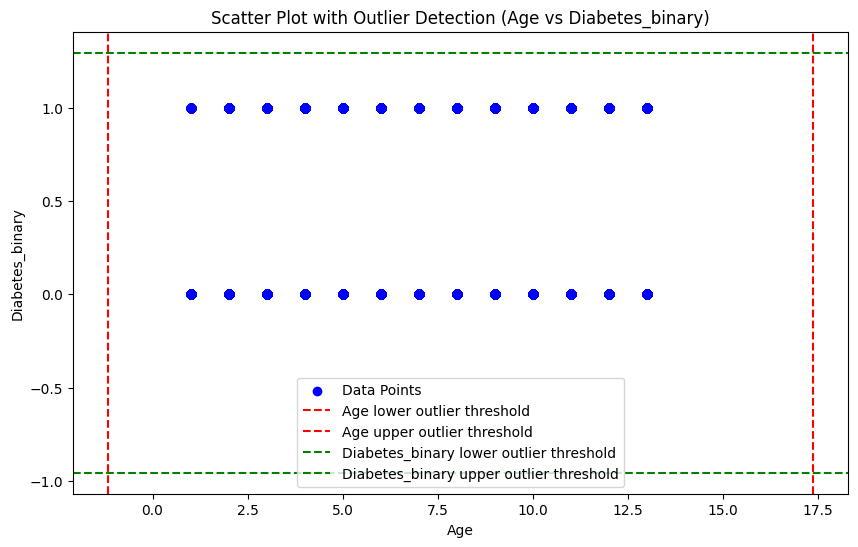


**General Correlation Observations:**

* **Health Conditions** (HighBP, HighChol, HeartDiseaseorAttack, Stroke) are moderately correlated with **Diabetes\_binary** and **Age**.
* **Physical Activity and Income** seem to be negatively correlated with health problems like high blood pressure, cholesterol, and diabetes.
* **General Health** is positively correlated with diabetes and health conditions, while negatively correlated with income and education, suggesting that people with lower income or education levels report poorer health outcomes.

**Conclusion on features correlation:**

* **Diabetes** is moderately correlated with factors like high blood pressure, cholesterol, poor general health, BMI, and age.
* **Higher income and education** seem to be protective against some of the health issues, while **physical activity** also shows a slight protective effect against diabetes and related conditions.
* Check out for outliers on the dataset



**Observation on Age and Diabetes\_binary:**

* There are no identified outliers between age and Diabetes\_binary.

# Conclusion

The dataset has been cleaned to some extent and some EDA performed. More checks need to be performed to ensure that the dataset is ready for the next course of action and future engineering.