Heart Disease Classification

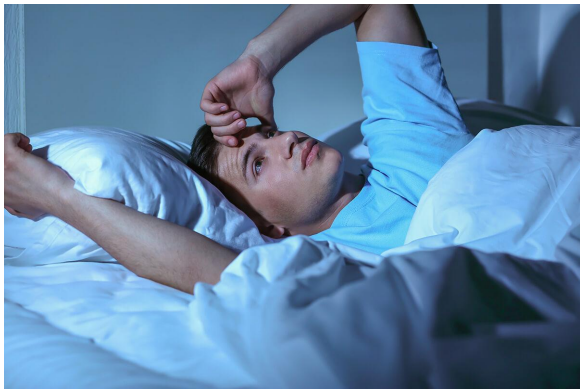
Junze He, Jun Ryu, Arya Patel, Giovanni Cinque

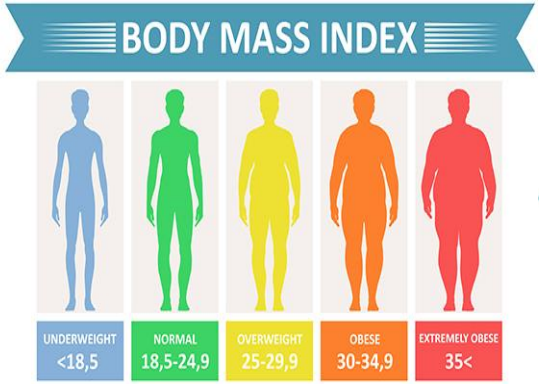
Background

Heart Disease is one of the top factors leading to death according to CDC.

About 50% of people in the United States have **at least one risk factor** for heart disease such as high BMI, poor sleep quality, or even mental health.

Risk factors are the **key indicators** of Heart Disease.





Picture Sources: https://www.redoakrecovery.com/addiction-blog/can-poor-sleep-lead-to-pessimism/ , https://www.cdc.gov/healthyweight/assessing/bmi/index.html,

https://www.health.harvard.edu/mind-and-mood/blasting-through-mental-health-misperception

Dataset & Goal

The dataset is a 2022 annual CDC survey data of 400k+ adults related to their health conditions.

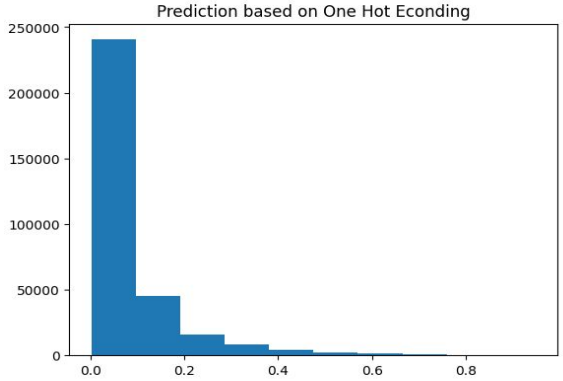
Cleaned Dataset

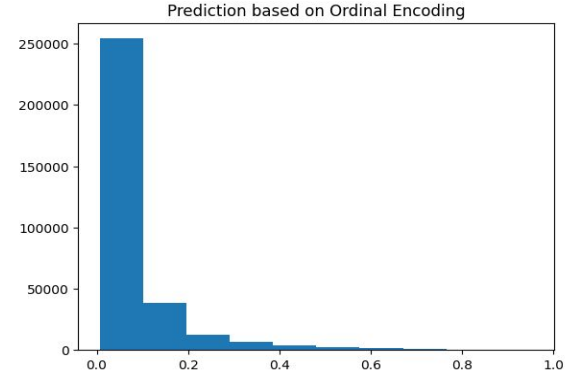
Goal: Develop a good **classification model** to detect whether an adult has heart disease.

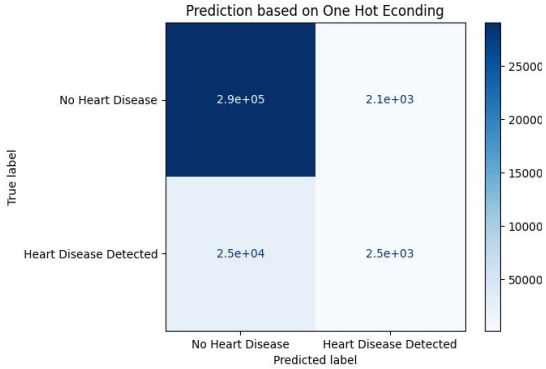
- Analyze the **Significance** of each predictor
- investigate the **Probability** of two prediction results occur
- Pipeline Models

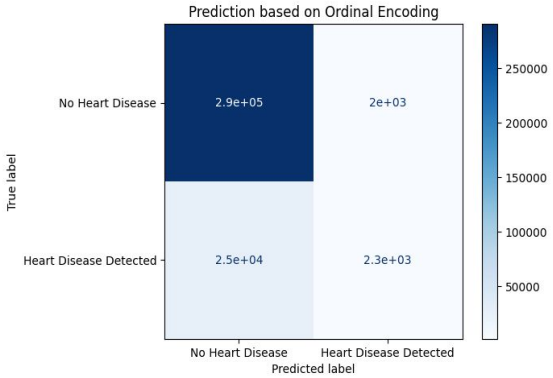
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes	Very good	5.0	Yes	No	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes	Very good	7.0	No	No	No
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes	Fair	8.0	Yes	No	No
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No	Good	6.0	No	No	Yes
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes	Very good	8.0	No	No	No

Simple EDA with Statsmodel.Logit









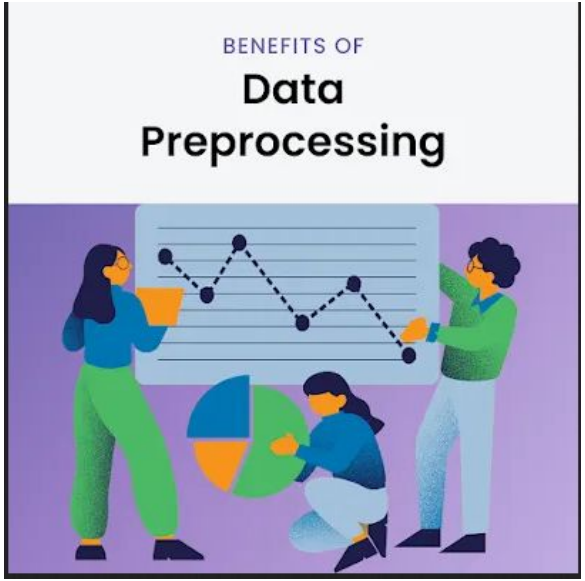
Logit Regression Results

Dep. Variable:	HeartDisease	**No. Observations:**	319795
Model:	Logit	**Df Residuals:**	319778
Method:	MLE	**Df Model:**	16
Date:	Fri, 01 Mar 2024	**Pseudo R-squ.:**	0.1445
Time:	03:16:04	**Log-Likelihood:**	-79953.
converged:	True	**LL-Null:**	-93453.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.0470	0.056	-72.583	0.000	-4.156	-3.938
BMI	-0.0020	0.001	-1.850	0.064	-0.004	0.000
PhysicalHealth	0.0217	0.001	28.943	0.000	0.020	0.023
MentalHealth	-0.0075	0.001	-8.844	0.000	-0.009	-0.006
SleepTime	0.0208	0.004	4.958	0.000	0.013	0.029
Smoking	0.5127	0.014	36.932	0.000	0.485	0.540
AlcoholDrinking	-0.4629	0.033	-14.204	0.000	-0.527	-0.399
Stroke	1.3787	0.022	61.540	0.000	1.335	1.423
DiffWalking	0.7608	0.018	42.948	0.000	0.726	0.795
Sex	0.6146	0.014	43.780	0.000	0.587	0.642
Race	0.1153	0.006	18.298	0.000	0.103	0.128
Diabetic	0.4136	0.008	52.904	0.000	0.398	0.429
PhysicalActivity	-0.2126	0.016	-13.593	0.000	-0.243	-0.182
GenHealth	-0.0076	0.005	-1.556	0.120	-0.017	0.002
Asthma	0.1733	0.019	9.324	0.000	0.137	0.210
KidneyDisease	0.8660	0.024	35.633	0.000	0.818	0.914
SkinCancer	0.6038	0.019	31.980	0.000	0.567	0.641

Data Preprocessing

- Split data into predicted and response variables
- Mapped Heart Disease and No Heart Disease values to 0 and 1 respectively
- Ordinal Encoding and One Hot Encoding
- Oversampling / Balanced Outcomes
- Divided predicted variables and response variable into train and test, Shuffled them
- Standardized predicted variables



Evaluation of Models

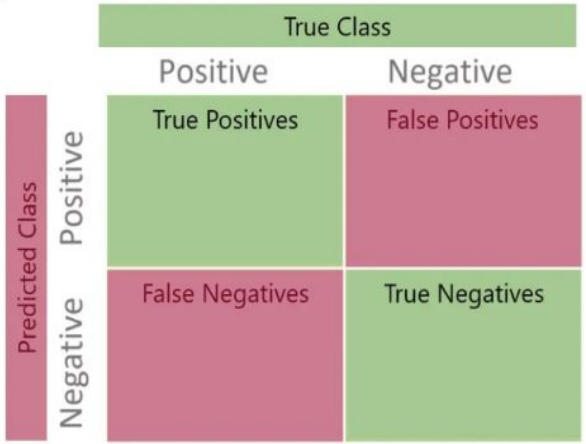
Confusion Matrix: represents how well a model can predict each category in a response variable in a matrix form

Precision: True Positive / (True Positive + False Positive)

Recall: True Positive / (True Positive + False Negative)

F1 Score: 2 *(Precision * Recall) / (Precision + Recall)

Validation Score: N Fold-Cross Validation / check overfitting



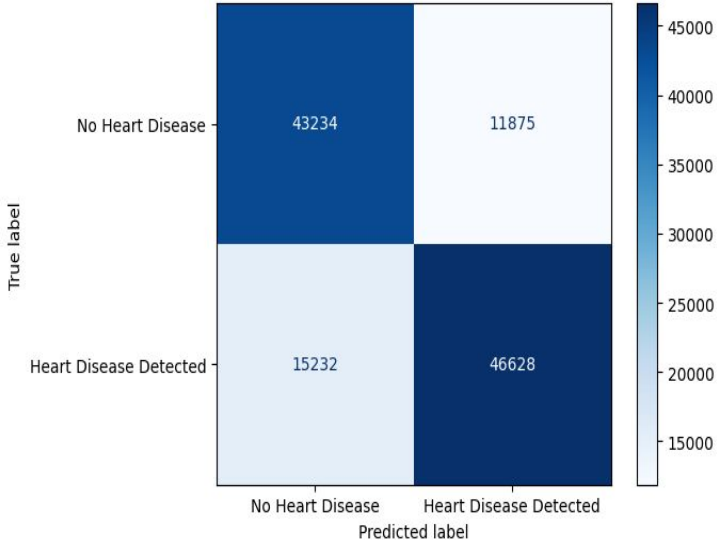
XGBClassifier (Ordinal Encoding)

```
classifier = XGBClassifier(
      n_estimators=100, learning_rate=0.05)

classifier.fit(
    train_scaled_x, train_y,
    verbose=False
)
```

Validation Score : 77.15%

	precision	recall	f1-score	support
0	0.74	0.78	0.76	55109
1	0.80	0.75	0.77	61860
accuracy			0.77	116969
macro avg	0.77	0.77	0.77	116969
weighted avg	0.77	0.77	0.77	116969



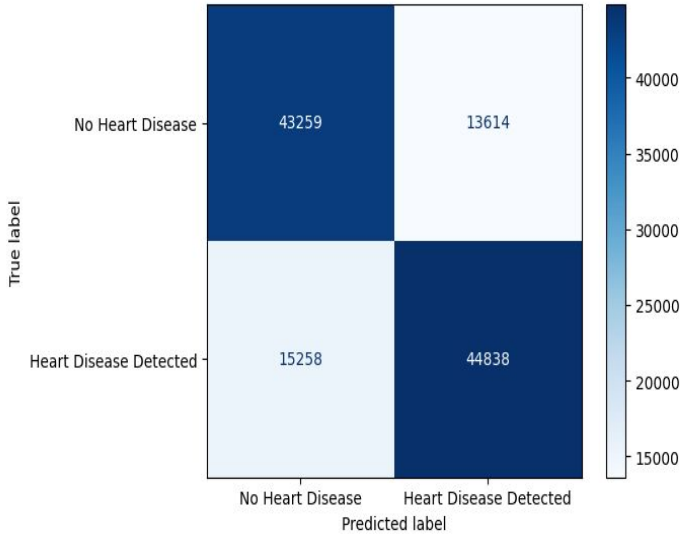
Logistic Regression (Ordinal Encoding)

```
model_lr = LogisticRegression(
      C=0.5, random_state=1234)

lr_pipeline, lr_cm, lr_report,
lr_validation_score = pipelineModel(
      model_lr, x_ros, y_ros)
```

Validation Score: 75.12%

	precision	recall	f1-score	support
0	0.74	0.76	0.75	56873
1	0.77	0.75	0.76	60096
accuracy			0.75	116969
macro avg	0.75	0.75	0.75	116969
weighted avg	0.75	0.75	0.75	116969



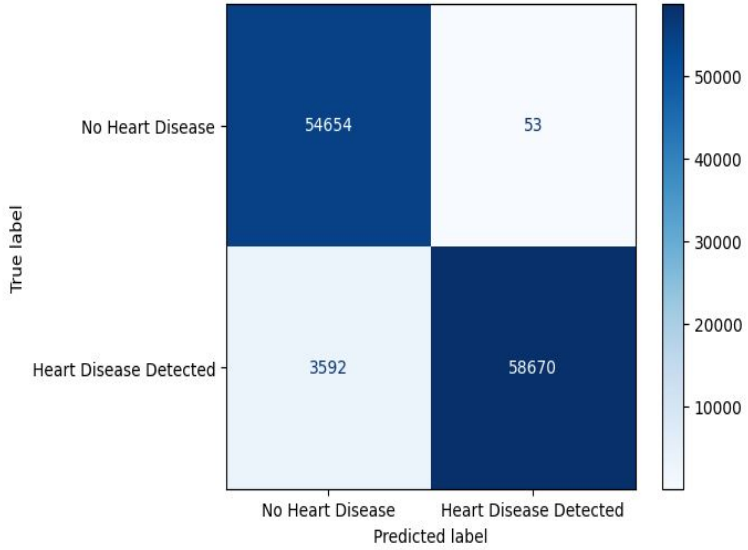
Decision Tree Classifier (Ordinal Encoding)

```
model_dt = DecisionTreeClassifier(
      max_depth=30, random_state=1234)

dt_pipeline, dt_cm, dt_report,
dt_validation_score = pipelineModel(
      model_dt, x_ros, y_ros)
```

Validation Score: 96.09%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	54707
1	1.00	0.94	0.97	62262
accuracy			0.97	116969
macro avg	0.97	0.97	0.97	116969
weighted avg	0.97	0.97	0.97	116969



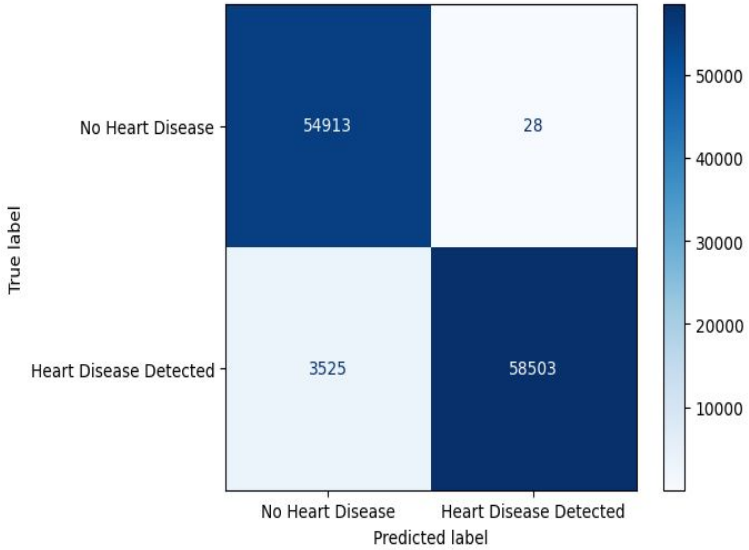
Random Forest Classifier (Ordinal Encoding)

```
rfc_model = RandomForestClassifier(
      n_estimators=30)

rfc_pipeline, rfc_cm,
rfc_report, rfc_validation_score = pipelineModel(
      rfc_model, x_ros, y_ros)
```

Validation Score : 96.10%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	54941
1	1.00	0.94	0.97	62028
accuracy			0.97	116969
macro avg	0.97	0.97	0.97	116969
weighted avg	0.97	0.97	0.97	116969



Model Selections

	XGBClassifier	Logistic Regression	Decision Tree Classifier	Random Forest Classifier
Precision	80%	77%	100%	100%
Recall	75%	75%	94%	94%
Validation Score	77%	76%	96.09%	96.10%
F1-Score	77.15%	75.12%	97%	97%

Feature Engineering

- Mean Age Category -> Mean Age
  - Given a value “20-30” in Mean Age Category variable
    - The mean of 20 and 30 is 25

- Mean Age -> General Walking Speed
  - Given a person who is between age 20 and age 29
    - Classifies his or her general walking speed as 1.35 Meters/Second

- BMI -> BMI Distinction
  - Given a person’s BIM is less than 18.5
    - Classifies that person as underweight

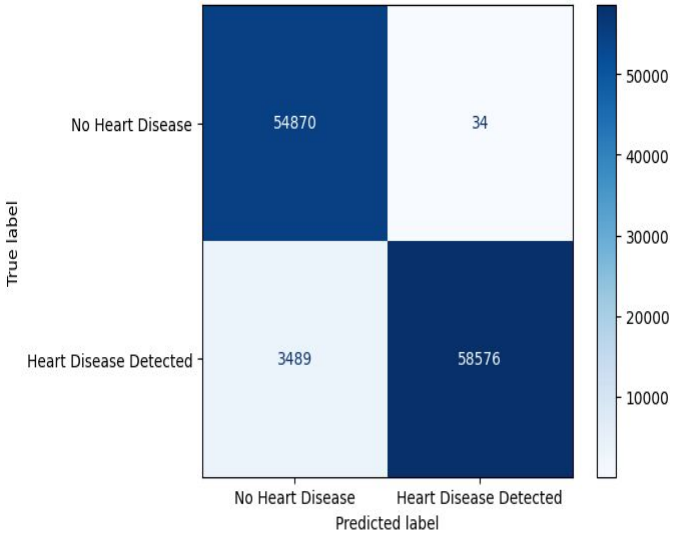
BMI_Distinction	MeanAge	Walking_Speed
Underweight	57.0	1.370
Normal Weight	80.0	0.955
Overweight	67.0	1.290
Normal Weight	77.0	1.195
Normal Weight	42.0	1.410

Add More Predictors in the Random Forest Classifier (Ordinal Encoding)

```
added_features_rfc_pipeline, added_features_rfc_cm,
added_features_rfc_report, added_features_rfc_validation
    RandomForestClassifier(n_estimators=30),
    x_added_features_ros,
    y_added_features_ros
)
```

Validation Score: 96.09%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	54904
1	1.00	0.94	0.97	62065
accuracy			0.97	116969
macro avg	0.97	0.97	0.97	116969
weighted avg	0.97	0.97	0.97	116969



Conclusion

- We achieved a decent overall model accuracy, precision, recall and validation score in the Random Forest Classifier.

- The added variables did not improve the model accuracy , but they enhanced the **running speed** of the model about 30 seconds faster.

- Learning Concepts:
  - Basic Machine Learning Concepts
  - Exploring Scikit-Learn packages
  - Oversampling
  - Feature Engineering
  - Model Selection

Reference

Understanding Precision, Sensitivity, and Specificity In Classification Modeling and How To Calculate Them With A Confusion Matrix | Towards Data Science

Indicators of Heart Disease (2022 UPDATE) (kaggle.com)

https://www.cdc.gov/heartdisease/index.htm

https://www.healthline.com/health/exercise-fitness/average-walking-speed#average-speed-by-age

https://medium.com/@vaibhav1822217/data-cleaning-and-preprocessing-techniques-in-data-analytics-351ee6e3dfa7