











Team 8 Guoshan Yu, Jiwoo Suh, Kyuri Kim







#### **CONTENTS**

- **01.** Dataset Information
- **02.** EDA Analysis
- **03.** Data preprocessing
- **04.** Modelling
- 05. Evaluation
- 06. Conclusion











## Data Information >

Data Size : *319,795* 

Target: Yes - 27,373 No: 292,422

Categorical Features: 13

Numerical Features: 4BMI, Physical Health,Mental Health, and Sleep Time

RangeIndex: 319795 entries, 0 to 319794 Data columns (total 18 columns):

#	Column	Non-Null Count	Dtype		
0	HeartDisease	319795 non-null	object		
1	BMI	319795 non-null	float64		
2	Smoking	319795 non-null	object		
3	AlcoholDrinking	319795 non-null	object		
4	Stroke	319795 non-null	object		
5	PhysicalHealth	319795 non-null	float64		
6	MentalHealth	319795 non-null	float64		
7	DiffWalking	319795 non-null	object		
8	Sex	319795 non-null	object		
9	AgeCategory	319795 non-null	object		
10	Race	319795 non-null	object		
11	Diabetic	319795 non-null	object		
12	PhysicalActivity	319795 non-null	object		
13	GenHealth	319795 non-null	object		
14	SleepTime	319795 non-null	float64		
15	Asthma	319795 non-null	object		
16	KidneyDisease	319795 non-null	object		
17	SkinCancer	319795 non-null	object		
dtyp	es: float64(4), ob	ject(14)			

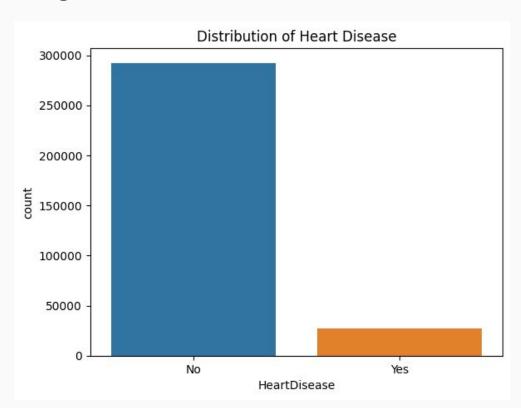








## **Target Variable - Heart Disease**



• Highly imbalanced dataset

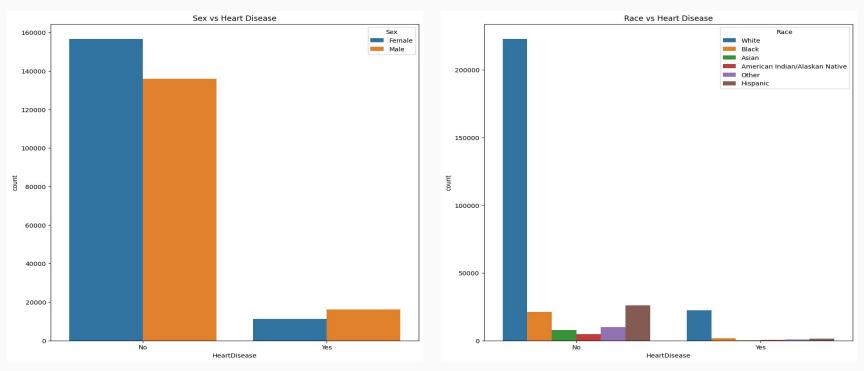








## **Sex/Race by Heart Disease**



- For Sex, male has higher proportion for having heart disease in the dataset
- For Race, White has higher proportion for having heart disease in the dataset but the data is imbalanced

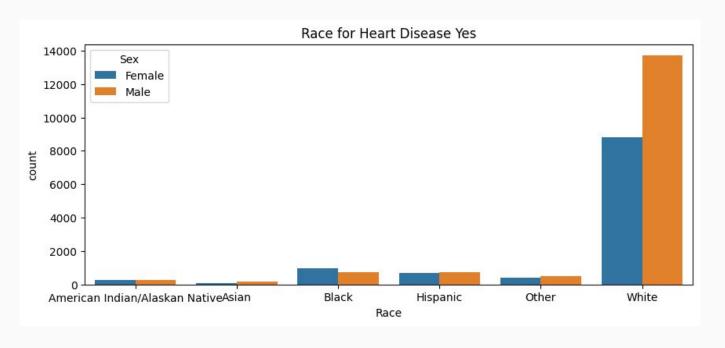








## **Race by Sex & Heart Disease**



By Race, Male still has higher portion for having heart disease for every race but Black

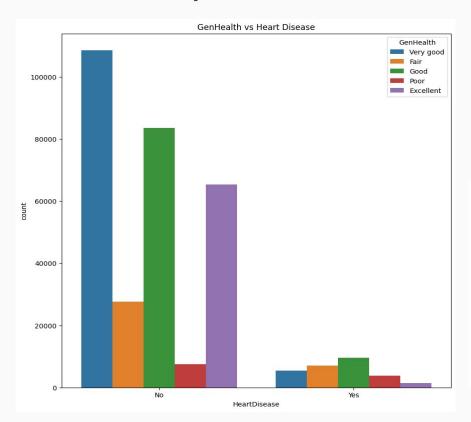




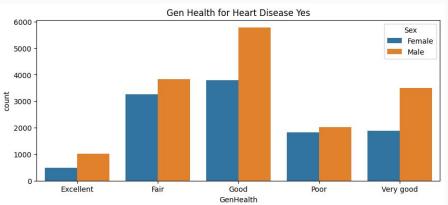




#### **General Health by Heart Disease**



- For General Health, respondents who reported having Excellent, Very good or Good health have bigger proportion for not having heart disease
- For having heart disease, respondents who have Good health are take the biggest part followed by Fair



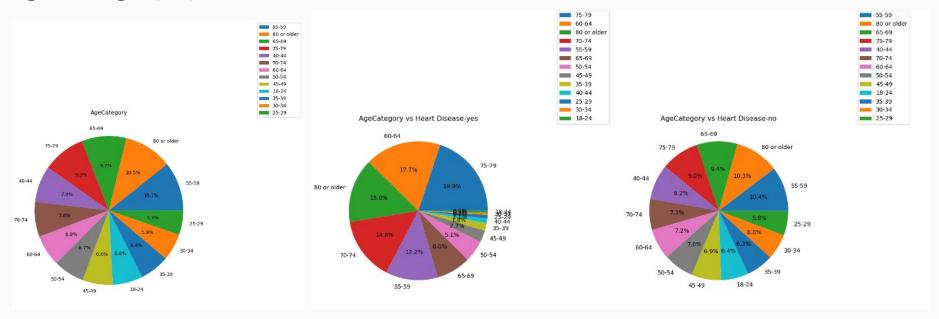








## **Age Category by Sex & Heart Disease**



- For Age category, the biggest one is age from 55-59, followed by 80 or older
- For Age category with having heart disease, age group 75-79 has the highest portion, followed by 60-64





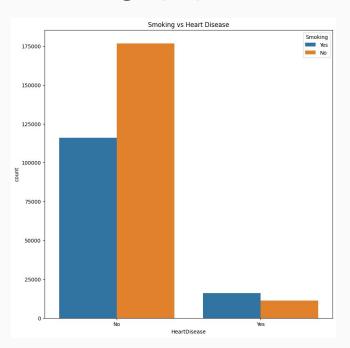


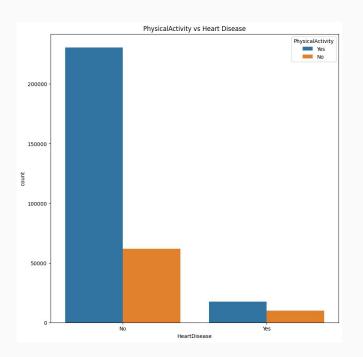






#### **Behavior Category by Heart Disease**





- Individuals who smoke have higher risk of getting a heart disease.
- Individuals who are doing physical activity or exercise other than their regular job have higher risk of getting a heart disease compared to those who doesn't.



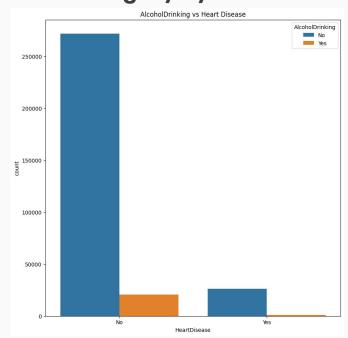


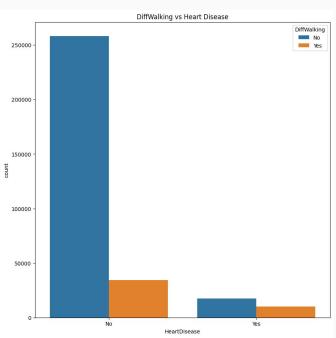






## **Behavior Category by Heart Disease**





- Alcohol drinking does not have huge impact on having a heart disease as individuals who drink alcohol do not tend to have heart disease.
- Difficult walking seem to not have much impact on heart disease in general, because individuals who have difficult walking do not tend to have heart disease.



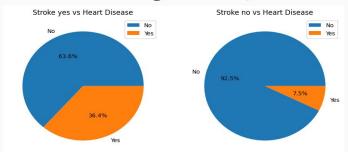




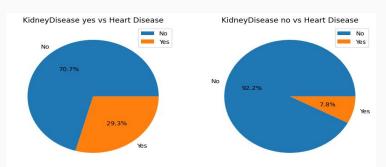




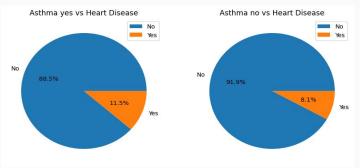
#### **Diseases Categories by Heart Disease**



 Individuals who have suffered from a stroke are at an elevated risk of developing heart disease



 Individuals with kidney disease have a higher risk of developing heart disease



 Individuals who have suffered from a Asthma does not have great impact on heart disease



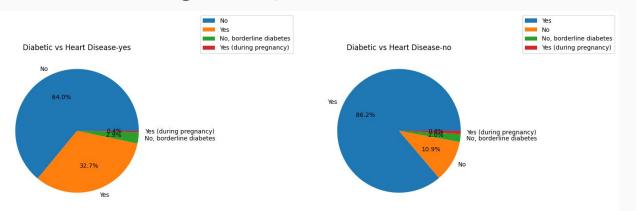






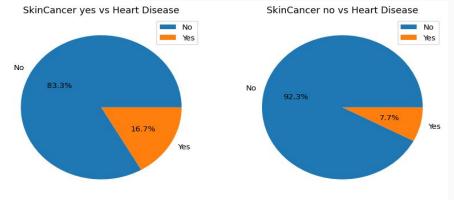


## **Diseases Categories by Heart Disease**



 Individuals with no diabetic have a higher risk of developing heart disease

 Individuals with Skin Cancer have a higher risk of developing heart disease



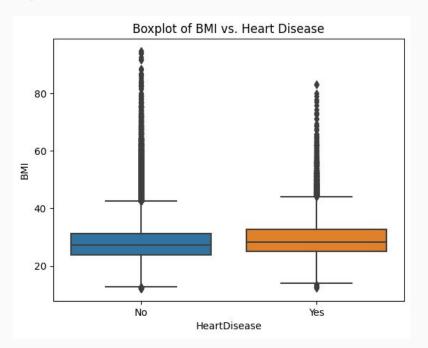


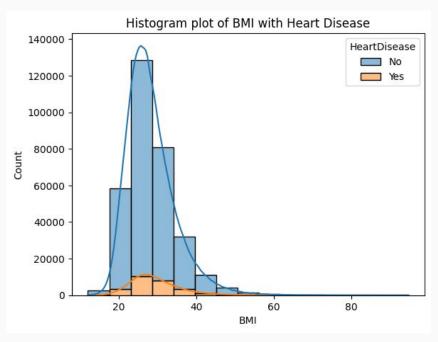






## **BMI by Heart Disease**





• For both groups, who are having heart disease yes and no, the distributions of BMI are similar.



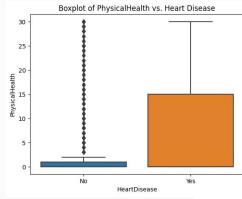




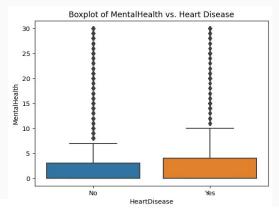




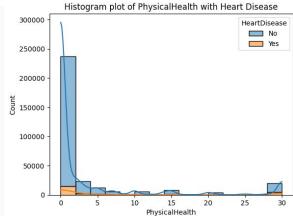
## **Health Condition by Heart Disease**

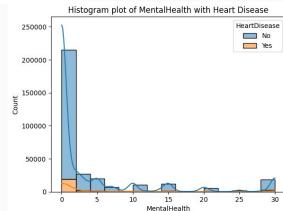


For both groups, who are having heart disease yes and no, have similar Physical Health condition during the past 30 days.



For both groups, who are having heart disease yes and no, have similar Mental Health condition during the past 30 days.





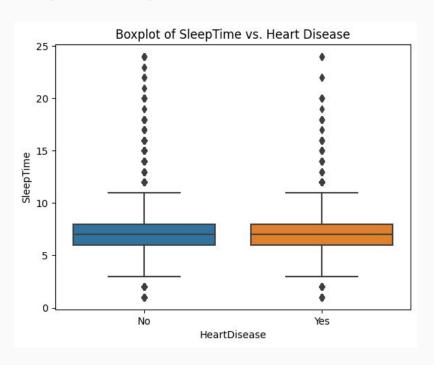


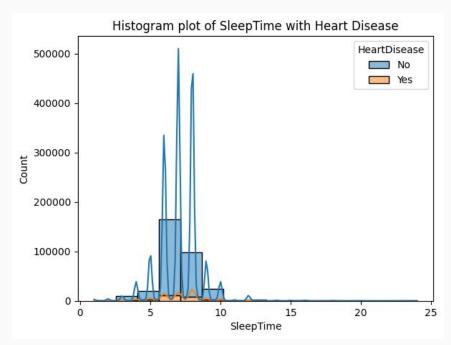






# **Sleep Time by Heart Disease**





For both groups, who have heart disease yes and no, have similar hours of sleep time in 24 hours.









#### **Exploratory Data Analysis - Correlation**



HeartDisease -	1.00	0.05	0.11	0.03	0.20	0.17	0.03	0.20	0.07	0.25	0.05	0.19	0.10	0.26	0.01	0.04	0.15	0.09
ВМІ -	0.05	1.00	0.02	0.04	0.02		0.06	0.18	0.03				0.15		-0.05	0.09	0.05	0.03
Smoking -		0.02	1.00	0.11	0.06		0.09		0.09	0.16		0.06	0.10		0.03	0.02	0.03	0.03
AlcoholDrinking -	0.03	0.04	0.11	1.00	0.02	0.02	0.05	0.04	0.00	0.07	0.04	0.06	0.02	0.03	0.01	0.00	0.03	0.01
Stroke -		0.02	0.06	0.02	1.00	0.14	0.05		0.00		0.04		0.08	0.19	0.01	0.04	0.09	0.05
PhysicalHealth -			0.12	0.02		1.00	0.29	0.43	0.04		0.05	0.16		0.59	-0.06	0.12	0.14	0.04
MentalHealth -	0.03	0.06	0.09	0.05	0.05		1.00	0.15	0.10	0.16	0.05	0.03	0.10		-0.12		0.04	0.03
DiffWalking -		0.18		0.04				1.00	0.07		0.06			0.46	0.02	0.10		0.06
Sex -	0.07	0.03	0.09	0.00	0.00	0.04	0.10	0.07	1.00	0.07	0.04	0.09	0.05	0.03	0.02	0.07	0.01	0.01
AgeCategory -			0.16	0.07	0.14		0.16		0.07	1.00	0.10		0.13	0.10		0.06		
Race -	0.05			0.04	0.04	0.05	0.05	0.06	0.04	0.10	1.00	0.05	0.08	0.06	0.05	0.04	0.02	0.15
Diabetic -			0.06	0.06			0.03		0.09		0.05	1.00	0.14		0.01	0.05	0.15	0.04
PhysicalActivity -	0.10		0.10	0.02	0.08	0.23	0.10		0.05		0.08		1.00	0.29	0.00	0.04	0.08	0.00
GenHealth -				0.03		0.59		0.46	0.03	0.10	0.06			1.00	0.06	0.14		0.05
SleepTime -		-0.05	0.03	0.01	0.01	-0.06	-0.12	0.02	0.02		0.05	0.01	0.00	0.06	1.00	0.05	0.01	0.04
Asthma -		0.09	0.02	0.00	0.04			0.10	0.07	0.06	0.04	0.05	0.04		0.05	1.00	0.04	0.00
KidneyDisease -		0.05	0.03	0.03	0.09		0.04		0.01		0.02		0.08		0.01	0.04	1.00	0.06
SkinCancer -		0.03	0.03	0.01	0.05	0.04	0.03	0.06	0.01	0.28	0.15	0.04	0.00	0.05	0.04	0.00	0.06	1.00
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer

HeartDisease	1.000000
GenHealth	0.259519
AgeCategory	0.245588
DiffWalking	0.201234
Stroke	0.196798
Diabetic	0.185101
PhysicalHealth	0.170721
KidneyDisease	0.145157
Smoking	0.107738
PhysicalActivity	0.100001
SkinCancer	0.093281
Sex	0.070007
BMI	0.051803
Race	0.051230
Asthma	0.041390
AlcoholDrinking	0.032009
MentalHealth	0.028591
SleepTime	0.008327











# Data Preprocessing >



#### **Outliers**

Outliers = 1,5 \* IQR

Total outliers: 994,324

Data Size: *319, 795 - 225, 471* 



#### **Data Encoding**

Target: yes: 1 No: 0

Cat Features : One hot Encoding



#### Standardization

Normalization



#### Split the Data

Test Size = 0.2























Step 01

Step 02

Step 03

Step 04

**Balance the Data** 

RandomUnderSampler

**Feature Reduction** 

**RFECV** 

**Train the models** 

- **Logistic Regression**
- **Decision Tree**
- Random Forest
- **Logistic Regression**

Fit in the test data

- **Decision Tree**
- Random Forest

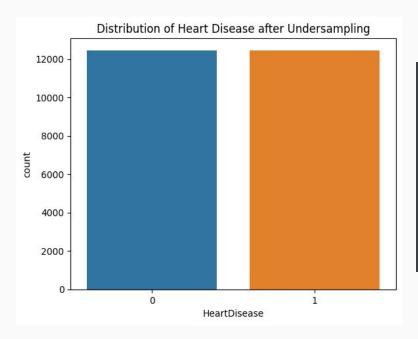








### **Undersampling**



```
Before Undersampling, counts of label '1': [12462]
Before Undersampling, counts of label '0': [167914]

After Undersampling, the shape of train_X: (24924, 50)
After Undersampling, the shape of train_y: (24924, 1)

After Undersampling, counts of label '1': [12462]

After Undersampling, counts of label '0': [12462]
```



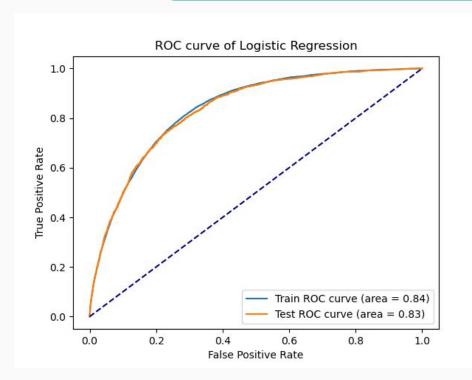












#### Number of features selected: 43

Target	Precision	Recall	f1-score
0	0.98	0.74	0.84
1	0.18	0.78	0.3
accuracy		0.74	
macro avg	0.58	0.76	0.57
weighed avg	0.92	0.74	0.8
balanced Accuracy	0.756	3	



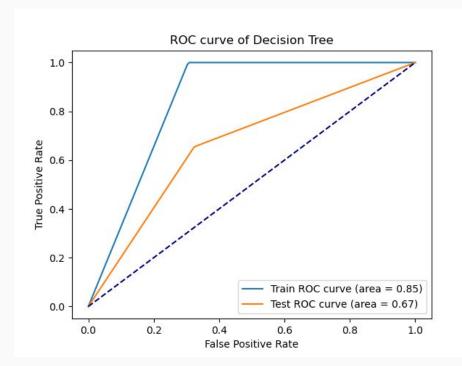






#### **Decision Tree**





#### Number of features selected: 49

Target	Precision	Recall	f1-score
0	0.96	0.68	0.80
1	0.13	0.65	0.22
accuracy		0.68	
macro avg	0.55	0.767	0.51
weighed avg	0.91	0.68	0.76
balanced Accuracy	0.664	5	



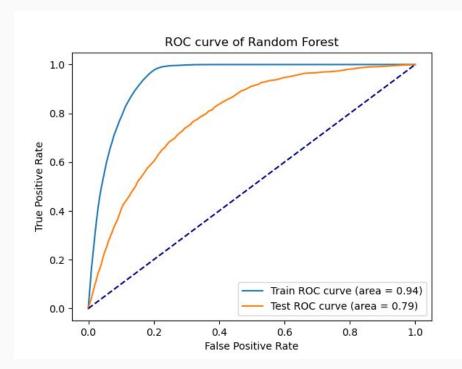












#### Number of features selected: 42

Target	Precision	Recall	f1-score
0	0.97	0.71	0.82
1	0.16	0.74	0.26
accuracy		0.71	
macro avg	0.57	0.72	0.54
weighed avg	0.92	0.71	0.78
balanced Accuracy	0.721	8	







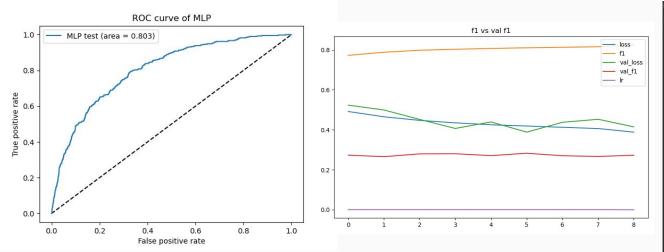






#### **Multi Layer Perceptron**





Model: "sequential_1"		
Layer (type)		Param #
flatten_1 (Flatten)		Θ
rtatten_1 (rtatten)	(Notic, 50)	
dense_1 (Dense)	(None, 256)	13056
danage (Danage)	(Name 25()	Θ
dropout (Dropout)	(None, 256)	U
dense_2 (Dense)	(None, 256)	65792
	(1)	
dropout_1 (Dropout)	(None, 256)	
dense_3 (Dense)	(None, 256)	65792
	(1) 05()	
dropout_2 (Dropout)	(None, 256)	
dense_4 (Dense)	(None, 128)	32896
<pre>batch_normalization (BatchN ormalization)</pre>	(None, 128)	512
or mactizaction)		
dense_5 (Dense)	(None, 1)	129
Total params: 178,177		
Trainable params: 177,921		

Total params: 178,177
Trainable params: 177,921
Non-trainable params: 256



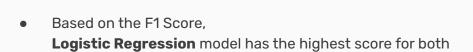








#### **Best Model - Logistic Regression**



Model	F1 Score
Logistic Regression	0.57
Decision Tree	0.51
Random Forest	0.54
Neural Networks	0.26





# Thanks!

<>企 ()