Predictive Model on the Likelihood of Diabetes based on the medical features.

BY EMEKA IKECHUKWU

Table of Contents:

INTRODUCTION

About the data

Detailed dataset comprising health and demographic data of 100,000 individuals, aimed at facilitating diabetes-related research and predictive modeling. This dataset includes information on gender, age, location, race, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and diabetes status.

This data set is available on Kaggle

EXPLORATORY DATA ANALYSIS

Data info

- The data have 10,000 observations and 16 columns with no missing values.
- After dropping the duplicate records, We have 99,986 observations left
- 3 columns are float, 10 int columns and 3 categorical columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 16 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   year                 100000 non-null  int64  
 1   gender               100000 non-null  object 
 2   age                  100000 non-null  float64
 3   location             100000 non-null  object 
 4   race:AfricanAmerican 100000 non-null  int64  
 5   race:Asian           100000 non-null  int64  
 6   race:Caucasian       100000 non-null  int64  
 7   race:Hispanic        100000 non-null  int64  
 8   race:Other           100000 non-null  int64  
 9   hypertension         100000 non-null  int64  
 10  heart_disease        100000 non-null  int64  
 11  smoking_history      100000 non-null  object 
 12  bmi                  100000 non-null  float64
 13  hbA1c_level          100000 non-null  float64
 14  blood_glucose_level  100000 non-null  int64  
 15  diabetes             100000 non-null  int64  
dtypes: float64(3), int64(10), object(3)
memory usage: 12.2+ MB
```

Summary statistics

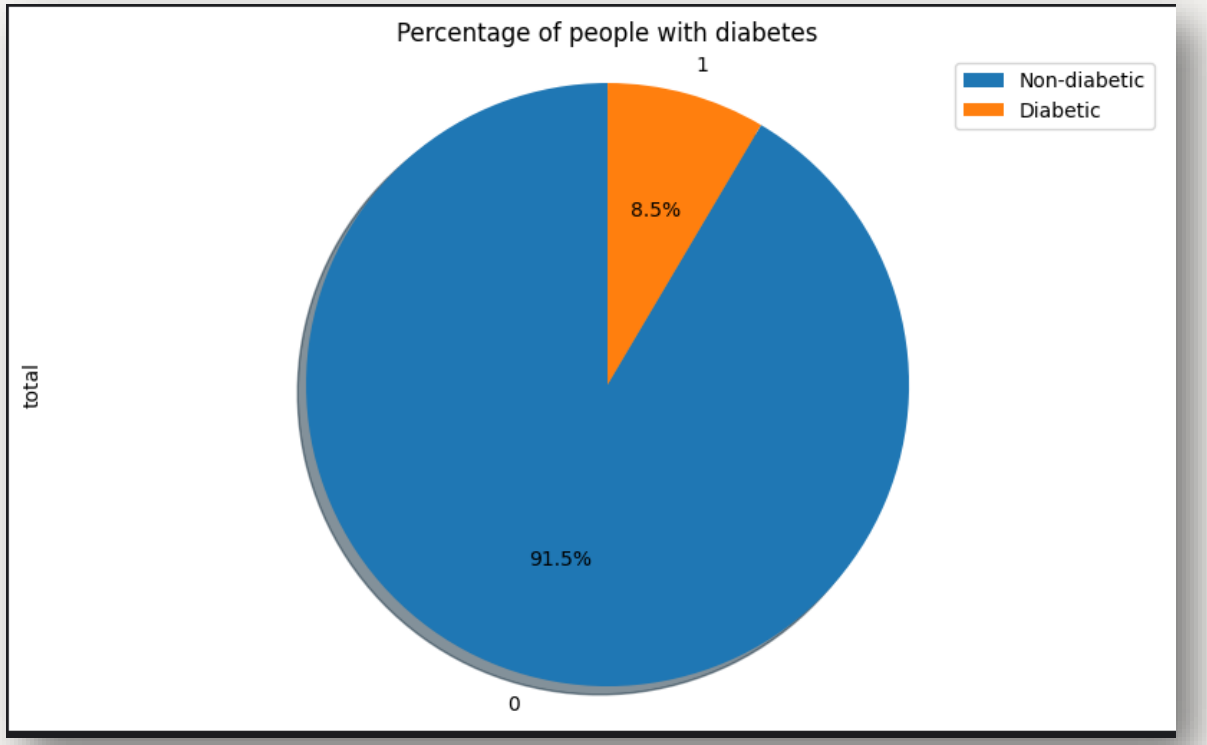
From the summary statistics;

- The average age of the individuals recorded in this dataset is approximately 42 years, with 80 being the maximum age
- Average BMI is approximately 27
- Average HbA1c level is 5.5
- Average blood glucose level is 138

	age	bmi	hbA1c_level	blood_glucose_level
count	99986.000000	99986.000000	99986.000000	99986.000000
mean	41.885930	27.320767	5.527517	138.059518
std	22.516409	6.637248	1.070674	40.708667
min	0.080000	10.010000	3.500000	80.000000
25%	24.000000	23.630000	4.800000	100.000000
50%	43.000000	27.320000	5.800000	140.000000
75%	60.000000	29.580000	6.200000	159.000000
max	80.000000	95.690000	9.000000	300.000000

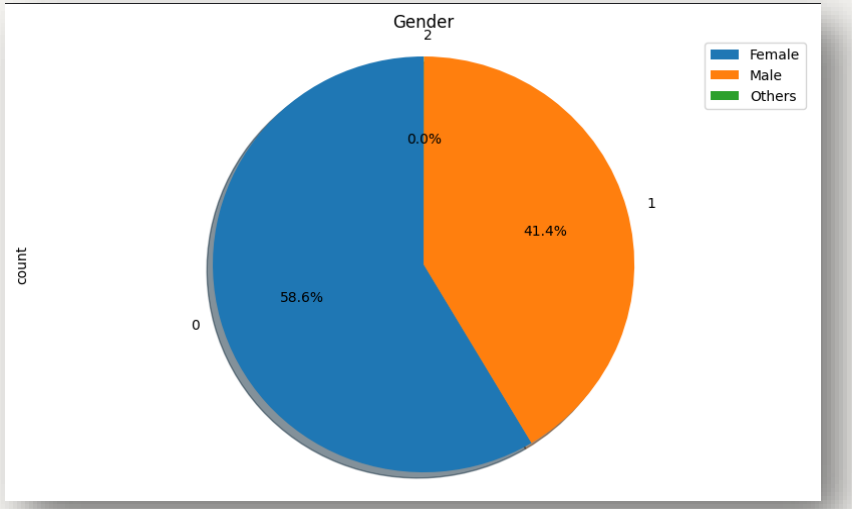
Number of individuals in the data set that have diabetes

From the pie chart above, 91.5% of the data do not have diabetes while 8.5% have diabetes



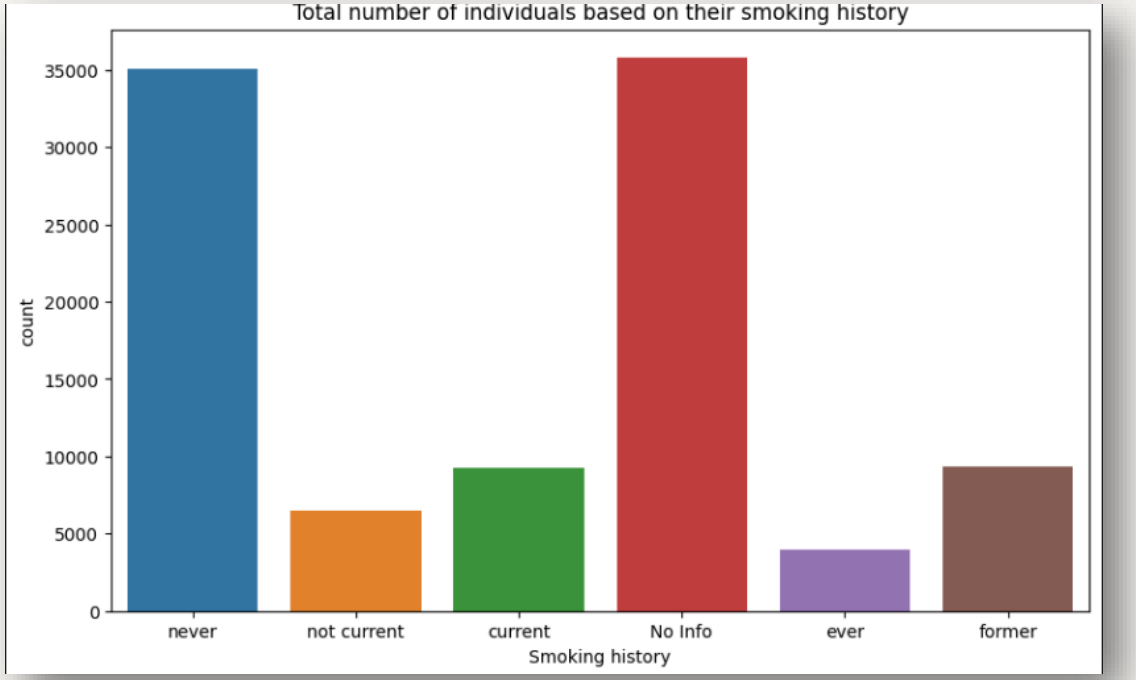
Gender distribution

- 58.6% of the individuals recorded in the data set are females, 41.4% are males while those that identified as others are very insignificant in comparison of with the former.
- Out of the 58,552 females in the data set, only 4461 have diabetes. 4,039 out of the 41430 males have diabetes. 18 individuals that identified as 'Other' do not have diabetes



	gender	diabetes	count
0	Female	0	54085
1	Female	1	4461
2	Male	0	37383
3	Male	1	4039
4	Other	0	18

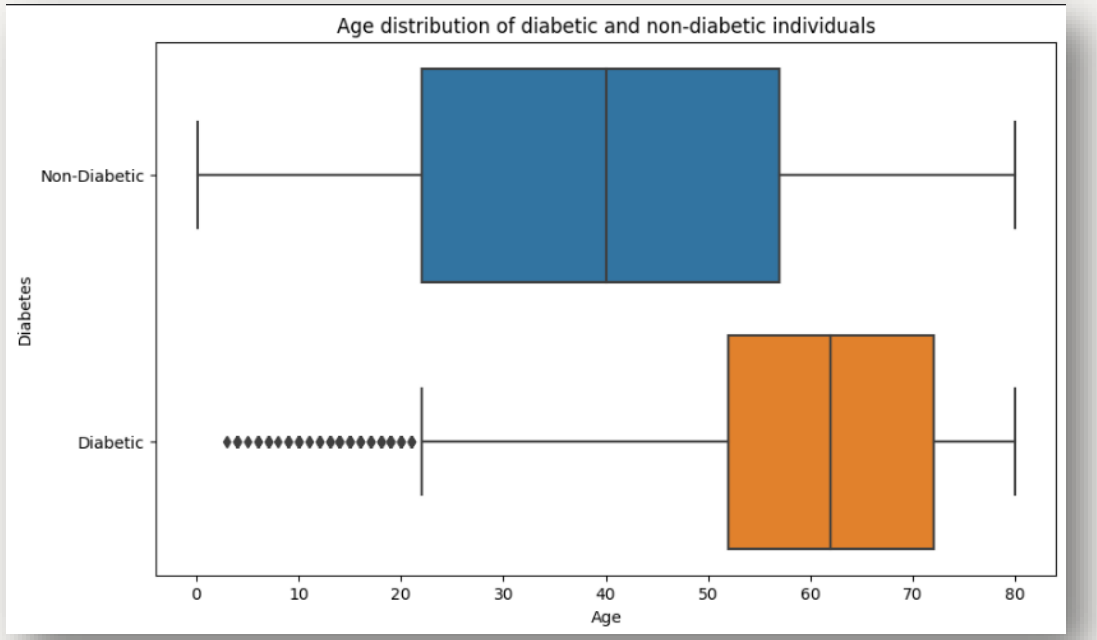
Total number of individuals in the data set based on their smoking history



	smoking_history	count
0	No Info	35806
1	never	35091
2	former	9352
3	current	9286
4	not current	6447
5	ever	4004

From the visualization above, the we have more individuals with 'No info' smoking history, followed by those that have 'Never' smoked

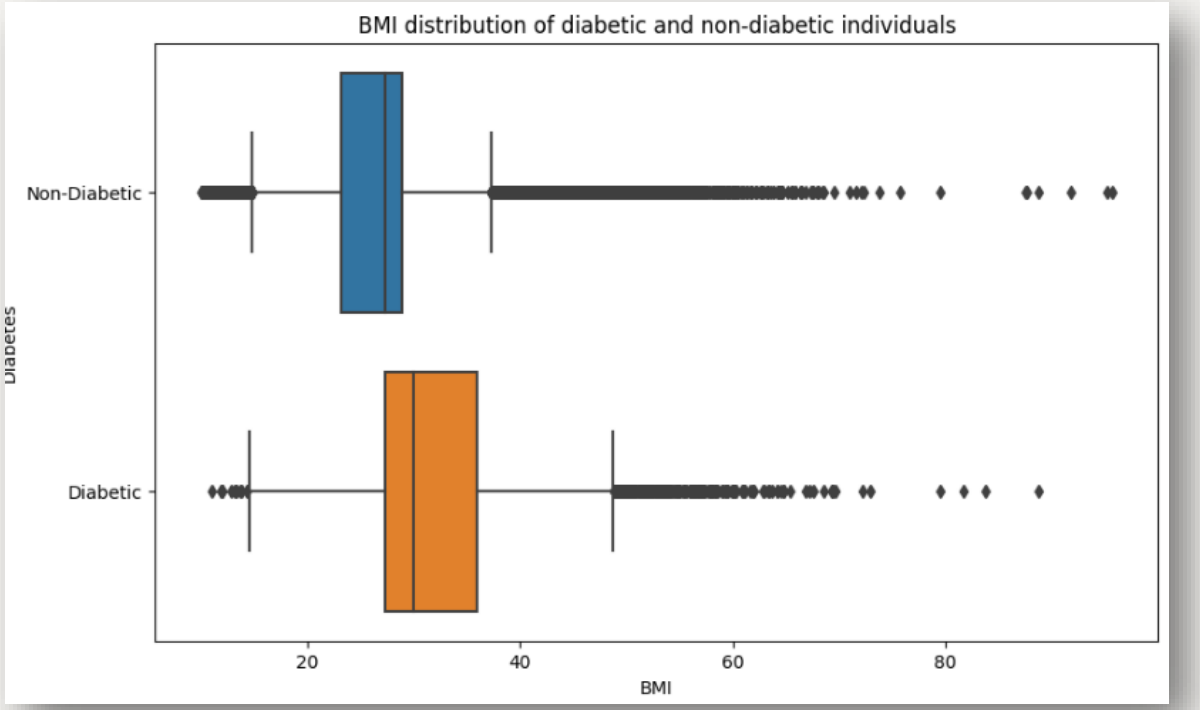
Age distribution of Diabetic and Non-diabetic individuals



From the boxplot above, the average age of non-diabetic individual is around 40 years of agee while that of diabetic individual is around 62 years of age. The age distribution of the individuals without diabetes is approximately normally distributed, whilst those with diabetes is skewed to the left.

BMI distributions for Diabetic and Non-diabetic individuals

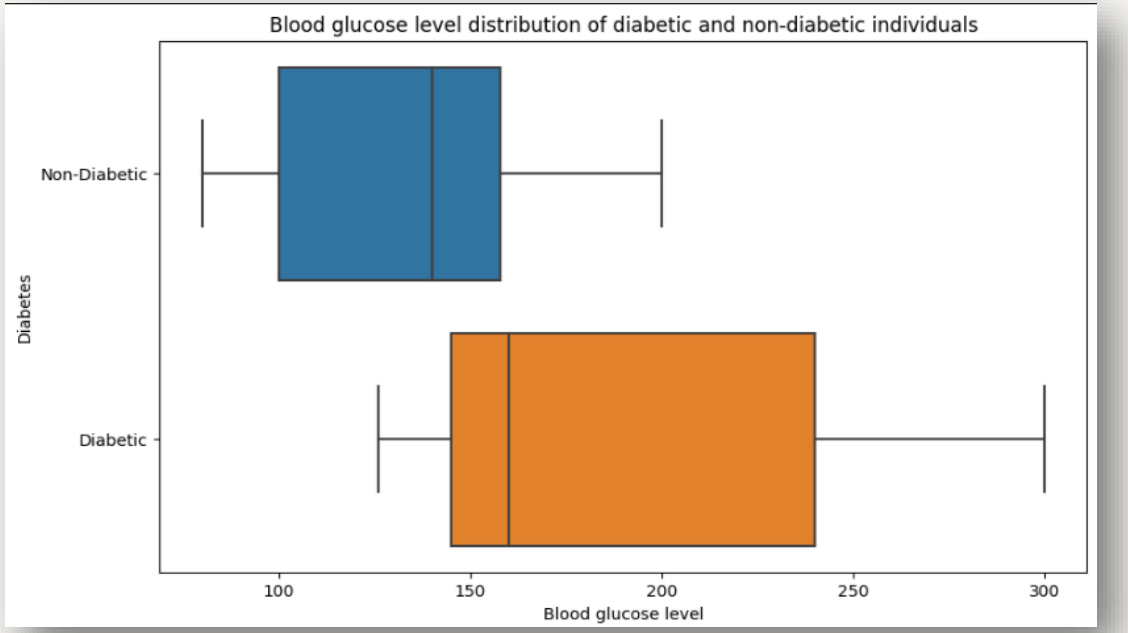
The distributions of the BMI of both individuals living with diabetes is rightly skewed due to the presence of extreme BMI values in the dataset suggesting there are many obese individuals recorded int the data set. The average BMI of individuals with diabetes is lesser than those with diabetes



Blood glucose level distributions of Diabetic and Non-diabetic individuals

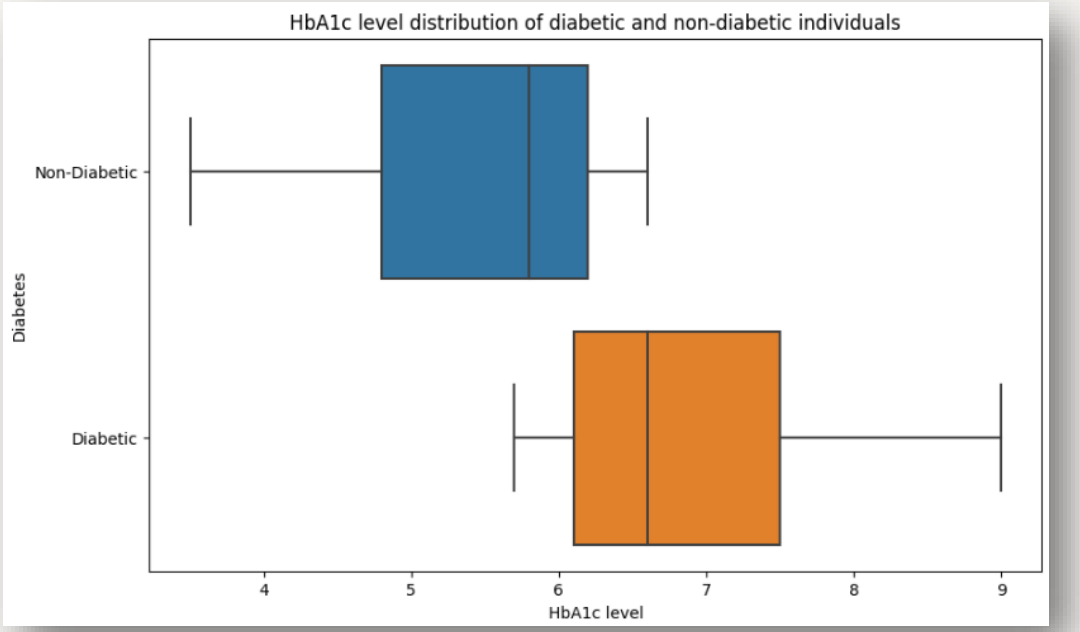
The blood glucose level of individuals with diabetes ranges from approximately 60mg/dl to 200mg/dl while those with diabetes ranges between 130mg/dl to 300mg/dl. This is expected because high blood glucose level is one of the major manifestation or sypmtom of diabetes.

The average blood glucose level of people without diabetes is a little below 150mg/dl and those with diabetes is around 160mg/dl



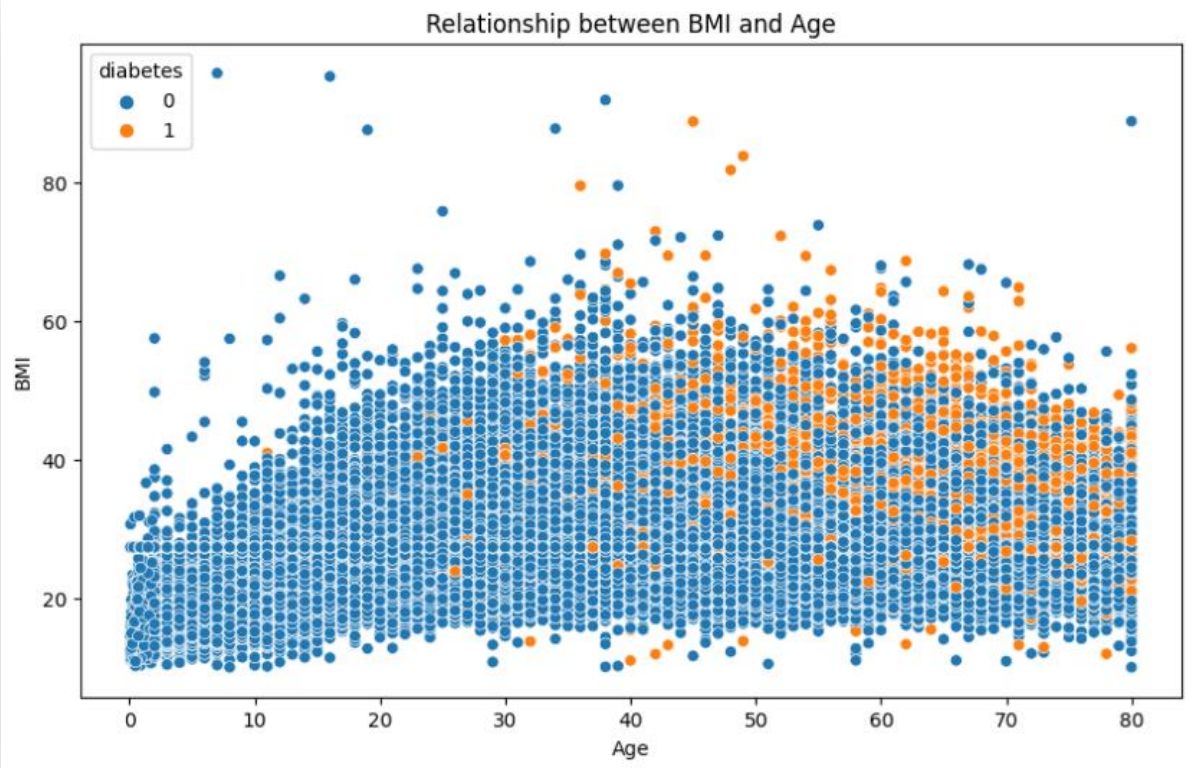
HbA1c level distribution of Diabetic and Non-diabetic individuals

HbA1c level of individuals with diabetes ranges between 2 to 6.6 while that of those without is from 5.8 to 9. This is equally expected as increased level of HbA1c is closely associated with diabetes. The average HbA1c level of the former category is 5.8 while that of the later 6.7



Relationship between Age and BMI

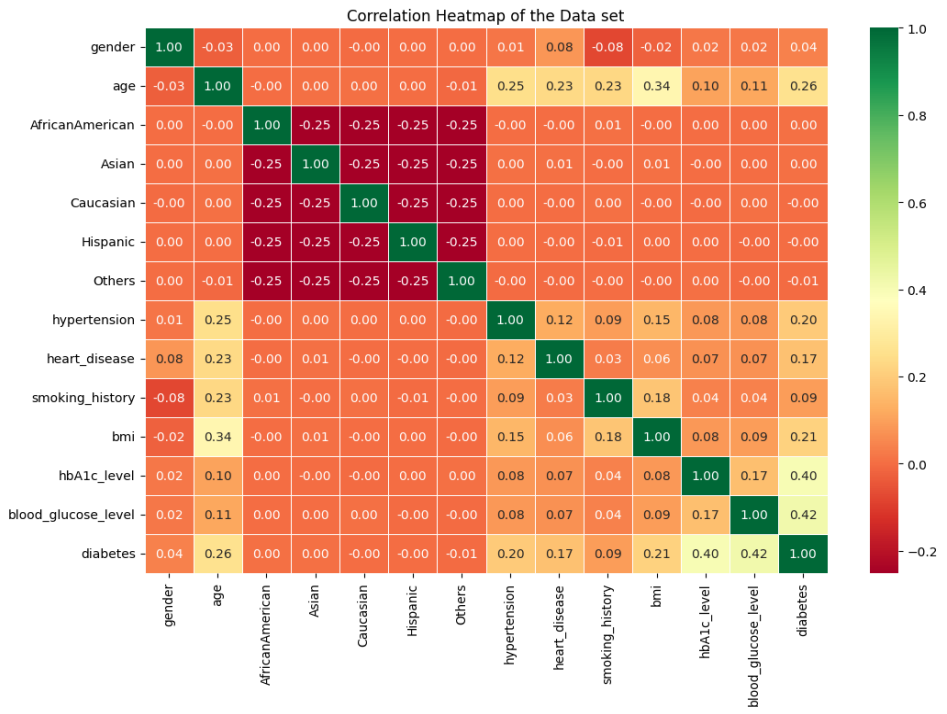
Age and BMI are weakly correlated, however, people with diabetes showed up more as age and BMI increases.



FEATURE SELECTION/ENGINEERING

Feature Selection/Engineering

- Used 'LabelEncoder()' to transform categorical columns, gender and smoking history, into numerical columns.
- Dropped columns that does not seem to improve the model' s predictability
- Used 'StandardScaler()' to normalize/standardize the data
- Built a correlation matrix to ensure that none of the variables are strongly correlated to eliminate multicollinearity

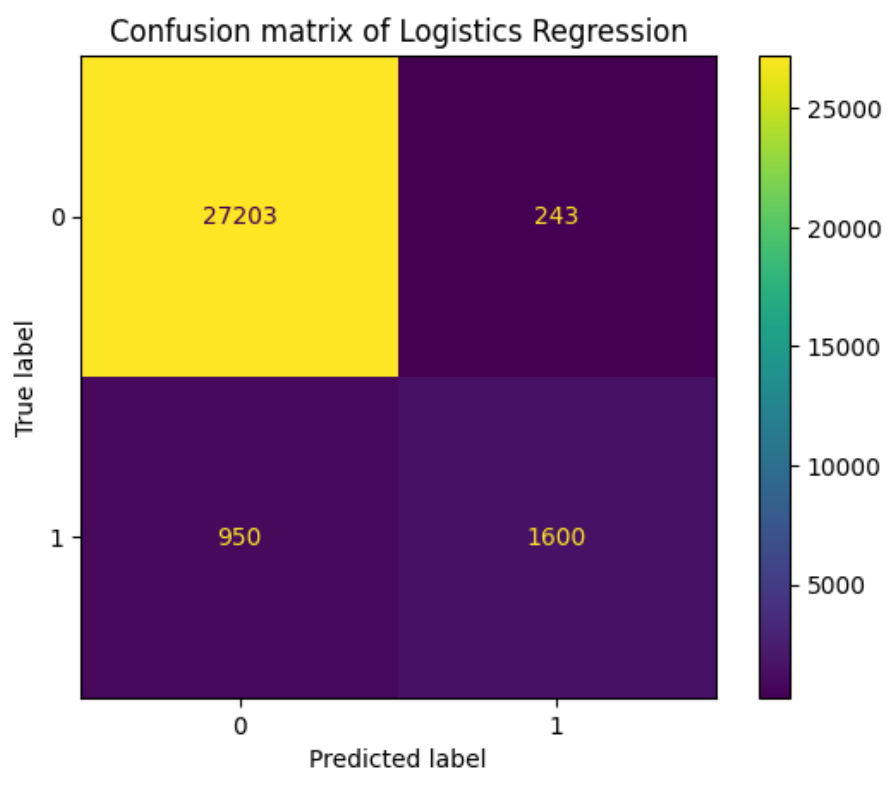


MODELING AND MODEL INTERPRETATION

Base model - Logistic Regression

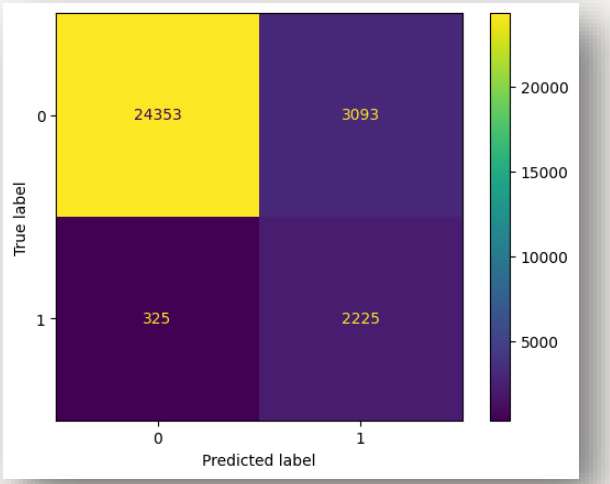
Logistic regression model performed considerably well for a baseline model with 96% overall accuracy. The recall for class 0 (individuals without diabetes) is 99% and 62% for class 1, 97% precision for class 0 and 86% for class 1.

This is mostly due to data the data being unbalanced with 8.5% of the records being positive outcomes and 91.5% being negative outcomes, and can affect the generalisabilty of our models. To fix this, I tried assigning weights to the classes.



Logistic Regression (Weighted classes)

```
{'accuracy': 0.886051473529804,
 'recall': array([0.88730598, 0.87254902]),
 'precision': array([0.98683038, 0.41839037]),
 'f1score': array([0.9344256, 0.5655821])}
```

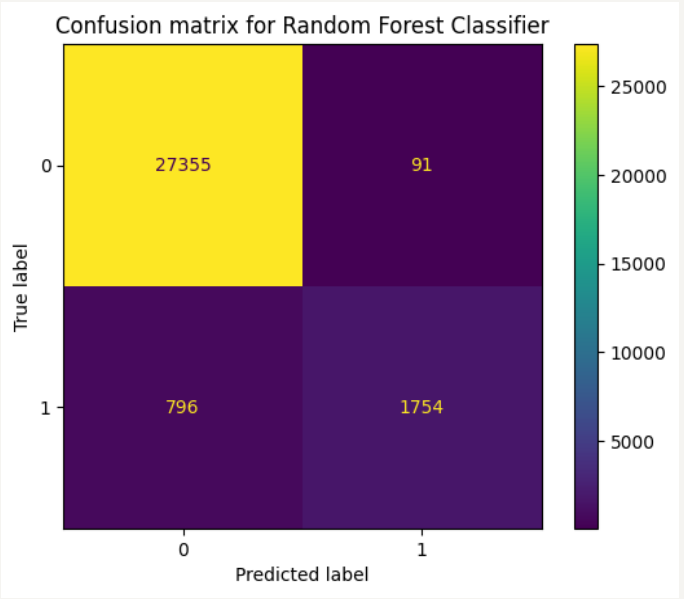


From the metrics and Confusion matrix above, you can see that assigning weights to the class do not improve the model

Random Forest Classifier

Random Forest classifier performed significantly better than our baseline model. There' s significant different in the false negative and positives instances, especially the later.

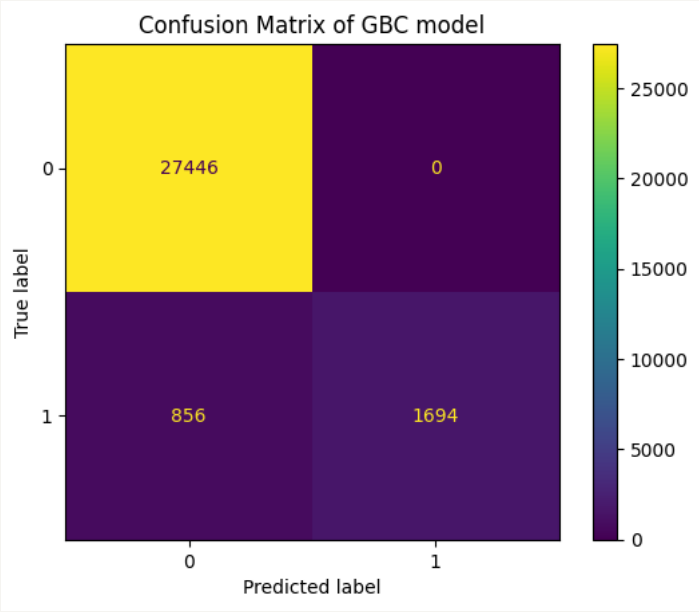
```
{'accuracy': 0.9704293905854113,
 'recall': array([0.9966844 , 0.68784314]),
 'precision': array([0.97172392, 0.95067751]),
 'f1score': array([0.9840459 , 0.79817975])}
```



Gradient Boosting Classifier

GBC performed better than Logistic regression and Random forest classifier

```
{'accuracy': 0.9714628617148953,
 'recall': array([1.        , 0.66431373]),
 'precision': array([0.96975479, 1.        ]),
 'f1score': array([0.98464519, 0.79830349])}
```



Conclusion

Gradient Boosting classifier performed better than Logistic Regression and Random Forest classifier. However, we need more positive instances (more data of people with diabetes to balance the classes) to improve our model's generalizability and performance.

Thank You!!!