SAMSUNG

Samsung Innovation Campus

PREDICTION OF DIABETES HEALTH INDICATORS

Artificial Intelligence Course

SAMSUNG

OUR TEAM

The project presented by:

Nourhan Mahmoud Ahmad Muhammad Abdullah Saad

Facilitator

Eng Ziad omer

Data used: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

SAMSUNG

AGENDA

1. Introduction
2. Problem and Solution
3. Data
4. Data Cleaning
5. Exploratory Data Analysis
6. Data Preprocessing
7. Data Modeling
8. Model Development
9. conclusion

SAMSUNG

INTRODUCTION

Diabetes is a serious condition where your blood glucose level is too high. It can happen when your body doesn't produce enough insulin or the insulin it produces isn't effective. Or, when your body can't produce any insulin at all.

It is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015.

The Behavioral Risk Factor Surveillance System (BRFSS) is a system of health-related telephone surveys that collect state data about U.S. residents the largest continuously conducted health survey system in the world.

The dataset originally has 330 features collected but only 22 features relevant to diabetes are used in our based for machine learning algorithms

SAMSUNG

PROBLEM

In this technological and competitive era, our daily lives have been greatly affected and influenced as we are busy with studying and working without concerning our health and fitness.

This resulted in the increase of pre-diabetes and diabetes patients over the years and the rate of prevalence of this disease has never been decreasing. Factors of this problem arise from the varieties of food available in our country, the convenience of ordering food via the super app such as Grab, as well as the lack of regular exercise due to the many workloads and laziness.

SOLUTION

Trianing Models for dataset

Logistic Regression

Decision Tree Classifier

Random Forest Classifier

KNN

To obtain the best model for predicting diabetes and thus reducing the diabetes prevalence rate

SAMSUNG

DATA

Overview

Overview   Alerts 6   Reproduction

Dataset statistics

Number of variables	22
Number of observations	253680
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	11369
Duplicate rows (%)	4.5%
Total size in memory	42.6 MiB
Average record size in memory	176.0 B

Variable types

Categorical	16
Numeric	6

SAMSUNG

DATA

Categorical features

- Diabetes binary
- HighBP
- HighChol
- CholCheck
- Smoker
- Stroke
- Fruits
- Veggies
- DiffWalk
- Sex
- HeartDiseaseorAttack
- PhysActivity
- AnyHealthCare
- NoDocbcCost
- HvyAlcoholConsump

Numerical features

- Age
- BMI
- Income
- MentHlth
- GenHlth
- PhysHlth
- Education

```
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
```

SAMSUNG

DATA CLEANING

Handling Missing values

Diabetes_binary	0
HighBP	0
HighChol	0
CholCheck	0
BMI	0
Smoker	0
Stroke	0
HeartDiseaseorAttack	0
PhysActivity	0
Fruits	0
Veggies	0
HvyAlcoholConsump	0
AnyHealthcare	0
NoDocbcCost	0
GenHlth	0
MentHlth	0
PhysHlth	0
DiffWalk	0
Sex	0
Age	0
Education	0
Income	0

There is no missing values in our dataset

Checking duplicate values

```
# Number of duplicates
```

24206

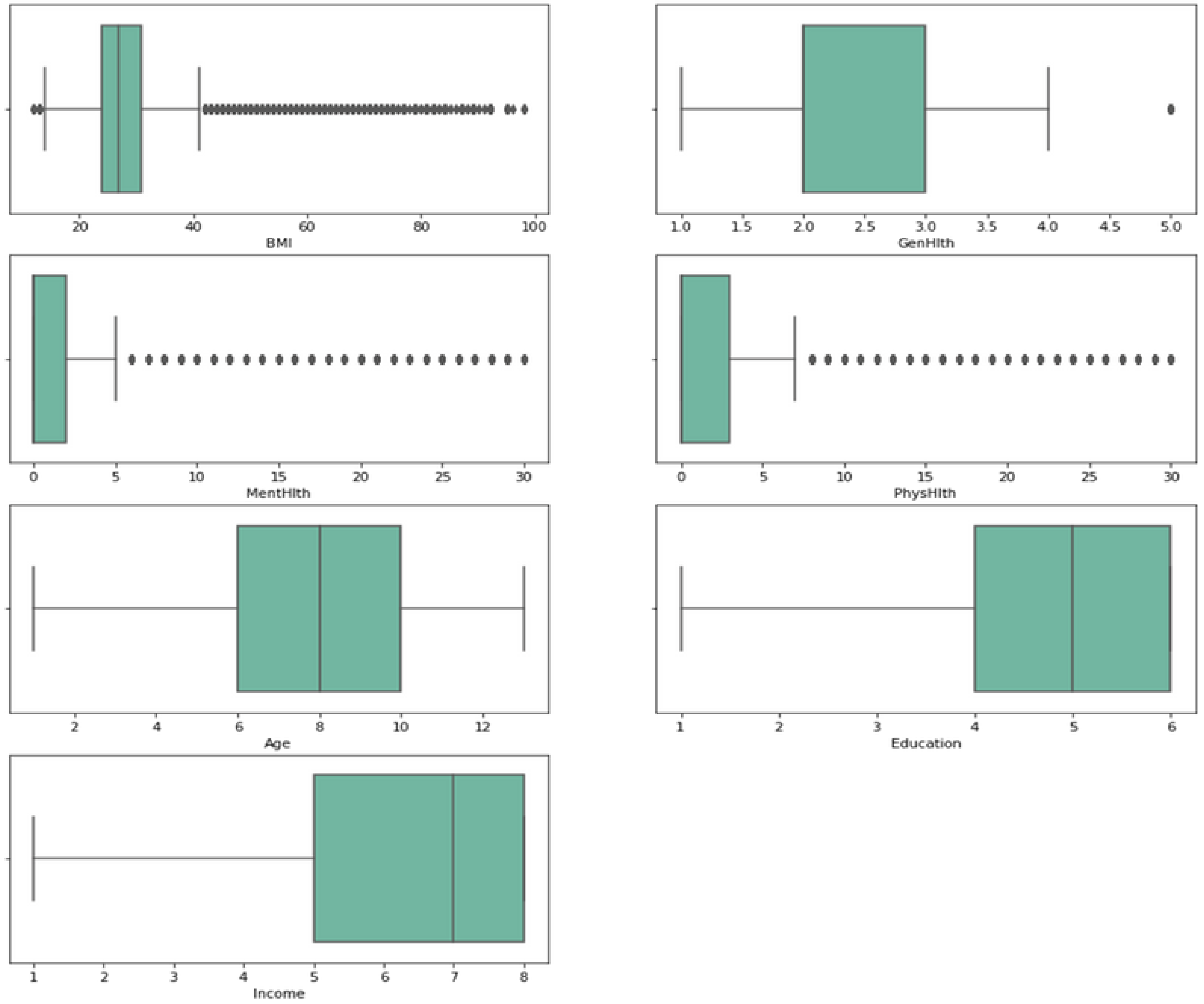
```
# drop duplicates
```

0

There are 24206 duplicated rows in our dataset, it could potentially introduce bias into the model. Therefore, it's a good idea to delete any duplicate rows before training the model.

DATA CLEANING

Checking outliers



The box plot shows us that there are outliers, but these values are real data and we must take them into account

transform the features type

```
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
```

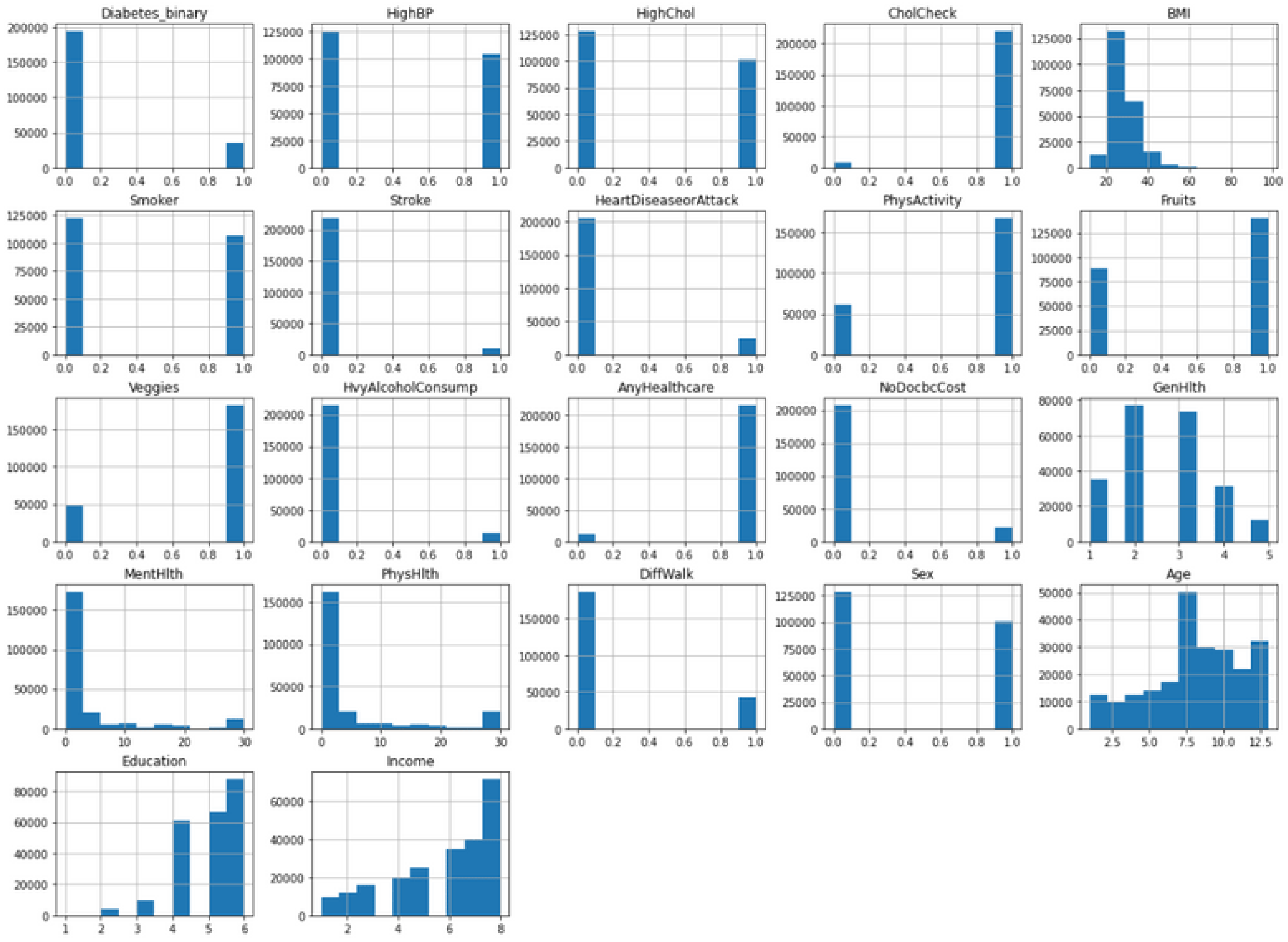
→

```
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  int32
 1   HighBP                253680 non-null  int32
 2   HighChol              253680 non-null  int32
 3   CholCheck             253680 non-null  int32
 4   BMI                   253680 non-null  int32
 5   Smoker                253680 non-null  int32
 6   Stroke                253680 non-null  int32
 7   HeartDiseaseorAttack  253680 non-null  int32
 8   PhysActivity          253680 non-null  int32
 9   Fruits                253680 non-null  int32
 10  Veggies               253680 non-null  int32
 11  HvyAlcoholConsump     253680 non-null  int32
 12  AnyHealthcare         253680 non-null  int32
 13  NoDocbcCost           253680 non-null  int32
 14  GenHlth               253680 non-null  int32
 15  MentHlth              253680 non-null  int32
 16  PhysHlth              253680 non-null  int32
 17  DiffWalk              253680 non-null  int32
 18  Sex                   253680 non-null  int32
 19  Age                   253680 non-null  int32
 20  Education             253680 non-null  int32
 21  Income                253680 non-null  int32
dtypes: int32(22)
```

Here we transform the features type to an integer to speed the model and the analysis

EXPLORATORY DATA ANALYSIS

Distribution of numerical features



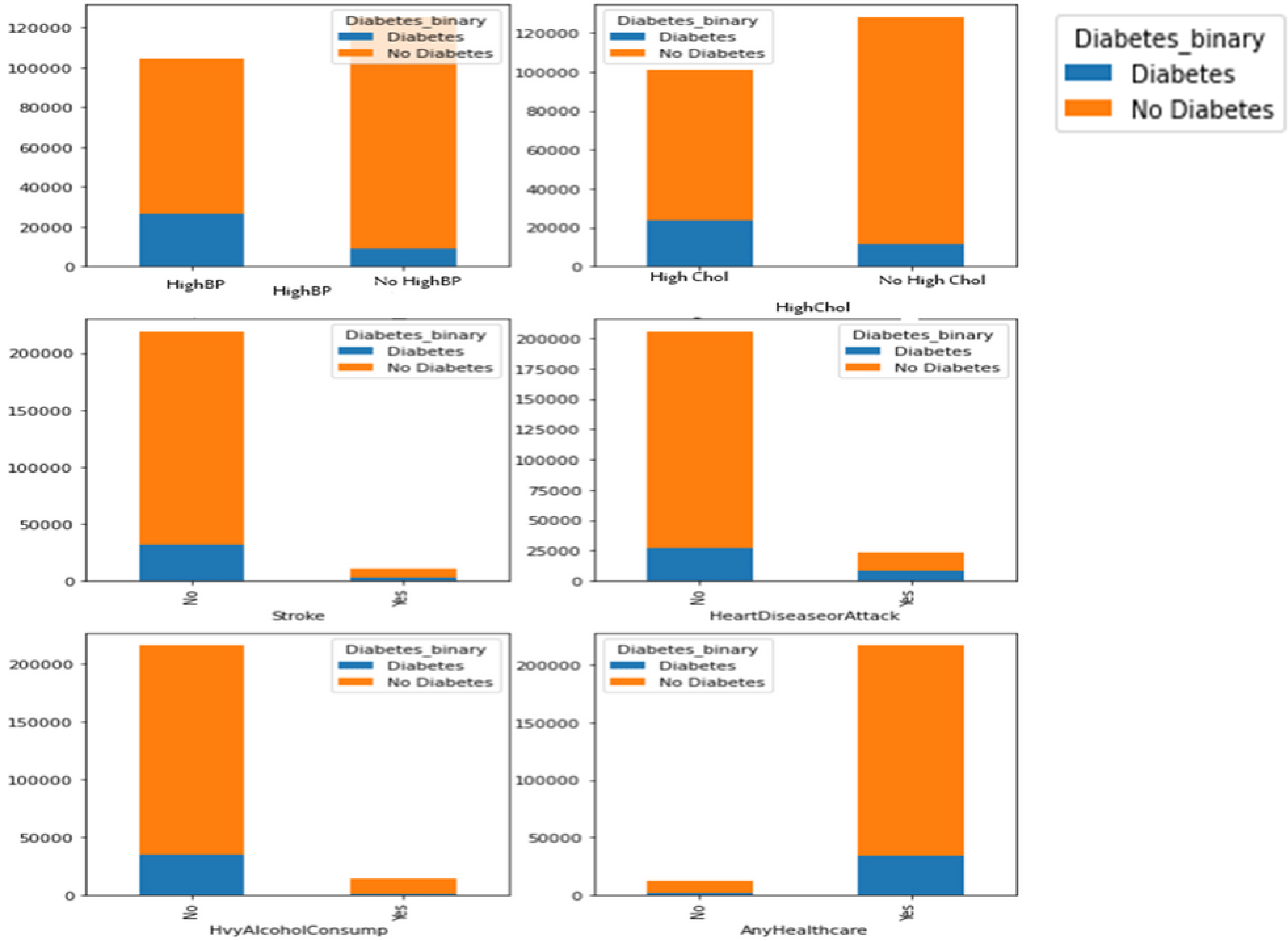
we can see here the value counts of all 22 columns some of them columns are continuous columns and some of them are discrete columns, here is the Frequency of values in different columns.

EXPLORATORY DATA ANALYSIS

Distribution of categorial features

People with HighBP and HighChol are diabetics.

In Other columns, according to the percentage of yes and no in each column, it is a normal amount.



SAMSUNG

EXPLORATORY DATA ANALYSIS

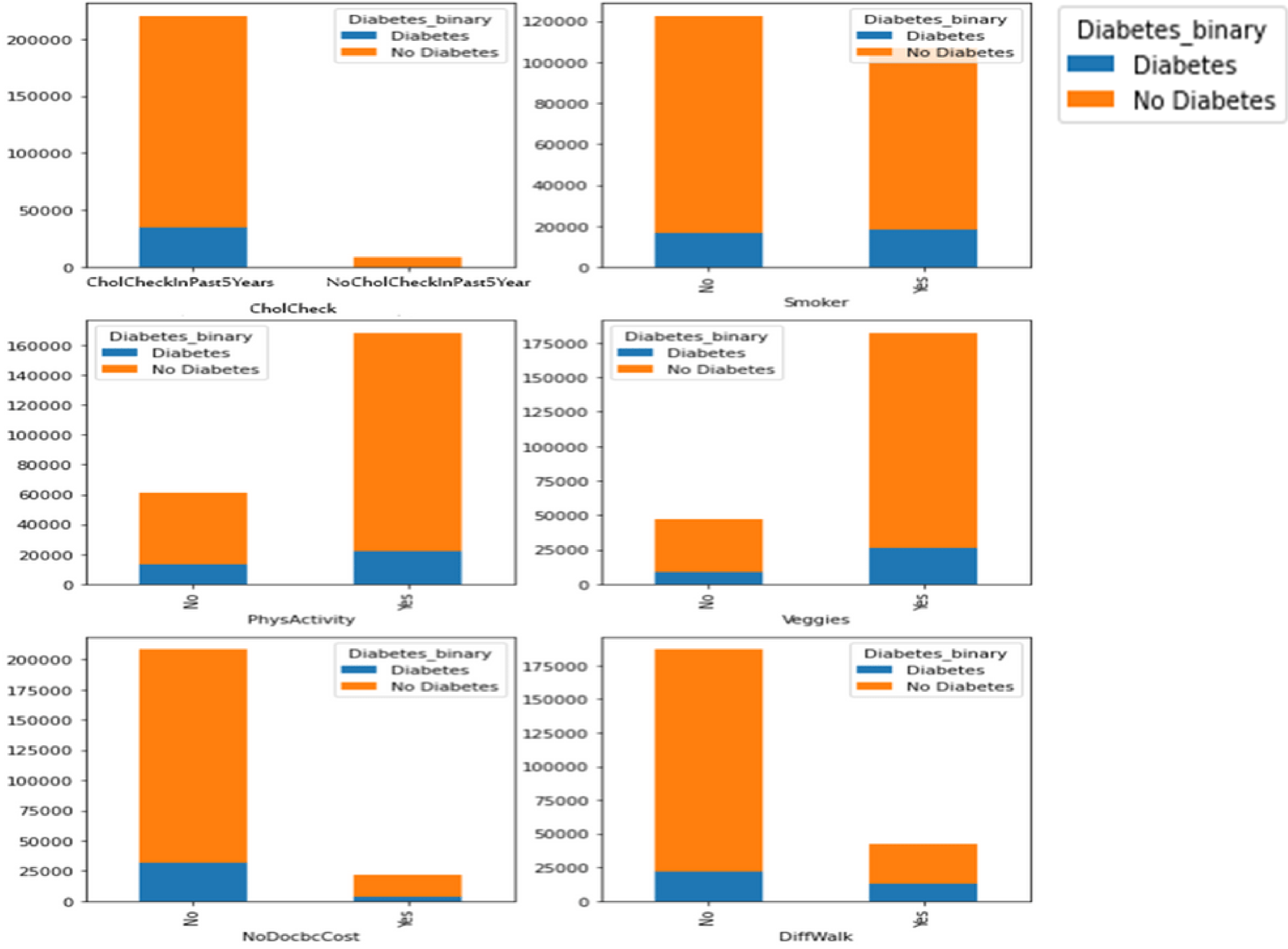
Distribution of categorial features

People who have cholesterol in the past 5 years are more likely to get diabetes

Smoking doesn’t affect diabetics

People who reported physical activity in the past 30 days are more likely to get diabetes, But considering the proportions it is normal

Considering the proportions of the other columns it is normal



SAMSUNG

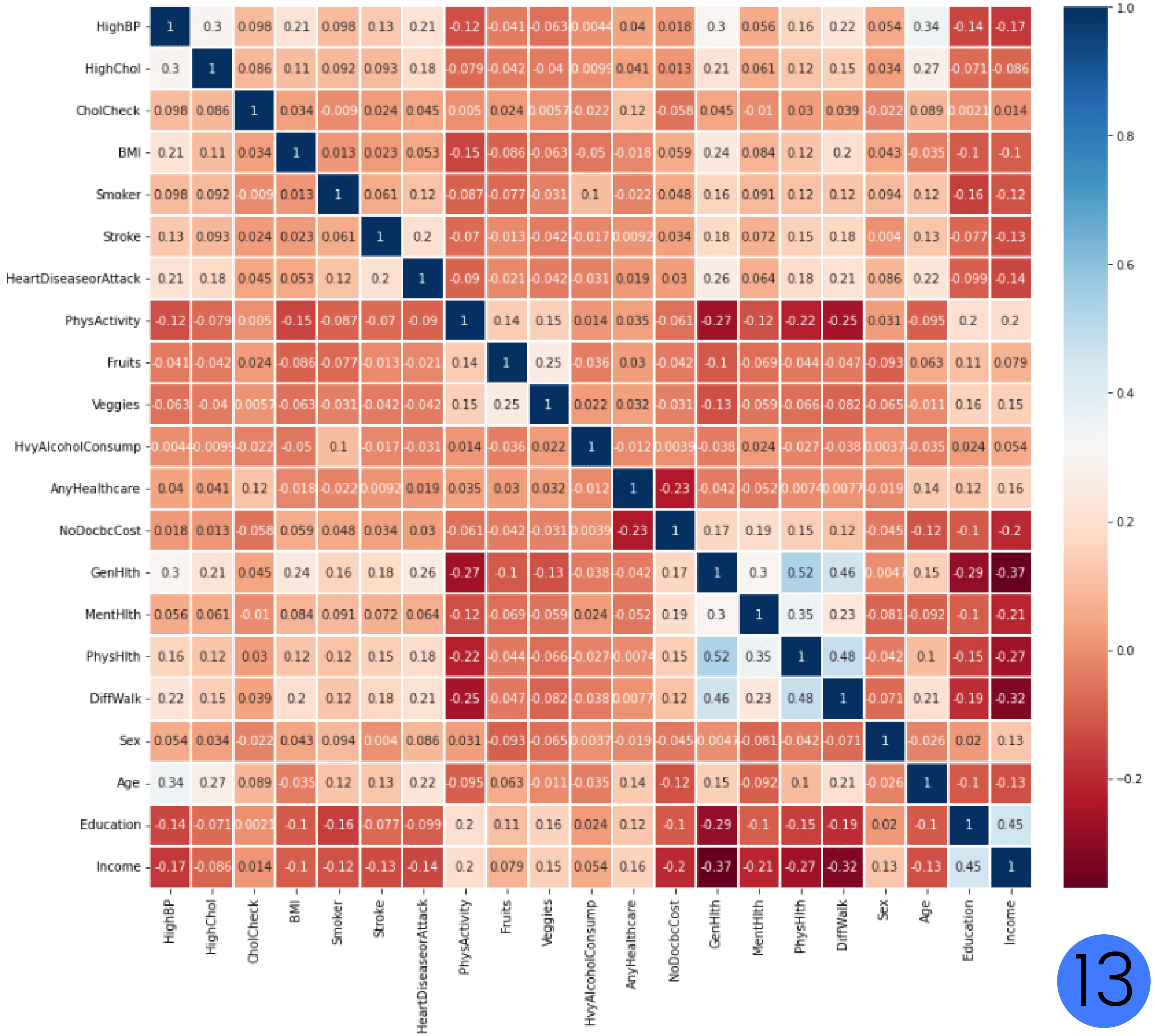
EXPLORATORY DATA ANALYSIS

Correlation of features

Correlation heatmap shows the relation between columns:

(GenHlth ,PhysHlth),(PhysHlth, DiffWalk),(GenHlth, DiffWalk)are highly correlated with each other
=> positive relation

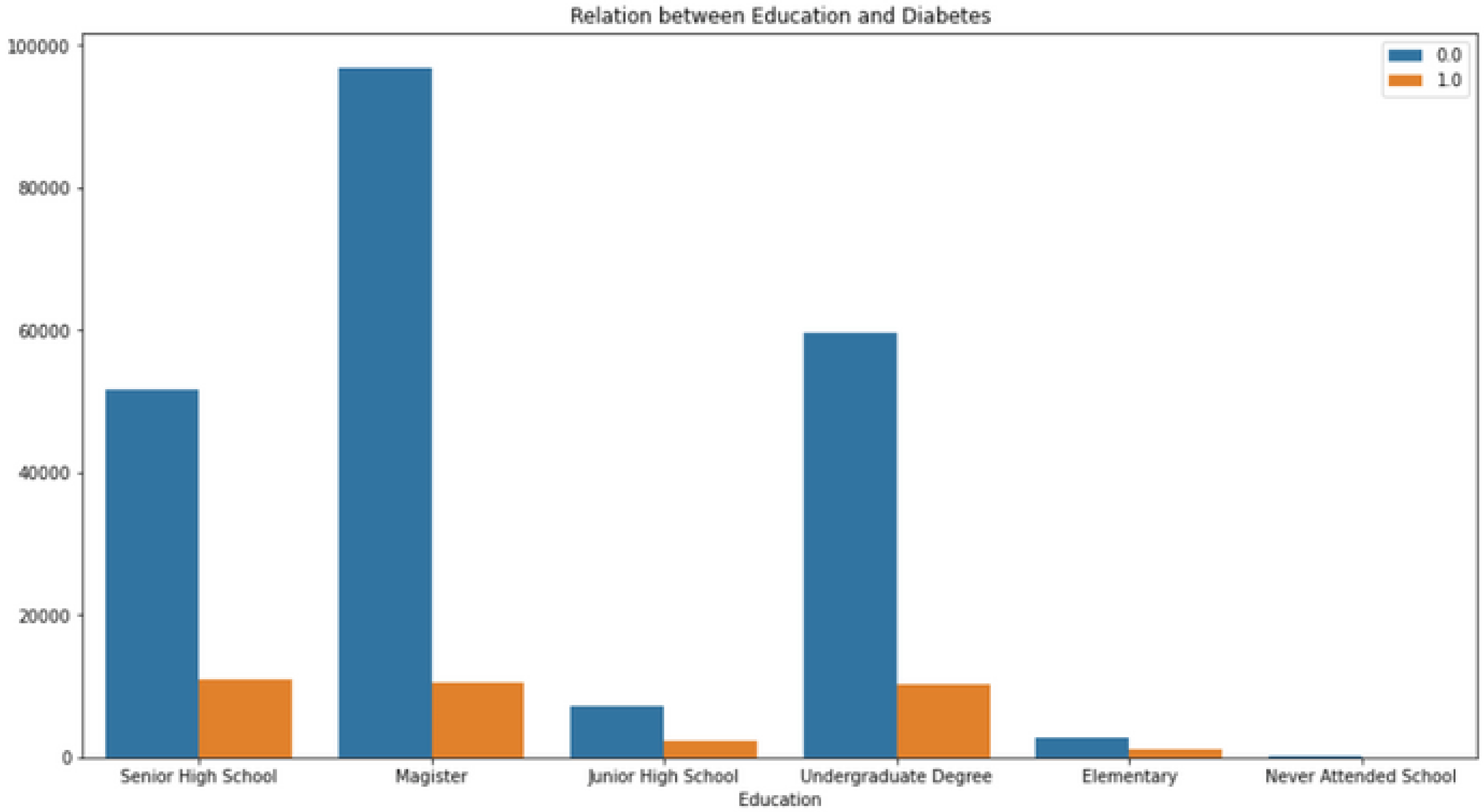
(GenHlth ,Income) , (DiffWalk, Income) are highly correlated with each other
=> Negative relation



EXPLORATORY DATA ANALYSIS

Relation between Education and diabetes

We can see that most people who have high level of education, healthy people are more than others

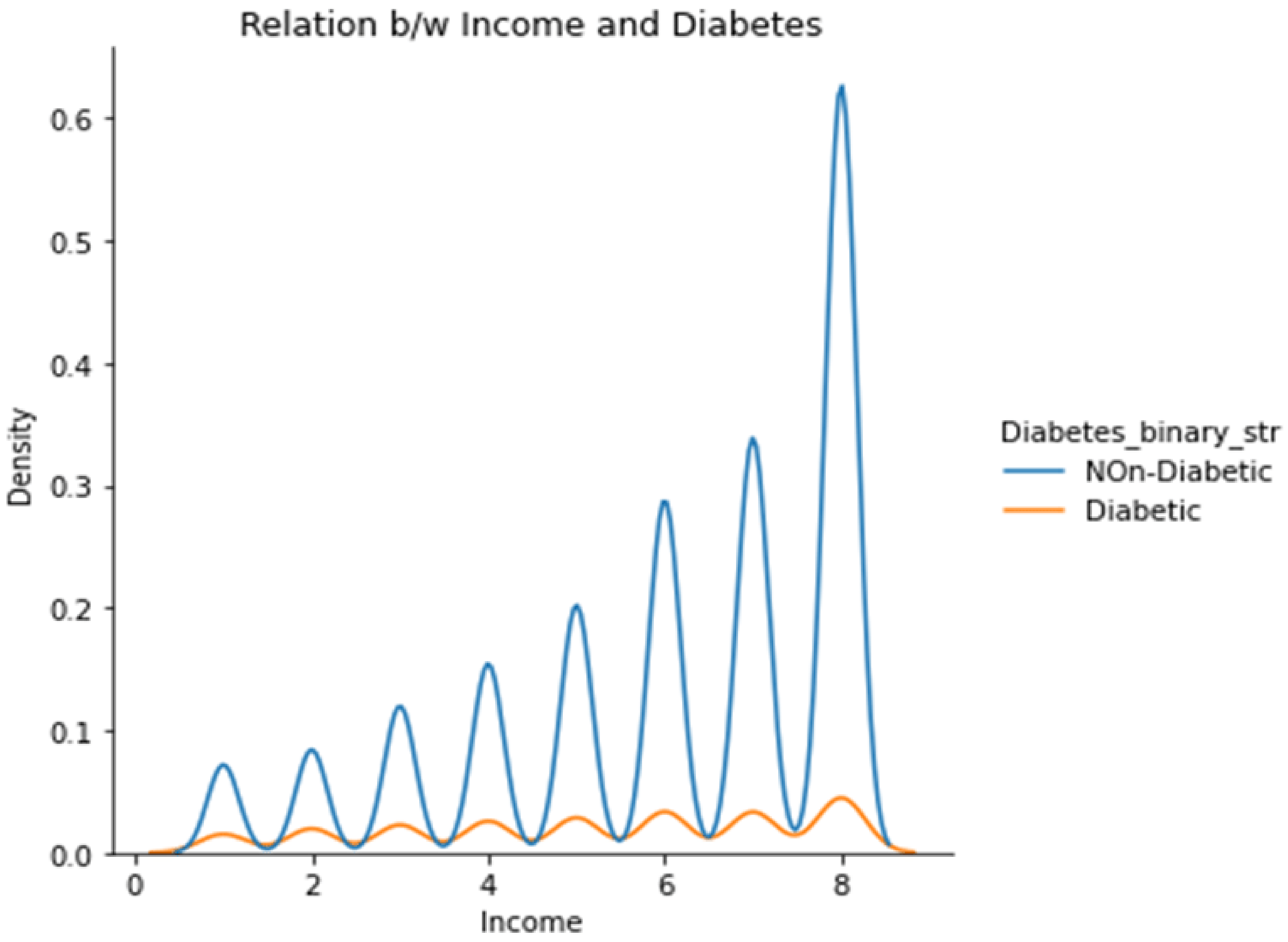


SAMSUNG

EXPLORATORY DATA ANALYSIS

Relation between Income and diabetes

We can see that most of people have high income and in the high level of income, healthy people are more than others

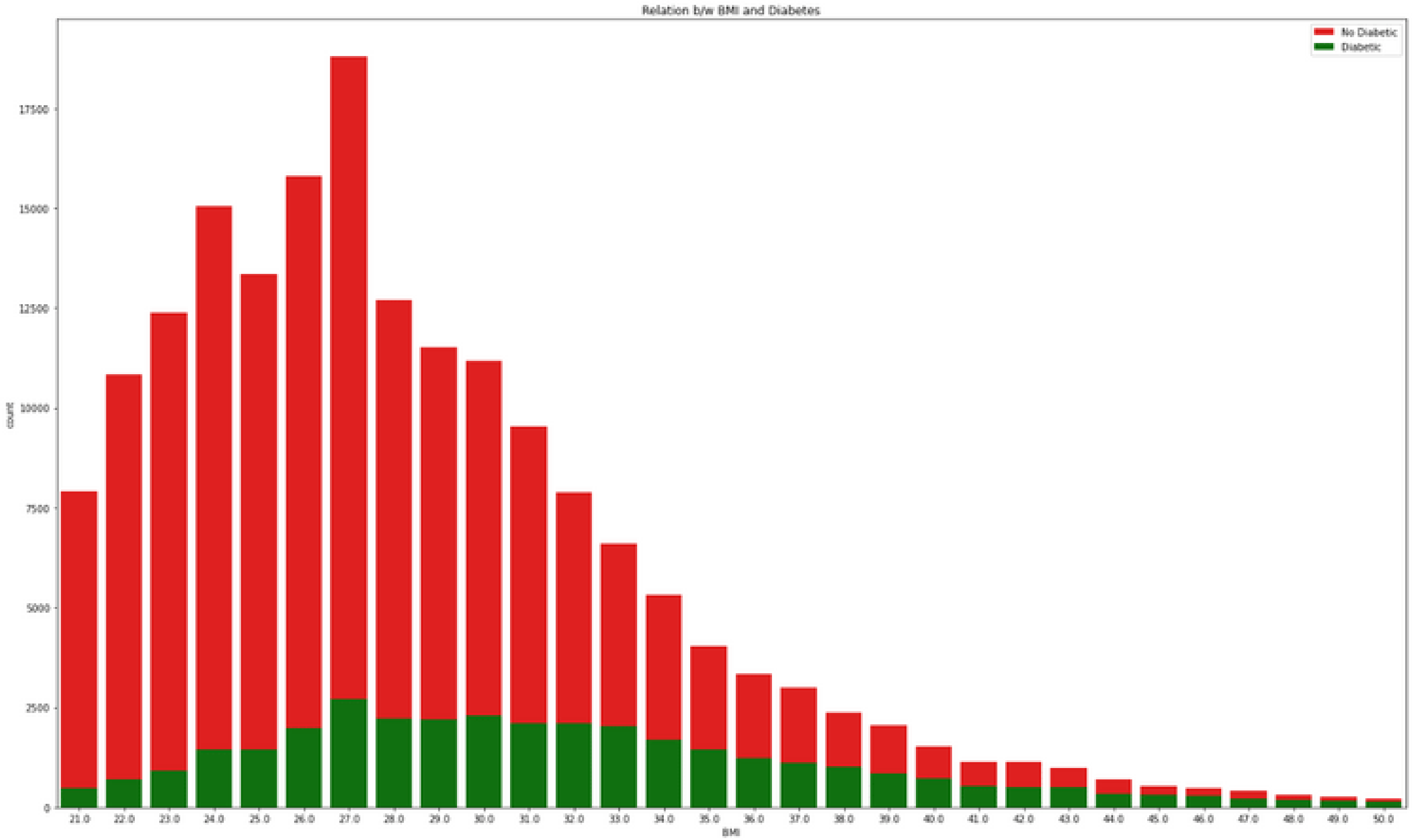


SAMSUNG

EXPLORATORY DATA ANALYSIS

Relation between BMI and diabetes

As we can see people range between 24-33 BMI have more likely to have Diabetic



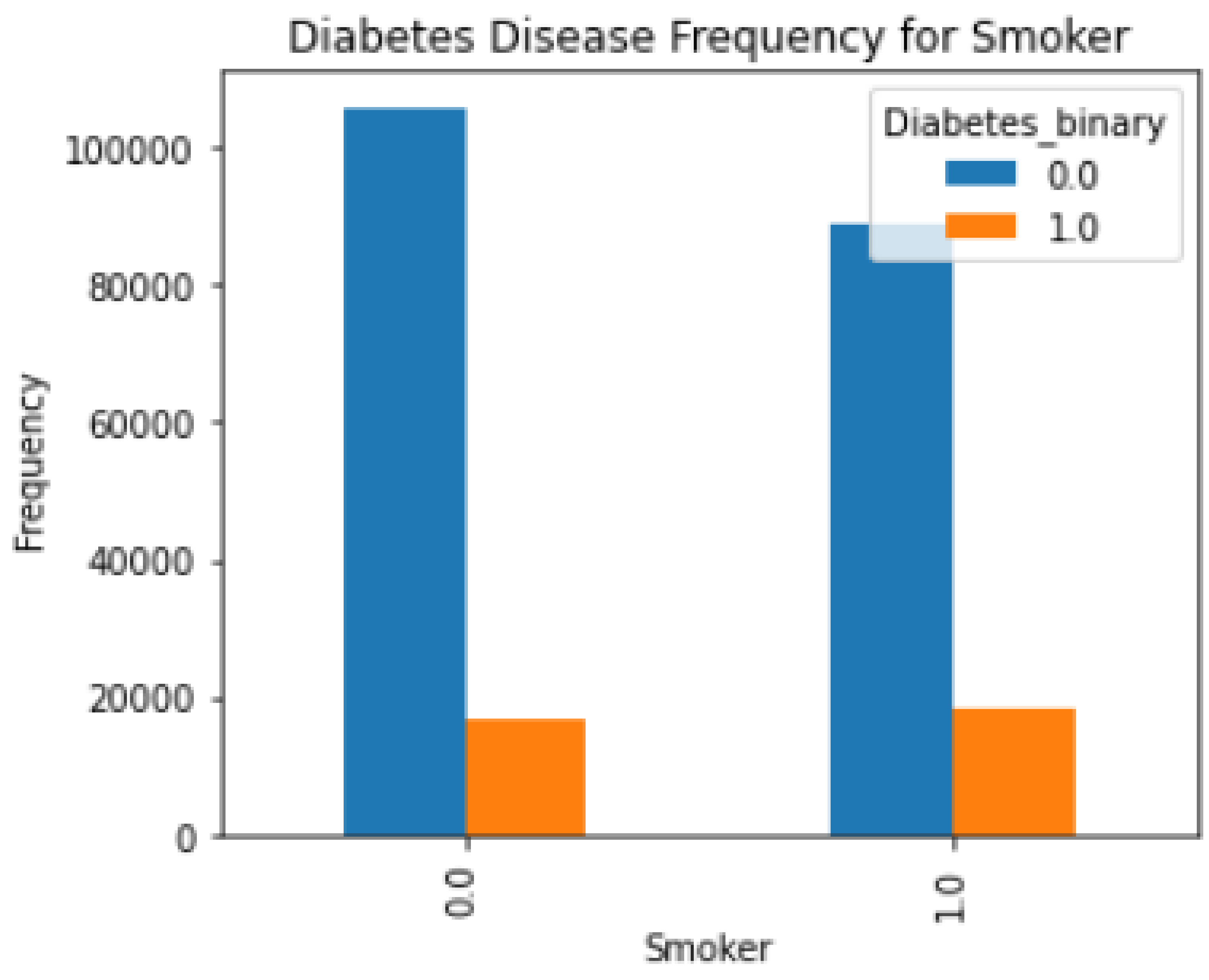
SAMSUNG

EXPLORATORY DATA ANALYSIS

Relation between Smoking and diabetes

Acording to this data, Only smoking has a minor effect on diabetes

Smoking is injurious to health



SAMSUNG

EXPLORATORY DATA ANALYSIS

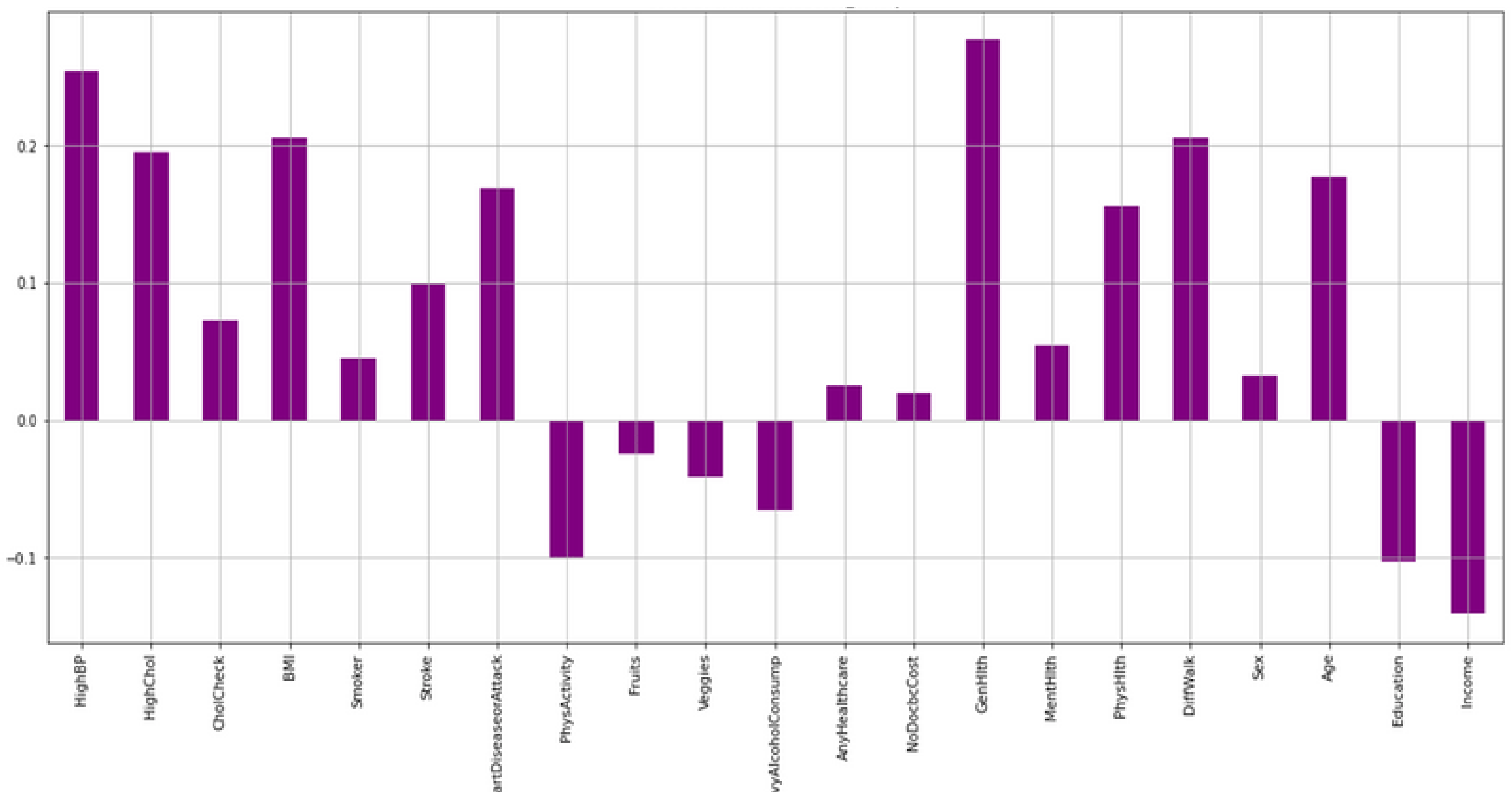
Business solution

- Attention to education among the different strata of society
- Working to raise the income and standard of living of citizens
- Spreading awareness among people of the need to pay attention to public health and physical health
- Establishing deterrent regulations for smoking and smokers and spreading awareness of its negative effects
- Taking care of the elderly and providing them with all their medical needs
- Providing means of treatment and providing permanent examinations and conducting therapeutic analyzes at close intervals for the general public

SAMSUNG

DATA PREPROCESSING

Feature Extraction & Feature Selection



Fruits, AnyHealthcare, NoDocbccost, and sex are least correlated with Diabetes binary

HighBP, HighChol, BMI, smoker, stroke, HeartDiseaseorAttack, PhysActivity, Veggies, MentHlth, HvyAlcoholconsump, GenHlth, PhysHlth, Age, Education, Income and DiffWalk have a significant correlation with Diabetes binary

"Fruits" , "Veggies" , "Sex" , "CholCheck" , " AnyHealthcare" will not be with us

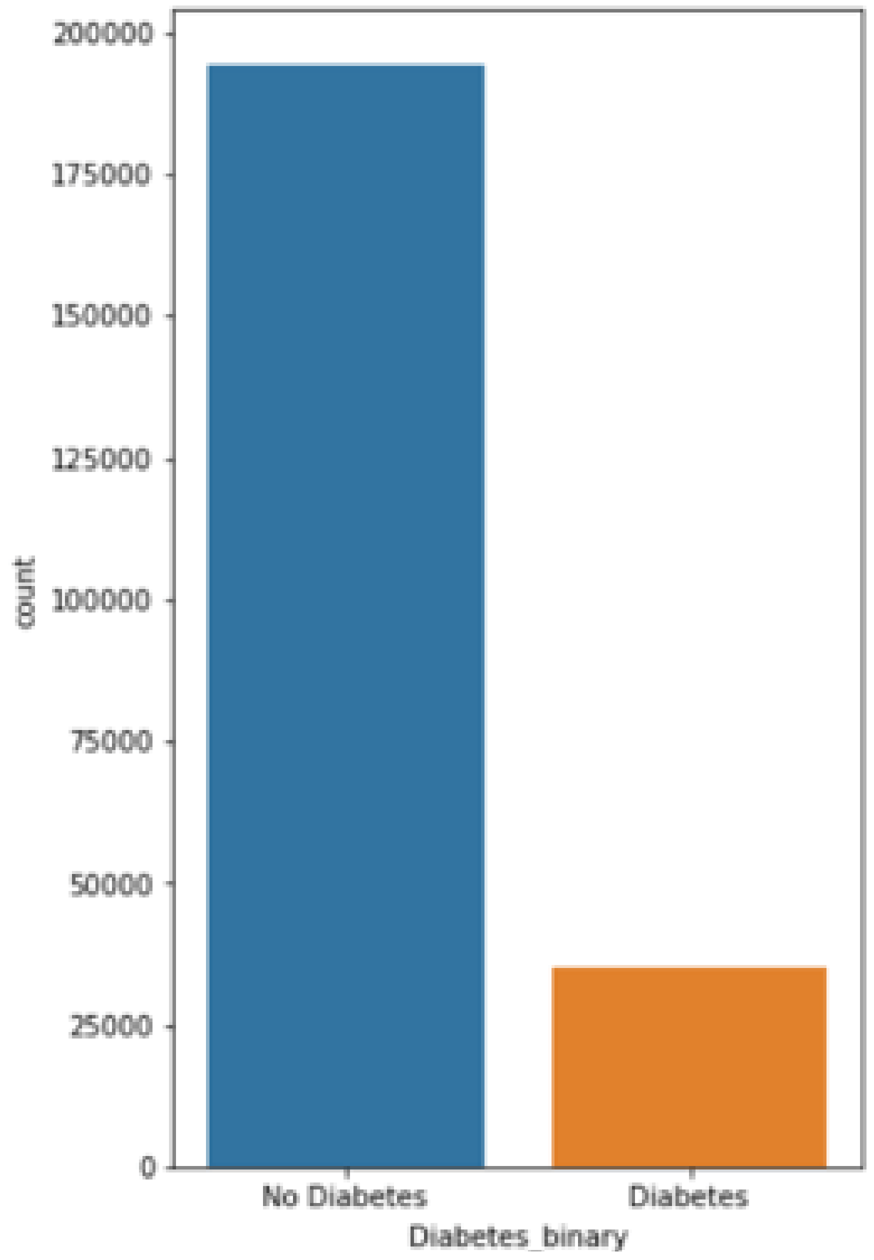
SAMSUNG

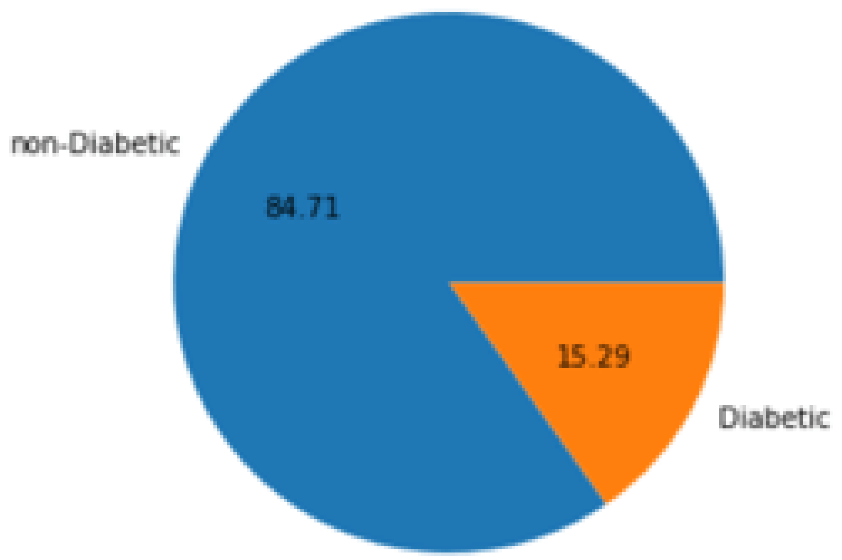
DATA PREPROCESSING

Handling Imbalanced

- ❑ **First** : we check the number of Diabetics and Nondiabetics in our target column
  - 0 means non-diabetics
  - 1 means diabetics

```
0    194377
1     35097
```





SAMSUNG

DATA PREPROCESSING

Handling Imbalanced

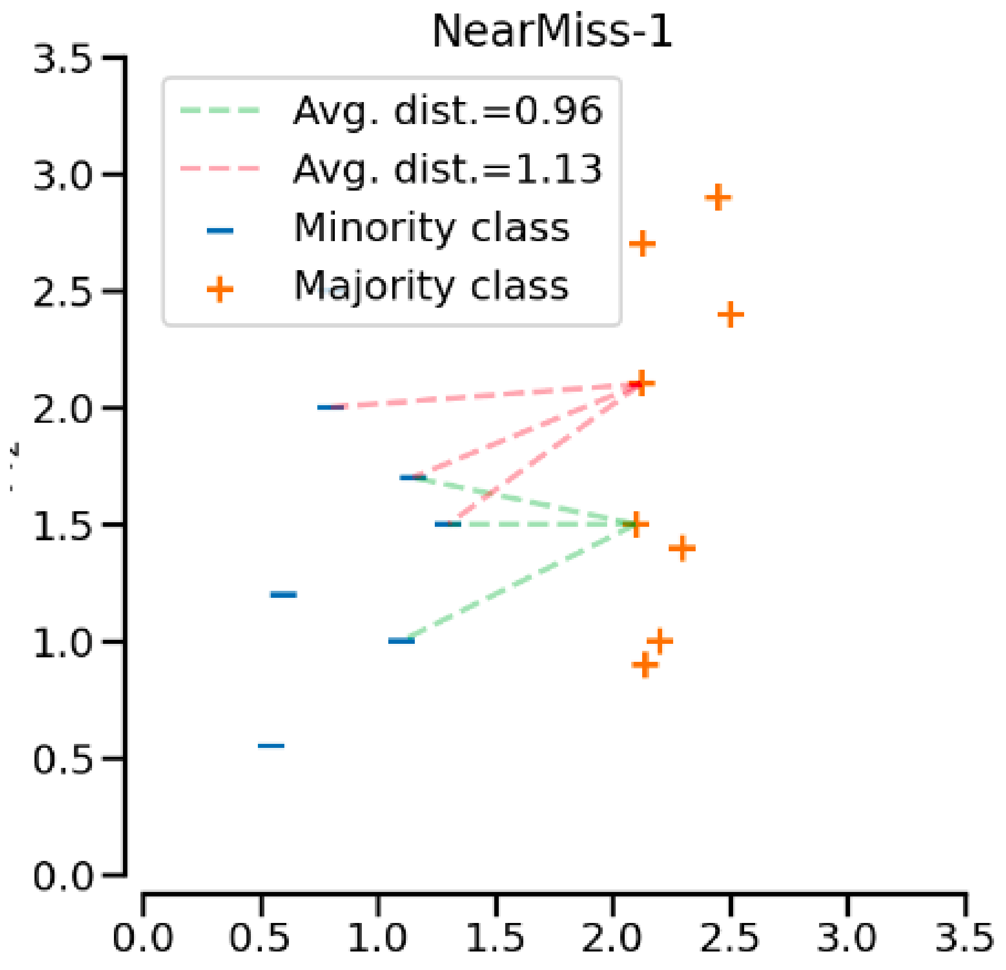
- **Second** :We apply under sample technic called “NearMiss”
  - 0 means non-diabetics
  - 1 means diabetics

```
0     35097
1     35097
```

NearMiss has 3 versions and we used version 1

n_neighbors refer to the size of the neighborhood to consider to compute the average distance to the minority point samples.

NearMiss-1 selects samples from the majority class for which the average distance of the k nearest samples of the minority class is the smallest.
NearMiss-2 selects the samples from the majority class for which the average distance to the farthest samples of the negative class is the smallest.
NearMiss-3 is a 2-step algorithm: first, for each minority sample, their m nearest-neighbors will be kept; then, the majority samples selected are the on for which the average distance to the k nearest neighbors is the largest.



SAMSUNG

DATA MODELING

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.8512	0.8472
Decision Tree	0.8657	0.8475
KNN	0.8424	0.8050
Random Forest	0.8713	0.8588
SVM	0.8687	0.8603
XGBoost	0.8770	0.8663

MODEL EVALUATION

Logistic Regrssion

	precision	recall	f1-score	support
0	0.80	0.93	0.86	10468
1	0.92	0.76	0.83	10591

Decision Tree

	precision	recall	f1-score	support
0	0.78	0.96	0.86	10468
1	0.95	0.74	0.83	10591

SAMSUNG

MODEL EVALUATION

K Neighbors

	precision	recall	f1-score	support
0	0.73	0.95	0.83	10468
1	0.93	0.66	0.77	10591

Random Forest

	precision	recall	f1-score	support
0	0.80	0.95	0.87	10468
1	0.94	0.77	0.85	10591

SAMSUNG

MODEL EVALUATION

SVM

	precision	recall	f1-score	support
0	0.80	0.96	0.87	10468
1	0.95	0.76	0.85	10591

XGBoost

	precision	recall	f1-score	support
0	0.81	0.95	0.88	10468
1	0.94	0.79	0.86	10591

SAMSUNG

CONCLUSION

Various machine learning algorithms are explored and compared to predict diabetes to further assist the medical healthcare sector. The highest accuracy of the machine learning algorithm model is the XGBoost with 86.63% in predicting diabetes based on health indicators

THANK YOU!