1. **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.

1. **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 |

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 |

1. **Step by step analysis of the data**

The following was done during my Data Analysis process that I conducted on three datasets that are aimed at predicting diabetes basing on lifestyle.

1. First and foremost, I loaded the dasets which were stored in csv files into dataframes for easy reading and manipulation.
2. I needed to work using one dataset and for that reason, I looked at the features of the different datasets to establish whether to merge or concatenate the datasets so that I can have a single file for my dataset.
3. By displaying the records of the different files, I found out that my data had similar features except for one dataset that had one feature different from the others. That is to say two of the datasets had a Diabetes\_binary and one had Diabetes\_012 instead.
4. I renamed Diabetes\_012 to Diabetes\_binary for uniformity and concatenated the three datasets in order to have one dataset which I named merged\_dataframe.
5. I checked the dataset information to establish the data types of the features and whether the dataset contains null values. I found out that the dataset had no missing values.
6. I identified the target feature as diabetes\_binary and using .nunique() I established that the target feature has three unique values [2,1,0] and I need two [1,0].
7. For easy predictions I renamed the target value 2 to 1.
8. Using .value\_counts() I discovered that there is an imbalance in the dataset the diabetes\_binary value [0] had **194,377** records and diabetes\_binary value [1] had **74,754** records causing a very huge imbalance in the datasdet.
9. I then checked to see if the merged dataframe had duplicates and my findings indicated that duplicates existed.
10. I had to drop all duplicate records and re-visualise the distribution but it still indicated a class imbalance.
11. In order to handle the imbalance, I had to do a resampling by udersampling the majority class. After the resampling, the class imbalance of the dataset was corrected and I had an evenly distributed dataset. Both diabetes\_binary value [0] and diabetes\_binary value [1] now have **39657** records.
12. I performed a feature correlation matrix and below are some of the findings;

**The four strongest Positive Correlations:**

* HighBP (0.34): High blood pressure is moderately positively correlated with diabetes. As blood pressure increases, the likelihood of having diabetes increases.
* GenHlth (0.26): General health rating is positively correlated with diabetes, meaning poorer general health is associated with a higher likelihood of diabetes.
* Age (0.26): Age shows a moderate positive correlation with diabetes. Older people are more likely to have diabetes.
* BMI (0.26): A positive correlation between Body Mass Index (BMI) and diabetes indicates that higher BMI is associated with a greater likelihood of having diabetes.
* The strongest negative correlation is Income (-0.20): Income shows a negative correlation, suggesting that individuals with lower income might be more prone to diabetes, possibly due to lifestyle factors or access to healthcare.

1. Performed a test on identifying outliers using BMI and age and below are my findings;

* **Outliers detected for BMI**: The plot clearly shows that there are several individuals with BMIs above the upper threshold (to the right of the red vertical line). These data points are considered BMI outliers.
* **No age outliers**: No data points lie outside the age boundaries, indicating that all individuals' ages fall within the expected range based on the outlier detection criteria.

1. **Conclusion**

Features such as HighBP, GenHlth, and Age are highly correlated to the diabetes\_binary which is the target featureand these features could be prioritized in my model.