FINAL PROJECT PRESENTATION

Diabetes

Habits

DS105
DIABETES HEALTH INDICATOR DATASET

Presented by Nurul Liyana

Symptoms

Health

AGENDA

1. **OBJECTIVE**

The Problem statement

2. **THE APPROACH**

Overall approach to the problem

3. **DATA PREPROCESSING**

Key steps in data preprocessing

4. **DATA UNDERSTANDING/ EXPLORATION**

Key findings/insights in data understanding/exploration

5. **TRAINING MODELS**

Training process of the machine learning model

6. **MODEL EVALUATION**

Evaluation metric(s), The chosen one

7. **MODEL IMPORTANCE**

Feature importance and Permutation importance, SHAP visual included.

8. **MODEL PERFORMANCE**

Model recommendations included

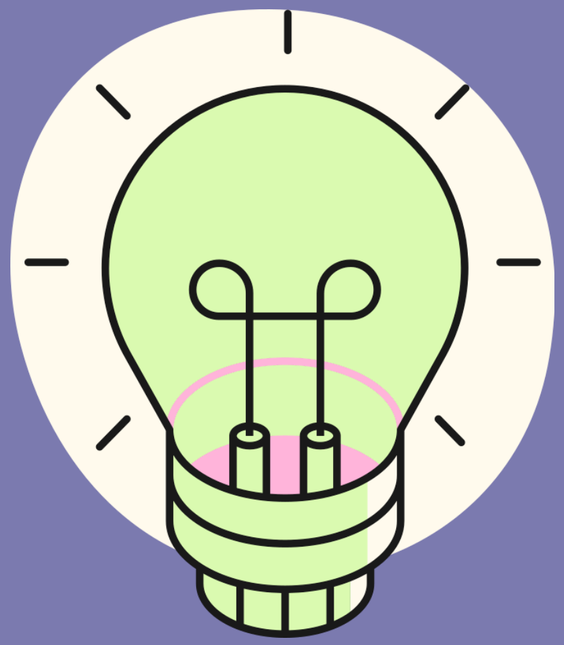
9. **CONCLUSION**

Conclusion on project

1. OBJECTIVE

THE PROBLEM STATEMENT

What are the major risk factors of getting Diabetes, does these factors alert an individual to cut down or to improve on their intakes/ activities, and using this Machine Learning to predict with these factors given to early detect if a person has Diabetes.



Supervised Learning:

1. Classification Algorithms
    a. K-Nearest Neighbour
    b. Naive Bayes
    c. Decision Tree
    d. Linear Support Vector Machine
    e. Logistic Regression

Ensemble Learning:

1. Bagging Algorithm
    a. Random Forest
2. Boosting Algorithm
    a. XGBoost
    b. CatBoost
    c. AdaBoost

Model Score:

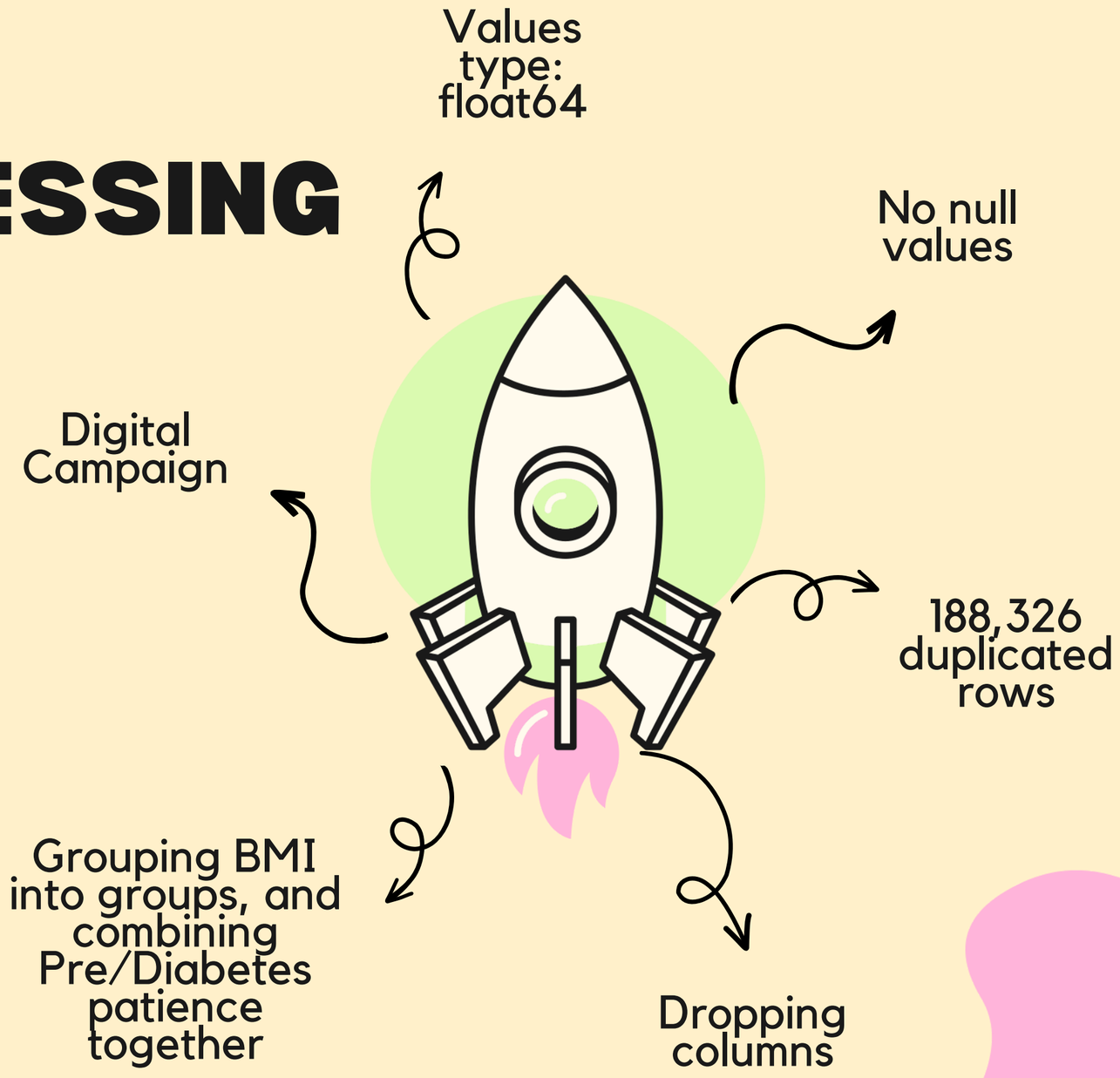
1. Receiver operating characteristic(ROC)
   Area under curve(AUC)
2. Accuracy Score
3. Confusion Matrix Score

Hyperparameter tuning:

1. GridsearchCV
2. Randomized search
3. Manual tuning

3. DATA PREPROCESSING

- Duplication rows remains as it may assume to be different patient with similar age, education, income group
- Combining Pre-Diabetes and Diabetes patients as they already been detect with diabetes.
- Putting the BMI feature to be group into 4 category as it has 84 unique values, underweight/ healthy weight/ overweight/ obese.
- Dropping columns that has not much relations to Diabetes.



To see if these factors are the major risk in getting Diabetes:

1. High Blood Pressure and High Cholesterol are risk factors of Diabetes

2. Which BMI group are most affected getting Diabetes and what did they claim for their general health

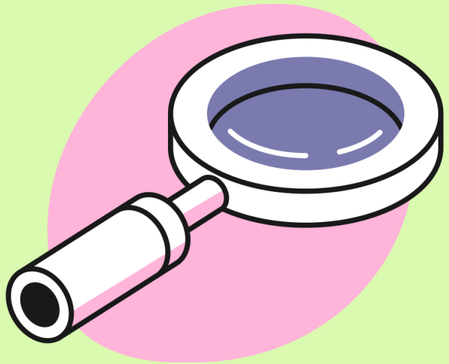
3. Does eating fruits and veggies lessen the risk of getting Diabetes

4. If getting stroke and heart conditions also risk of Diabetes

5. Does Diabetes patients do physical activity and has walking problem

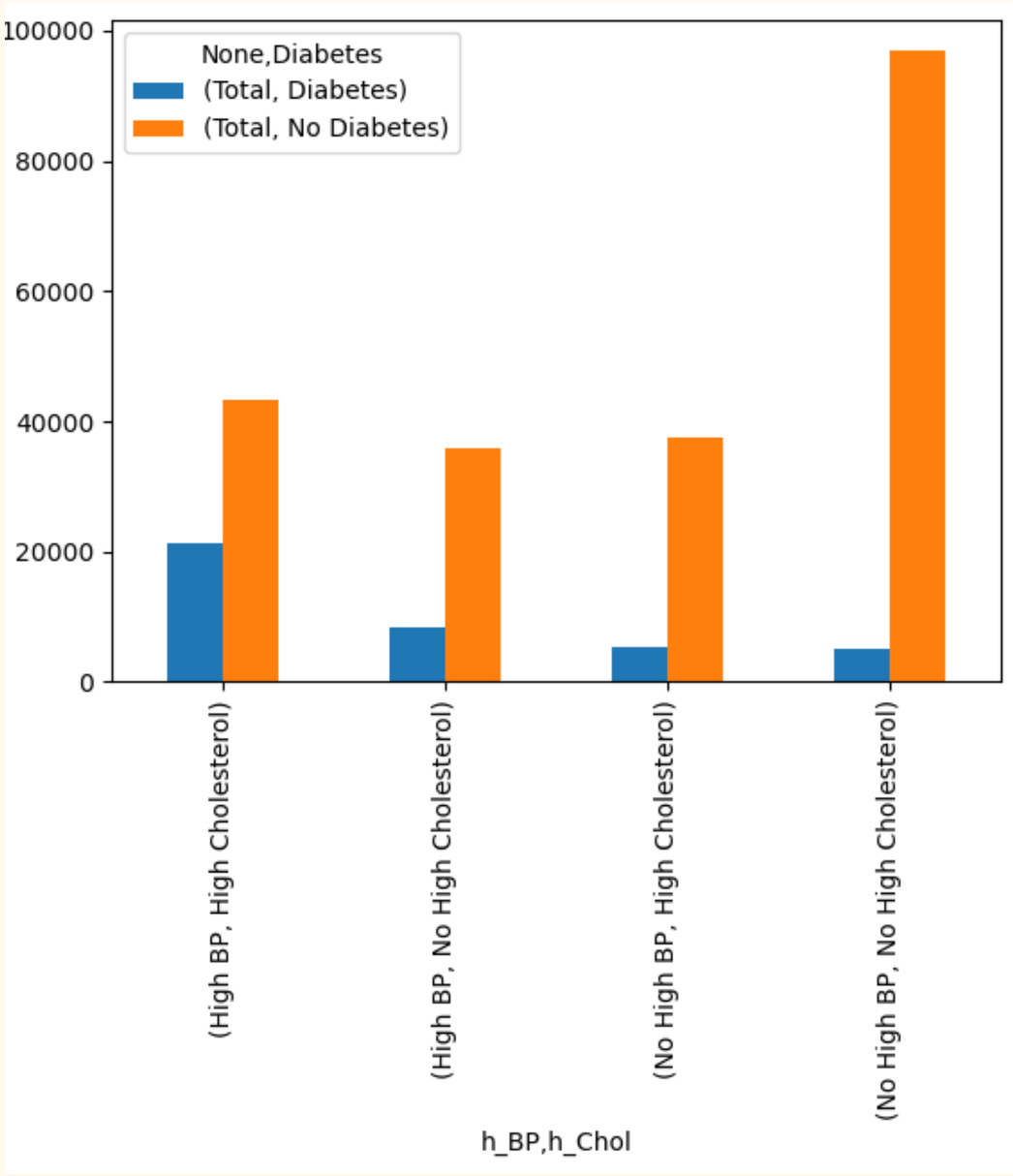
6. Being a Heavy alcohol consumption and Smoking are risk factors of Diabetes

7. Diabetes patient's most affected in which gender and which group of Age

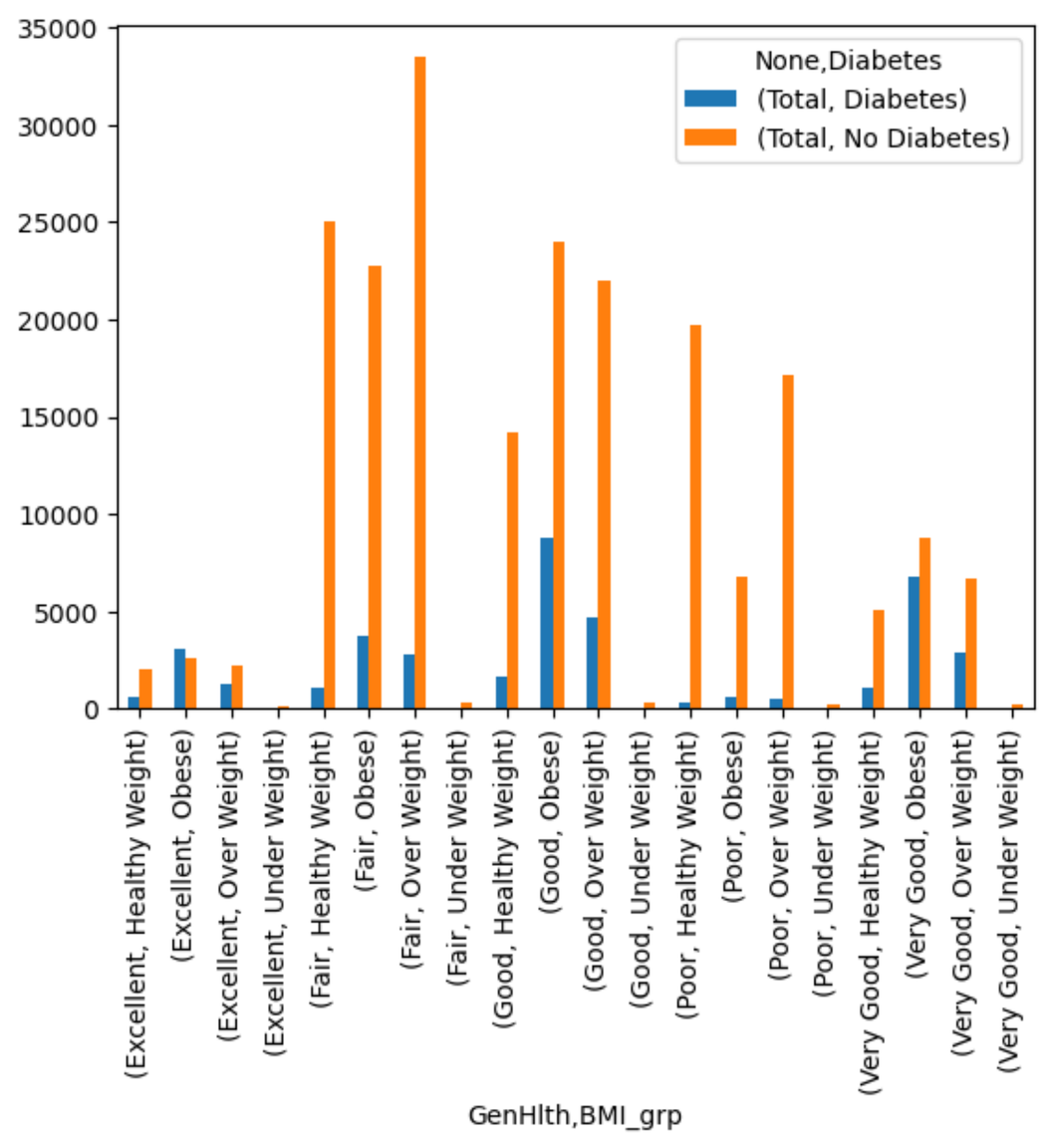


4.1 HIGH BLOOD PRESSURE AND HIGH CHOLESTEROL ARE RISK FACTORS OF DIABETES

h_BP	h_Chol	Diabetes	Total
High BP	High Cholesterol	Diabetes	21233
		No Diabetes	43427
	No High Cholesterol	Diabetes	8284
		No Diabetes	35885
No High BP	High Cholesterol	Diabetes	5328
		No Diabetes	37603
	No High Cholesterol	Diabetes	5132
		No Diabetes	96788



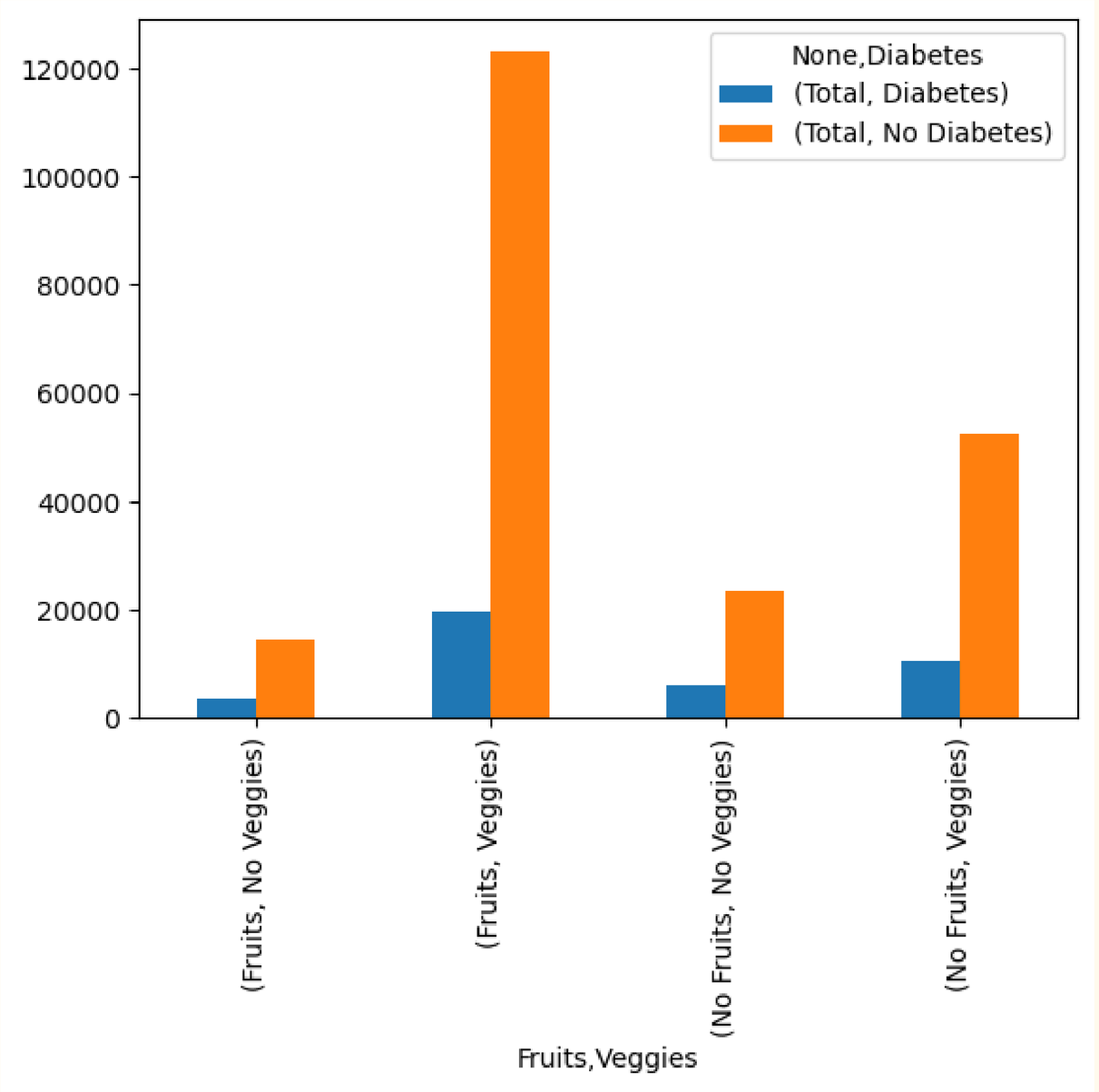
4.2 WHICH BMI GROUP ARE MOST AFFECTED GETTING DIABETES AND WHAT DID THEY CLAIM FOR THEIR GENERAL HEALTH



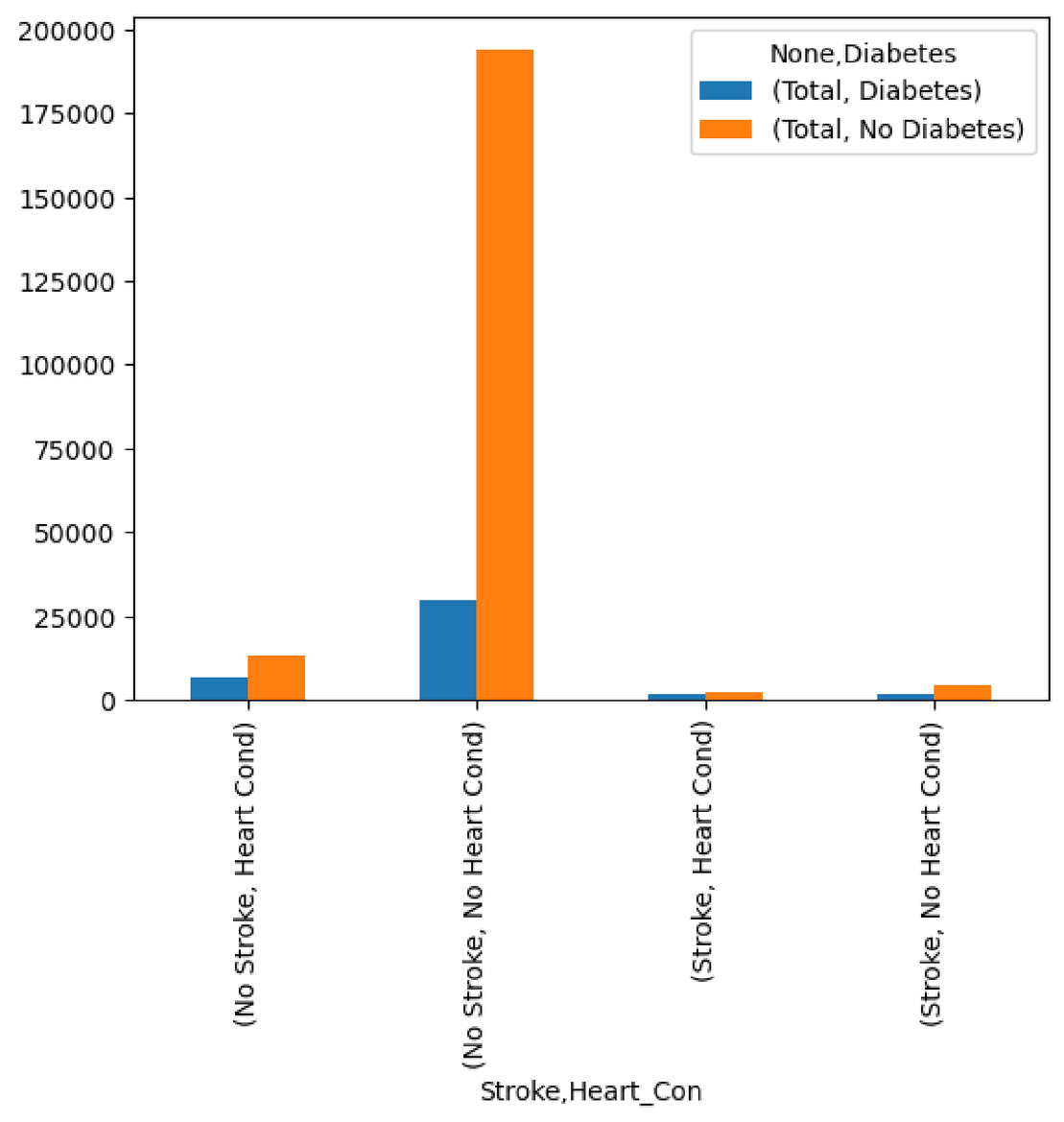
			Total
GenHlth	**BMI_grp**	**Diabetes**	
Excellent	Healthy Weight	Diabetes	574
		No Diabetes	2083
	Obese	Diabetes	3060
		No Diabetes	2653
	Over Weight	Diabetes	1256
		No Diabetes	2254
	Under Weight	Diabetes	39
		No Diabetes	162
Fair	Healthy Weight	Diabetes	1072
		No Diabetes	25013
	Obese	Diabetes	3711
		No Diabetes	22710
	Over Weight	Diabetes	2800
		No Diabetes	33451
	Under Weight	Diabetes	12
		No Diabetes	315
Good	Healthy Weight	Diabetes	1652
		No Diabetes	14200
	Obese	Diabetes	8828
		No Diabetes	23960
	Over Weight	Diabetes	4686
		No Diabetes	21975
	Under Weight	Diabetes	19
		No Diabetes	326
Poor	Healthy Weight	Diabetes	303
		No Diabetes	19708
	Obese	Diabetes	586
		No Diabetes	6781
	Over Weight	Diabetes	563
		No Diabetes	17132
	Under Weight	Diabetes	1
		No Diabetes	225
Very Good	Healthy Weight	Diabetes	1114
		No Diabetes	5037
	Obese	Diabetes	6760
		No Diabetes	8802
	Over Weight	Diabetes	2913
		No Diabetes	6719
	Under Weight	Diabetes	28
		No Diabetes	197

4.3 DOES EATING FRUITS AND VEGGIES LESSEN THE RISK OF GETTING DIABETES

			Total
Fruits	**Veggies**	**Diabetes**	
Fruits	No Veggies	Diabetes	3734
		No Diabetes	14452
	Veggies	Diabetes	19748
		No Diabetes	122964
No Fruits	No Veggies	Diabetes	5946
		No Diabetes	23707
	Veggies	Diabetes	10549
		No Diabetes	52580



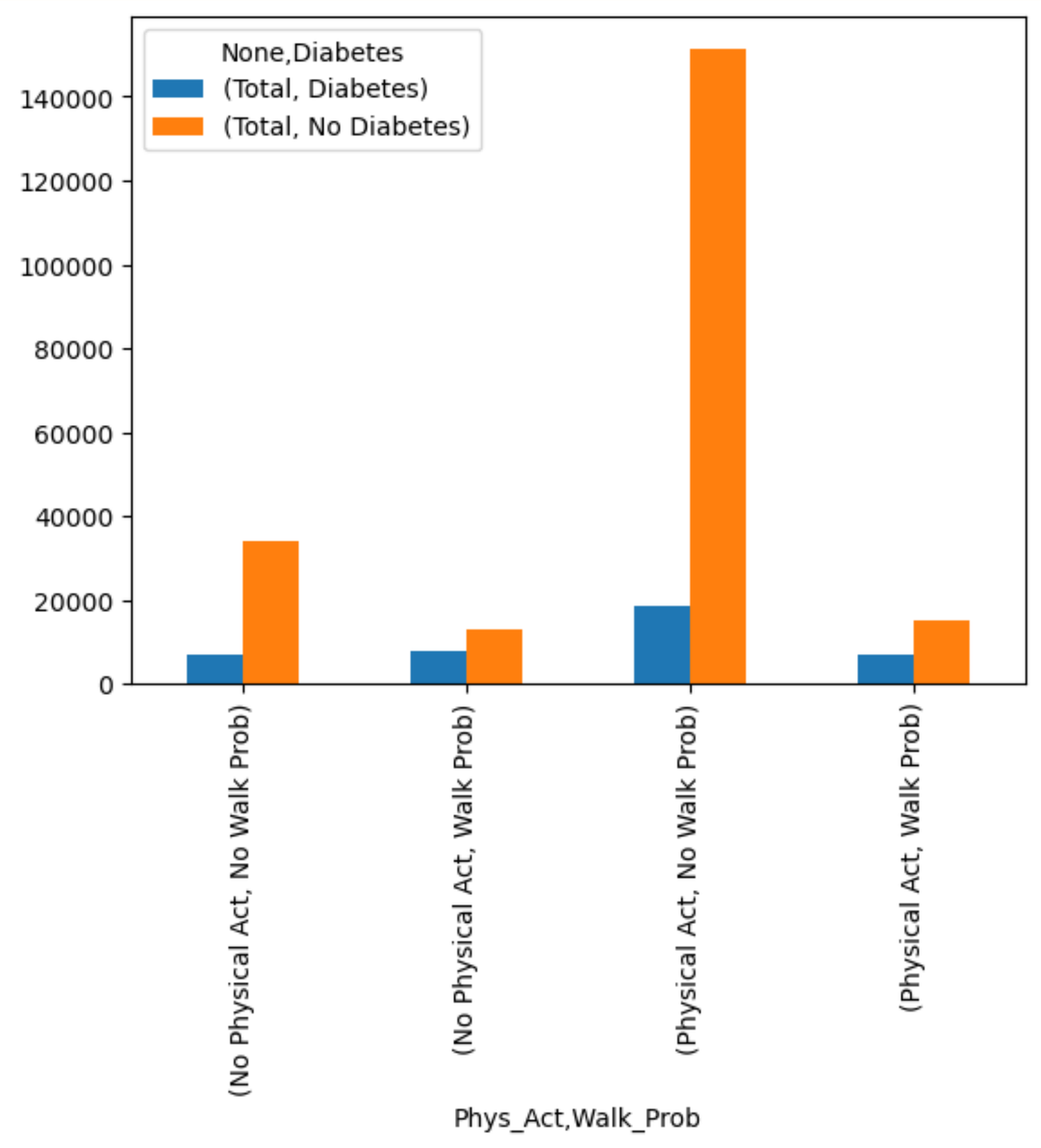
4.4 IF GETTING STROKE AND HEART CONDITIONS ALSO RISK OF DIABETES



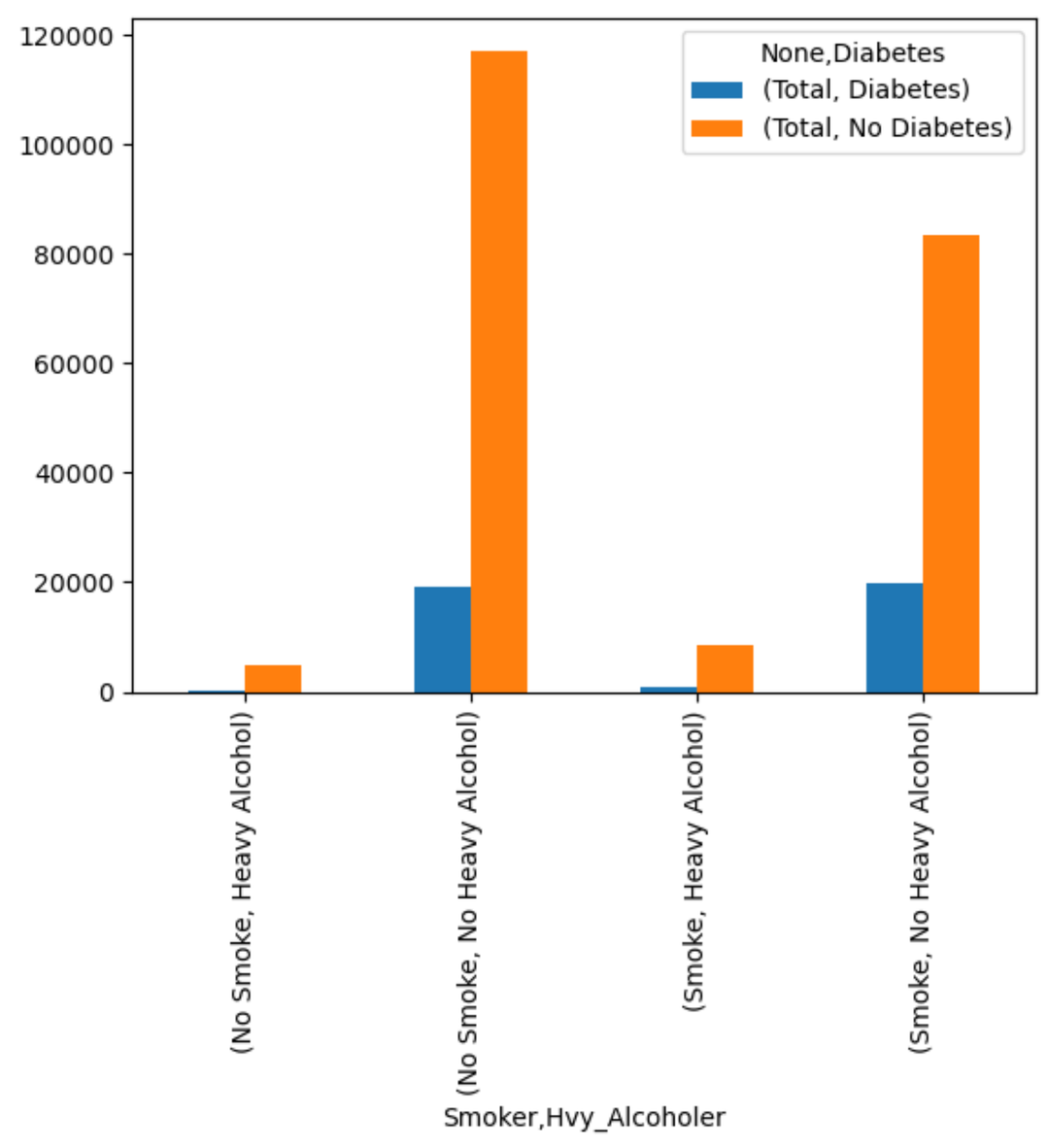
			Total
Stroke	**Heart_Con**	**Diabetes**	
No Stroke	Heart Cond	Diabetes	6805
		No Diabetes	13151
	No Heart Cond	Diabetes	29639
		No Diabetes	193793
Stroke	Heart Cond	Diabetes	1737
		No Diabetes	2200
	No Heart Cond	Diabetes	1796
		No Diabetes	4559

4.5 DOES DIABETES PATIENTS DO PHYSICAL ACTIVITY AND HAS WALKING PROBLEM

			Total
Phys_Act	**Walk_Prob**	**Diabetes**	
No Physical Act	No Walk Prob	Diabetes	6980
		No Diabetes	34080
	Walk Prob	Diabetes	7568
		No Diabetes	13132
Physical Act	No Walk Prob	Diabetes	18591
		No Diabetes	151354
	Walk Prob	Diabetes	6838
		No Diabetes	15137



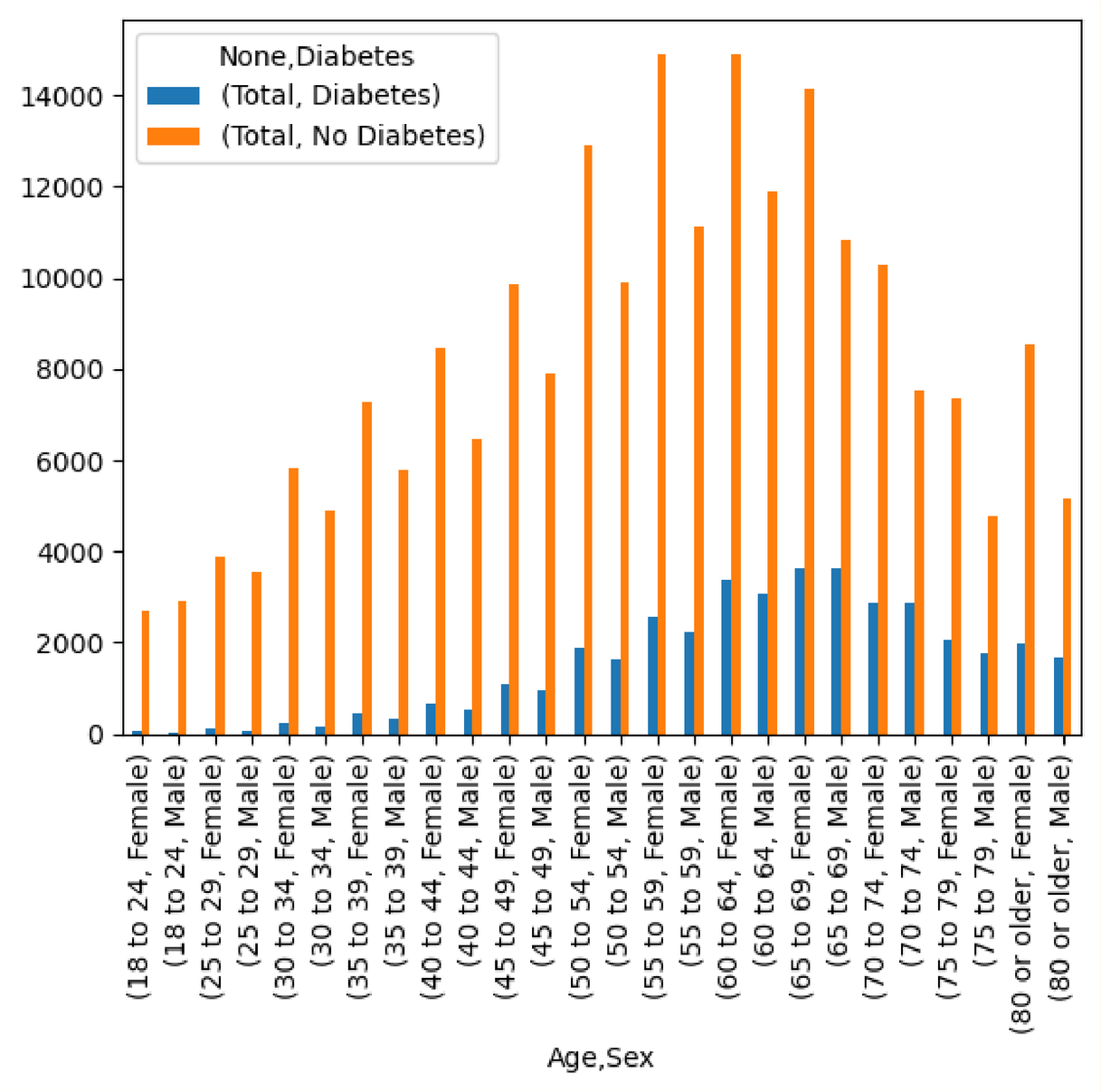
4.6 BEING A HEAVY ALCOHOL CONSUMPTION AND SMOKING ARE RISK FACTORS OF DIABETES



			Total
Smoker	**Hvy_Alcoholer**	**Diabetes**	
No Smoke	Heavy Alcohol	Diabetes	266
		No Diabetes	4723
	No Heavy Alcohol	Diabetes	19112
		No Diabetes	117156
Smoke	Heavy Alcohol	Diabetes	774
		No Diabetes	8493
	No Heavy Alcohol	Diabetes	19825
		No Diabetes	83331

4.7. DIABETES PATIENT'S MOST AFFECTED IN WHICH GENDER AND WHICH GROUP OF AGE

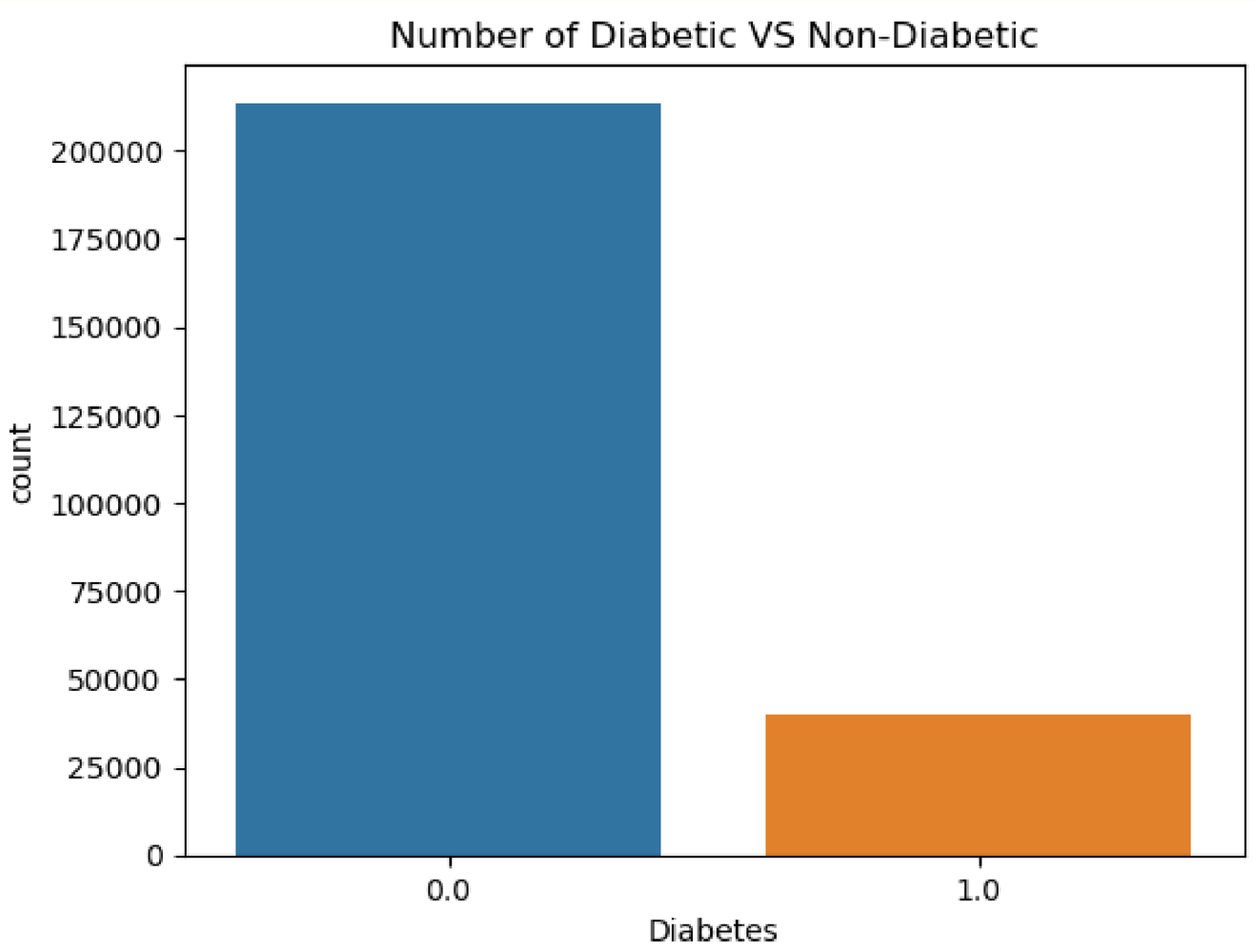
Age	Sex	Diabetes	Total
18 to 24	Female	Diabetes	61
		No Diabetes	2684
	Male	Diabetes	38
		No Diabetes	2917
25 to 29	Female	Diabetes	126
		No Diabetes	3865
	Male	Diabetes	68
		No Diabetes	3539
30 to 34	Female	Diabetes	243
		No Diabetes	5819
	Male	Diabetes	143
		No Diabetes	4918
35 to 39	Female	Diabetes	446
		No Diabetes	7279
	Male	Diabetes	322
		No Diabetes	5776
40 to 44	Female	Diabetes	672
		No Diabetes	8464
	Male	Diabetes	542
		No Diabetes	6479
45 to 49	Female	Diabetes	1087
		No Diabetes	9841
	Male	Diabetes	967
		No Diabetes	7924
50 to 54	Female	Diabetes	1876
		No Diabetes	12929
	Male	Diabetes	1630
		No Diabetes	9879
50 to 54	Female	Diabetes	1876
		No Diabetes	12929
	Male	Diabetes	1630
		No Diabetes	9879
55 to 59	Female	Diabetes	2582
		No Diabetes	14887
	Male	Diabetes	2231
		No Diabetes	11132
60 to 64	Female	Diabetes	3353
		No Diabetes	14918
	Male	Diabetes	3082
		No Diabetes	11891
65 to 69	Female	Diabetes	3623
		No Diabetes	14120
	Male	Diabetes	3632
		No Diabetes	10819
70 to 74	Female	Diabetes	2877
		No Diabetes	10282
	Male	Diabetes	2866
		No Diabetes	7508
75 to 79	Female	Diabetes	2075
		No Diabetes	7343
	Male	Diabetes	1773
		No Diabetes	4789
80 or older	Female	Diabetes	1994
		No Diabetes	8528
	Male	Diabetes	1668
		No Diabetes	5173



The features are not correlated to one another

```
SMOTE(random_state=28)
StandardScaler()

 StandardScaler()
```

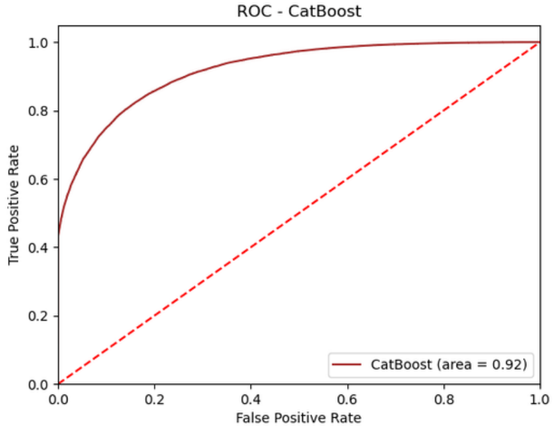


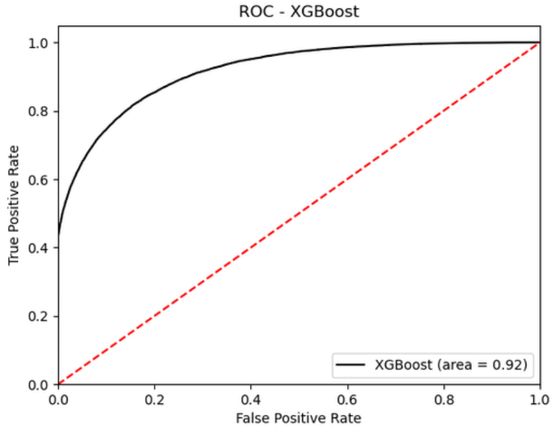
Before sampling class distribution: Counter({0.0: 213703, 1.0: 39977})
After sampling class distribution: Counter({0.0: 213703, 1.0: 213703})

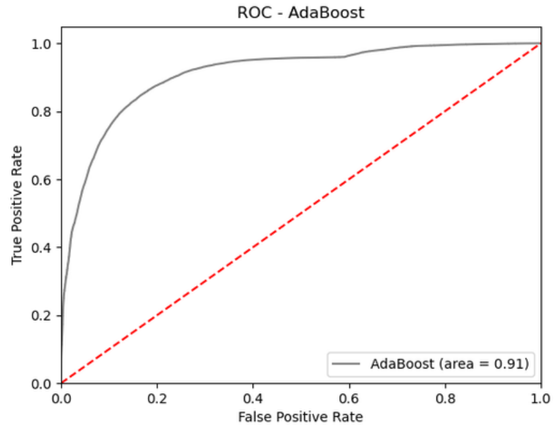
```
X_train , X_test , Y_train , Y_test = train_test_split(smote_X_train , smote_Y_train , test_size = 0.2,
random_state = 28)
```

5. TRAINING MODELS

	Model	ROC AUC Score	Accuracy Score	CM Score	Confusion Matrix	Time taken(s)
6	Catboost	0.919845	0.831543	0.831543	[[35936, 6805], [7595, 35146]]	0:00:16.904670
5	Xgboost	0.918872	0.830315	0.830315	[[35903, 6838], [7667, 35074]]	0:00:23.606730
7	AdaBoost	0.910717	0.839896	0.839896	[[35159, 7582], [6104, 36637]]	0:17:54.186838
3	K-Nearest Neighbour	0.846917	0.770794	0.770794	[[33494, 9247], [10346, 32395]]	0:00:00.357229
4	RandomForest	0.831260	0.754170	0.754170	[[30977, 11764], [9250, 33491]]	0:01:01.167985
2	Decision_Tree	0.814371	0.739945	0.739945	[[28970, 13771], [8459, 34282]]	0:00:00.248422
0	Logistic_Regression	0.813933	0.741337	0.741337	[[30648, 12093], [10018, 32723]]	0:00:00.203615
1	Guassian	0.782302	0.721731	0.721731	[[31134, 11607], [12180, 30561]]	0:00:00.073837
8	ExtraTree	0.767945	0.708137	0.708137	[[26218, 16523], [8426, 34315]]	0:00:00.290814
9	SVClinear	0.741080	0.741080	0.741080	[[30351, 12390], [9743, 32998]]	0:00:50.767865





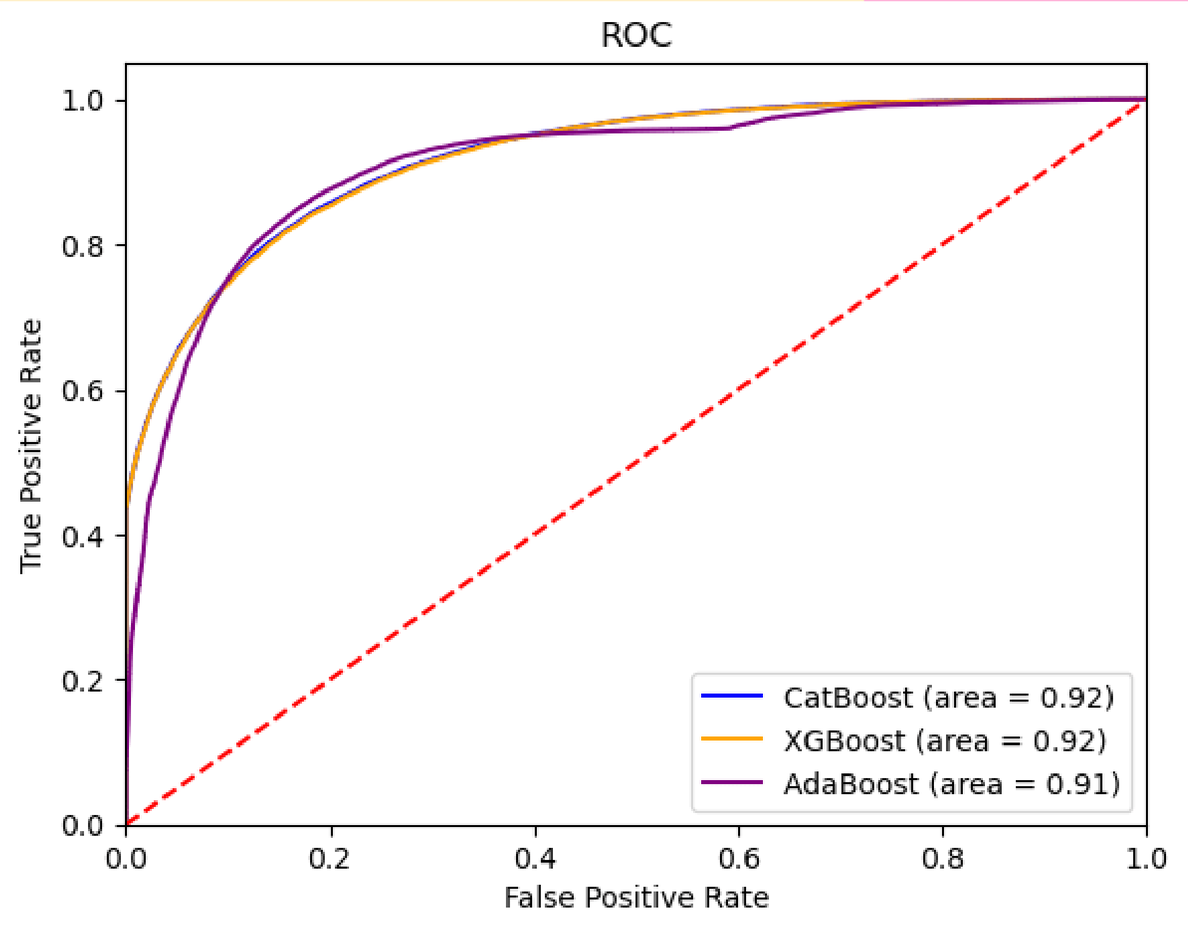


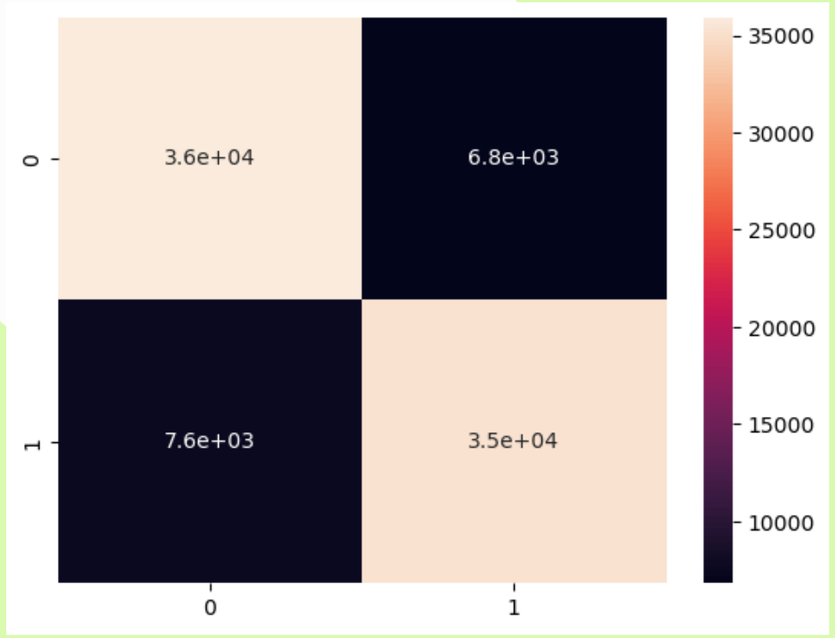
6. MODEL EVALUATION

ROC AUC score: 0.919845
Accuracy score: 0.831543
CM score: 0.831543

ROC AUC score: 0.918872
Accuracy score: 0.830315
CM score: 0.830315

ROC AUC score: 0.910717
Accuracy score: 0.839896
CM score: 0.839896

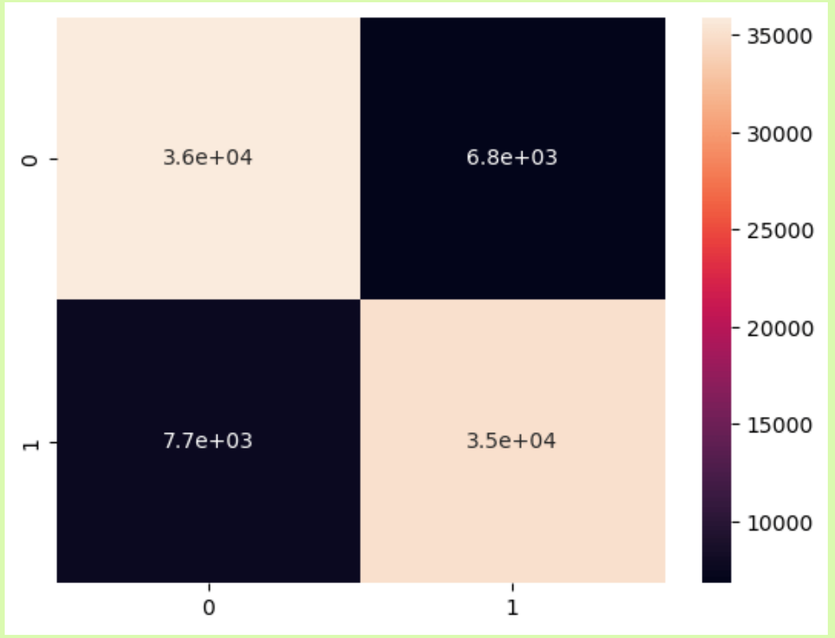




[35936 6805]
[7595 35146]

CatBoost

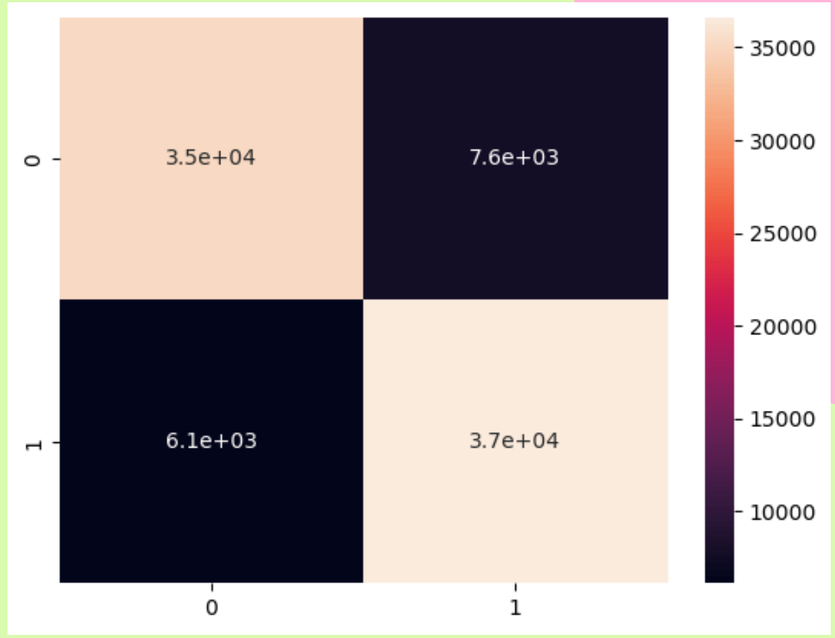
	Actual	Predicted
3468	1.0	0.767885
207261	0.0	0.628550
147729	0.0	0.052851
401174	1.0	0.782375
307769	1.0	0.999726



[35903 6838]
[7667 35074]

XGBoost

	Actual	Predicted
3468	1.0	0.787497
207261	0.0	0.651620
147729	0.0	0.051549
401174	1.0	0.765803
307769	1.0	0.999959

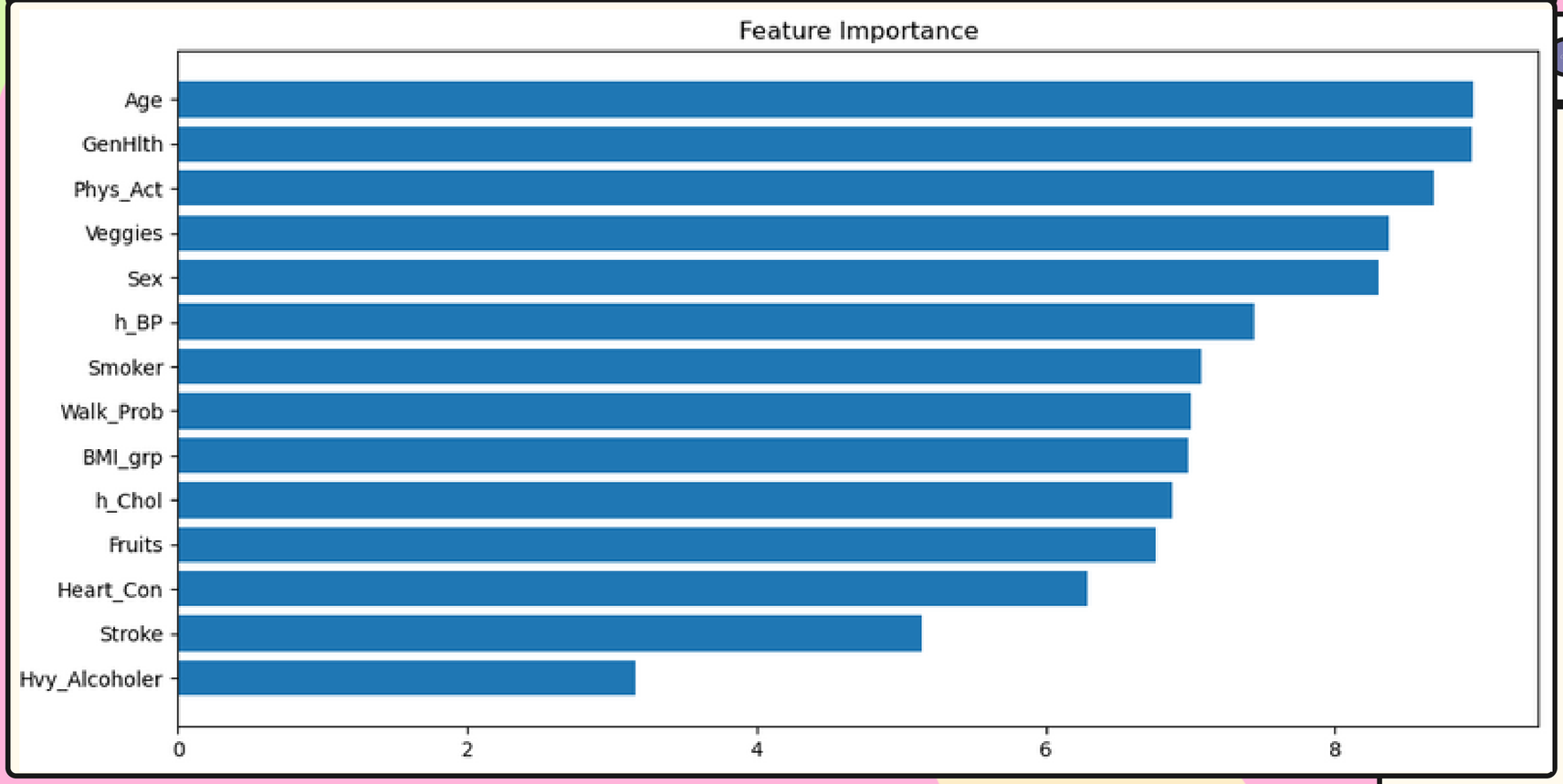


[35159 7582]
[6104 36637]

AdaBoost

	Actual	Predicted
3468	1.0	0.906227
207261	0.0	0.503509
147729	0.0	0.479146
401174	1.0	0.508842
307769	1.0	0.874657

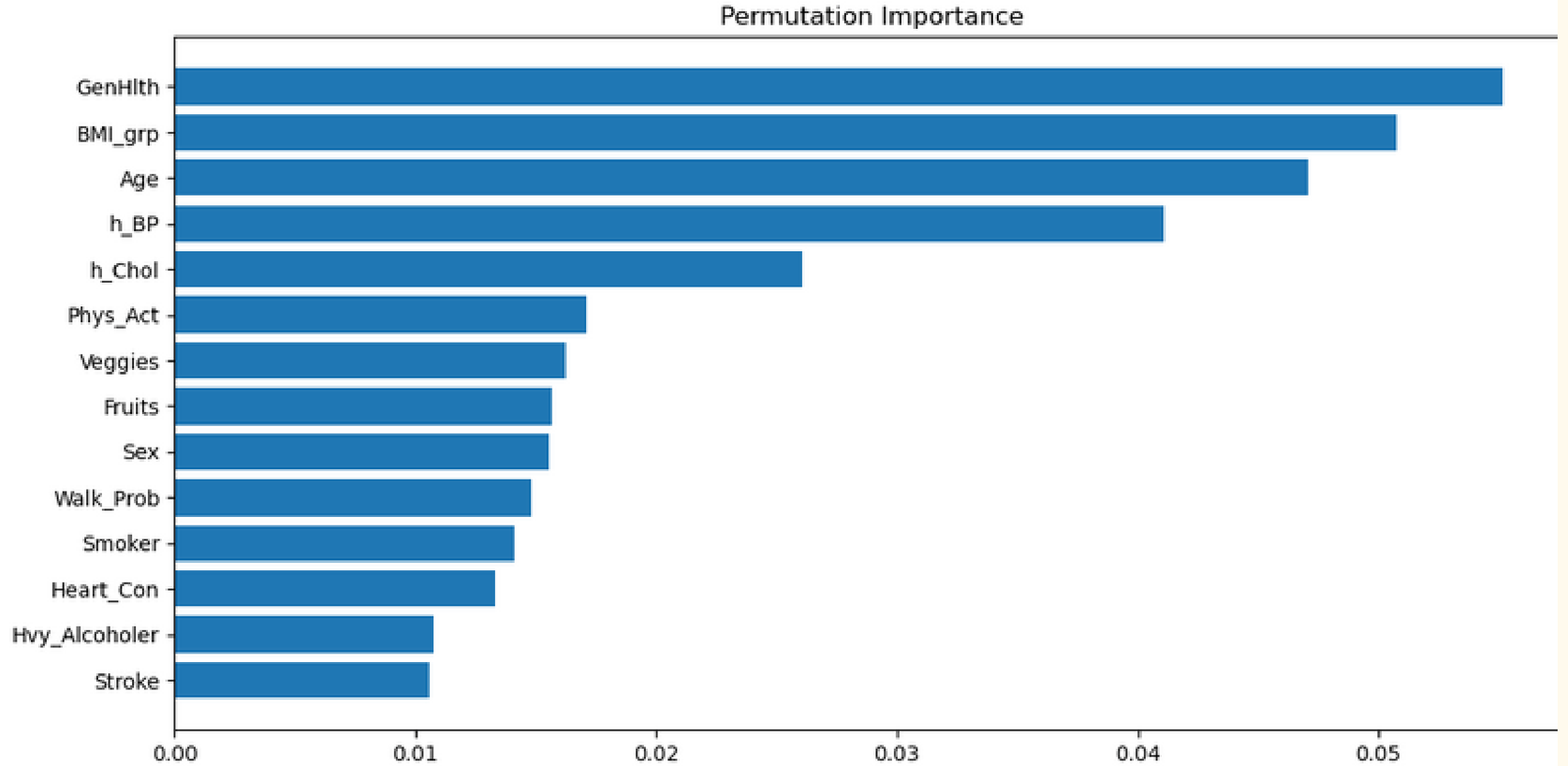
7.1 FEATURE IMPORTANCE



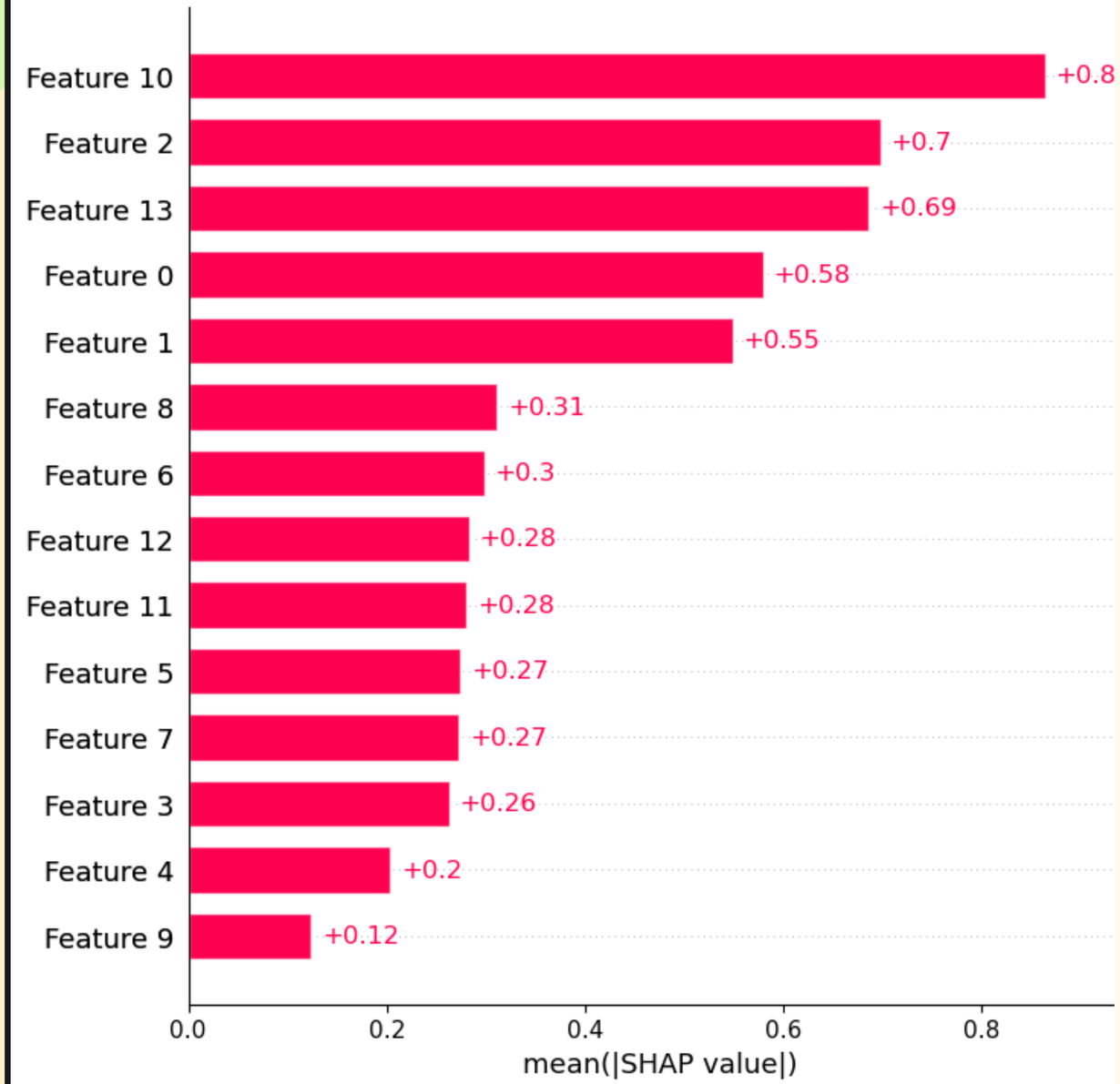
Feature: Hvy_Alcoholer, Score: 3.15649
Feature: Stroke, Score: 5.14738
Feature: Heart_Con, Score: 7.07972
Feature: Fruits, Score: 6.76520
Feature: h_Chol, Score: 6.29016
Feature: BMI_grp, Score: 7.00418
Feature: Walk_Prob, Score: 8.30285
Feature: Smoker, Score: 8.68247
Feature: h_BP, Score: 8.37071
Feature: Sex, Score: 6.87174
Feature: Veggies, Score: 7.43648
Feature: Phys_Act, Score: 8.95977
Feature: GenHlth, Score: 6.98144
Feature: Age, Score: 8.95141

Feature: Stroke, Score: 0.01075
Feature: Hvy_Alcoholer, Score: 0.01055
Feature: Heart_Con, Score: 0.01409
Feature: Smoker, Score: 0.01567
Feature: Walk_Prob, Score: 0.01330
Feature: Sex, Score: 0.01482
Feature: Fruits, Score: 0.01557
Feature: Veggies, Score: 0.01709
Feature: Phys_Act, Score: 0.01624
Feature: h_Chol, Score: 0.02610
Feature: h_BP, Score: 0.04112
Feature: Age, Score: 0.04710
Feature: BMI_grp, Score: 0.05077
Feature: GenHlth, Score: 0.05518

7. PERMUTATION IMPORTANCE

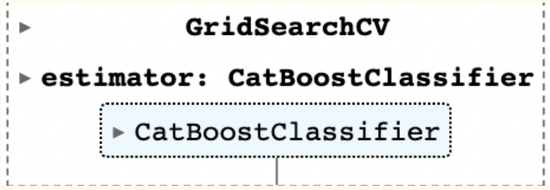


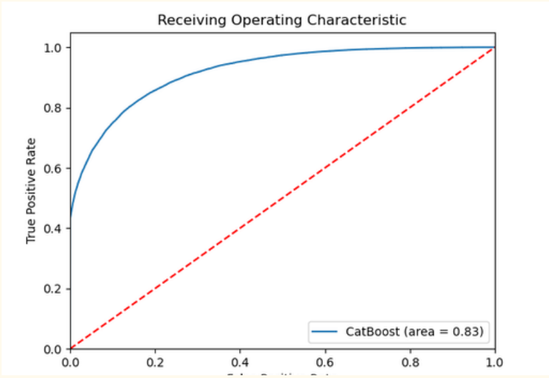
7.3 SHAP VALUES



Feature: 0, Score: 0.57985
Feature: 1, Score: 0.54851
Feature: 2, Score: 0.69756
Feature: 3, Score: 0.26319
Feature: 4, Score: 0.20372
Feature: 5, Score: 0.27460
Feature: 6, Score: 0.29834
Feature: 7, Score: 0.27278
Feature: 8, Score: 0.31158
Feature: 9, Score: 0.12419
Feature: 10, Score: 0.86384
Feature: 11, Score: 0.28094
Feature: 12, Score: 0.28253
Feature: 13, Score: 0.68626

8. MODEL PERFORMANCE





Grid:
'max_depth': [5,7,9],
'n_estimators':[100, 200, 300]

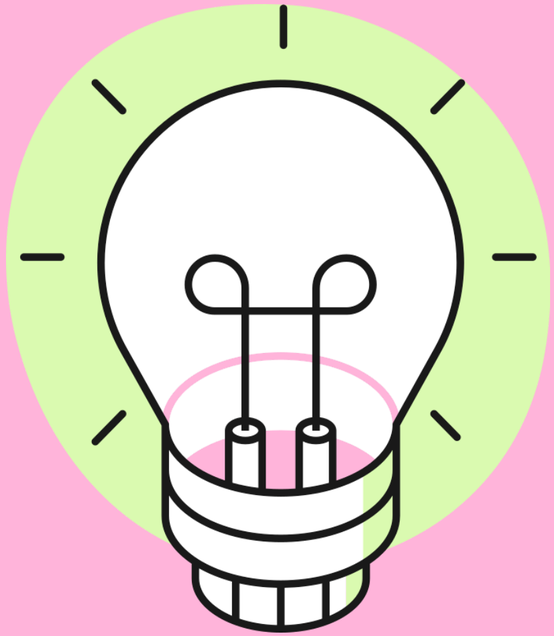
Method:
GridSearchCV (
estimator = cbc,
param_grid = grid,
scoring ='accuracy',
cv = 5)

Best score:
0.8334629865355009

Best Parameters:
{'max_depth': 9,
'n_estimators': 300}

	precision	recall	f1-score	support
0.0	0.76	0.71	0.73	42741
1.0	0.73	0.77	0.75	42741
...				
accuracy			0.74	85482
macro avg	0.74	0.74	0.74	85482
weighted avg	0.74	0.74	0.74	85482

	precision	recall	f1-score	support
0.0	0.83	0.84	0.84	42741
1.0	0.84	0.83	0.84	42741
...				
accuracy			0.84	85482
macro avg	0.84	0.84	0.84	85482
weighted avg	0.84	0.84	0.84	85482



9. CONCLUSION

1. I would've studied more on what happens to the data after SMOTE was introduce to it.

2. Do more research on understanding the data when it's plotted.

3. Understand the formulas in each models so to understand how it fitted in the data.

4. Even when codes are everywhere online, using it blindly won't bring you anywhere.

5. More time to try every models as some models takes more than an hour.

DIABETES HEALTH INDICATOR DATASET

THANK YOU

Github link:
(ENTER GITHUB LINK HERE)