csca5622-final-NR

May 21, 2024

1 Analysis of risk factors and their interactions in predicting diabetes and pre-diabetes

The Diabetes Health Indicators Dataset encompasses a comprehensive array of healthcare metrics and lifestyle-related survey data, broadly representing the populace in conjunction with their respective diabetes diagnoses. It is composed of 35 distinct attributes, which include demographic details, laboratory test outcomes, and responses to a series of survey inquiries pertinent to each subject. The primary variable designated for classification purposes delineates the health status of the subjects into three categories: diabetic, pre-diabetic, or non-diabetic.

Data Set Information: The dataset 'CDC Diabetes Health Indicators' [https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators] contains detailed information about lifestyle and diabetes in the United States. The dataset was created to explore the relationship between lifestyle factors and diabetes in the U.S. The data was collected by the Centers for Disease Control and Prevention (CDC) and funded by the CDC. The dataset contains 35 attributes including demographics, lab test results, and survey responses. The dataset is suitable for classification tasks and is available for use under specific licensing conditions and acknowledgment policies.

- **Dataset Overview**: A collection of healthcare statistics and lifestyle survey data related to diabetes diagnosis in the U.S., featuring 35 attributes including demographics, lab test results, and survey responses.
- **Purpose and Creation**: Created to explore the relationship between lifestyle factors and diabetes in the U.S., funded by the CDC.
- **Data Characteristics**: The dataset is tabular and multivariate, suitable for classification tasks, with each instance representing an individual participant.
- **Sensitive Content**: Contains potentially sensitive information such as gender, income, and education level.
- **Licensing and Access**: Available for use under specific licensing conditions and acknowledgment policies.
- **Dataset Source**: Centers for Disease Control and Prevention (CDC).
- **Dataset Version**: 1 (2021).

Exploratory Data Analysis (EDA) and Data Preprocessing Steps: 1. Understand the dataset by examining its structure and contents. 2. Identify the data types of the attributes and check for missing values. 3. Perform descriptive statistics to summarize the dataset. 4. Visualize the data to gain insights and identify patterns. 5. Preprocess the data by handling missing values,

encoding categorical variables, and scaling numerical features. 6. Split the data into training and testing sets for model development and evaluation.

```
   Diabetes_binary  HighBP  HighChol  CholCheck   BMI  Smoker  Stroke  \
0              0.0     1.0       1.0        1.0  40.0     1.0     0.0
1              0.0     0.0       0.0        0.0  25.0     1.0     0.0
2              0.0     1.0       1.0        1.0  28.0     0.0     0.0
3              0.0     1.0       0.0        1.0  27.0     0.0     0.0
4              0.0     1.0       1.0        1.0  24.0     0.0     0.0

   HeartDiseaseorAttack  PhysActivity  Fruits  …  AnyHealthcare  \
0                   0.0           0.0     0.0  …            1.0
1                   0.0           1.0     0.0  …            0.0
2                   0.0           0.0     1.0  …            1.0
3                   0.0           1.0     1.0  …            1.0
4                   0.0           1.0     1.0  …            1.0

   NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk  Sex   Age  Education  \
0          0.0      5.0      18.0      15.0       1.0  0.0   9.0        4.0
1          1.0      3.0       0.0       0.0       0.0  0.0   7.0        6.0
2          1.0      5.0      30.0      30.0       1.0  0.0   9.0        4.0
3          0.0      2.0       0.0       0.0       0.0  0.0  11.0        3.0
4          0.0      2.0       3.0       0.0       0.0  0.0  11.0        5.0

   Income
0     3.0
1     1.0
2     8.0
3     6.0
4     4.0

[5 rows x 22 columns]

(253680, 22)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   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)
memory usage: 42.6 MB
```

```
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
dtype: int64
```

Data Description The dataset contains 22 attributes, including both numerical and categorical variables. The target variable, 'Diabetes_binary', is binary and represents the health status of the subjects: diabetic (1), pre-diabetic (2), or non-diabetic (0). The features include demographic details, laboratory test results, and survey responses. The dataset has 253680 instances, and there are no missing values in the dataset.

Columns like HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActiv-

ity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex, and Diabetes_binary are binary variables, indicating the presence or absence of a specific condition or habit (1 means condition is present, and 0 means it's absent).

The BMI column represents Body Mass Index values with a range from 24.0 to 40.0. This is a measure of body fat based on height and weight that applies to adult men and women.

The GenHlth, MentHlth, and PhysHlth are health score variables or the number of days a person felt unhealthy (either physically or mentally). Their higher standard deviations (up to 13.4 for PhysHlth) could point to a diversity of health conditions in the dataset.

The Age column does not seem to represent the actual age in years (since the Max is 11), it might be encoded differently or represent age groups.

Education and Income could denote categories rather than exact values, considering the given range.

- Diabetes_binary: target variable representing whether the individual has diabetes or not (1 for Yes/Prediabetic and 0 for No).
- HighBP: represents whether the individual has high blood pressure or not.
- HighChol: indicates whether the individual has high cholesterol or not.
- CholCheck: represent if the individual has checked their cholesterol level.
- BMI: This is Body Mass Index of the individual.
- Smoker: indicates if the individual is a smoker or not.
- Stroke: show if the person has had a stroke.
- HeartDiseaseorAttack: indicates if the individual has heart disease or has had a heart attack in the past.
- PhysActivity: represent the amount or frequency of physical activity of the person.
- Fruits: represent the intake of fruits.
- Veggies: represent the intake of vegetables.
- HvyAlcoholConsump: shows if the person is a heavy alcohol consumer.
- AnyHealthcare: indicates if the individual has any healthcare or not (like insurance).
- NoDocbcCost: represent if the person could not see a doctor due to cost.
- GenHlth: represent the general health condition of the individual.
- MentHlth: show the mental health condition/statistics of the person.
- PhysHlth: illustrate physical health status of the person.
- DiffWalk: show if the person has any difficulty while walking.
- Sex: gender of the individual.
- Age: age of the individual.
- Education: level of education of individual.
- Income: income level of the individual.

	Diabetes_binary	HighBP	HighChol	CholCheck	\
count	253680.000000	253680.000000	253680.000000	253680.000000	
mean	0.139333	0.429001	0.424121	0.962670	
std	0.346294	0.494934	0.494210	0.189571	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	1.000000	
50%	0.000000	0.000000	0.000000	1.000000	
75%	0.000000	1.000000	1.000000	1.000000	

```
max           1.000000       1.000000       1.000000       1.000000

                BMI         Smoker         Stroke  HeartDiseaseorAttack  \
count  253680.000000  253680.000000  253680.000000         253680.000000
mean       28.382364       0.443169       0.040571              0.094186
std         6.608694       0.496761       0.197294              0.292087
min        12.000000       0.000000       0.000000              0.000000
25%        24.000000       0.000000       0.000000              0.000000
50%        27.000000       0.000000       0.000000              0.000000
75%        31.000000       1.000000       0.000000              0.000000
max        98.000000       1.000000       1.000000              1.000000

       PhysActivity         Fruits  …  AnyHealthcare    NoDocbcCost  \
count  253680.000000  253680.000000  …  253680.000000  253680.000000
mean        0.756544       0.634256  …       0.951053       0.084177
std         0.429169       0.481639  …       0.215759       0.277654
min         0.000000       0.000000  …       0.000000       0.000000
25%         1.000000       0.000000  …       1.000000       0.000000
50%         1.000000       1.000000  …       1.000000       0.000000
75%         1.000000       1.000000  …       1.000000       0.000000
max         1.000000       1.000000  …       1.000000       1.000000

            GenHlth       MentHlth       PhysHlth       DiffWalk  \
count  253680.000000  253680.000000  253680.000000  253680.000000
mean        2.511392       3.184772       4.242081       0.168224
std         1.068477       7.412847       8.717951       0.374066
min         1.000000       0.000000       0.000000       0.000000
25%         2.000000       0.000000       0.000000       0.000000
50%         2.000000       0.000000       0.000000       0.000000
75%         3.000000       2.000000       3.000000       0.000000
max         5.000000      30.000000      30.000000       1.000000

                Sex            Age      Education         Income
count  253680.000000  253680.000000  253680.000000  253680.000000
mean        0.440342       8.032119       5.050434       6.053875
std         0.496429       3.054220       0.985774       2.071148
min         0.000000       1.000000       1.000000       1.000000
25%         0.000000       6.000000       4.000000       5.000000
50%         0.000000       8.000000       5.000000       7.000000
75%         1.000000      10.000000       6.000000       8.000000
max         1.000000      13.000000       6.000000       8.000000

[8 rows x 22 columns]

24206
```

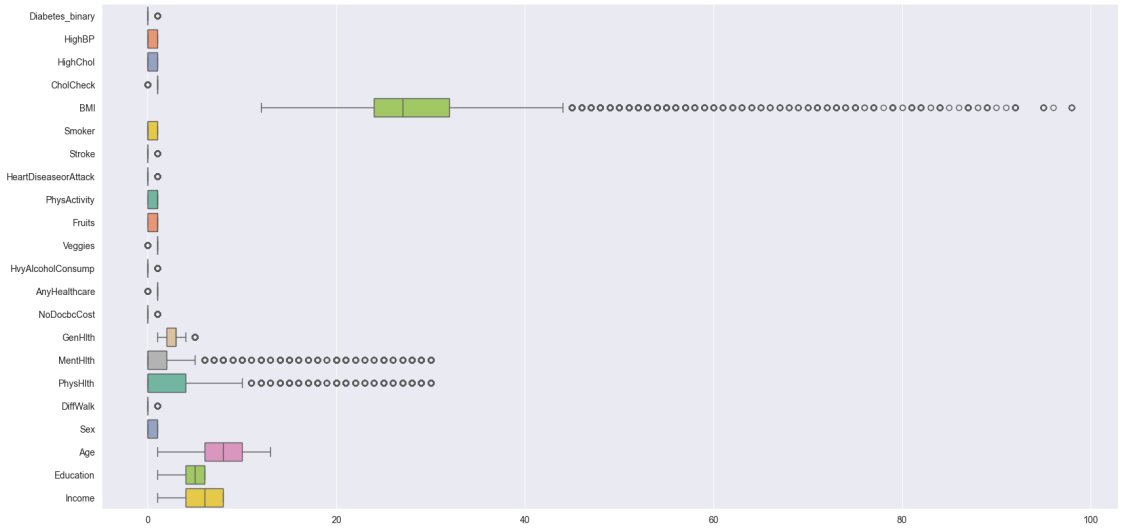
1.0.1 Data Preprocessing

There are no missing values in the dataset. However, there are duplicate rows.

```
(229474, 22)
```

```
array([[<Axes: title={'center': 'Diabetes_binary'}>,
        <Axes: title={'center': 'HighBP'}>,
        <Axes: title={'center': 'HighChol'}>,
        <Axes: title={'center': 'CholCheck'}>,
        <Axes: title={'center': 'BMI'}>],
       [<Axes: title={'center': 'Smoker'}>,
        <Axes: title={'center': 'Stroke'}>,
        <Axes: title={'center': 'HeartDiseaseorAttack'}>,
        <Axes: title={'center': 'PhysActivity'}>,
        <Axes: title={'center': 'Fruits'}>],
       [<Axes: title={'center': 'Veggies'}>,
        <Axes: title={'center': 'HvyAlcoholConsump'}>,
        <Axes: title={'center': 'AnyHealthcare'}>,
        <Axes: title={'center': 'NoDocbcCost'}>,
        <Axes: title={'center': 'GenHlth'}>],
       [<Axes: title={'center': 'MentHlth'}>,
        <Axes: title={'center': 'PhysHlth'}>,
        <Axes: title={'center': 'DiffWalk'}>,
        <Axes: title={'center': 'Sex'}>, <Axes: title={'center': 'Age'}>],
       [<Axes: title={'center': 'Education'}>,
        <Axes: title={'center': 'Income'}>, <Axes: >, <Axes: >, <Axes: >]],
      dtype=object)
```


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



While the boxplot for BMI shows some outliers, generally High BMI is a good indicator for health issues and in particular its a good indicator for diabetes. However, PhysHlth and MentHlth also shows some outliers. There is some positive correlation between the variables, such as GenHlth and PhysHlth, PhysHlth and DiffWalk, and GenHlth and DiffWalk which is expected since they are related to health conditions. There is negative correlation between GenHlth and Income, and Diffwalk and Income. The correlation matrix provides insights into the relationships between the features, which can be useful for feature selection and model building.

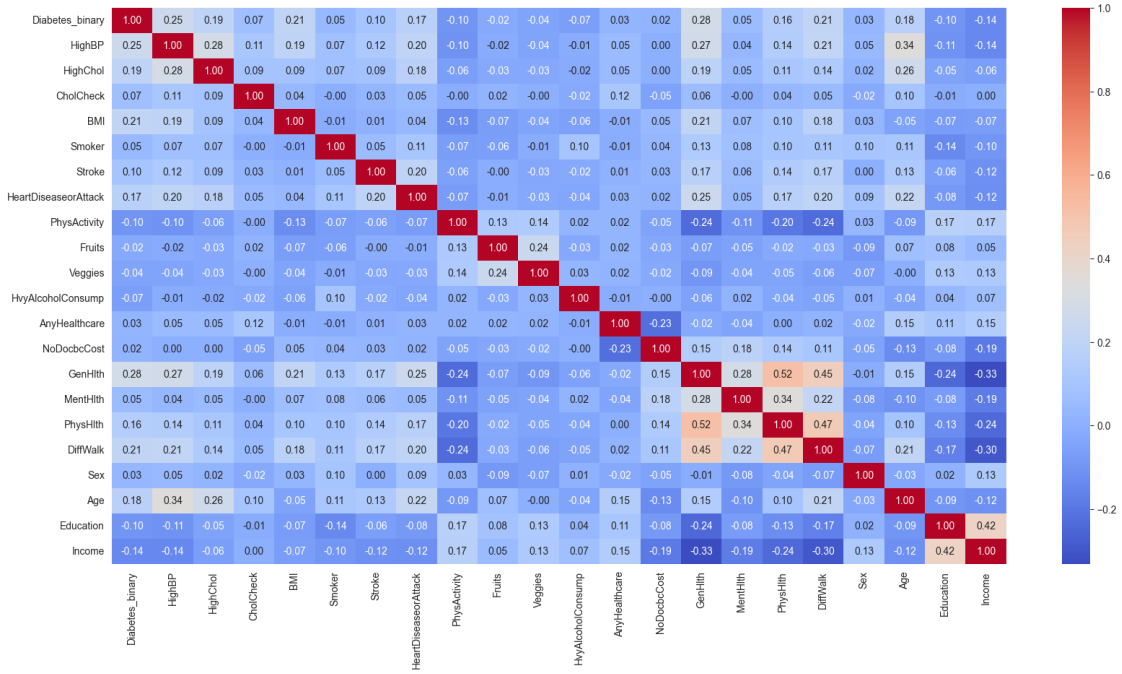
Since most of the features are binary, the histograms show the distribution of the binary variables. The Age column seems to be encoded differently, as the values are not continuous and is most likely to be categorical. The Education and Income columns also appear to be categorical, with a limited number of distinct values. The target variable, Diabetes_binary, is imbalanced, with more non-diabetic individuals than diabetic or pre-diabetic individuals. This imbalance should be considered when building predictive models. The skewed distribution of the age, education, income and BMI columns is interesting and may require further preprocessing, such as binning or normalization, depending on the modeling approach.

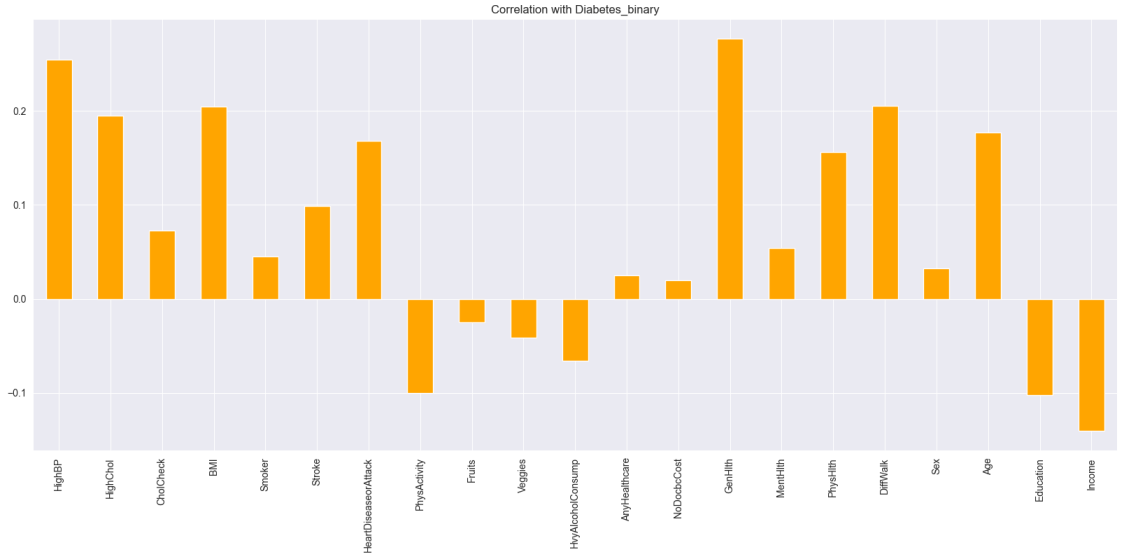
To help us relate to the data, we will convert the binary variables to categorical variables and add columns with more descriptive names. - We will add a diabetes_str column that will represent the diabetes status in a more human-readable format. 0 will be 'Non-Diabetic', 1 will be 'Diabetic'. - To show relation between features clearly during feature selection we will categorize the data.

We can now visualise the columns with their relation to the target variable.

1.0.2 Feature Selection

Feature selection is a critical step in the machine learning pipeline, as it helps identify the most relevant features for building predictive models. In this section, we will explore different feature selection techniques to identify the most important features for predicting diabetes. We will use the following methods: - We will use a correlation matrix to identify the features that are most correlated with the target variable. - We will look at individual features and their importance in predicting the target variable. - We will use a machine learning model to identify the most important features for predicting diabetes.



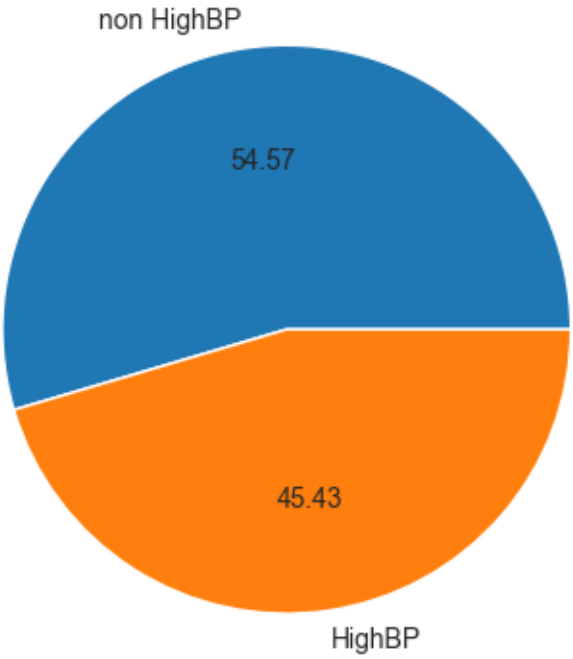


The correlation matrix and the correlation bar plot show the relationships between the features and the target variable, 'Diabetes_binary'. The correlation matrix provides a visual representation of the correlation coefficients between the features, while the bar plot displays the correlation of each feature with the target variable. We can see that Fruits, AnyHealthCare, NoDocbcCost and sex are the least correlated with the target variable. The features that are most correlated with the target variable are HighBP, HighChol, GenHlth and BMI. These features are likely to be important for predicting diabetes. We will now look at individual features and their importance in predicting

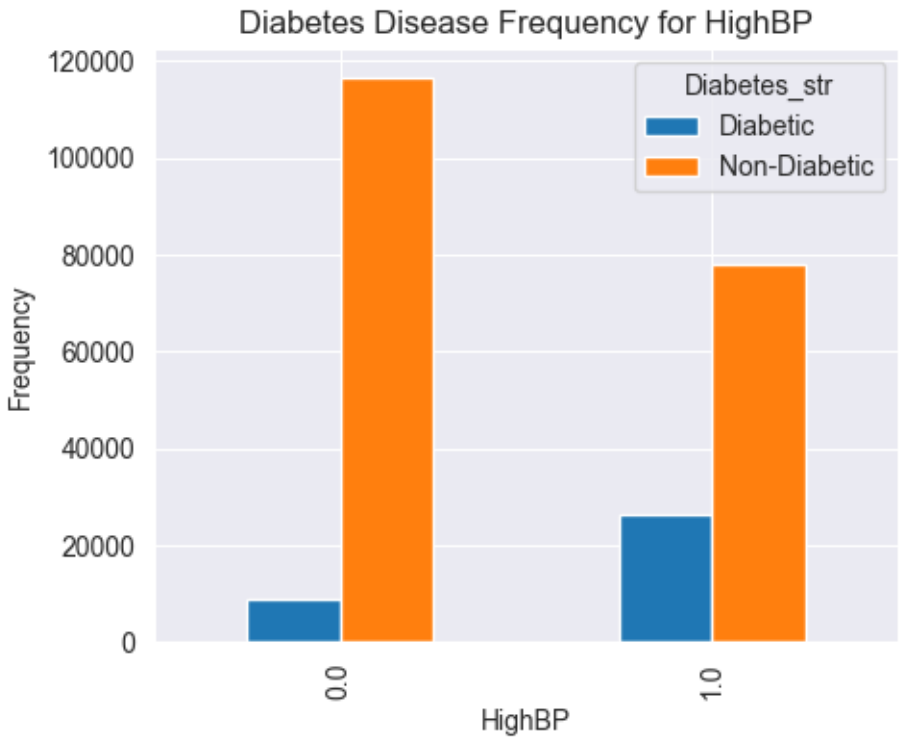
diabetes.

Effects of Physiological Factors on Diabetes We will now explore the relationship between physiological factors such as HighBP, HighCol, BMI, stroke, HeartDiseaseorAttack and Diabetes. We will visualize the distribution of these factors for diabetic and non-diabetic individuals to identify any patterns or trends.

Diabetes_str	Diabetic	Non-Diabetic
HighBP		
0.0	8692	116522
1.0	26405	77855



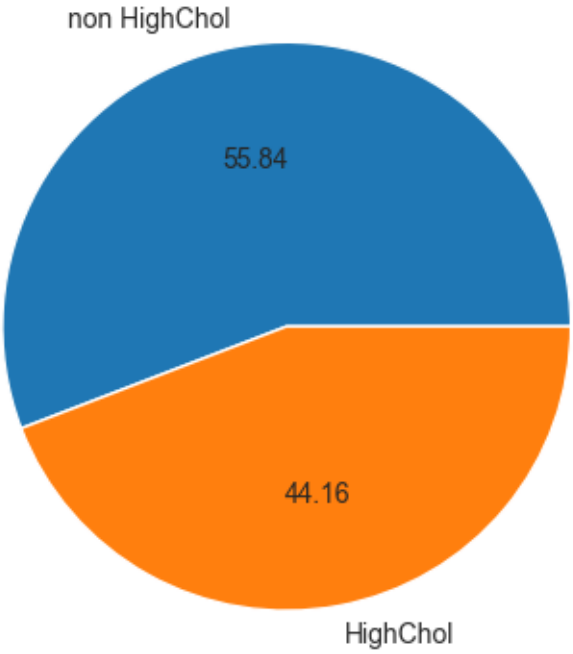
The pie chart shows the distribution of individuals with high blood pressure (HighBP) and those without high blood pressure. The table below the pie chart displays the number of diabetic and non-diabetic individuals based on their high blood pressure status. The majority of individuals in the dataset do not have high blood pressure, and there are more non-diabetic individuals than diabetic individuals in both categories. However, we have seen before that the data is imbalanced. We will look at the frequency of Diabetes in individuals with high blood pressure and also take an average.

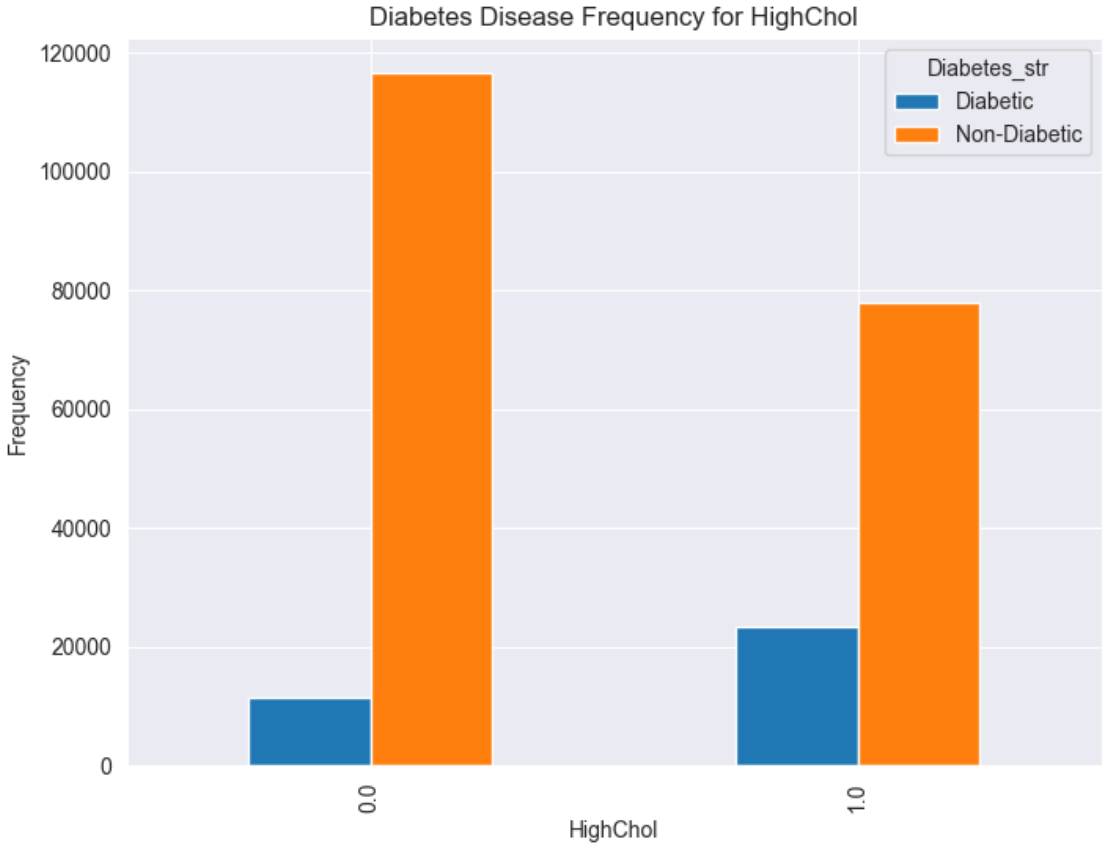



```
Diabetes_str  HighBP
Diabetic      1.0       75.234351
              0.0       24.765649
Non-Diabetic  0.0       59.946393
              1.0       40.053607
dtype: float64
```

We can see that the percentage of individuals with high blood pressure is higher in the diabetic group compared to the non-diabetic group. This indicates that high blood pressure may be a risk factor for diabetes. We will now explore the relationship between high cholesterol (HighChol) and diabetes.

```
Diabetes_str  Diabetic  Non-Diabetic
HighChol
0.0              11601        116528
1.0              23496         77849
```

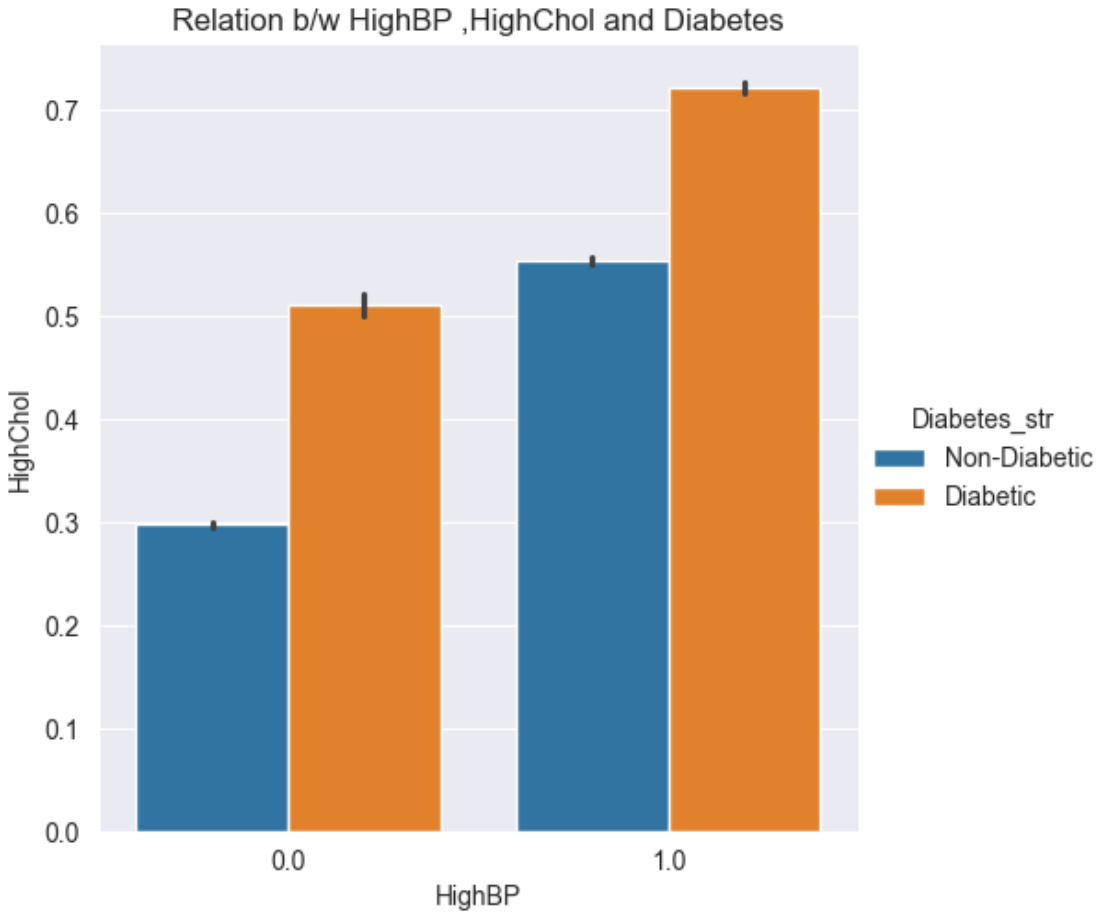

non HighChol
55.84
44.16
HighChol


Diabetes Disease Frequency for HighChol
Diabetes_str
Diabetic
Non-Diabetic
Frequency
120000
100000
80000
60000
40000
20000
0
0.0
1.0
HighChol

```
Diabetes_str  HighChol
Diabetic      1.0         66.945893
              0.0         33.054107
Non-Diabetic  0.0         59.949480
              1.0         40.050520
dtype: float64
```

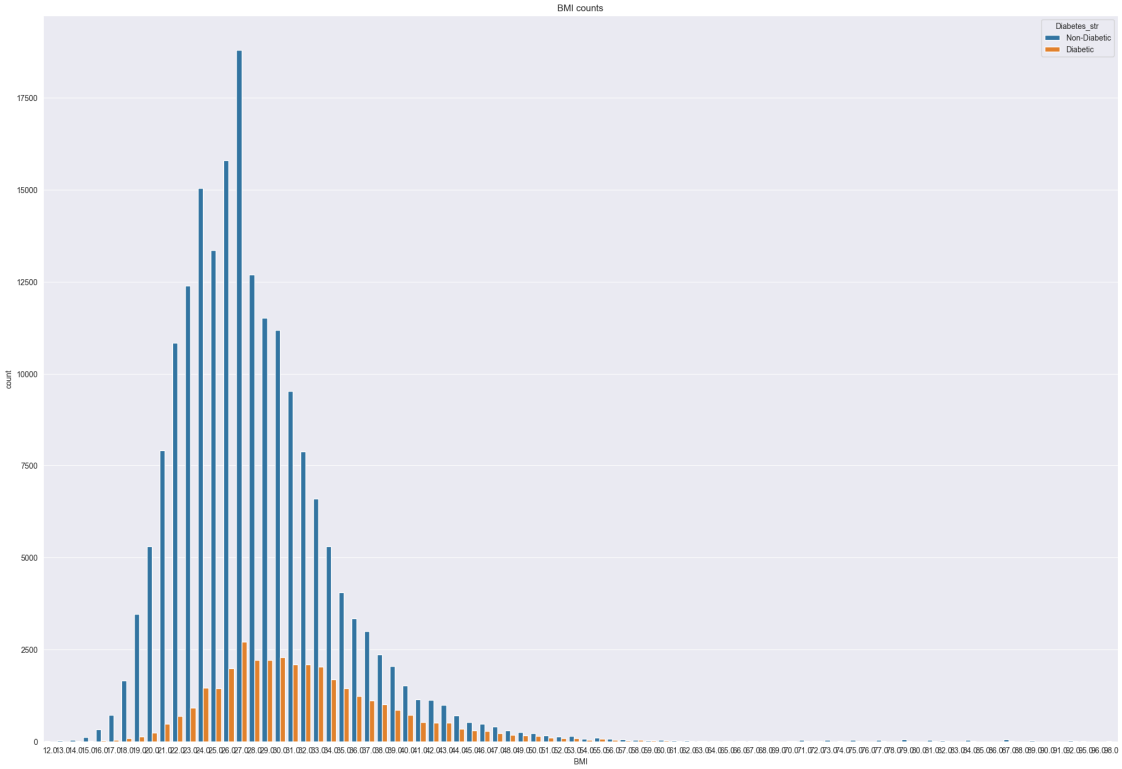
Similarly, the percentage of individuals with high cholesterol is higher in the diabetic group compared to the non-diabetic group. This suggests that high cholesterol may also be a risk factor for diabetes. It is likely that the combined impact of HighBP and HighChol on diabetes is significant.

```
Text(0.5, 1.0, 'Relation b/w HighBP ,HighChol and Diabetes')
```



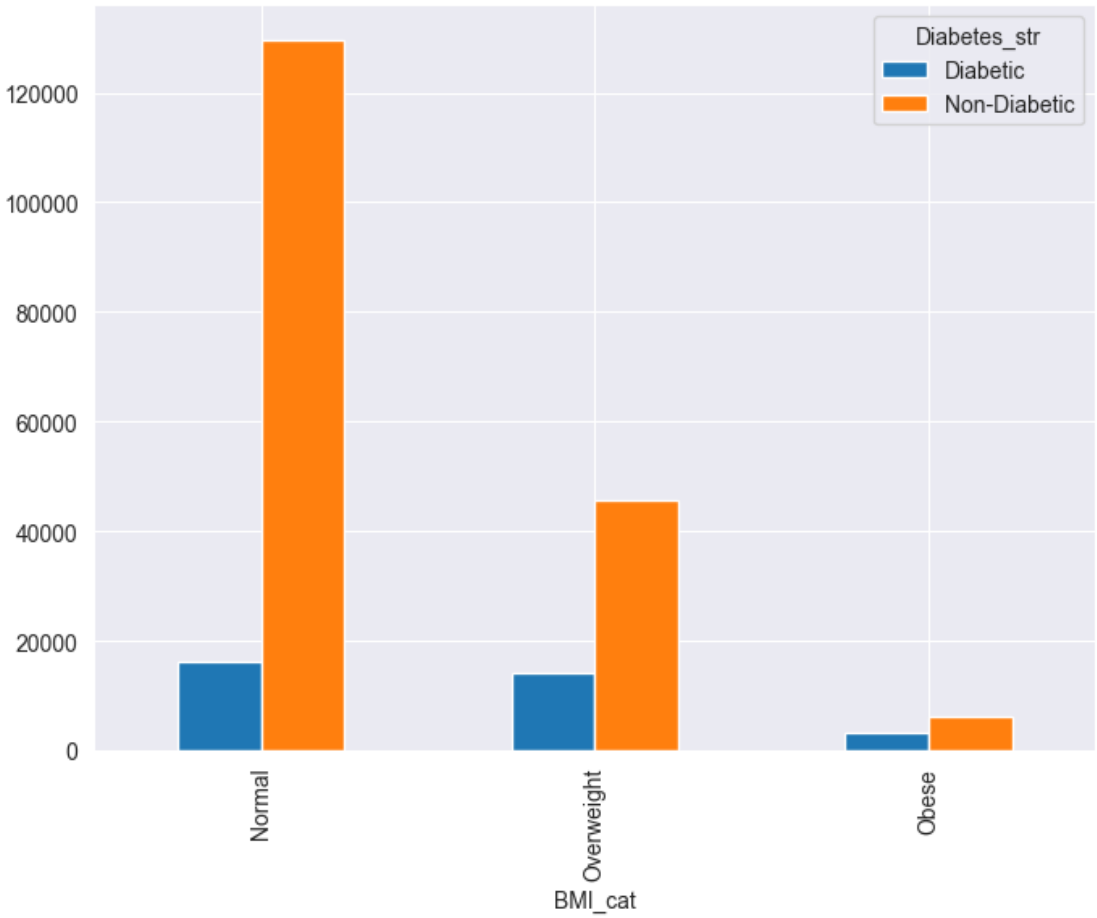
As expected, the percentage of diabetic individuals is higher in the group with both high blood pressure and high cholesterol compared to the other groups. This indicates that the combination of high blood pressure and high cholesterol may be a significant risk factor for diabetes. We will

now explore the relationship between BMI and diabetes. As BMI is a categorical variable, we will start with visualizing the distribution of BMI categories in the dataset.



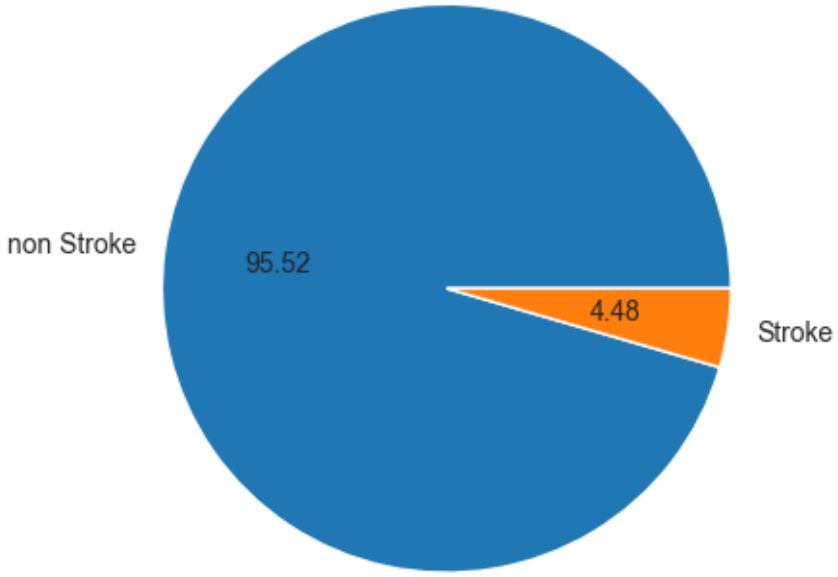
The BMI data has outliers below 16 and above 50. We will bin the data into categories to better understand the relationship between BMI and diabetes.

```
<Axes: xlabel='BMI_cat'>
```

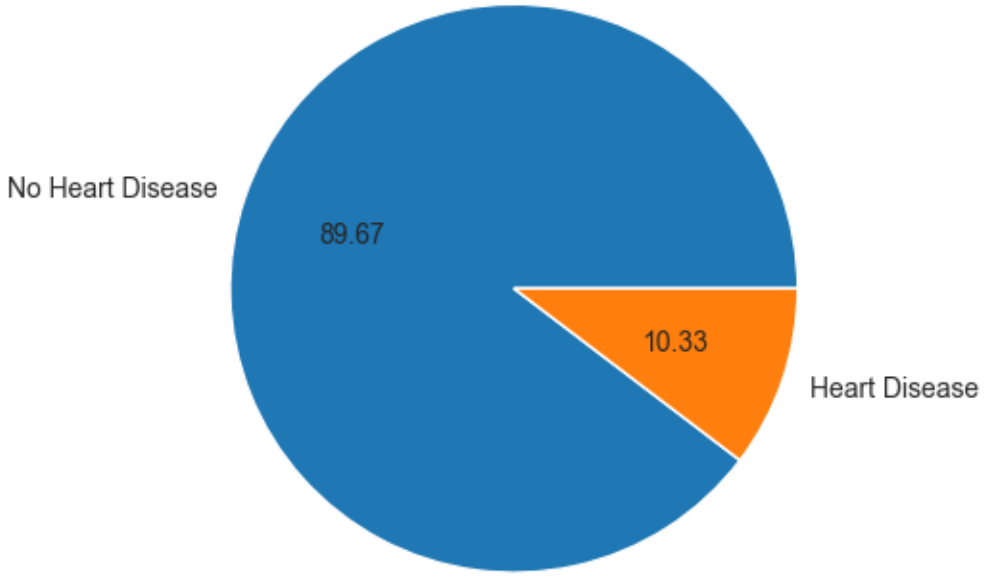


We can now see a more clear relationship between BMI and diabetes. The percentage of diabetic individuals is higher in the obese category compared to the overweight and normal categories. This indicates that obesity may be a significant risk factor for diabetes. We will now explore the relationship between stroke, heart disease and diabetes.

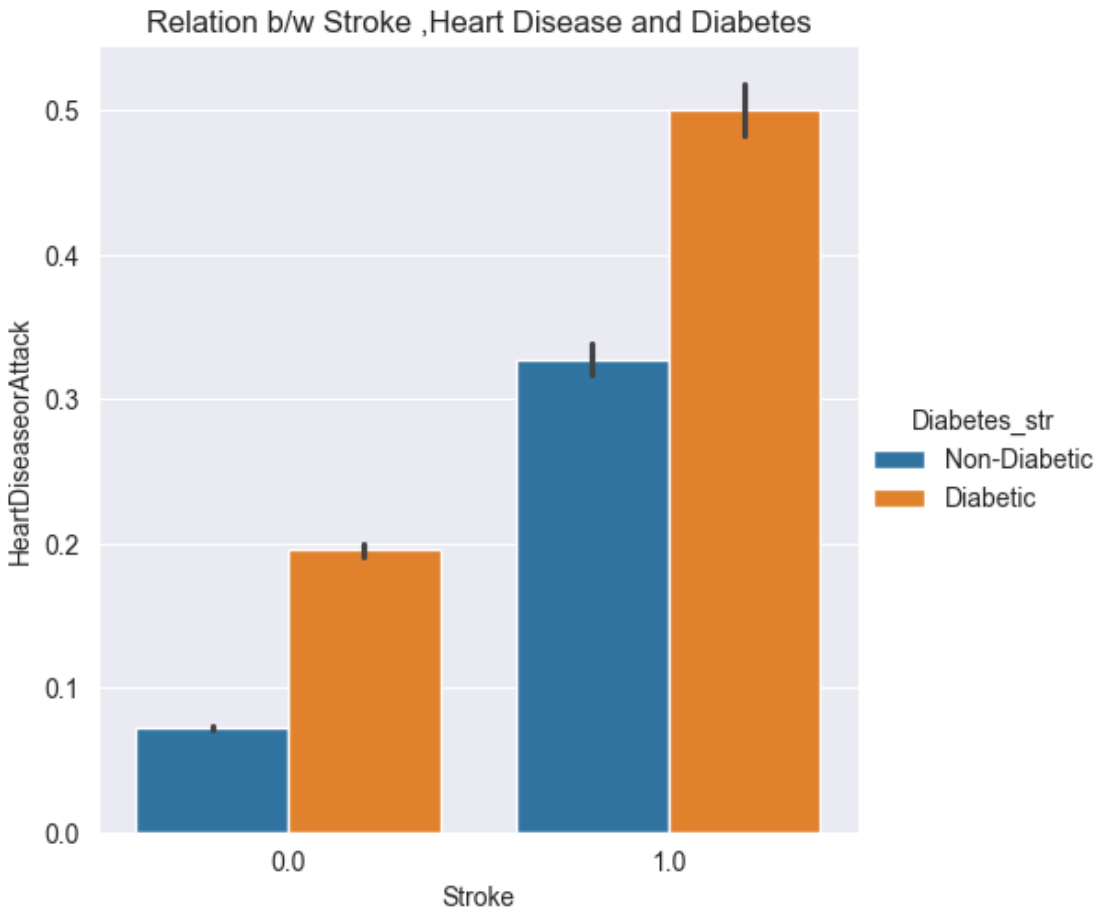
Diabetes_str	Diabetic	Non-Diabetic
Stroke		
0.0	31829	187361
1.0	3268	7016



Diabetes_str	Diabetic	Non-Diabetic
HeartDiseaseorAttack		
0.0	27241	178520
1.0	7856	15857



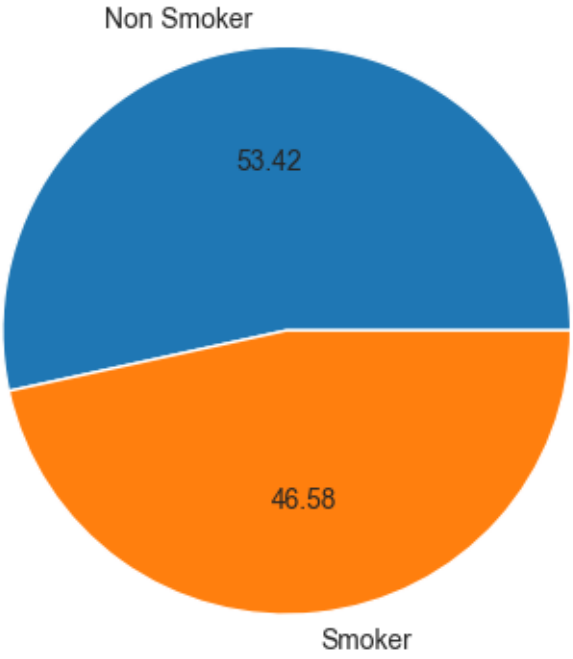

```
Stroke  HeartDiseaseorAttack  Diabetes_str
0.0     0.0                   Non-Diabetic    87.157808
                              Diabetic        12.842192
        1.0                   Non-Diabetic    68.549325
                              Diabetic        31.450675
1.0     0.0                   Non-Diabetic    74.291115
                              Diabetic        25.708885
        1.0                   Non-Diabetic    58.434959
                              Diabetic        41.565041
dtype: float64

Text(0.5, 1.0, 'Relation b/w Stroke ,Heart Disease and Diabetes')
```



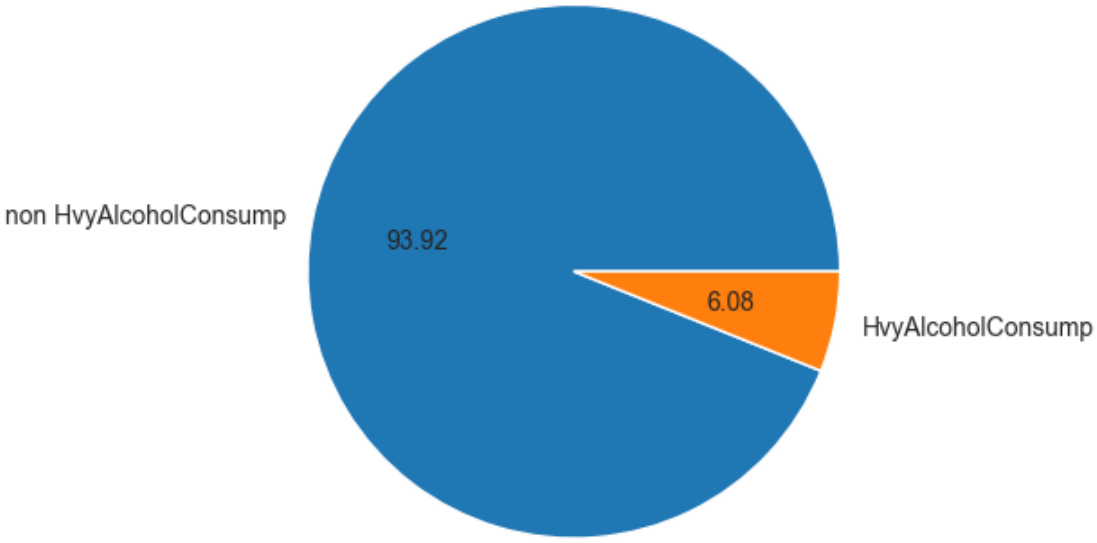
While the incidence of stroke and heart disease is relatively low in the dataset, the percentage of diabetic individuals is higher in the groups with stroke and heart disease compared to the groups without these conditions. This suggests that stroke and heart disease may be risk factors for diabetes. We will now explore the relationship between lifestyle factors such as smoking, physical activity, and dietary habits with diabetes.

Effects of Lifestyle Factors on Diabetes We will now explore the relationship between lifestyle factors such as smoking, physical activity, fruit and vegetable intake, heavy alcohol consumption, and diabetes. We will visualize the distribution of these factors for diabetic and non-diabetic individuals to identify any patterns or trends.

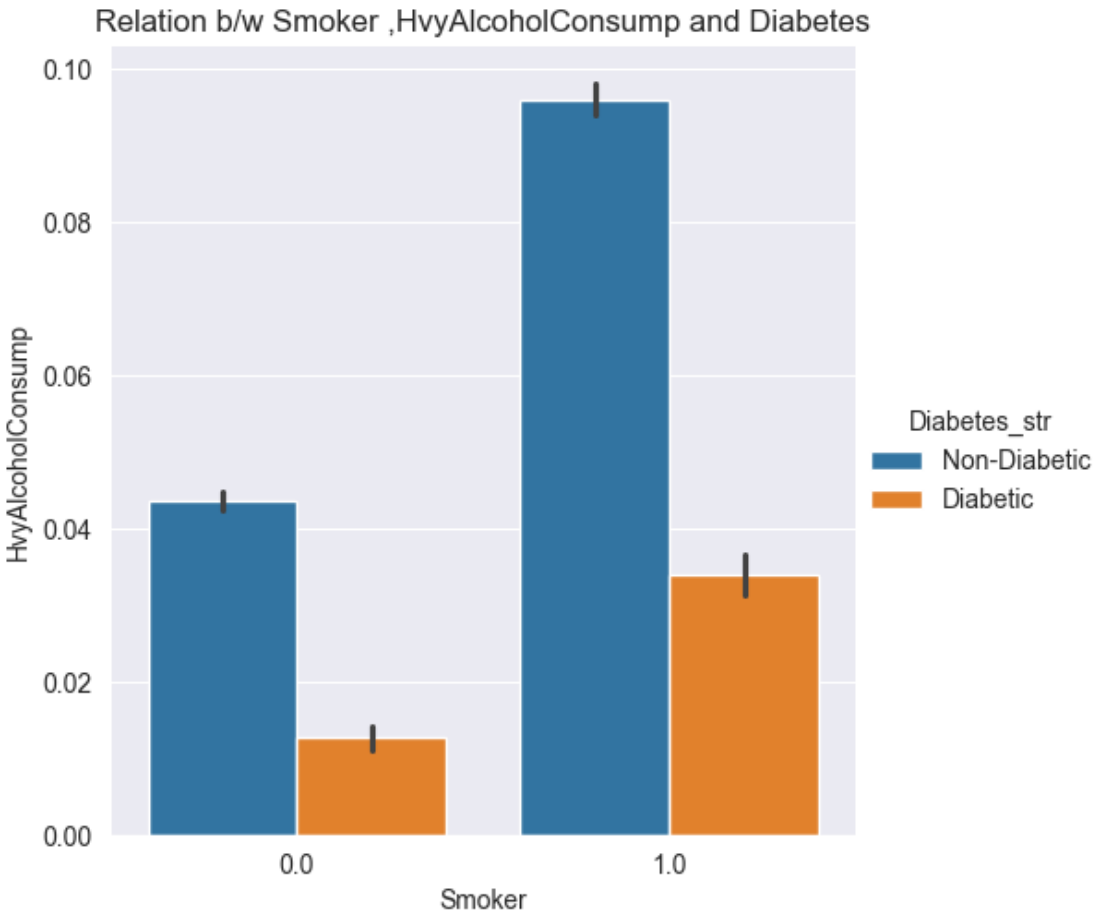
Diabetes_str	Diabetic	Non-Diabetic
Smoker		
0.0	16874	105711
1.0	18223	88666



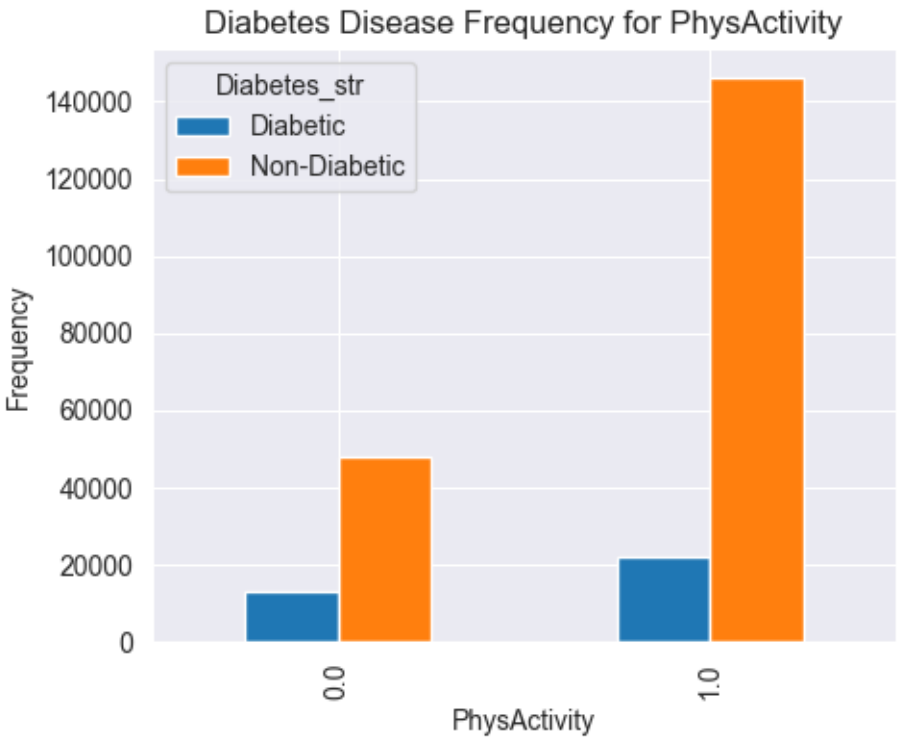
Diabetes_str	Diabetic	Non-Diabetic
HvyAlcoholConsump		
0.0	34265	181259
1.0	832	13118



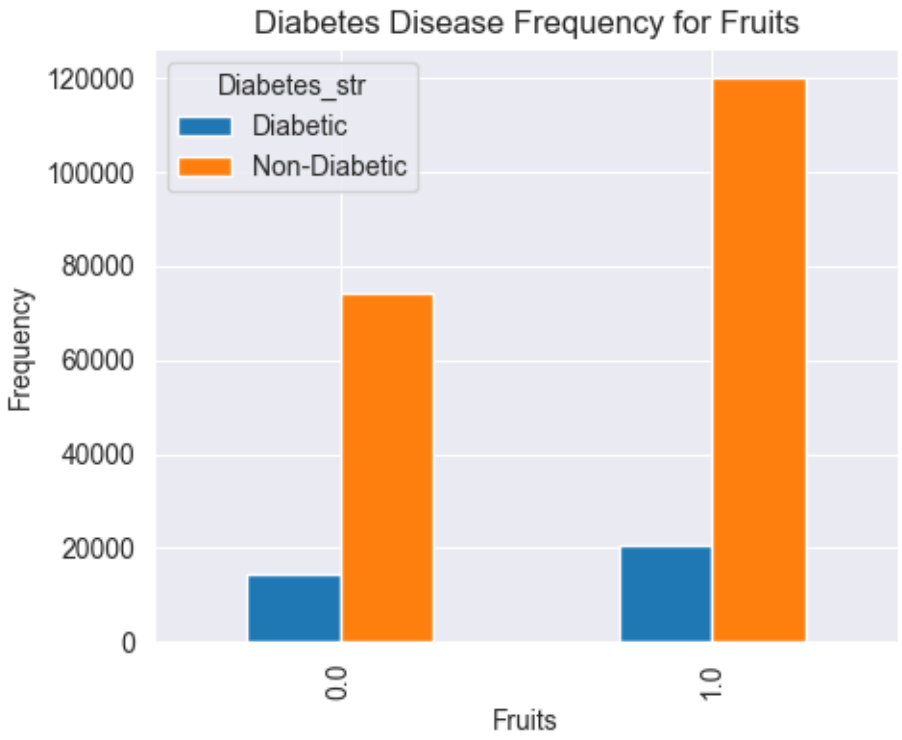

```
Text(0.5, 1.0, 'Relation b/w Smoker ,HvyAlcoholConsump and Diabetes')
```



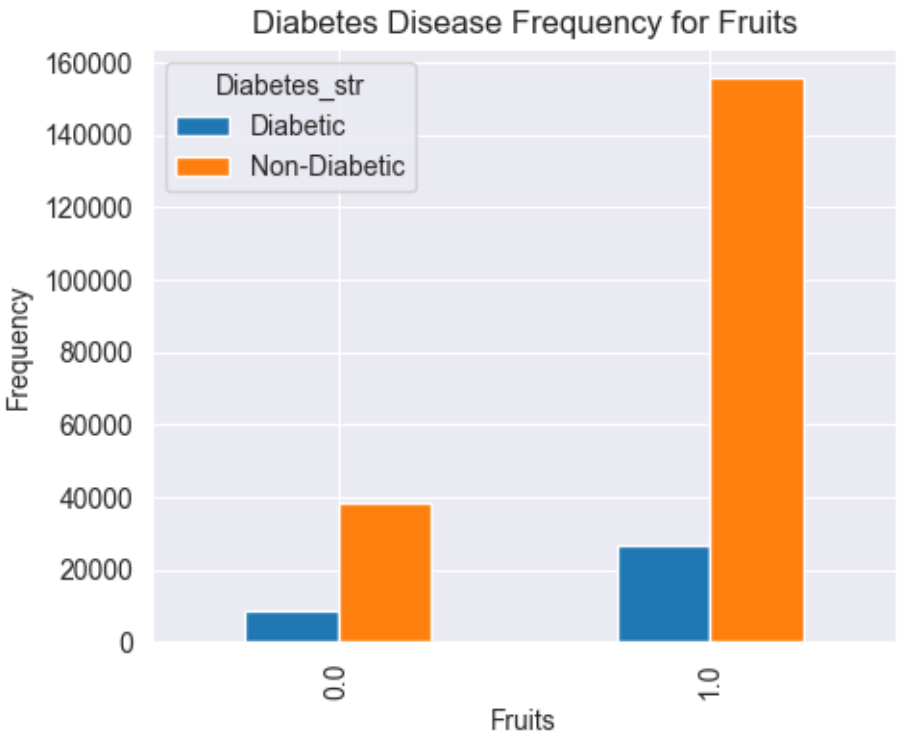
While, individually, smoking and heavy alcohol consumption do not show a strong correlation with diabetes, the combination of smoking and heavy alcohol consumption may be a risk factor for diabetes. We will now explore the relationship between physical activity, fruit and vegetable intake, and diabetes.



Diabetes_str	Diabetic	Non-Diabetic
PhysActivity		
0.0	13038	48222
1.0	22059	146155



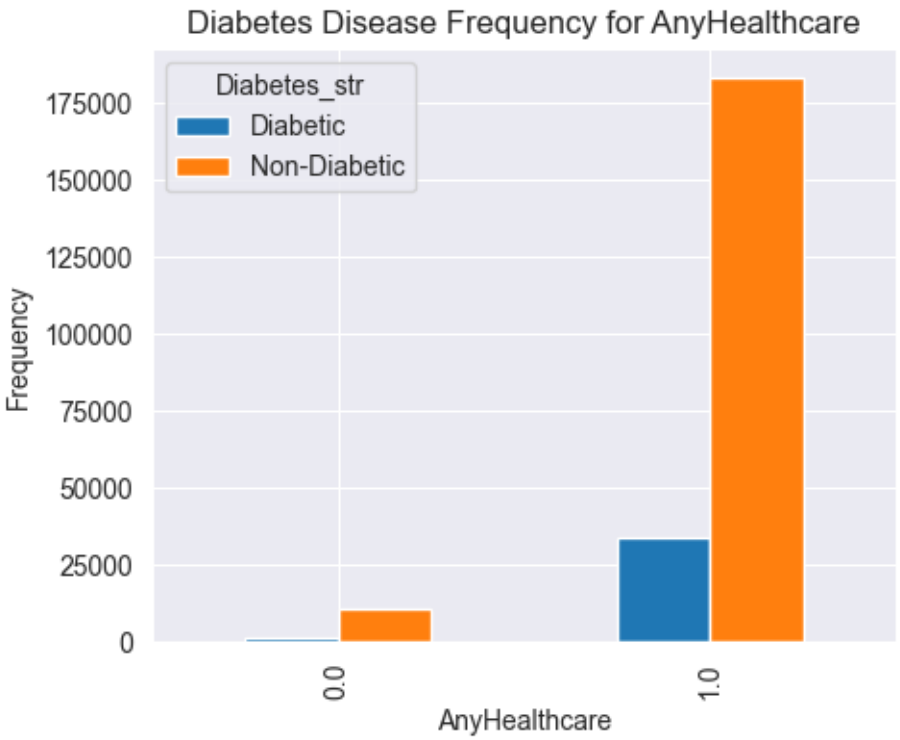
Diabetes_str	Diabetic	Non-Diabetic
Fruits		
0.0	14592	74289
1.0	20505	120088



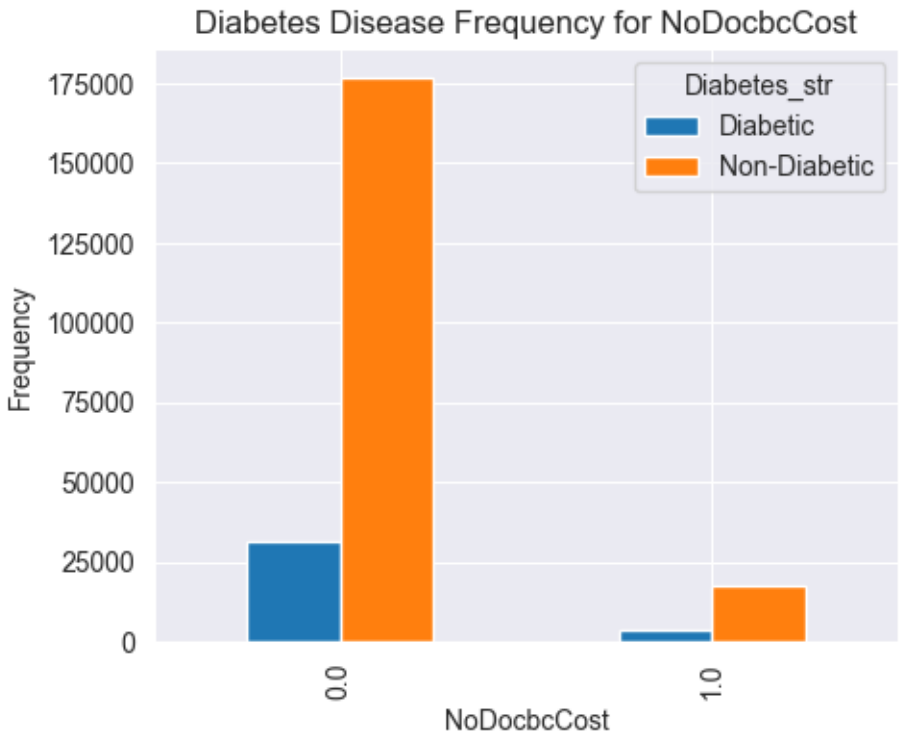
Diabetes_str	Diabetic	Non-Diabetic
Veggies		
0.0	8602	38535
1.0	26495	155842

While physical activity, fruit intake and vegetable intake demonstrate a positive correlation with diabetes, the relationship is not as strong as with physiological factors. We will now explore the relationship between healthcare access and diabetes.

Effects of Healthcare Access on Diabetes We will now explore the relationship between healthcare access factors such as healthcare coverage, cost barriers, and diabetes. We will visualize the distribution of these factors for diabetic and non-diabetic individuals to identify any patterns or trends.



Diabetes_str	Diabetic	Non-Diabetic
AnyHealthcare		
0.0	1422	10967
1.0	33675	183410



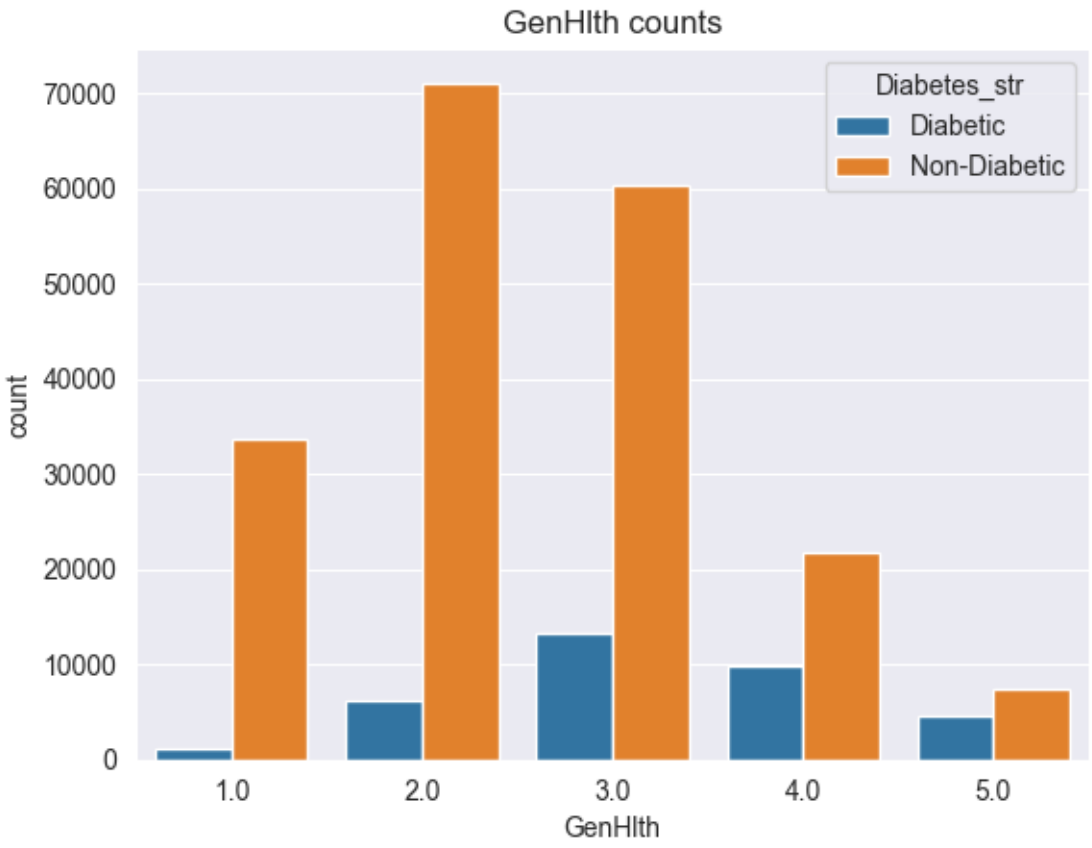
Diabetes_str	Diabetic	Non-Diabetic
NoDocbcCost		
0.0	31355	176796
1.0	3742	17581

There isn't a strong correlation between healthcare access factors and diabetes. We will now explore the relationship between general health, mental health, and physical health with diabetes.

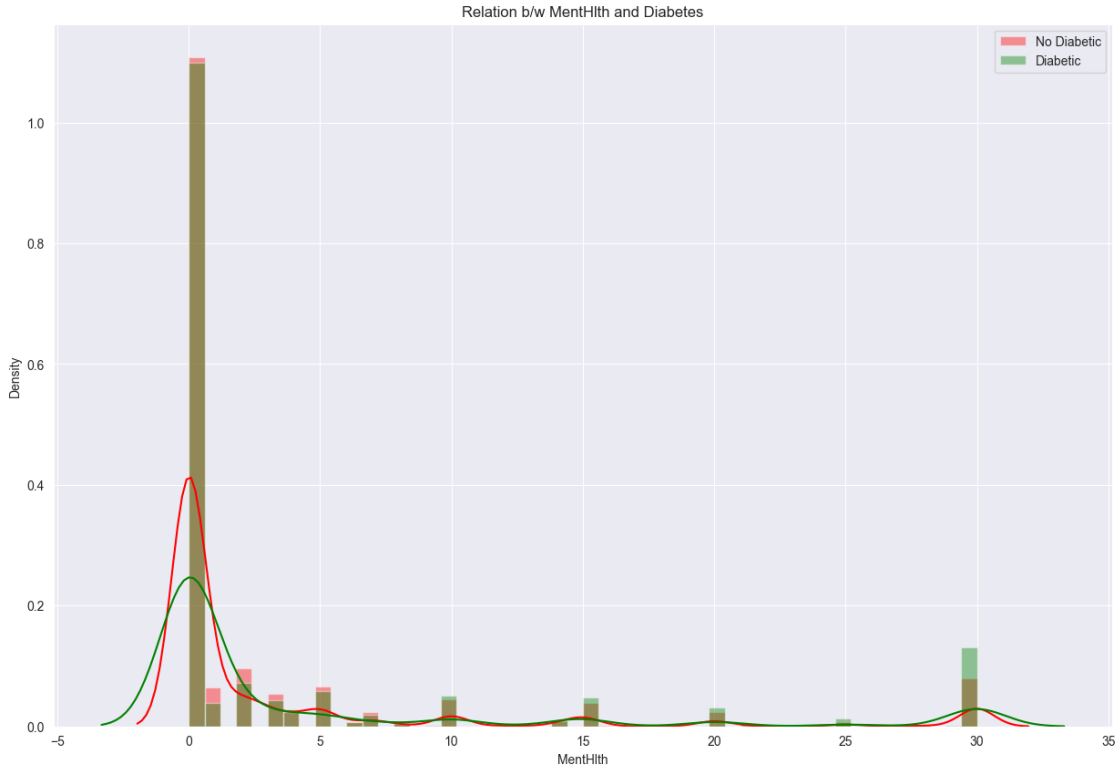
Effects of Health Status on Diabetes We will now explore the relationship between general health, mental health, physical health, and diabetes. We will visualize the distribution of these factors for diabetic and non-diabetic individuals to identify any patterns or trends. Since these are categorical variables, we will start by visualizing the distribution of the categories.

Diabetes_str	Diabetic	Non-Diabetic
GenHlth		
1.0	1135	33719
2.0	6280	71085
3.0	13324	60308
4.0	9781	21764
5.0	4577	7501

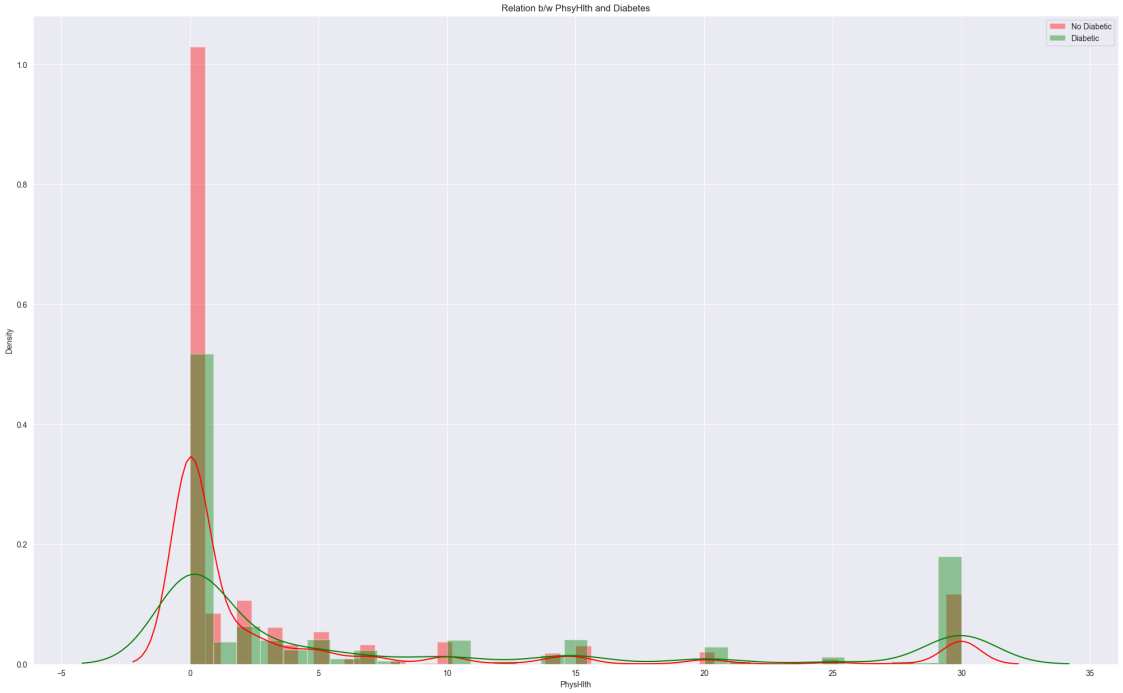
```
Text(0.5, 1.0, 'GenHlth counts')
```

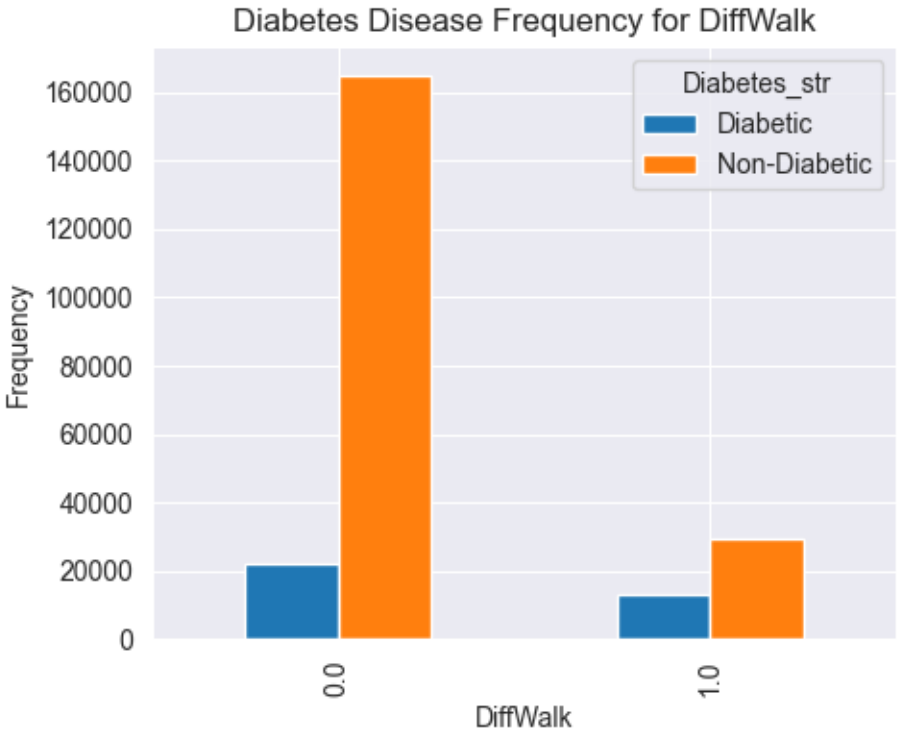



```
MentHlth        0.0    1.0    2.0   3.0   4.0   5.0   6.0   7.0   8.0   9.0   \
Diabetes_str
Diabetic      23155    812   1507   924   489  1223   164   405   110    13
Non-Diabetic  129170  7495  11185  6377  3285  7690   824  2685   529    78

MentHlth      …  21.0  22.0  23.0  24.0  25.0  26.0  27.0  28.0  29.0  30.0
Diabetes_str  …
Diabetic      …    48    11     8     6   273     7    12    57    30  2768
Non-Diabetic  …   179    52    30    27   915    38    67   270   128  9311

[2 rows x 31 columns]

<matplotlib.legend.Legend at 0x2364eddab40>
```




```
PhysHlth        0.0   1.0    2.0   3.0   4.0   5.0   6.0   7.0   8.0   9.0   \
Diabetes_str
Diabetic       16497  1187   2027  1289   763  1314   282   742   159    36
Non-Diabetic  120081  9886  12464  7146  3758  6281  1046  3789   650   143

PhysHlth      …  21.0  22.0  23.0  24.0  25.0  26.0  27.0  28.0  29.0  30.0
Diabetes_str  …
Diabetic      …   139    23    16    15   394    22    21   143    74  5724
Non-Diabetic  …   524    47    40    57   942    47    78   379   141  13661

[2 rows x 31 columns]

<matplotlib.legend.Legend at 0x23652eb6450>
```

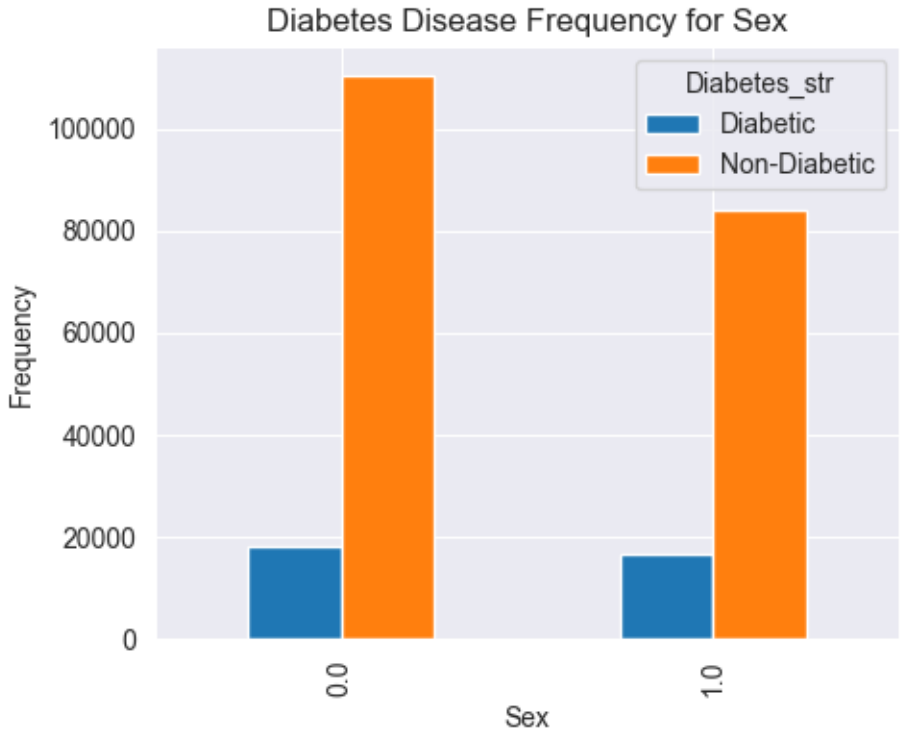

Relation b/w PhsyHlth and Diabetes
No Diabetic
Diabetic
Density
PhysHlth


Diabetes Disease Frequency for DiffWalk
Diabetes_str
Diabetic
Non-Diabetic
Frequency
DiffWalk

Diabetes_str	Diabetic	Non-Diabetic
DiffWalk		
0.0	21983	164866
1.0	13114	29511

While general health and physical health show a positive correlation with diabetes, mental health does not show a strong correlation. However, the relationship between these factors and diabetes is not as strong as with physiological factors. Interestingly, while there is no causal link, the percentage of diabetic individuals is higher in the group with difficulty walking compared to the group without difficulty walking. This suggests that mobility issues may be a risk factor from diabetes. We will now explore the relationship between demographic factors such as Sex, Age, Education, and Income with diabetes.

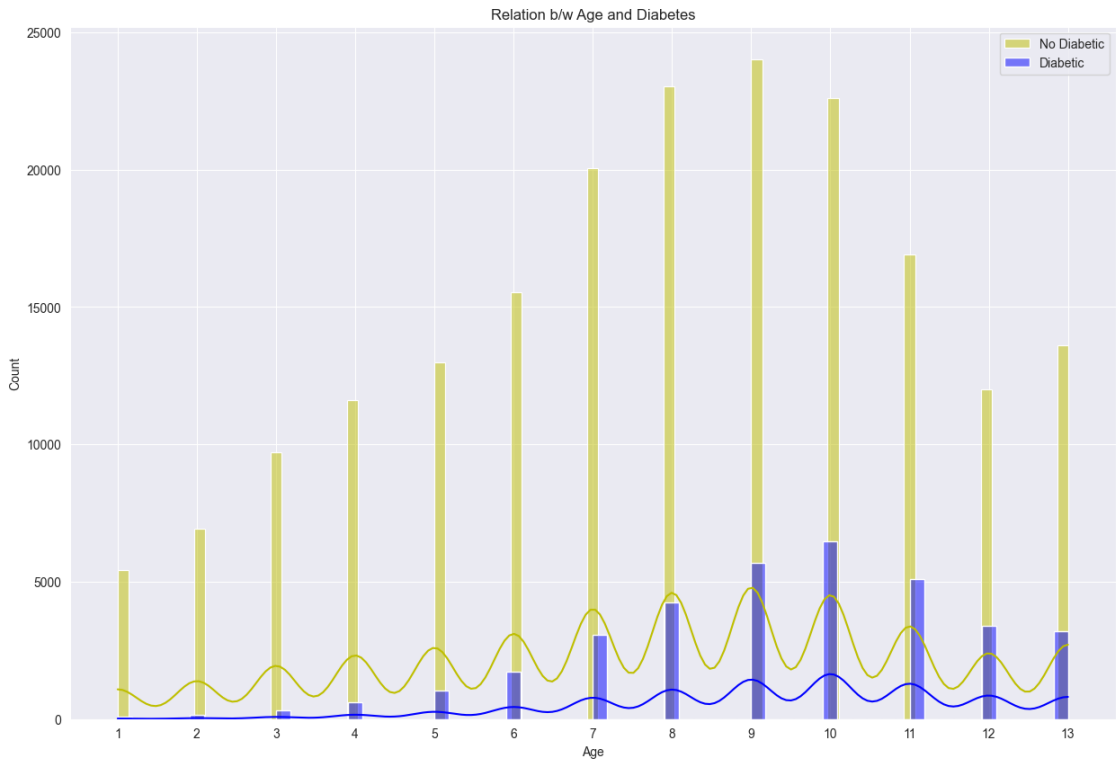
Effects of Demographic Factors on Diabetes We will now explore the relationship between demographic factors such as Sex, Age, Education and Income. We will visualize the distribution of these factors for diabetic and non-diabetic individuals to identify any patterns or trends.



Diabetes_str	Diabetic	Non-Diabetic
Sex		
0.0	18345	110370
1.0	16752	84007

Diabetes_str	Diabetic	Non-Diabetic
Age		
1.0	78	5433
2.0	140	6924
3.0	314	9709
4.0	625	11604
5.0	1049	12991
6.0	1741	15539
7.0	3072	20049
8.0	4241	23031
9.0	5681	23997
10.0	6483	22610
11.0	5090	16903
12.0	3383	11996
13.0	3200	13591

```
<matplotlib.legend.Legend at 0x236521bd340>
```



We can see that the distribution of diabetic and non-diabetic indivisuals is similar across the sexes. Therefore, we can state that sex is not a causal factor for diabetes. However, the percentage of diabetic individuals is higher in the older age groups compared to the younger age groups. This suggests that age may be a risk factor for diabetes. In particular the risk of diabetes increases in the age groups from 45 to 79. (Groups 6 to 12). - 1 Age 18 to 24 - 2 Age 25 to 29 - 3 Age 30 to 34

- 4 Age 35 to 39 - 5 Age 40 to 44 - 6 Age 45 to 49 - 7 Age 50 to 54 - 8 Age 55 to 59 - 9 Age 60 to 64 - 10 Age 65 to 69 - 11 Age 70 to 74 - 12 Age 75 to 79 - 13 Age 80 or older

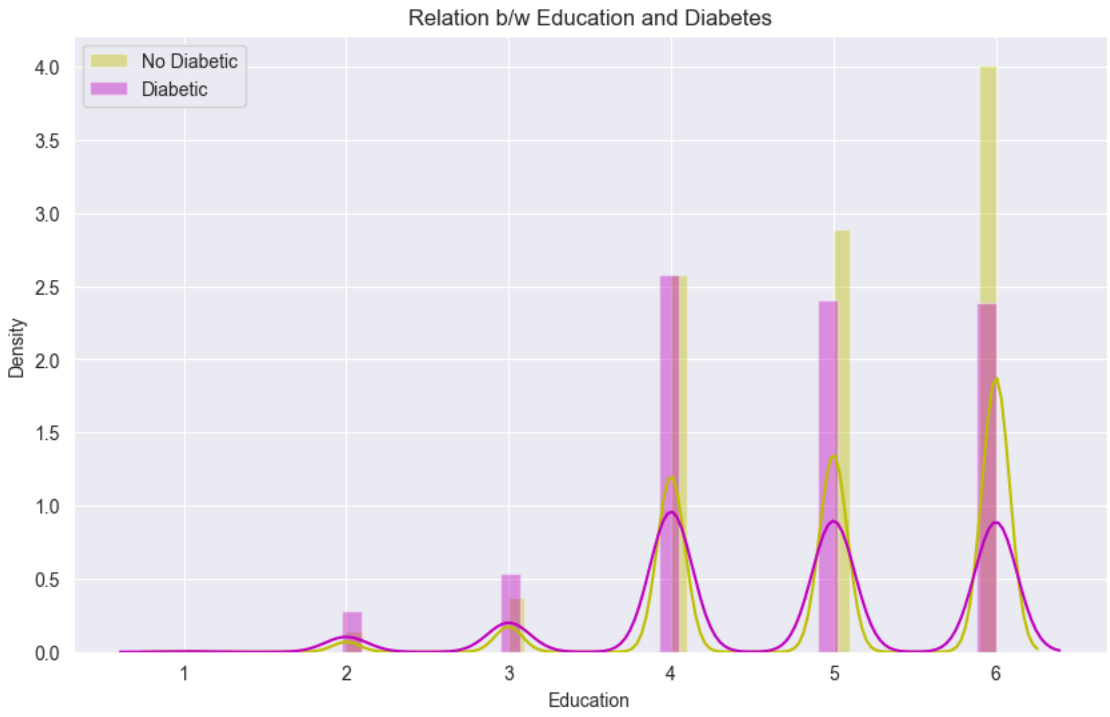
We will now explore the relationship between education and income with diabetes. Education and Income are categorical variables. We will start by visualizing the distribution of the categories.

Education Categories: - 1 Never attended school or only kindergarten - 2 Grades 1 through 8 (Elementary) - 3 Grades 9 through 11 (Some high school) - 4 Grade 12 or GED (High school graduate) - 5 College 1 year to 3 years (Some college or technical school) - 6 College 4 years or more (College graduate)

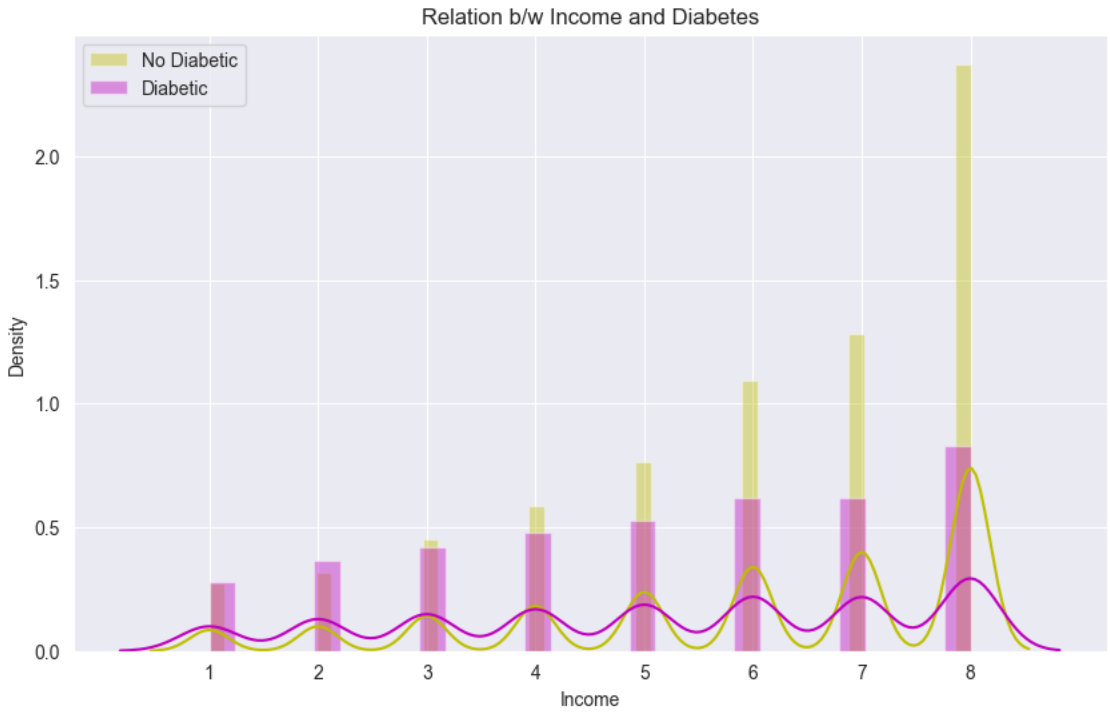
Income Categories:

Diabetes_str	Diabetic	Non-Diabetic
Education		
1.0	47	127
2.0	1183	2857
3.0	2296	7171
4.0	11032	50092
5.0	10311	56133
6.0	10228	77997

```
<matplotlib.legend.Legend at 0x236521f3650>
```



Diabetes_str	Diabetic	Non-Diabetic
Income		
1.0	2383	7408
2.0	3086	8670
3.0	3564	12356
4.0	4047	15906
5.0	4489	20837
6.0	5260	29697
7.0	5226	34905
8.0	7042	64598



We can see that increases in education and income shows a positive correlation with diabetes. This suggests that higher education and income levels may be risk factors for diabetes. This is an interesting finding and may require further investigation. We will now proceed with feature selection using machine learning models.

1.0.3 Feature Selection Using Machine Learning Models:

We have seen earlier through the correlation plots that HighBP , HighChol , BMI , smoker , stroke , HeartDiseaseorAttack , PhysActivity , Veggies , MentHlth , HvyAlcoholconsump , GenHlth , PhysHlth , Age , Education , Income and DiffWalk have a significant correlation with Diabetes_binary. We will calculate the variance inflation factor (VIF) to check for multicollinearity among the features.

```
const                   116.856706
Diabetes_binary           1.193120
HighBP                    1.344502
HighChol                  1.180932
CholCheck                 1.033501
BMI                       1.160280
Smoker                    1.091872
Stroke                    1.081612
HeartDiseaseorAttack      1.175776
PhysActivity              1.157396
Fruits                    1.112540
Veggies                   1.112397
HvyAlcoholConsump         1.025418
AnyHealthcare             1.113209
NoDocbcCost               1.144200
GenHlth                   1.821914
MentHlth                  1.239497
PhysHlth                  1.623288
DiffWalk                  1.536636
Sex                       1.075748
Age                       1.354954
Education                 1.326495
Income                    1.505649
dtype: float64
```

The VIF values are less than 5, which indicates that there is no multicollinearity among the features. We will use the ANOVA F-test to identify the most important features for predicting diabetes. We will use the F-test to calculate the p-values for each feature and select the features with the lowest p-values.

```
(253680, 10)
```

	Feature	Score	P-Value
0	HighBP	10029.013935	0.000000e+00
1	HighChol	5859.710582	0.000000e+00
2	CholCheck	39.716825	2.935854e-10
3	BMI	18355.166400	0.000000e+00
4	Smoker	521.978858	1.570423e-115
5	Stroke	2725.225194	0.000000e+00
6	HeartDiseaseorAttack	7221.975378	0.000000e+00
7	PhysActivity	861.887532	1.893271e-189
8	Fruits	154.291404	2.000073e-35
9	Veggies	153.169215	3.517963e-35
10	HvyAlcoholConsump	779.424807	1.605281e-171
11	AnyHealthcare	3.280938	7.008884e-02
12	NoDocbcCost	229.542412	7.501278e-52
13	GenHlth	9938.507776	0.000000e+00
14	MentHlth	21029.632228	0.000000e+00

15	PhysHlth	133424.406534	0.000000e+00
16	DiffWalk	10059.506391	0.000000e+00
17	Sex	140.248274	2.349212e-32
18	Age	9276.141199	0.000000e+00
19	Education	756.035496	1.954675e-166
20	Income	4829.816361	0.000000e+00

	Feature	Score	P-Value
15	PhysHlth	133424.406534	0.000000e+00
14	MentHlth	21029.632228	0.000000e+00
3	BMI	18355.166400	0.000000e+00
16	DiffWalk	10059.506391	0.000000e+00
0	HighBP	10029.013935	0.000000e+00
13	GenHlth	9938.507776	0.000000e+00
18	Age	9276.141199	0.000000e+00
6	HeartDiseaseorAttack	7221.975378	0.000000e+00
1	HighChol	5859.710582	0.000000e+00
20	Income	4829.816361	0.000000e+00
5	Stroke	2725.225194	0.000000e+00
7	PhysActivity	861.887532	1.893271e-189
10	HvyAlcoholConsump	779.424807	1.605281e-171
19	Education	756.035496	1.954675e-166
4	Smoker	521.978858	1.570423e-115
12	NoDocbcCost	229.542412	7.501278e-52

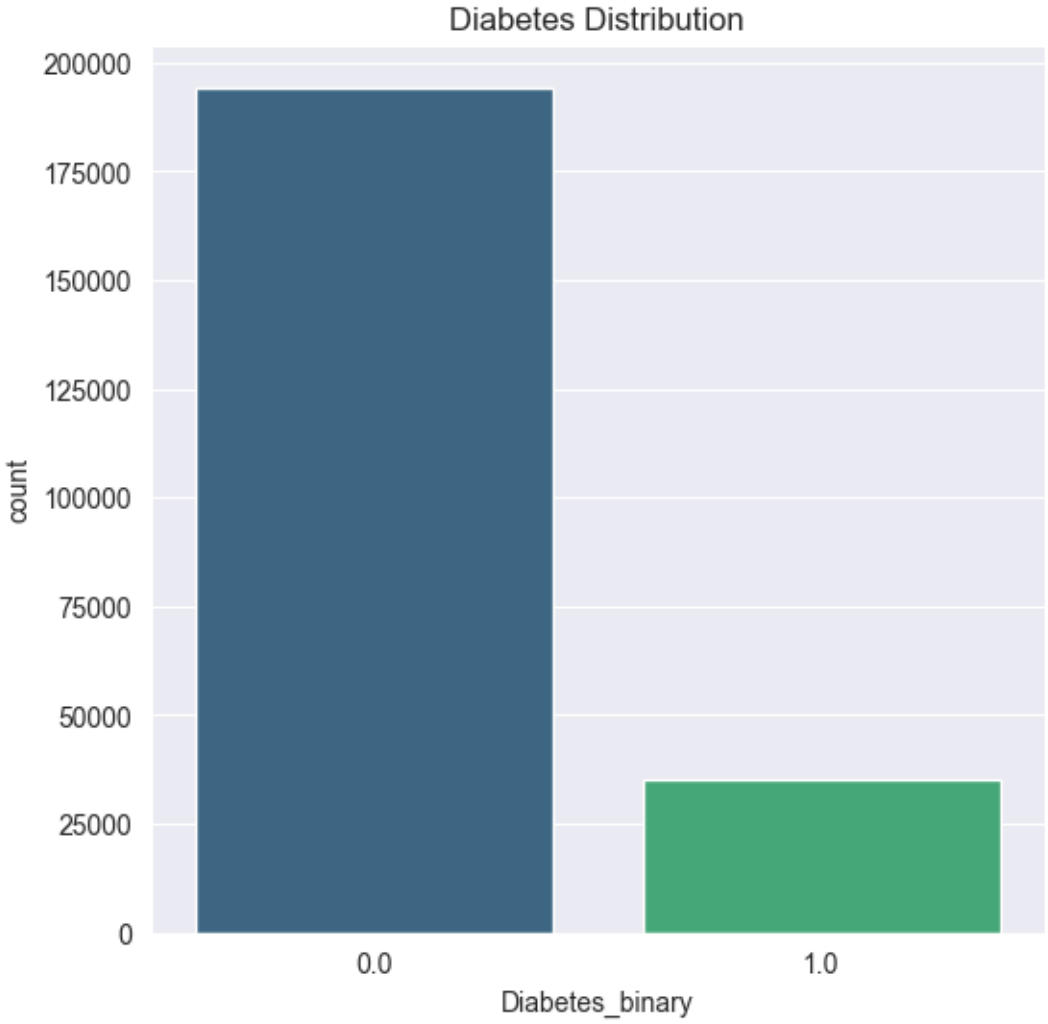
The ANOVA F-test results show the top 10 features with the highest scores for predicting diabetes. Based on the ANOVA F-score output and the provided p-values, we can summarize the significant variables as follows:

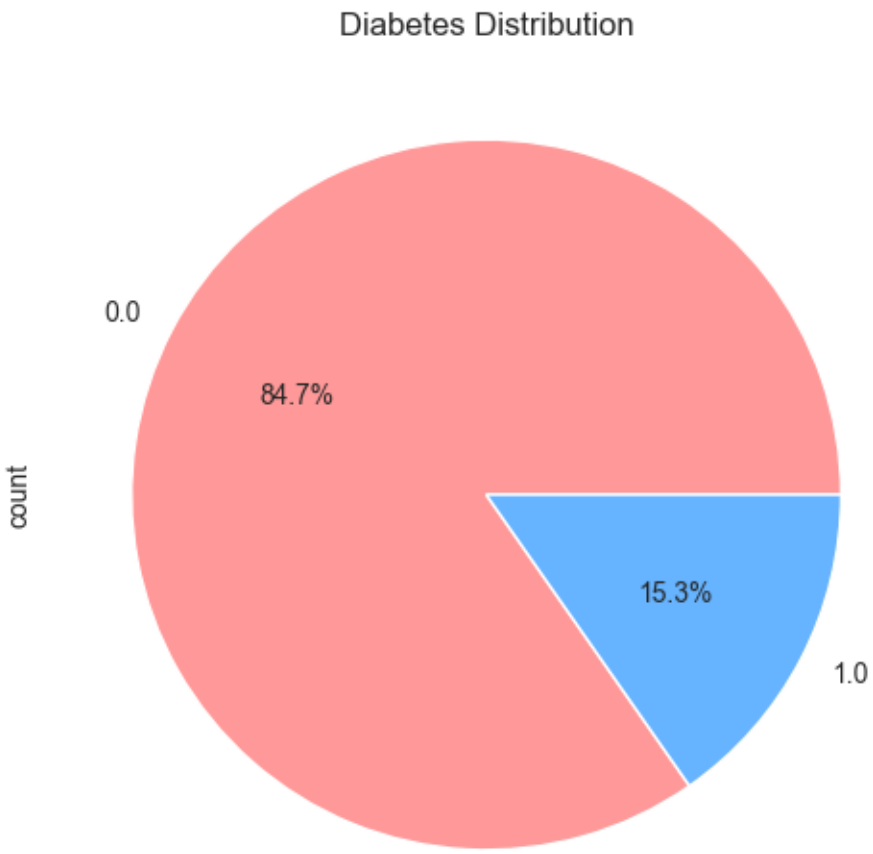
- **PhysHlth**: With the highest F-score of **133424.406534** and a p-value of **0.000000e+00**, Physical Health is the most significant variable influencing the dependent variable.
- **MentHlth**: Mental Health also shows a high level of significance with an F-score of **21029.632228** and a p-value of **0.000000e+00**.
- **BMI**: Body Mass Index is another significant predictor, having an F-score of **18355.166400** and a p-value of **0.000000e+00**.
- **DiffWalk**: Difficulty Walking has an F-score of **10059.506391** and a p-value of **0.000000e+00**, indicating its significance.
- **HighBP**: High Blood Pressure has an F-score of **10029.013935** with a p-value of **0.000000e+00**.
- **GenHlth**: General Health has an F-score of **9938.507776** and a p-value of **0.000000e+00**.
- **Age**: Age has an F-score of **9276.141199** and a p-value of **0.000000e+00**.
- **HeartDiseaseorAttack**: This variable has an F-score of **7221.975378** and a p-value of **0.000000e+00**.
- **HighChol**: High Cholesterol has an F-score of **5859.710582** and a p-value of **0.000000e+00**.
- **Income**: Income level has an F-score of **4829.816361** and a p-value of **0.000000e+00**.
- **Stroke**: Stroke has an F-score of **2725.225194** and a p-value of **0.000000e+00**.
- **PhysActivity**: Physical Activity has a lower F-score of **861.887532** but still a significant p-value of **1.893271e-189**.

- **HvyAlcoholConsump**: Heavy Alcohol Consumption has an F-score of **779.424807** and a p-value of **1.605281e-171**.

All these variables have p-values that are effectively zero, which suggests that they are statistically significant predictors of the dependent variable in your model. The variables are listed in descending order of their F-scores, which generally indicates the relative influence they have on the dependent variable, with higher scores suggesting a greater influence. However, it's important to consider other factors such as the context of the data and the model used when interpreting these results.

We will drop the following features from the dataset as they are not significant in predicting diabetes: - Fruits - Veggies - Sex - CholCheck - AnyHealthcare - NoDocbcCost - Smoker - Education





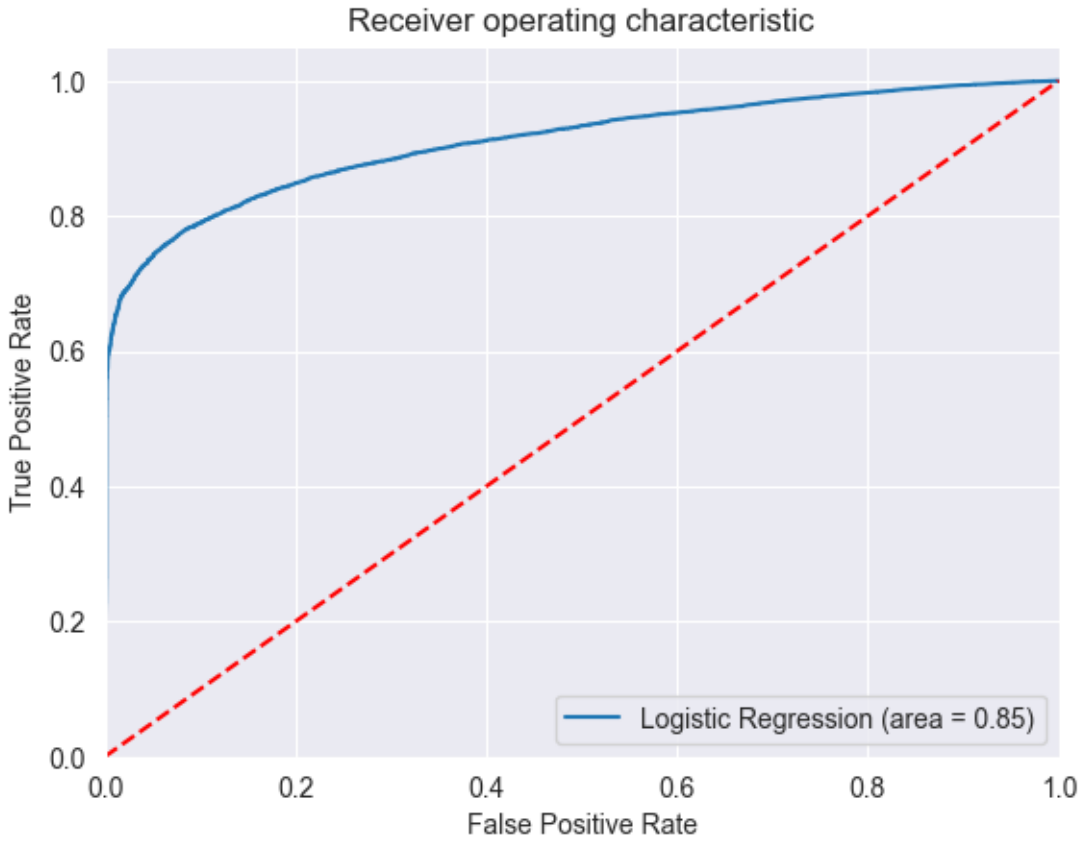
Balancing the Dataset The dataset is imbalanced as only 15.3% of the data shows diabetic individuals. We will address this imbalance by using the NearMiss algorithm to undersample the majority class and balance the dataset. This will help improve the performance of the machine learning models by reducing the bias towards the majority class. We will then take a test-train split of the data and scale the features before building the models.

```
((70194, 13), (70194,))
```

Model Building We will build the following machine learning models to predict diabetes and compare performance: - Logistic Regression - Decision Tree Classifier - Gradient Boosting Classifier

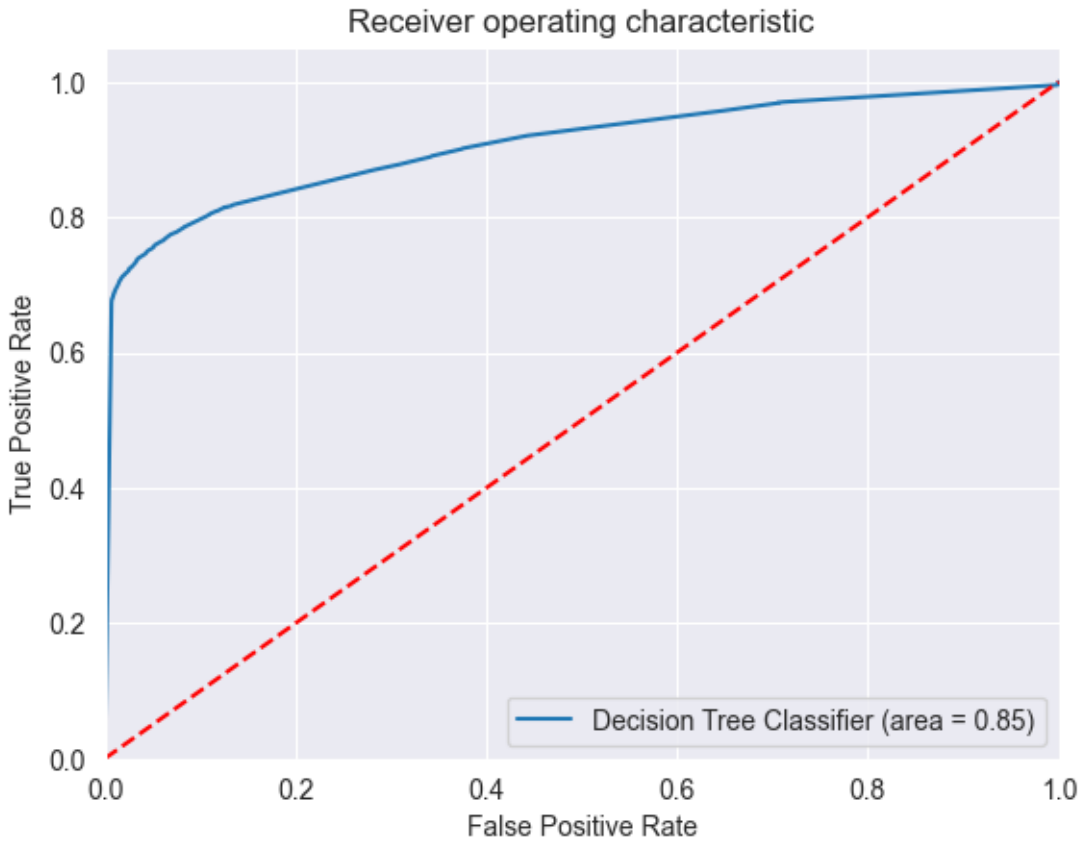
```
Logistic Regression
Accuracy: 0.8460040837646612
MSE: 0.15399591623533881
RMSE: 0.39242313417450153
Classification Report:
              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

         0.0       0.79      0.94      0.86     10468
         1.0       0.93      0.75      0.83     10591

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

Confusion Matrix:
[[9850  618]
 [2625 7966]]
```



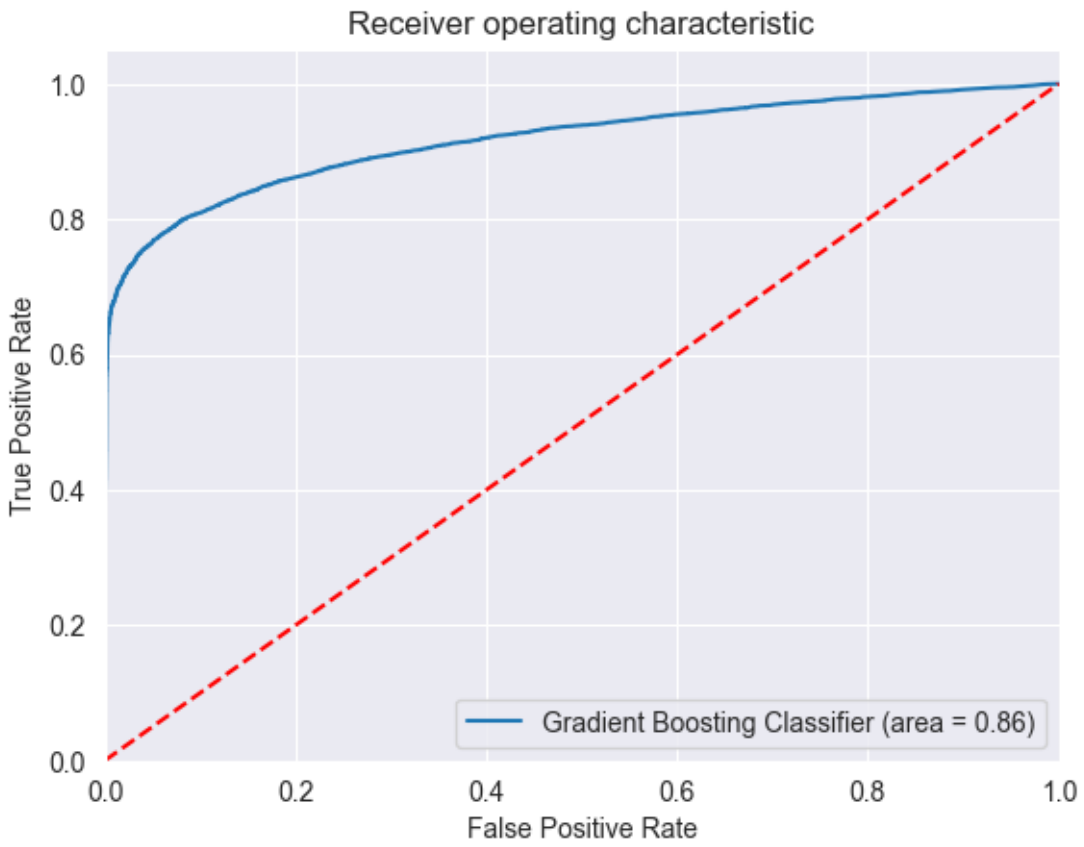

```
Decision Tree Classifier
Accuracy: 0.8510375611377559
MSE: 0.14896243886224417
RMSE: 0.3859565245753
Classification Report:
              precision    recall  f1-score   support

         0.0       0.79      0.96      0.86     10468
         1.0       0.95      0.75      0.83     10591
```

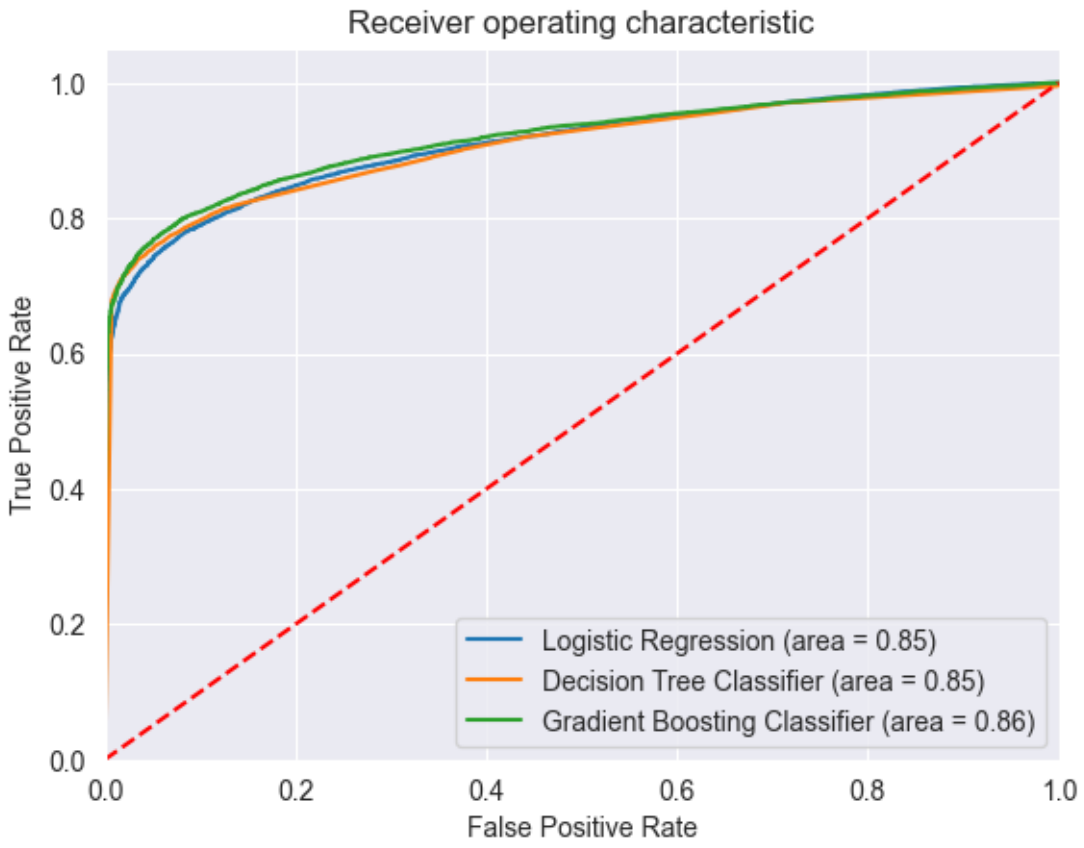
```
    accuracy                           0.85     21059
   macro avg       0.87      0.85      0.85     21059
weighted avg       0.87      0.85      0.85     21059

Confusion Matrix:
[[10023   445]
 [ 2692  7899]]
```




```
Gradient Boosting Classifier
Accuracy: 0.8582553777482311
MSE: 0.14174462225176884
RMSE: 0.376489870051996
Classification Report:
              precision    recall  f1-score   support

         0.0       0.81      0.94      0.87     10468
         1.0       0.92      0.78      0.85     10591

    accuracy                           0.86     21059
   macro avg       0.87      0.86      0.86     21059
```

```
weighted avg       0.87      0.86      0.86      21059

Confusion Matrix:
[[9788  680]
 [2305 8286]]
```





The results of the three models are as follows: * Logistic Regression: - Accuracy: 84.6% - MSE (Mean Squared Error): 0.154 - RMSE (Root Mean Squared Error): 0.392 - Key Points: - Performs well overall. - Strong at predicting the absence of diabetes (class 0), with 94% recall. - Less effective at predicting the presence of diabetes (class 1), with 75% recall. - Misclassifies a significant number of actual diabetes cases as non-diabetes.

- Decision Tree Classifier:
  - Accuracy: 85.1%
  - MSE: 0.149
  - RMSE: 0.386
  - Key Points:
    - Slightly higher accuracy than logistic regression.
    - Excellent at predicting the absence of diabetes (class 0) with 96% recall.
    - Still struggles with predicting the presence of diabetes (class 1), with 75% recall.
    - Makes fewer false negatives (predicting diabetes as absent) than logistic regression.
- Gradient Boosting Classifier:
  - Accuracy: 85.8%
  - MSE: 0.142
  - RMSE: 0.376
  - Key Points:
    - Highest accuracy among the three models.

    - Good balance between predicting both classes.
    - Performs slightly better than the decision tree in predicting the presence of diabetes (class 1).
    - Overall, the best model in terms of accuracy and error metrics.
- Comparative Assessment:

Metric	Logistic Regression	Decision Tree	Gradient Boosting
Accuracy	84.6%	85.1%	85.8%
MSE	0.154	0.149	0.142
RMSE	0.392	0.386	0.376
Recall (Class 0)	94%	96%	94%
Recall (Class 1)	75%	75%	78%

Overall:

The Gradient Boosting Classifier is the top performer in terms of overall accuracy and error metrics. It also demonstrates the best balance in predicting both the presence and absence of diabetes. All models struggle to some extent with predicting the presence of diabetes, highlighting a potential area for improvement in future model iterations. While these metrics provide valuable insights, the choice of the "best" model depends on the specific goals of the prediction task. If minimizing false negatives (missing diabetes cases) is crucial, the decision tree or gradient boosting model might be preferred, even with a slight trade-off in accuracy. On the other hand, if overall accuracy is the primary concern, the gradient boosting model is the clear winner.

Discussion and Conclusion Identifying the factors that affect the prevalence of diabetes is crucial for public health interventions and policy decisions. In this analysis, we explored the relationship between various health indicators and diabetes using data from the CDC. We found that several factors, including physiological, lifestyle, healthcare access, and demographic factors, are associated with diabetes. We used feature selection techniques to identify the most important features for predicting diabetes and built machine learning models to predict diabetes and evaluate their performance. Within the features, we found that physiological factors such as HighBP, HighChol, BMI, stroke, HeartDiseaseorAttack, and lifestyle factors such as smoking, physical activity, and dietary habits are strongly correlated with diabetes. Healthcare access factors and general health indicators also play a role in predicting diabetes. A further study is required to understand the relationship between features such as education and income with diabetes. We used the ANOVA F-test to identify the most important features for predicting diabetes and found that PhysHlth, MentHlth, BMI, HighBP, GenHlth, Age, HeartDiseaseorAttack, HighChol, Income, Stroke, PhysActivity, and HvyAlcoholConsump are the most significant predictors of diabetes.

We also used machine learning models to predict diabetes and evaluate their performance. Three classification models were built: Logistic Regression, Decision Tree Classifier, and Gradient Boosting Classifier. The Gradient Boosting Classifier performed the best in terms of overall accuracy and error metrics, demonstrating the highest accuracy and the best balance in predicting both the presence and absence of diabetes. The Decision Tree Classifier also showed promising results, with high accuracy and recall for predicting the absence of diabetes. The Logistic Regression model performed well overall but struggled with predicting the presence of diabetes. These findings can help inform public health strategies and interventions to reduce the prevalence of diabetes and

improve health outcomes. Future research could explore additional factors and develop more advanced machine learning models to further improve the prediction of diabetes and inform targeted interventions.

The present study had some limitations such as the imbalanced dataset, which was addressed using the NearMiss algorithm to undersample the majority class. The dataset was also limited in terms of the number of features and the sample size. Future studies could explore additional features and use larger datasets to improve the prediction of diabetes. Further study is required to establish causal relationships between the identified factors and diabetes and to develop more accurate prediction models. Overall, this analysis provides valuable insights into the factors influencing diabetes and demonstrates the potential of machine learning models in predicting and preventing the disease.