CAPSTONE PROJECT

YUMEMI KINSELLA

HEART DISEASE PREDICTION

THE SUBJECT AREA / THE PROBLEM STATEMENT & IMPACT OF THE SOLUTION

- Leading cause of death in the US
- 1 in every 5 deaths

- The official poverty rate in the US: 11.4% in 2020
- 22% of Americans have avoided of medical care

- Able to have some idea of the probability of heart disease without incurring any expense
- Might be able to get necessary treatment
- Able to improve their lifestyle or habits

OVERVIEW OF THE DATASET & PREPROCESSING PROCEDURES

Original Dataset
Rows: 319,072
Columns: 14

Target "Yes" value:
27,269 rows (8.55%)
Target "No" value:
291,804 rows (91.45%)

Imbalanced data

Train set
Rows: 223,351 rows

Test set
Rows: 95722 rows

Train set target value balance:
"Yes" : 19,091
"No" : 204,260

Sampling Methods

- **Under sampling**

Reduce "No" values to 19,091

- **Over sampling**

Increase "Yes" values to 204,260 by duplicating instances randomly

- **SMOTE**

Increase "Yes" values to 204,260 by generates synthetic samples

ORIGINAL & SAMPLED DATASET

Original Dataset
Train set
Rows: 223,351
"Yes" 19,091 rows
"No" 204,260 rows

Test set
Rows: 95,722
"Yes" 8,178 rows
"No" 87,544 rows

Under sampled Dataset
Train set
Rows: 38,182
"Yes" 19,091 rows
"No" 19,091 rows

Test set
Rows: 95,722
"Yes" 8,178 rows
"No" 87,544 rows

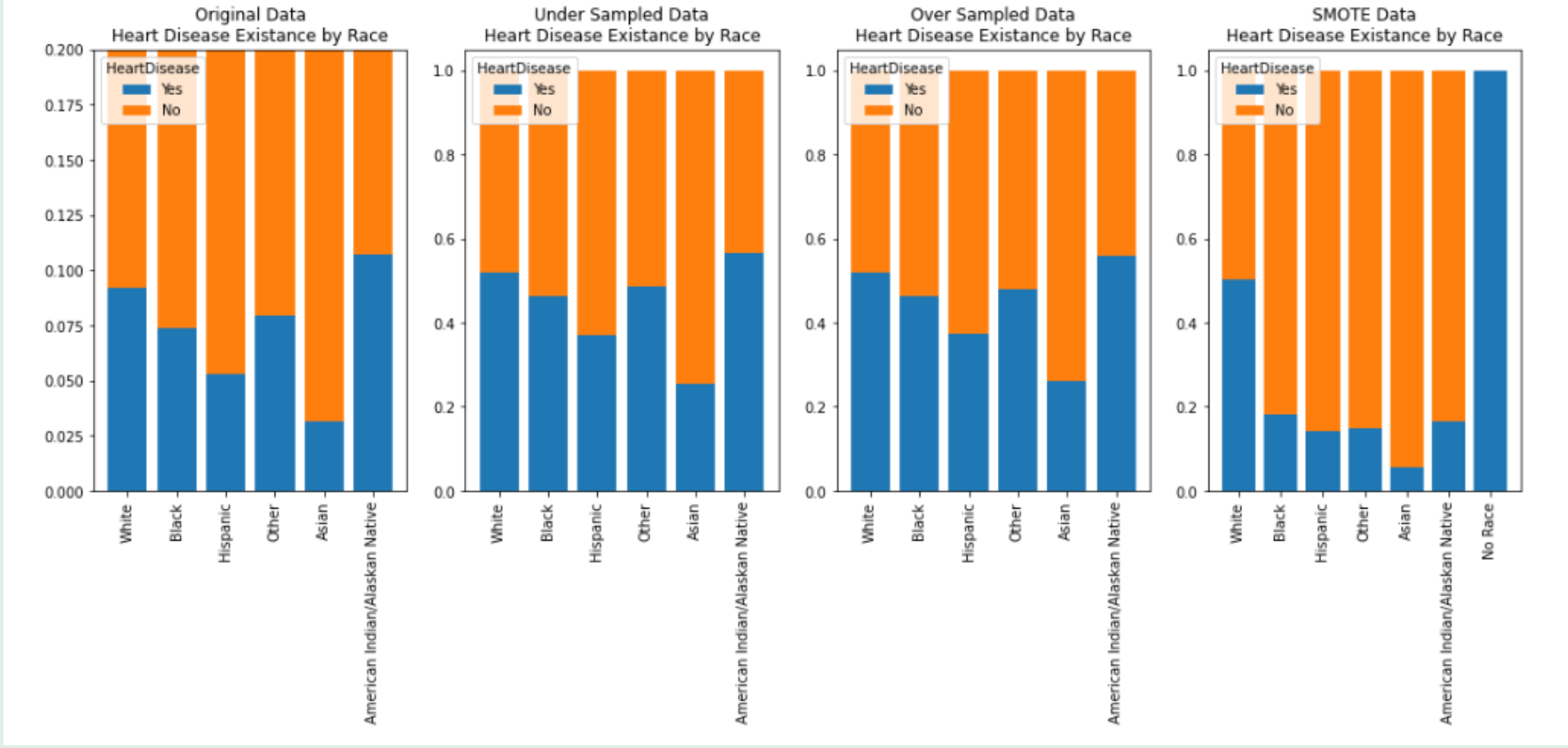
Over sampled Dataset
Trainset
Rows: 408,520
"Yes" 204,260 rows
"No" 204,260 rows

Test set
Rows: 95,722
"Yes" 8,178 rows
"No" 87,544 rows

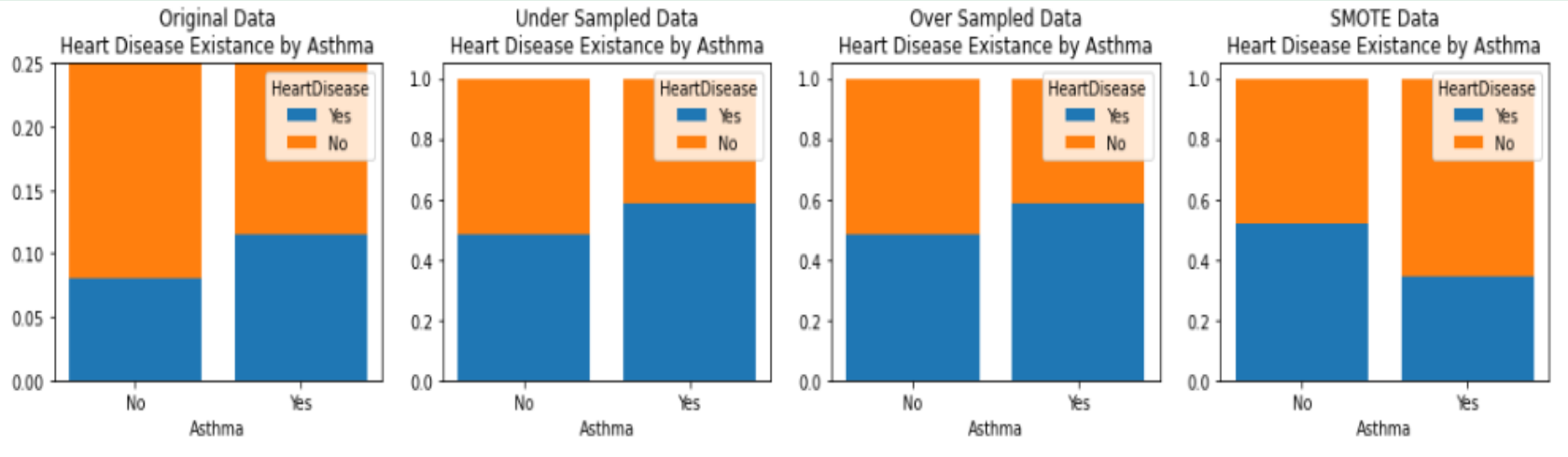
SMOTE Dataset
Trainset
Rows: 408,520
"Yes" 204,260 rows
"No" 204,260 rows

Test set
Rows: 95,722
"Yes" 8,178 rows
"No" 87,544 rows

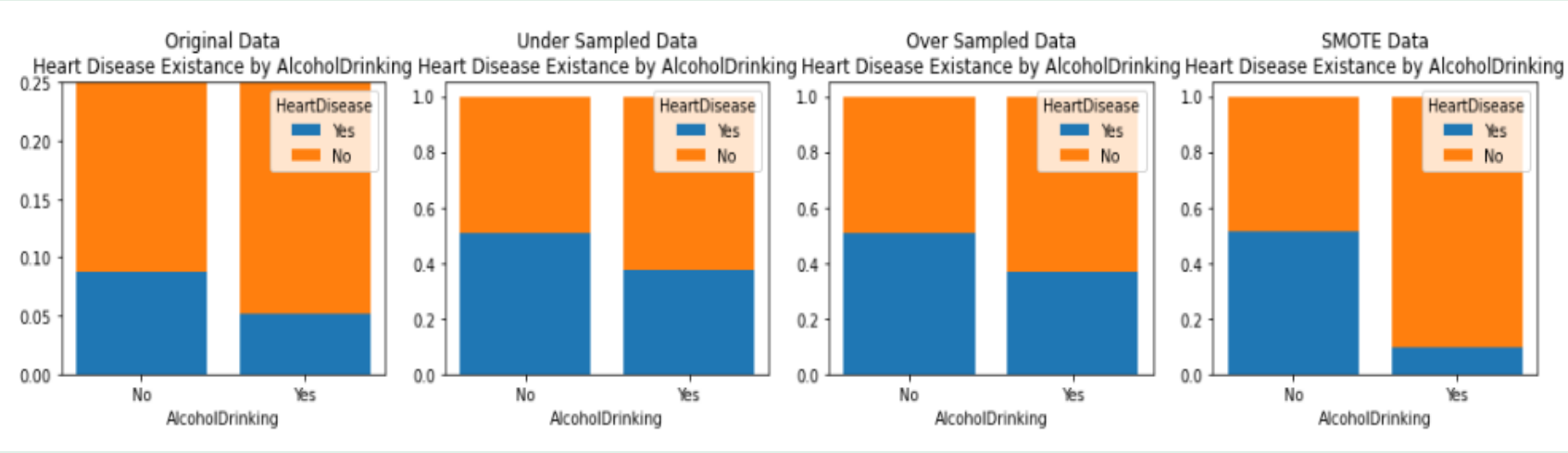
IMPORTANT FINDINGS FROM EDA



IMPORTANT FINDINGS FROM EDA



SMOTE data has opposite distribution of heart disease existence



In generally alcohol drinking is not good for health...

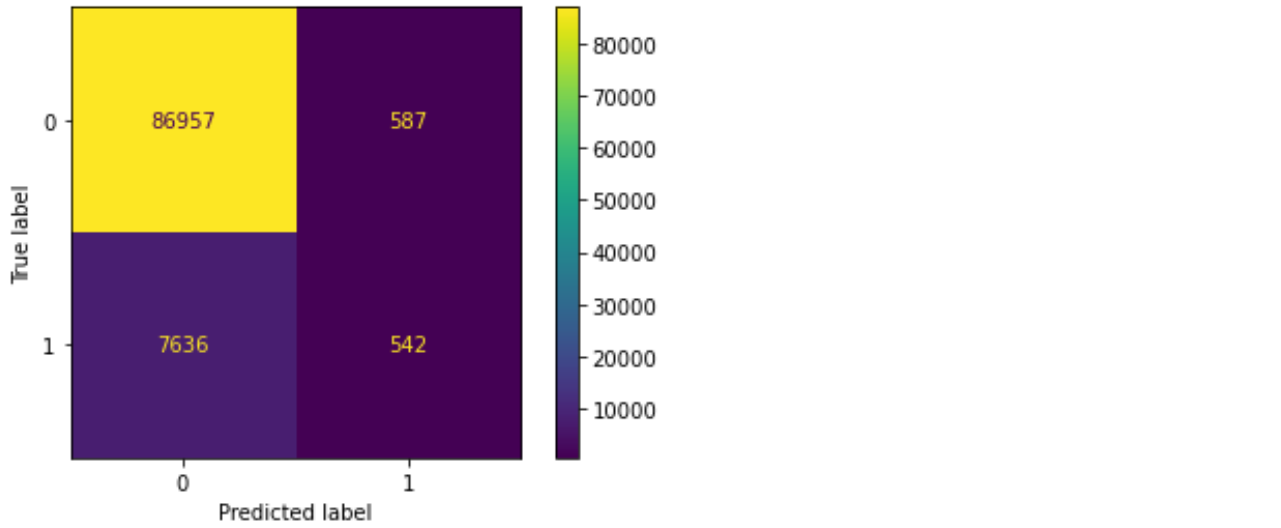
BASELINE MODELS & EVALUATION METRICS

ORIGINAL DATA

```
Original data
train score: 91.44306495157846
test score: 91.40949833893984
```

	precision	recall	f1-score	support
0	0.92	0.99	0.95	87544
1	0.48	0.07	0.12	8178
accuracy			0.91	95722
macro avg	0.70	0.53	0.54	95722
weighted avg	0.88	0.91	0.88	95722

```
Out[164]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1f1051df340>
```



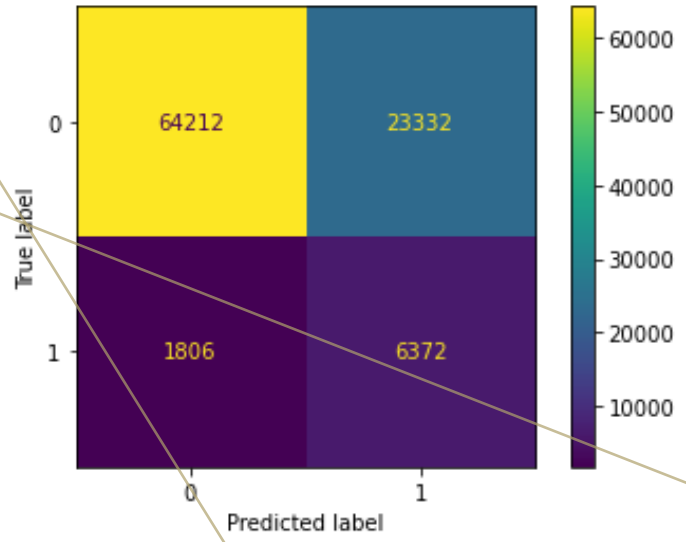
BASELINE MODELS & EVALUATION METRICS

UNDER SAMPLED DATA

```
Under sampled data
train score: 75.64559216384684
test score: 73.73853450617412
```

	precision	recall	f1-score	support
0	0.97	0.73	0.84	87544
1	0.21	0.78	0.34	8178
accuracy			0.74	95722
macro avg	0.59	0.76	0.59	95722
weighted avg	0.91	0.74	0.79	95722

Out[167]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at

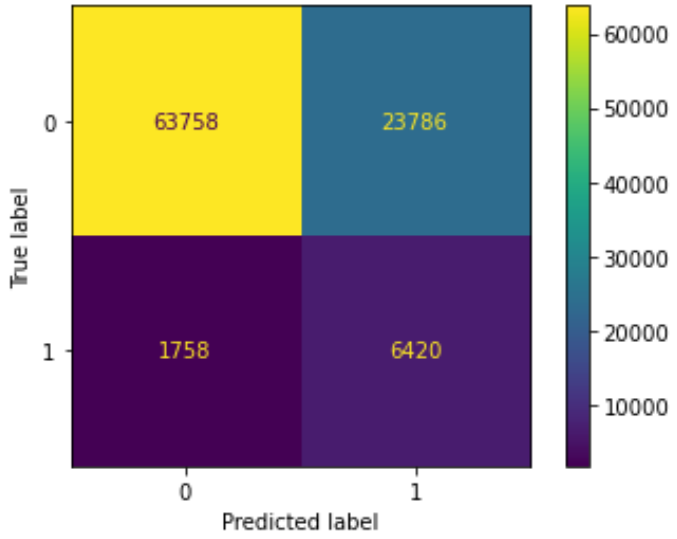


OVER SAMPLED DATA

```
Over sampled data
train score: 75.63448545970822
test score: 73.31438958651094
```

	precision	recall	f1-score	support
0	0.97	0.73	0.83	87544
1	0.21	0.79	0.33	8178
accuracy			0.73	95722
macro avg	0.59	0.76	0.58	95722
weighted avg	0.91	0.73	0.79	95722

Out[169]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDispla



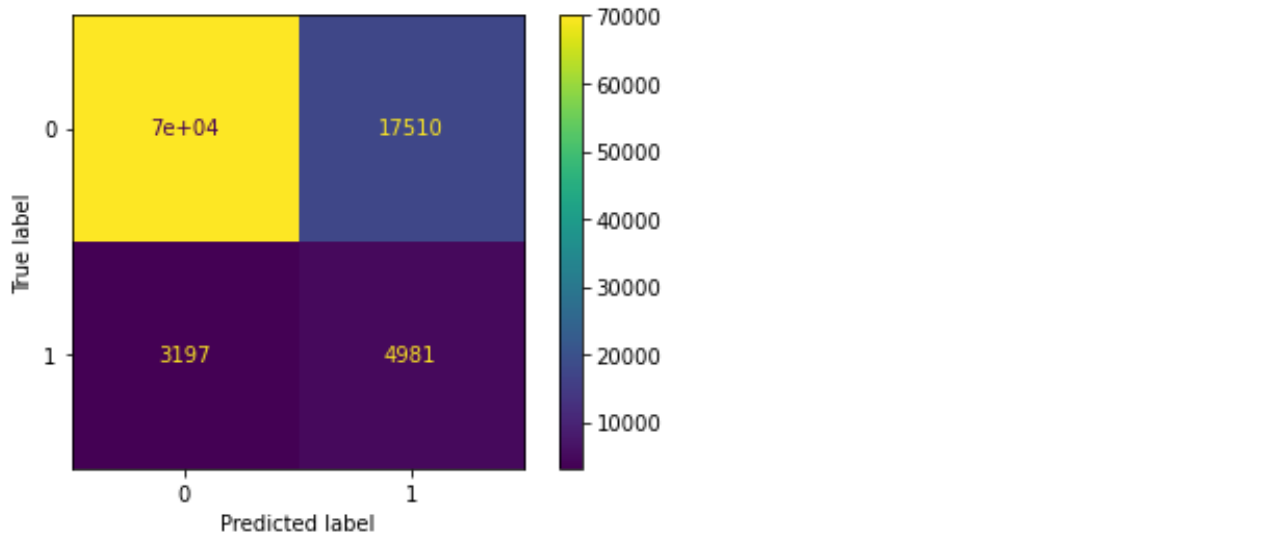
BASELINE MODELS & EVALUATION METRICS

SMOTE DATA

```
SMOTE data
train score: 80.12165867032213
test score: 78.36756440525689
```

	precision	recall	f1-score	support
0	0.96	0.80	0.87	87544
1	0.22	0.61	0.32	8178
accuracy			0.78	95722
macro avg	0.59	0.70	0.60	95722
weighted avg	0.89	0.78	0.82	95722

```
Out[171]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1f106e6f850>
```



NEXT STEPS

- Optimize hyperparameter
- Create pipelines
- Create models
- Evaluation
- Create an application

THANK YOU FOR YOUR ATTENTION