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

Under sampled data

Train score: 0.756534492692892 Test score: 0.7352019389482042 Over sampled data Train score: 0.756489278370704 Test score: 0.7338647332901528

	precision	recall	f1-score	support		precision	recall	f1-score	support
0 1	0.97 0.21	0.73 0.78	1	87544 8178	0 1	0.97 0.21	0.73 0.79	0.83 0.34	87544 8178
accuracy macro avg weighted avg	0.59 0.91	0.76 0.74	0.74 0.59 0.79	95722 95722 95722	accuracy macro avg weighted avg	0.59 0.91	0.76 0.73	0.73 0.58 0.79	95722 95722 95722

SMOTE data

Train score: 0.8012165867032214 Test score: 0.7836756440525688

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

IMPORTANT FINDINGS FROM EDA

Under sampled data

Train score: 0.756534492692892 Test score: 0.7352019389482042 Over sampled data

Train score: 0.756489278370704 Test score: 0.7338647332901528

	precision	recall	f1-score	support		precision	recall	f1-score	support
0 1	0.97 0.21	0.73 0.78		87544 8178	0 1	0.97 0.21	0.73 0.79	0.83 0.34	87544 8178
accuracy macro avg weighted avg	0.59 0.91	0.76 0.74	0.74 0.59 0.79	95722 95722 95722	accuracy macro avg weighted avg	0.59 0.91	0.76 0.73	0.73 0.58 0.79	95722 95722 95722

Original data

Train score: 0.9144261722580154 Test score: 0.914126324147009

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



Overfitting

MODEL COMPARISON

Logistic Regression: Over sampled data

Train score: 0.756489278370704 Test score: 0.7338647332901528 F1 score: 0.335879454626033

	precision	recall	f1-score	support
0 1	0.97 0.21	0.73 0.79	0.83 0.34	87544 8178
accuracy macro avg weighted avg	0.59 0.91	0.76 0.73	0.73 0.58 0.79	95722 95722 95722

Logistic Regression: SMOTE data

Train score: 0.8012165867032214 Test score: 0.7836756440525688 F1 score: 0.3248231112850109

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

Naive Bayes

Train score: 0.7496719866836385 Test score: 0.7213075364075134 F1 score: 0.3254696705352853

	precision	recall	f1-score	support
0 1	0.97 0.21	0.72	0.82 0.33	87544 8178
accuracy macro avg weighted avg	0.59 0.91	0.75 0.72	0.72 0.57 0.78	95722 95722 95722

Dec:	LSION	iree

Train score: 0.7661509840399491 Test score: 0.7124903365997367 F1 score: 0.32142416845427424

	precision	recall	f1-score	support
(0.97	0.70 0.80	0.82 0.32	87544 8178
accuracy macro avę weighted avę	/ g 0.59	0.75 0.71	0.71 0.57 0.78	95722 95722 95722

Random Forest

Train score: 0.7811010423759887 Test score: 0.7131171517519483 F1 score: 0.31981769994798503

	precision	recall	f1-score	support
0	0.97	0.71	0.82	87544
1	0.20	0.79	0.32	8178
				- 1
accuracy			0.71	95722
macro avg	0.59	0.75	0.57	95722
weighted avg	0.91	0.71	0.78	95722

LinearSVC

Train score: 0.756313032409674 Test score: 0.7281711623242306 F1 score: 0.3324781939456131

	precision	recall	f1-score	support
0 1	0.97 0.21	0.72 0.79	0.83 0.33	87544 8178
accuracy macro avg weighted avg	0.59 0.91	0.76 0.73	0.73 0.58 0.79	95722 95722 95722

MODEL COMPARISON

Logistic Regr	ression: Ove	r sampled	data		Decision Tree					Random Forest	:			
Original Test					Original Test					Original Test				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.97	0.73	0.83	87544	0	0.97	0.70	0.82	87544	0	0.97	0.71	0.82	87544
1	0.21	0.79	0.34	8178	1	0.20	0.80	0.32	8178	1	0.20	0.79	0.32	8178
accuracy			0.73	95722	accuracy			0.71	95722	accuracy			0.71	95722
macro avg	0.59	0.76	0.58	95722	macro avg	0.59	0.75	0.57	95722	macro avg	0.59	0.75	0.57	95722
weighted avg	0.91	0.73	0.79	95722	weighted avg	0.91	0.71	0.78	95722	weighted avg	0.91	0.71	0.78	95722
Test set 1	precision	recall	f1-score	support	Test set 1	precision	recall	f1-score	support	Test set 1	precision	recall	f1-score	support
	precision	recarr	11-30016	зиррог с		precision	recuir	11 30010	Suppor C		precision	recuir	11 30010	Suppor c
0	0.97	0.73	0.83	87548	0	0.98	0.71	0.82	87548	0	0.97	0.71	0.82	87548
1	0.21	0.78	0.33	8174	1	0.21	0.81	0.33	8174	1	0.20	0.80	0.33	8174
accuracy			0.73	95722	accuracy			0.72	95722	accuracy			0.72	95722
macro avg	0.59	0.75	0.58	95722	macro avg	0.59	0.76	0.57	95722	macro avg	0.59	0.76	0.57	95722
weighted avg	0.91	0.73	0.79	95722	weighted avg	0.91	0.72	0.78	95722	weighted avg	0.91	0.72	0.78	95722
					Test set 2					Toot oot 2				
Test set 2	precision	recall	f1-score	support	rest set 2	precision	recall	f1-score	support	Test set 2	precision	recall	f1-score	support
											·			
0	0.97 0.22	0.73 0.79	0.83 0.34	87484 8238	0 1	0.98 0.21	0.71 0.81	0.82 0.33	87484 8238	0	0.98 0.21	0.71 0.82	0.82 0.33	87484 8238
1	0.22	0.79	0.34	0238	1	0.21	0.01	0.55	0230	1	0.21	0.32	0.33	0230
accuracy			0.74	95722	accuracy			0.72	95722	accuracy			0.72	95722
macro avg weighted avg	0.59 0.91	0.76 0.74	0.59 0.79	95722 95722	macro avg weighted avg	0.59 0.91	0.76 0.72	0.58 0.78	95722 95722	macro avg weighted avg	0.59 0.91	0.76 0.72	0.58 0.78	95722 95722
weighten avg	0.91	0.74	0.79	93/22	weighten avg	0.51	0.72	0.76	33122	weighten avg	0.91	0.72	0.78	33122

MODEL COMPARISON

	Logistic Reg	ression				
	Threshold = !	50 precision	recall	f1-score	support	
	0	0.97 0.21	0.73 0.79	0.83 0.34	87544 8178	
	accuracy macro avg weighted avg	0.59 0.91	0.76 0.73	0.73 0.58 0.79	95722 95722 95722	
1	Threshold = !	52 precision	recall	f1-score	support	
	0 1	0.97 0.22	0.75 0.76	0.85 0.34	87544 8178	
	accuracy macro avg weighted avg	0.60 0.91	0.76 0.75	0.75 0.60 0.80	95722 95722 95722	
•	Threshold = !	53 precision	recall	f1-score	support	
	0 1	0.97 0.23	0.76 0.75	0.85 0.35	87544 8178	
	accuracy macro avg weighted avg	0.60 0.91	0.76 0.76	0.76 0.60 0.81	95722 95722 95722	

Decision Tree	1			
Threshold = 5	10			
THI COHOLU - D	precision	recall	f1-score	support
	precion	recult	11 3001 0	опррот с
0	0.97	0.70	0.82	87544
1	0.20	0.80	0.32	8178
accuracy			0.71	95722
macro avg	0.59	0.75	0.57	95722
weighted avg	0.91	0.71	0.78	95722
Threshold = 5	841 0 0 0 0 0			
	precision	recall	f1-score	support
0	0.97	0.74	0.84	87544
1	0.22	0.76	0.34	8178
accuracy			0.74	95722
macro avg	0.59	0.75	0.59	95722
weighted avg	0.91	0.74	0.80	95722
Threshold = 5	6			
Till Eshold = 3	precision	recall	f1-score	support
0	0.97	0.75	0.84	87544
1	0.22	0.75	0.34	8178
accuracy			0.75	95722
macro avg	0.59	0.75	0.59	95722
weighted avg	0.91	0.75	0.80	95722

Random Forest									
Threshold = 50									
	precision	recall	f1-score	support					
0	0.97	0.71	0.82	87544					
1	0.20	0.79	0.32	8178					
accuracy			0.71	95722					
macro avg	0.59	0.75	0.57	95722					
weighted avg	0.91	0.71	0.78	95722					
Threshold = 54									
	precision	recall	f1-score	support					
0	0.97	0.73	0.83	87544					
0	0.21	0.73	0.83	8178					
1	0.21	0.76	0.33	81/8					
accuracy			0.73	95722					
macro avg	0.59	0.75	0.58	95722					
weighted avg	0.91	0.73	0.79	95722					
Threshold = 5	5								
TIII ESHOTU - 3	precision	recall	f1-score	support					
	Parameter and a second			CONTRACTOR					
0	0.97	0.75	0.85	87544					
1	0.22	0.73	0.33	8178					
accuracy			0.75	95722					
macro avg	0.59	0.74	0.59	95722					
weighted avg	0.90	0.75	0.80	95722					



APPLICATION

Heart Disease Risk Prediction
with Streamlit

