DATA SCIENCE PROJECT 1

Diabetes Survey OPTIMIZATION

EDA, Data Preprocessing, Lazy Classification Assessment, Machine Learning Modelling (XGBC, LGBMC, SVC, Random Tree Regressor)

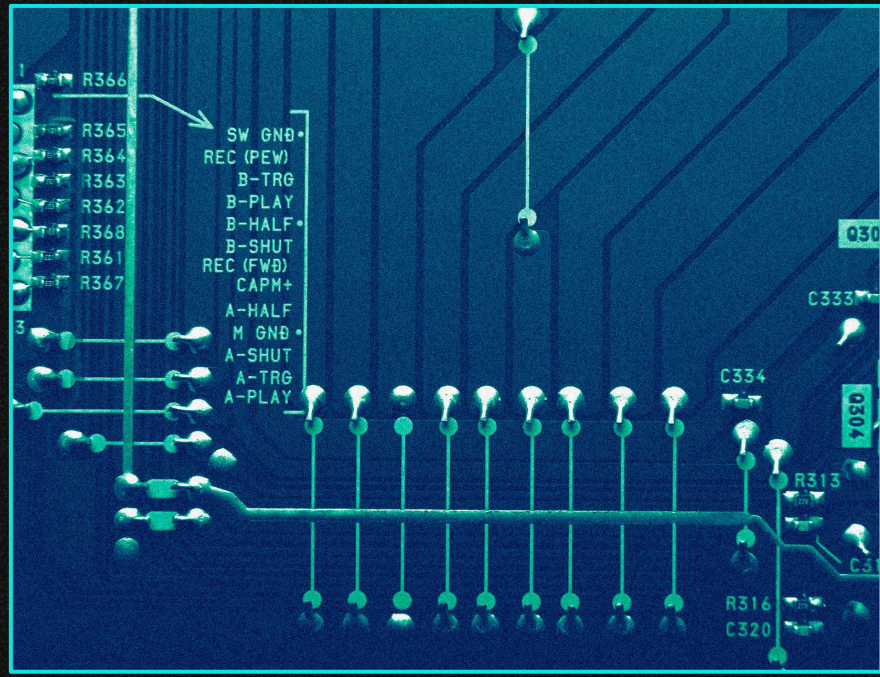
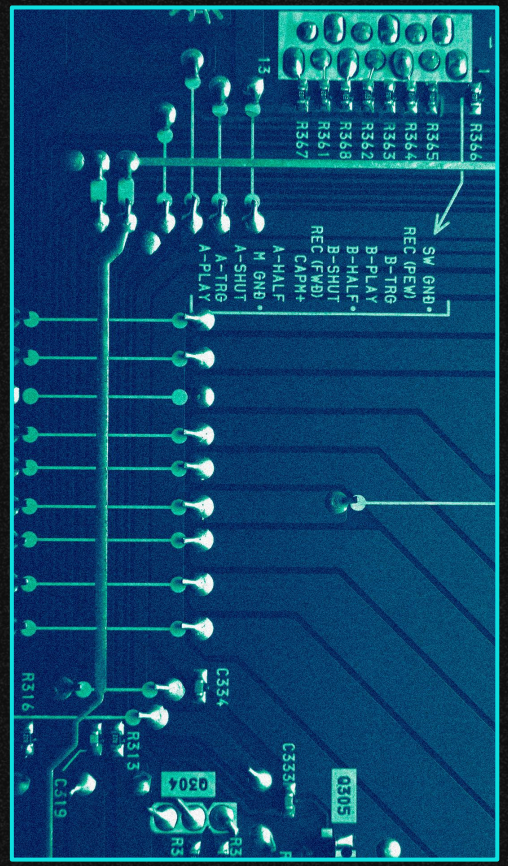
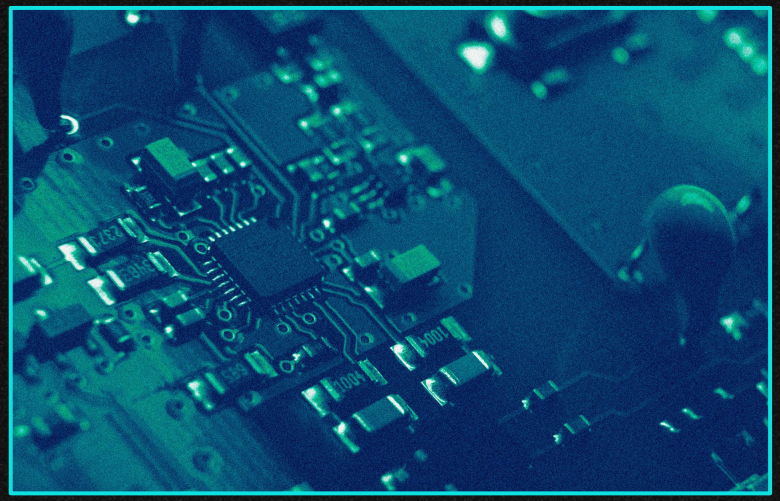


Table of contents





Introduction

Diabetes survey dataset is a Diabetes Health Indicator dataset obtained from Kaggle.

The dataset contains a result from Behavioral Risk Factor Surveillance System (BRFSS) survey about diabetes. BRFSS survey is a health-related telephone survey that is collected annually by the CDC.

For this project, a csv of the dataset available on Kaggle for the year 2015 was used.

What do we want to know?

Accuracy

Can survey questions from the BRFSS provide accurate predictions of whether an individual has diabetes?

Risk Factor

What risk factors are most predictive of diabetes risk?

Efficiency

Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?

Shorter Form

Can we create a short form of questions from the BRFSS using feature selection to accurately predict if someone might have diabetes or is at high risk of diabetes?

General Dataset Information

Import the .csv as a dataframe named diabetes_1

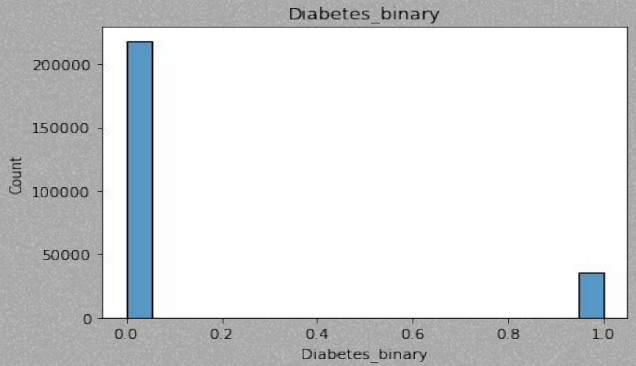
Examine the data head(), describe(), info(). There are around 253680 float data and 21 columns. The data shall be converted into integers.

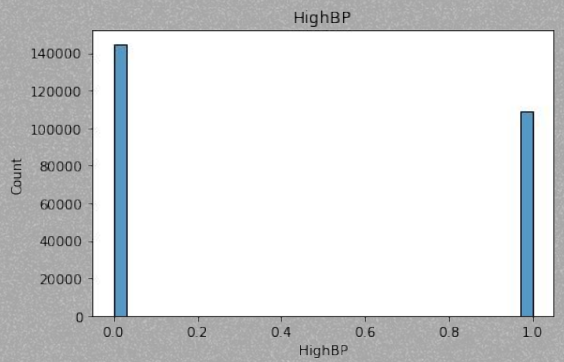
Examine the amount of null data. There are no null data.

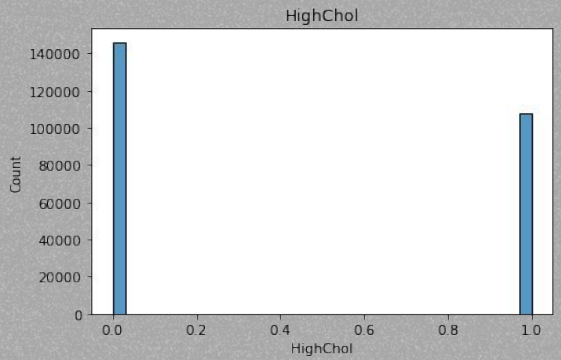
Examine the amount of duplicates and dropped them.

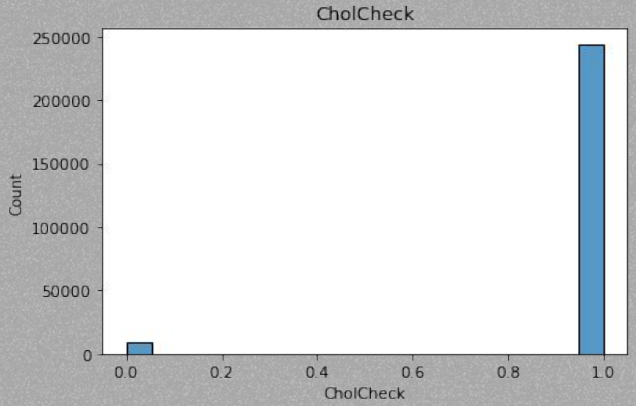
```
In [4]: print(diabetes_1.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          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)
memory usage: 42.6 MB
None
```

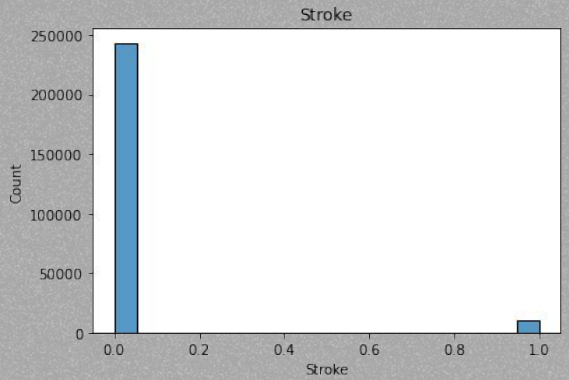
General Dataset Information

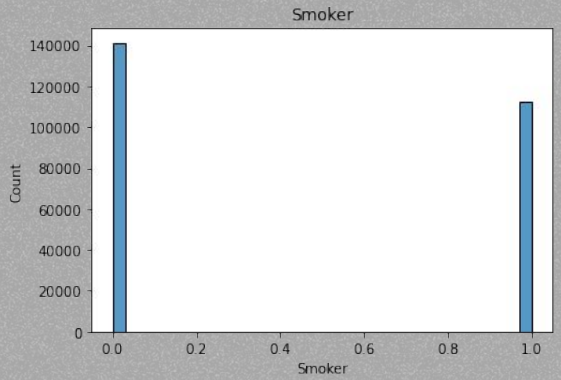




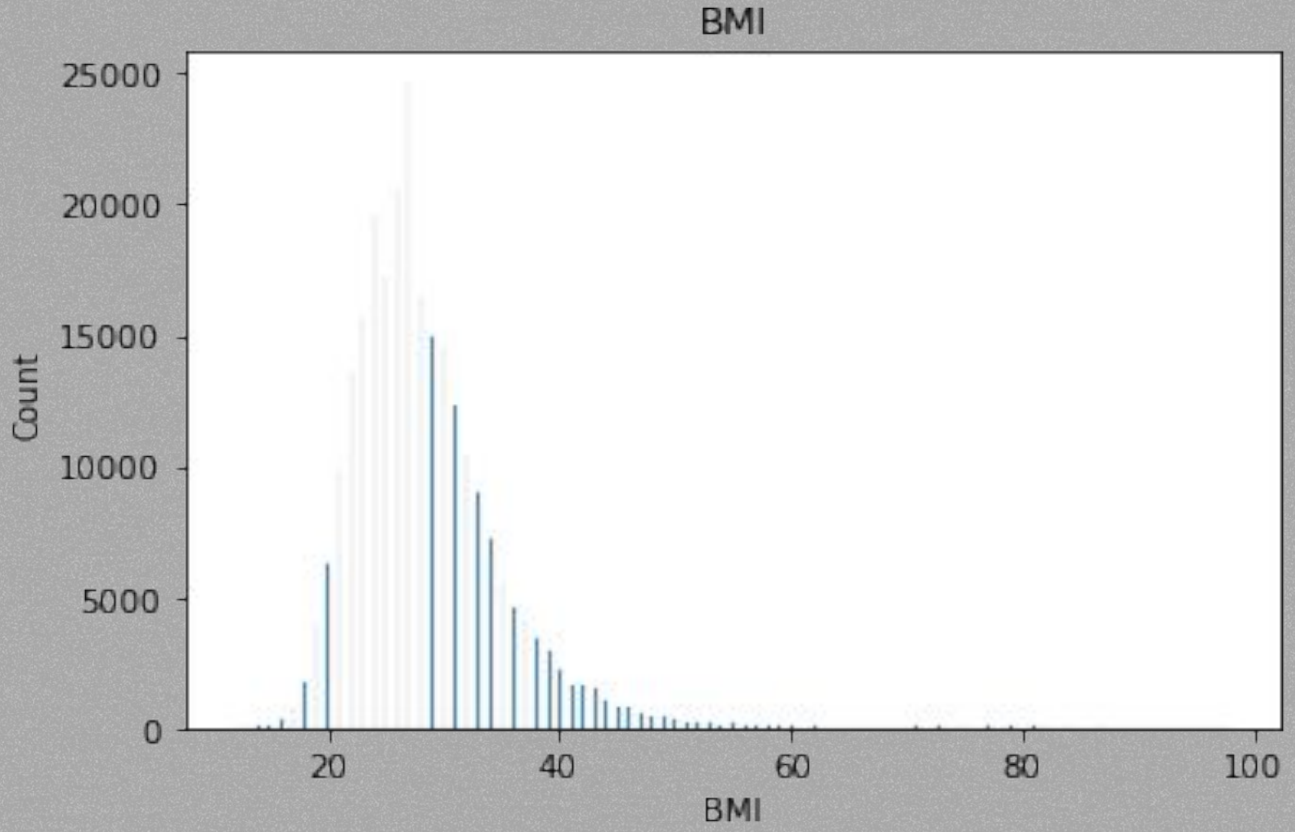




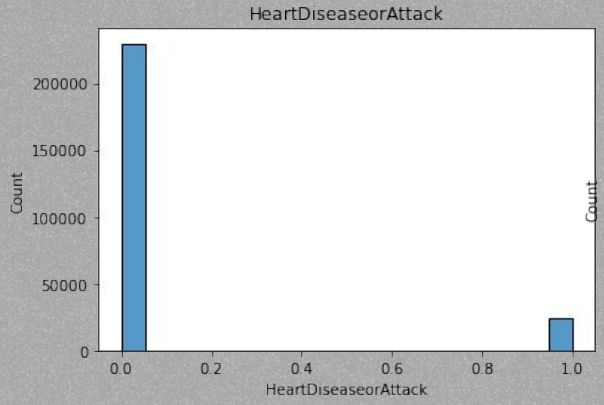


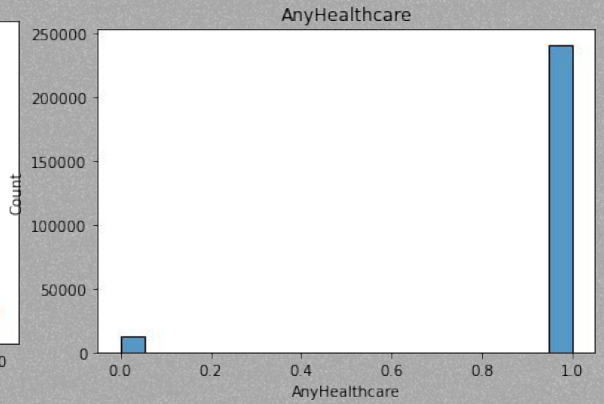


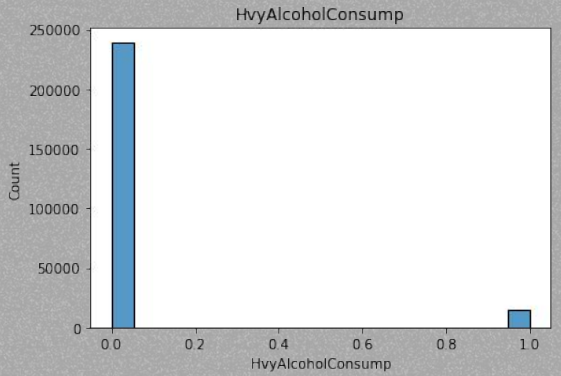
General Dataset Information

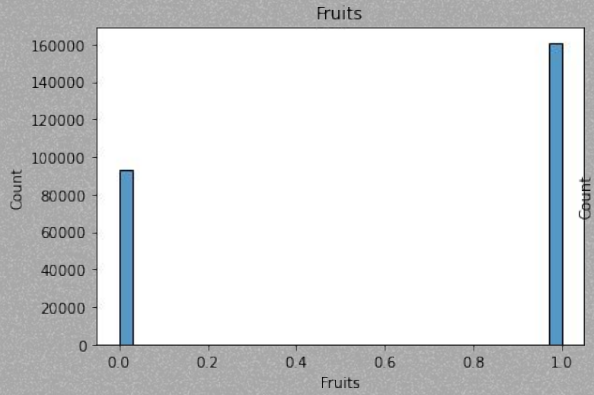


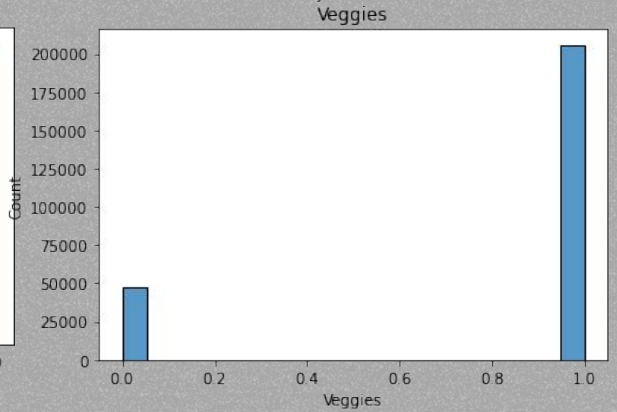
General Dataset Information

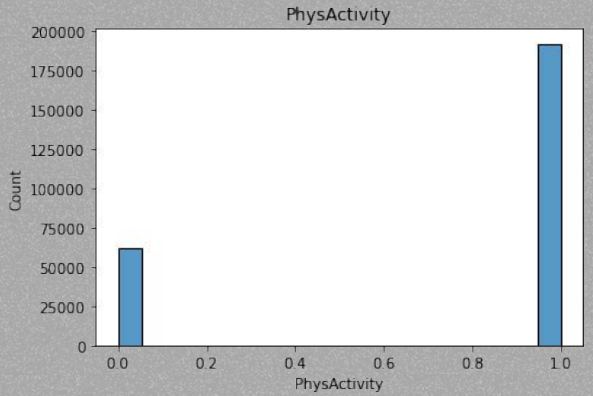




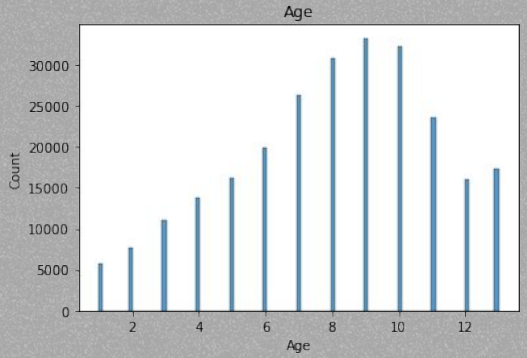


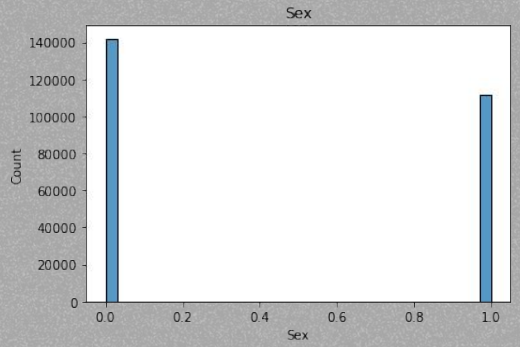


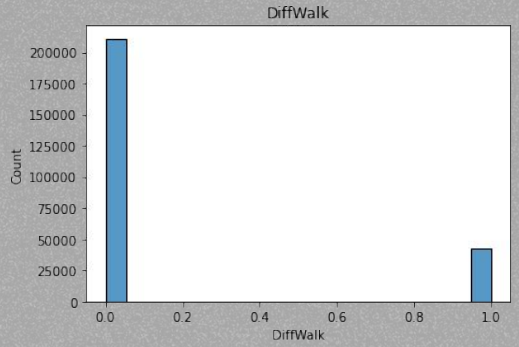


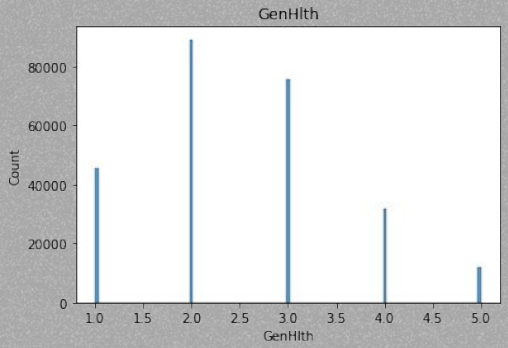


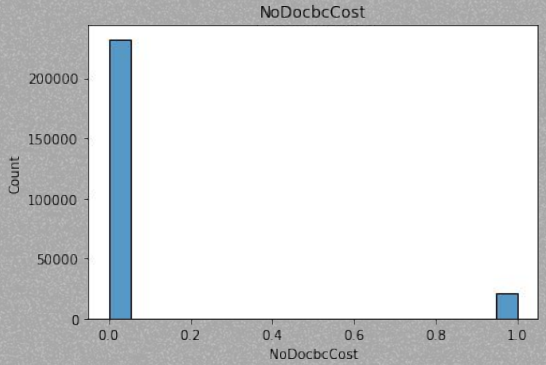
General Dataset Information

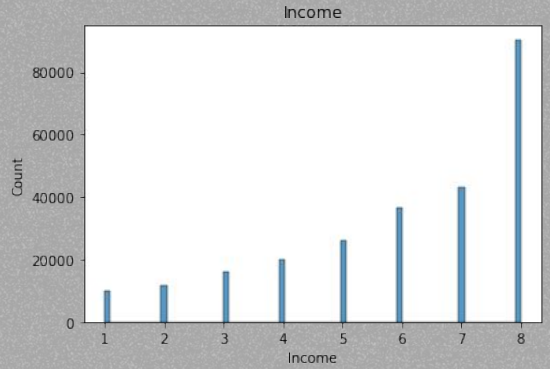




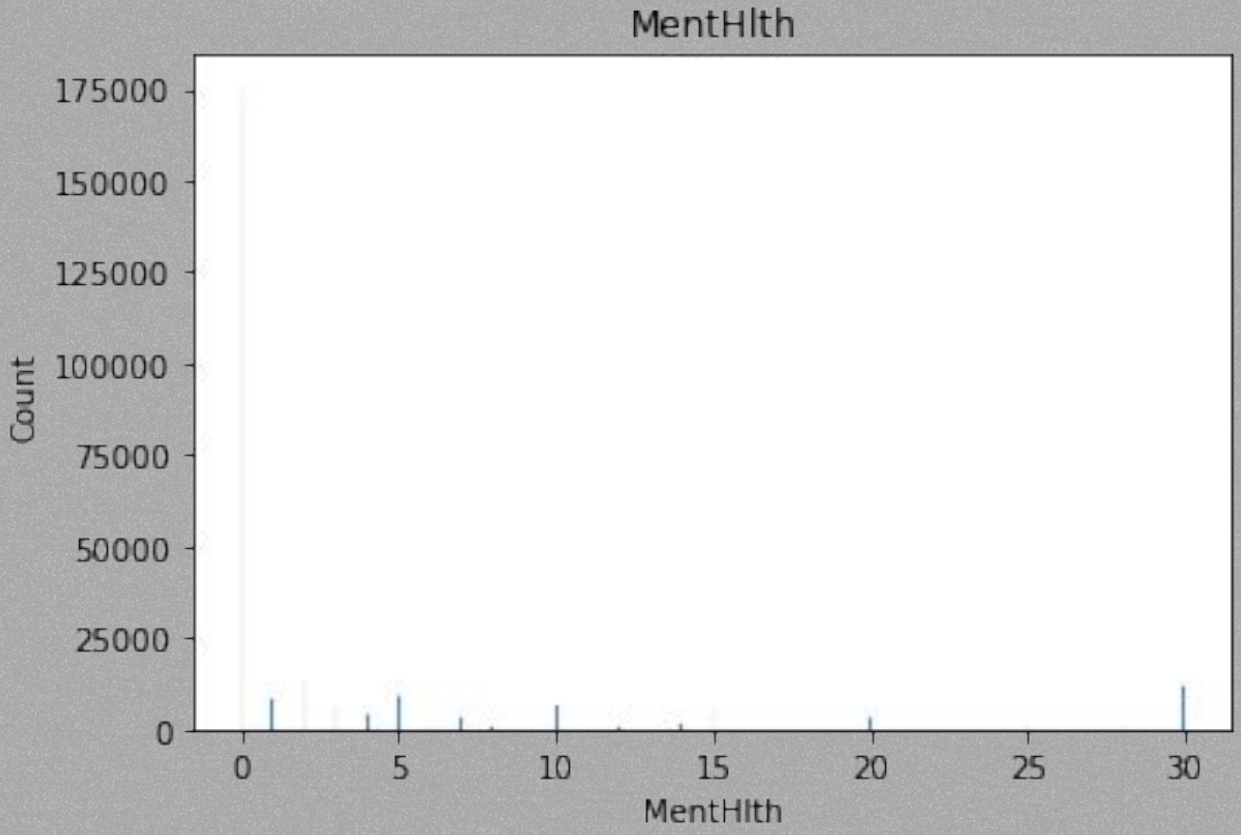




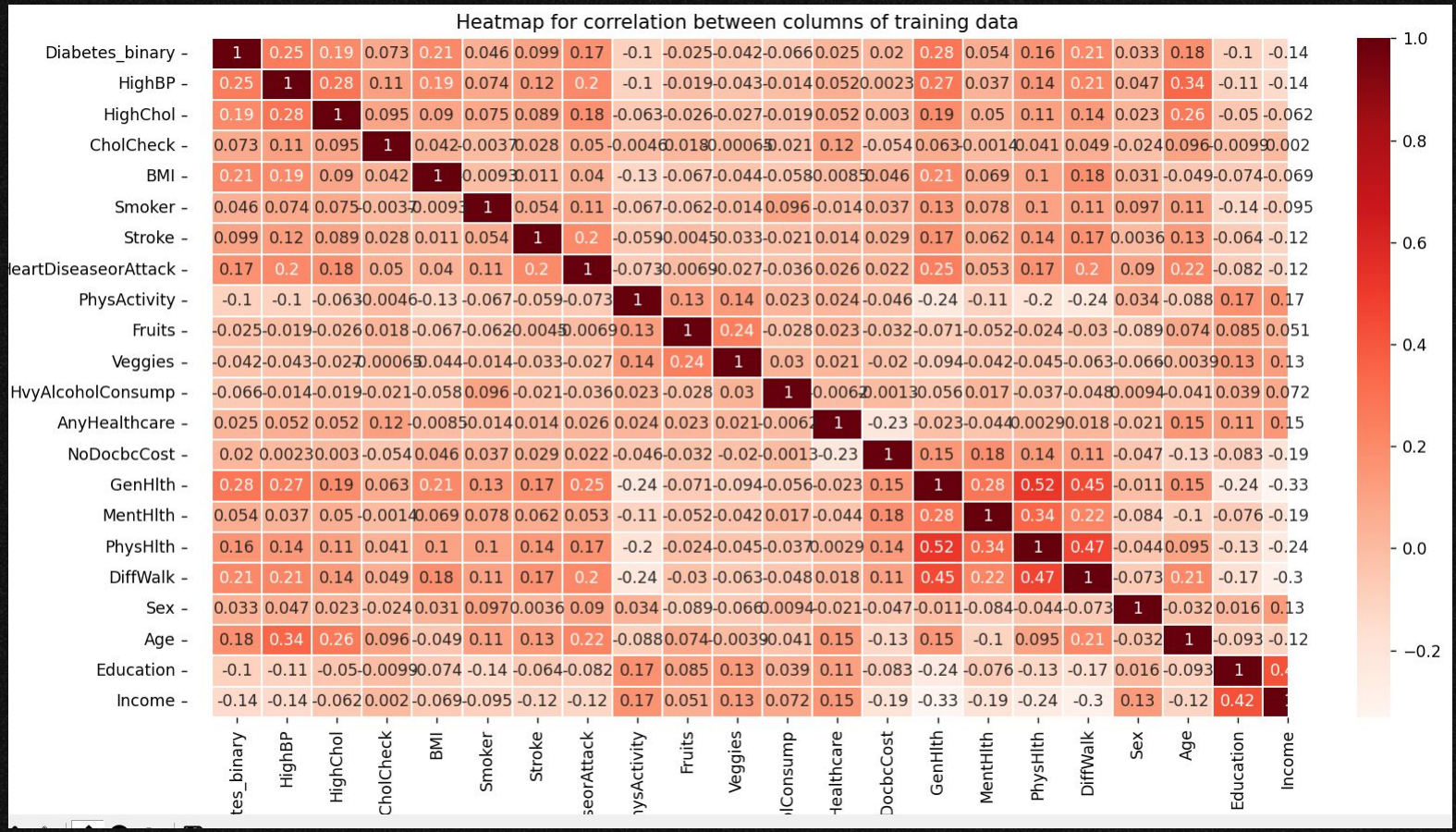




General Dataset Information



Correlations Between Variables



None

	Feature	Score
0	HighBP	8098.55
1	HighChol	4869.31
2	CholCheck	48.90
3	BMI	15507.74
4	Smoker	253.83
5	Stroke	2156.68
6	HeartDiseaseorAttack	5822.15
7	PhysActivity	617.56
8	Fruits	54.69
9	Veggies	82.10
10	HvyAlcoholConsump	937.40
11	AnyHealthcare	7.95
12	NoDocbcCost	83.66
13	GenHlth	7671.73
14	MentHlth	11419.58
15	PhysHlth	97988.76
16	DiffWalk	7875.50
17	Sex	137.84
18	Age	8539.91
19	Education	479.11
20	Income	3377.10

Correlations Between Variables

Based on the hotmap analysis :

- The most correlated parameter with whether or not the person has diabetes is general health, high BP, BMI, diffwalk, and high cholesterol.
- While parameters with the lowest correlation (under 0.05) are fruits, veggies, sex, smoker, and NoDocbcCost.

Based on the Chi Square Test:

- Parameter with highest significance are PhysHlth, BMI, MentHlth, Age, and HighBP.
- Parameter with lowest significance are anyhealthcare, cholcheck, fruits, veggies, NoDocbcCost, Sex, and Smoker.

Based on this we will remove ”Fruits” , ”Veggies” , ”Sex” , ”CholCheck” , ”AnyHealthcare” , ”NoDocbcCost”, and “Smoker” from the dataset.

Data Preprocessing and Lazy Classification

Fix the imbalance in the dataset and scale the dataset using StandardScaler.

```
Name: Diabetes_binary, dtype: int64
100%|██████████| 29/29 [07:23<00:00, 15.30s/it]
```

Model	Accuracy	Balanced Accuracy	...	F1 Score	Time Taken
			...		
LGBMClassifier	0.87	0.87	...	0.87	0.26
XGBClassifier	0.87	0.87	...	0.87	1.08
SVC	0.87	0.87	...	0.86	112.31
AdaBoostClassifier	0.86	0.86	...	0.86	1.80
LogisticRegression	0.85	0.85	...	0.85	0.16
SGDClassifier	0.85	0.85	...	0.84	0.29
CalibratedClassifierCV	0.84	0.84	...	0.84	31.99
LinearSVC	0.84	0.84	...	0.84	9.46
RandomForestClassifier	0.84	0.84	...	0.84	4.99
ExtraTreesClassifier	0.84	0.84	...	0.84	5.37

Machine Learning Model

```
Training set score: 0.8931
Test set score: 0.8637
Mean Squared Error : 0.13633449675902842
Root Mean Squared Error : 0.36923501561881755
              precision    recall  f1-score   support

           0       0.80      0.97      0.88      7012
           1       0.96      0.76      0.85      7027

    accuracy                           0.86     14039
   macro avg       0.88      0.86      0.86     14039
weighted avg       0.88      0.86      0.86     14039
```

Random Forest Regression

```
Training set score: 0.8963
Test set score: 0.8722
Mean Squared Error : 0.1277868794073652
Root Mean Squared Error : 0.3574729072354508
              precision    recall  f1-score   support

           0       0.82      0.96      0.88      7012
           1       0.95      0.79      0.86      7027

    accuracy                           0.87     14039
   macro avg       0.88      0.87      0.87     14039
weighted avg       0.88      0.87      0.87     14039
```

LGBMC

Machine Learning Model

```
Training set score: 0.8843
Test set score: 0.8726
Mean Squared Error : 0.13633449675902842
Root Mean Squared Error : 0.3568746257998487
              precision    recall  f1-score   support

           0       0.82      0.95      0.88      7012
           1       0.94      0.79      0.86      7027

    accuracy                           0.87     14039
   macro avg       0.88      0.87      0.87     14039
weighted avg       0.88      0.87      0.87     14039
```

XGBC

```
Training set score: 0.8492
Test set score: 0.8458
Mean Squared Error : 0.15418585877771973
Root Mean Squared Error : 0.39266507201140227
              precision    recall  f1-score   support

           0       0.78      0.96      0.86     10468
           1       0.95      0.73      0.83     10591

    accuracy                           0.85     21059
   macro avg       0.86      0.85      0.84     21059
weighted avg       0.87      0.85      0.84     21059


Process finished with exit code 0
```

SVC

From these results, it can be concluded that LGBMC is the most optimal model, just like Lazy Classifier predictions. But, Random Forest Regression with hyperparameter tuning can be more optimal than SVC. This happens because I am more familiar with Random Forest Regression hyperparameters than SVC, so I can optimize the model better.

Conclusion

1. **Can survey questions from the BRFSS provide accurate predictions of whether an individual has diabetes?**
   By creating machine learning models without dropping any data, it is certain that the questions can provide accurate predictions with similar accuracy with dataset that contains the data with less correlation.

2. **What risk factors are most predictive of diabetes risk?**
   High blood pressure, BMI, high cholesterol, age,

3. **Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?**
   Yes because some risk factors have very low correlation to whether an individual has diabetes such as fruits, veggies, sex, smoker, and NoDocbcCost.

4. **Can we create a short form of questions from the BRFSS using feature selection to accurately predict if someone might have diabetes or is at high risk of diabetes?**
   Yes we can. We can drop and still obtain enough data to predict whether someone have diabetes or not by 88%.