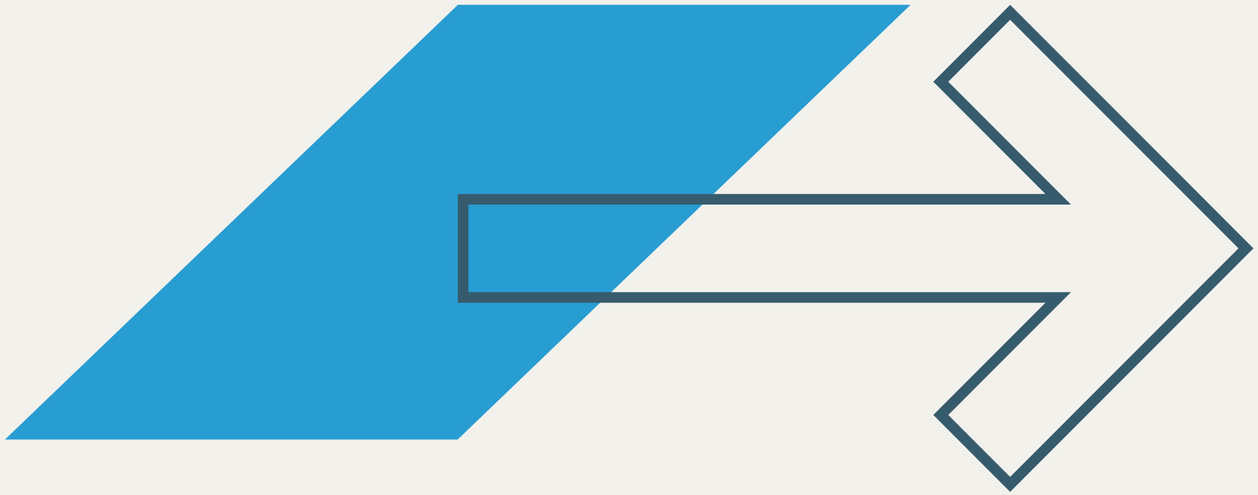
SC1015 Mini Project SC7 Group 5

Can indicators of heart diseases also indicate diabetes?

Prepared by
Lee Yu Xuan
Ler Lian Ping
Ng Teng Hian



Why choose diabetes?

Accordingly to World Health Organization

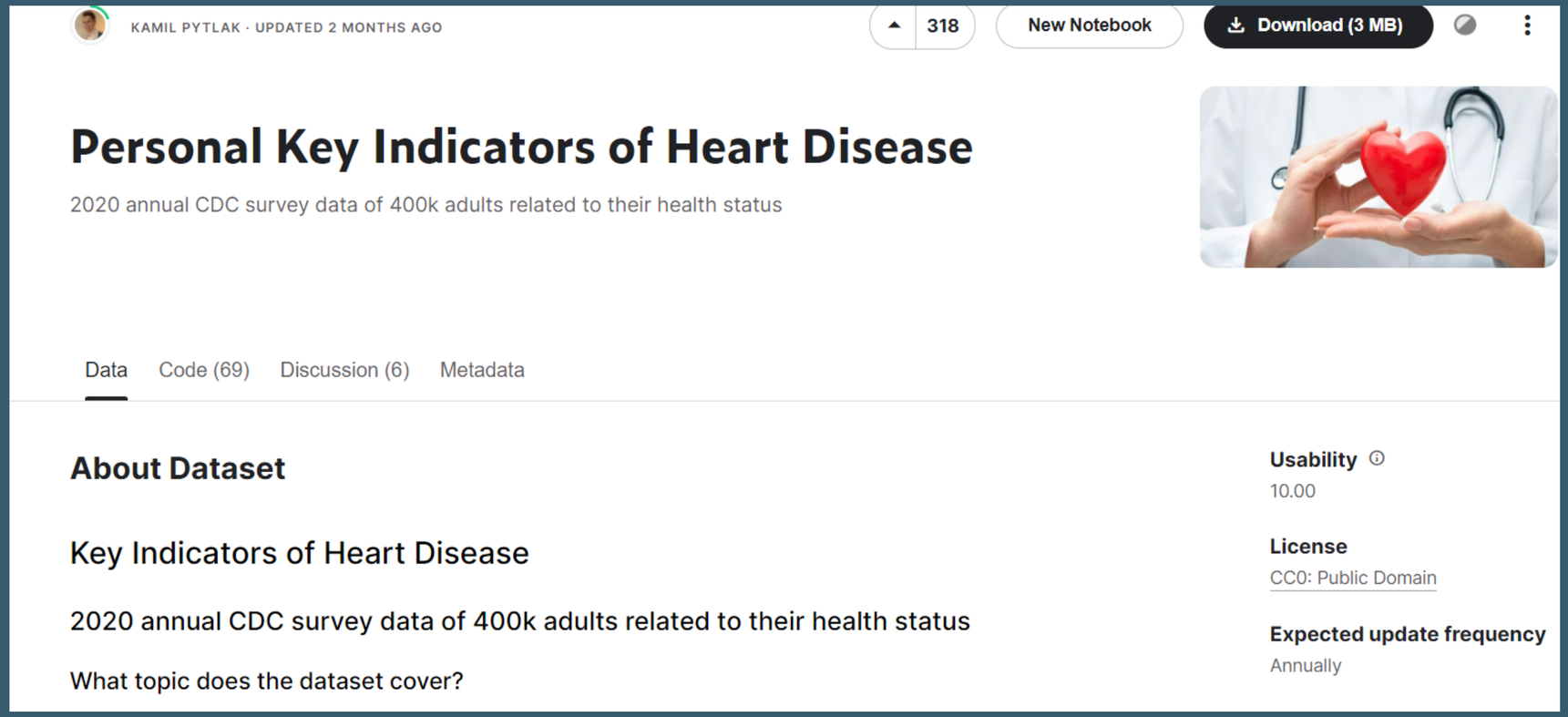
9th LEADING CAUSE OF DEATH

Estimated 1.5 MILLION Deaths caused directly by diabetes

48%

Occurred before the age of 70

Personal Key Indicators of Heart Disease



https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Flow of Sequence

01

Data Cleaning/ Filtering

Removing irrelevant samples and outliers, Under-sampling

02

EDA

03

Machine Learning

Classification Tree and Random Forest

01 - Data Cleaning/Filtering

Removing samples that were
1) No, borderline diabetes
2) Yes (during pregnancy

```
: #To check how many diabetics patients were stated
  print(dataset["Diabetic"].value_counts())
```

```
No                             269653
Yes                             40802
No, borderline diabetes          6781
Yes (during pregnancy)           2559
Name: Diabetic, dtype: int64
```

01 - Data Cleaning/Filtering

Quantified categorical columns into 0 to represent "No" and 1 to represent "Yes"

1) Diabetes
2) Kidney Disease
3) Heart Disease
4) Difficulty Walking
5) Smoking
6) Alcohol

Diabetes	KidneyDisease_Quantified	HeartDisease_Quantified	DiffWalking_Quantified	Smoking_Quantified	Alcohol	agecategory
1	0	0	0	1	0	7
0	0	0	0	0	0	12
1	0	0	0	1	0	9
0	0	0	0	0	0	11
0	0	0	1	0	0	4
...	...	...	...	...	...	...
1	0	1	1	1	0	8
0	0	0	0	1	0	3
0	0	0	0	0	0	5
0	0	0	0	0	0	1
0	0	0	0	0	0	12

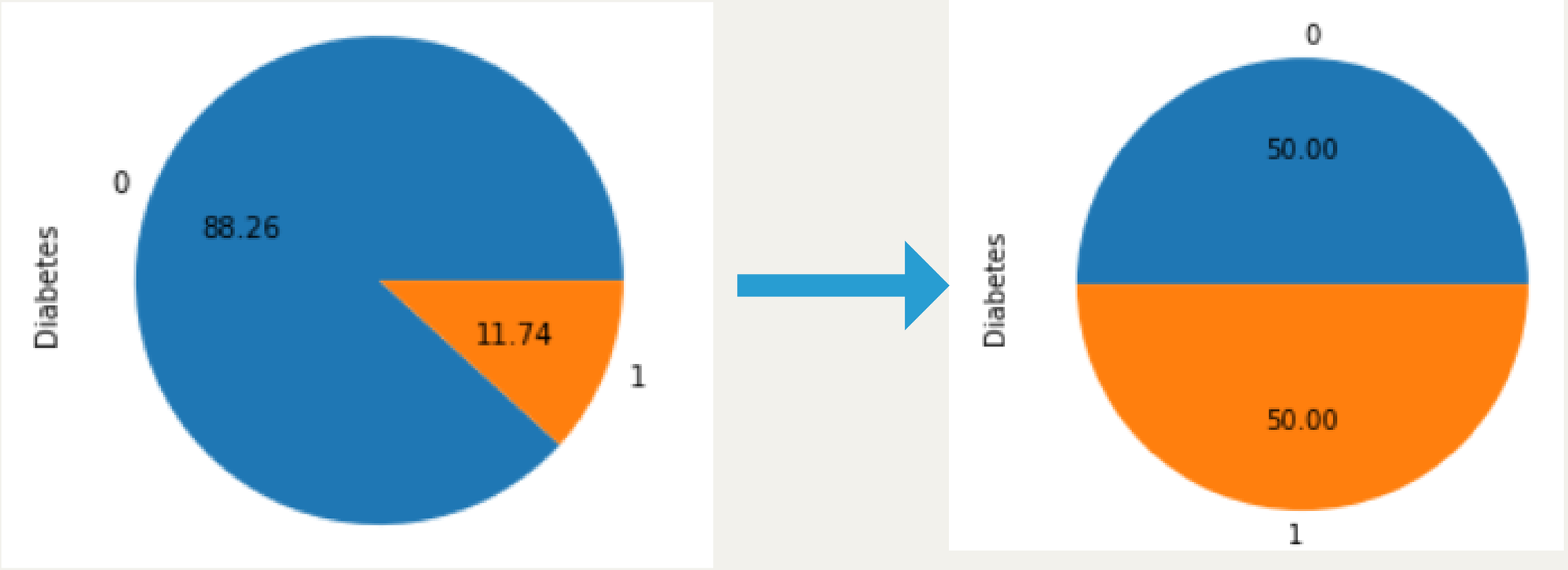
01 - Data Cleaning/Filtering

Quantified age groups with numbers from 0-12, with largest number representing oldest age group.

18-24	0
25-29	1
30-34	2
35-39	3
40-44	4
45-49	5
50-54	6
55-59	7
60-64	8
65-69	9
70-74	10
75-79	11
80 or older	12

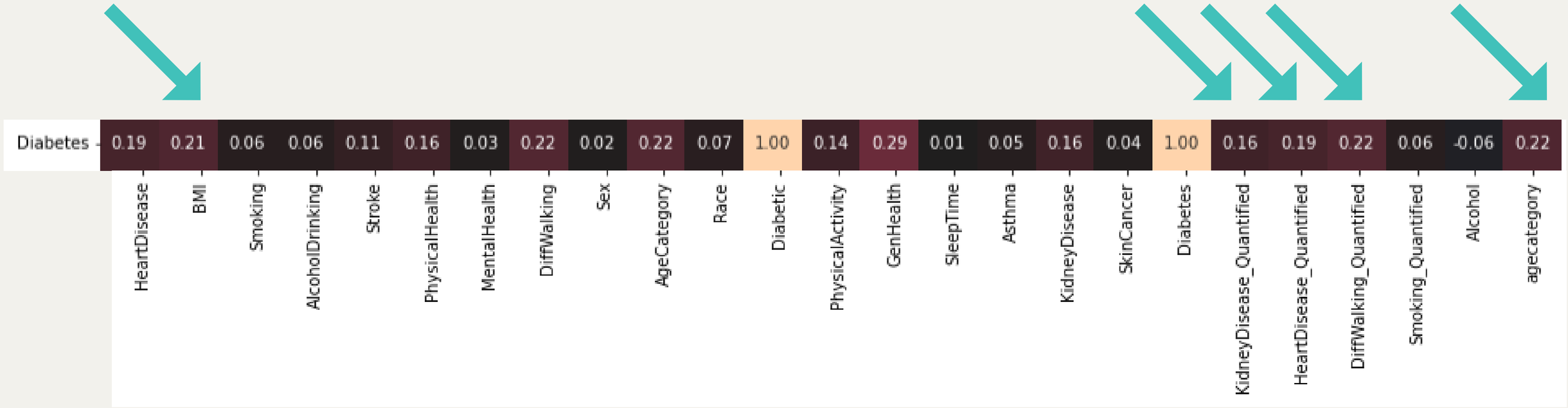
01 - Data Cleaning/Filtering (Under-Sampling)

Ensure that the number of samples who have Diabetes is equal to the number of samples who do not have Diabetes



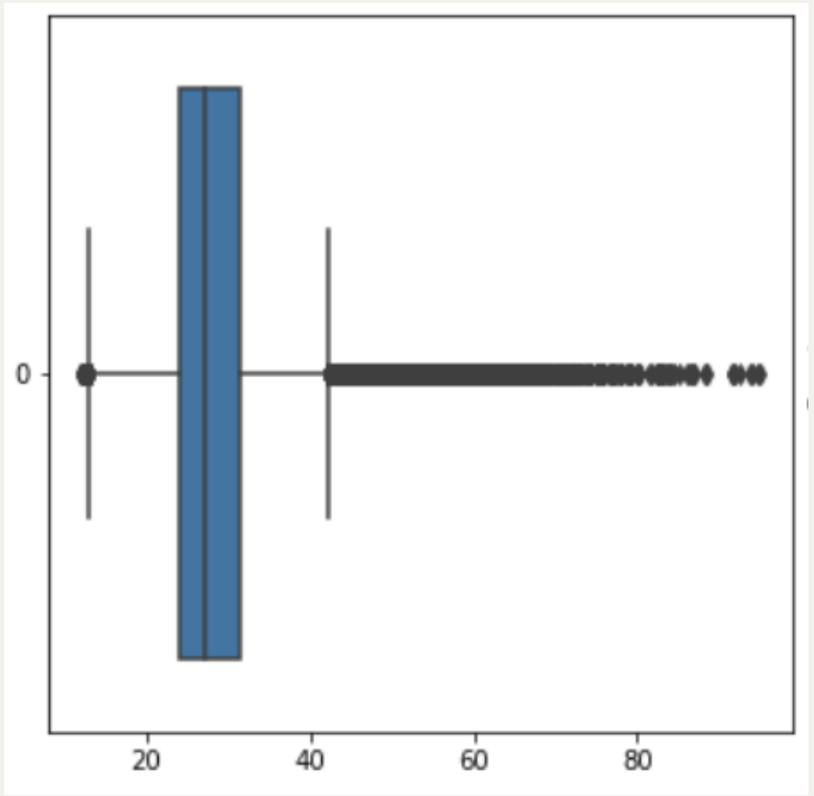
02 - EDA

Compute the Categorical Correlation with Dython library and display Correlation Matrix accordingly.



02 - EDA

Removing outliers via BMI

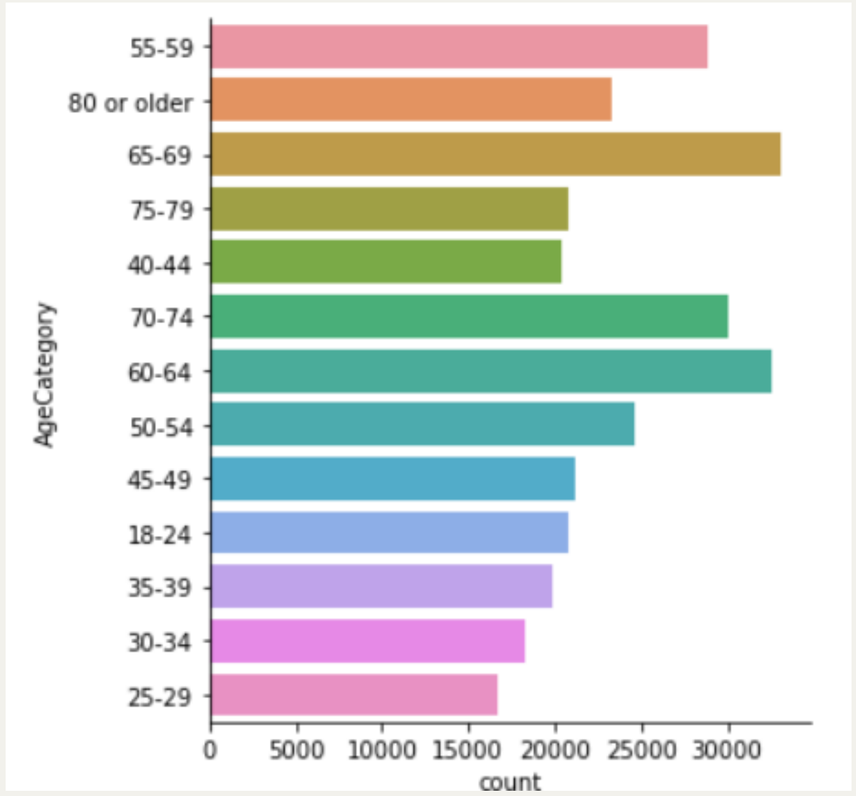


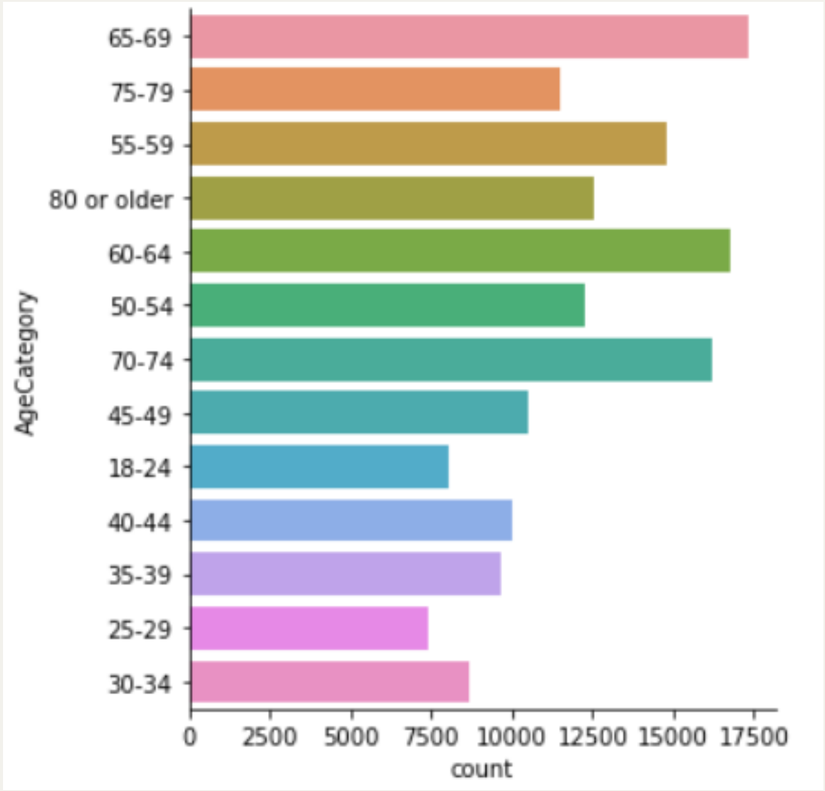
	BMI
count	310455.000000
mean	28.274180
std	6.328361
min	12.020000
25%	23.990000
50%	27.280000
75%	31.320000
max	94.850000

	BMI
count	155633.000000
mean	27.369929
std	2.061895
min	23.990000
25%	25.700000
50%	27.290000
75%	29.050000
max	31.320000

02 - EDA

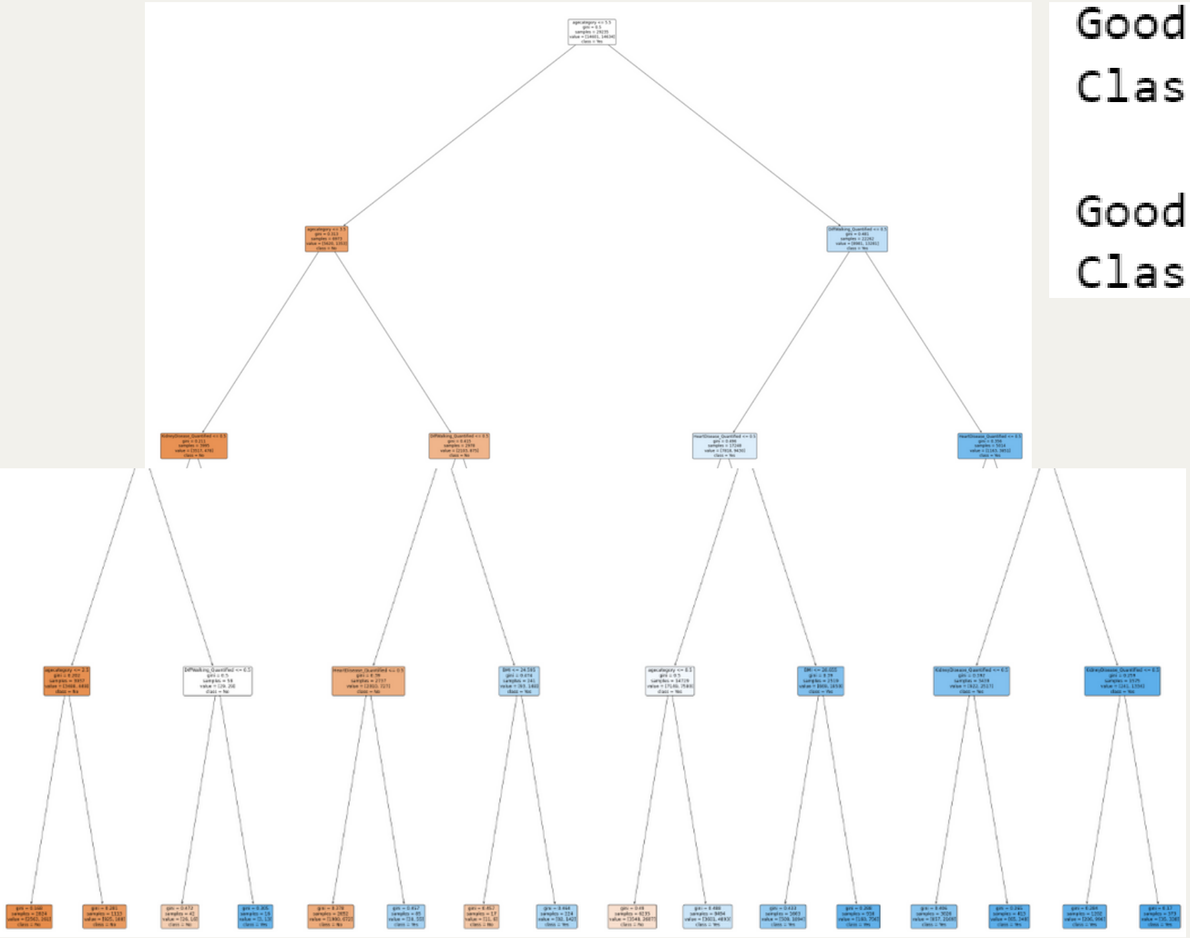
Show the Age Catergory after removing outliers





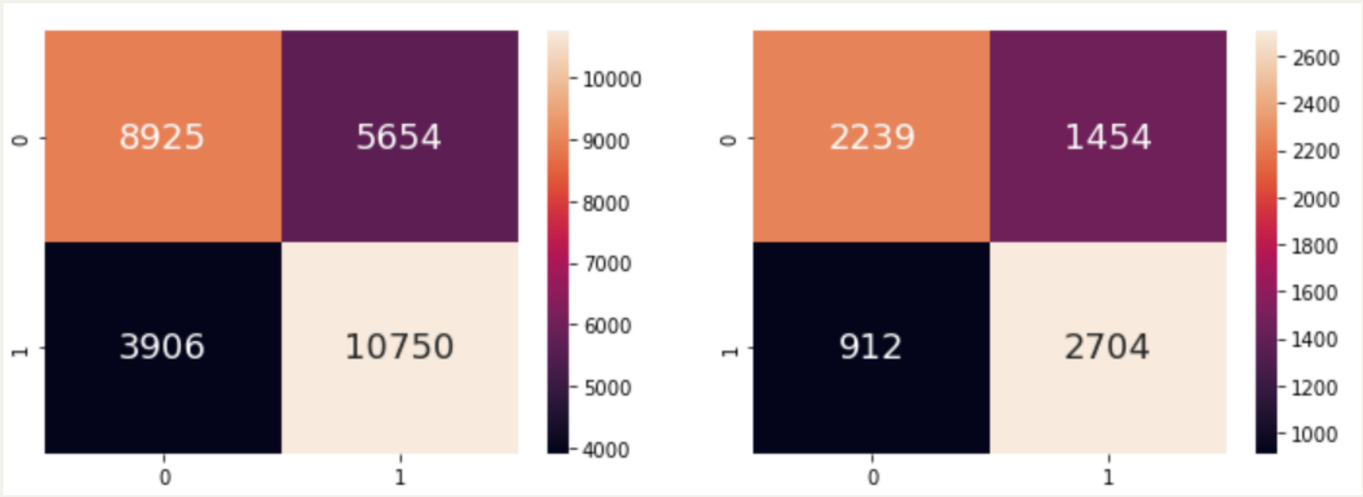
03 - Machine Learning

Classification Tree for Classification Accuracy and Confusion Matrix



```
Goodness of Fit of Model          Train Dataset
Classification Accuracy           : 0.6729946981357962

Goodness of Fit of Model          Test Dataset
Classification Accuracy           : 0.6762895060883842
```



03 - Machine Learning

Use Random Forest Classification Tree for better accuracy

```
print("Accuracy = ", metrics.accuracy_score(B_test, prediction_test))
```

```
Accuracy =  0.8473208677586124
```

```
feature_list = list(A_train.columns)
feature_imp = pd.Series(model.feature_i
print(feature_imp)
```

```
BMI                          0.754509
agecategory                  0.108113
DiffWalking_Quantified       0.059584
HeartDisease_Quantified      0.039513
KidneyDisease_Quantified     0.024846
Alcohol                      0.006893
Smoking_Quantified           0.006542
dtype: float64
```

```
feature_list2 = list(A_test.columns)
feature_imp2 = pd.Series(model.feature_
print(feature_imp2)
```

```
BMI                          0.754509
agecategory                  0.108113
DiffWalking_Quantified       0.059584
HeartDisease_Quantified      0.039513
KidneyDisease_Quantified     0.024846
Alcohol                      0.006893
Smoking_Quantified           0.006542
dtype: float64
```

Conclusion + Recommendation

Indicators of Heart Diseases can be used to help detect Diabetes but it should not be fully relied on

You can use these indicators for specifically testing for diabetes but it should be combined with other factors as well

Reference

Diabetes. (2021, November 10). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

PYTLAK, K. A. M. I. L. (2022, February 18). Personal Key Indicators of Heart Disease. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease