## **Heart Disease Dataset**

Group 7
Stephan Dupoux, Allie Schneider & Francis Michael Villamater



#### Contents

**Dataset Overview** Hyper Parameter Tuning

**Project Objective** Model Performance Reports

Dataset Variables Proposed solution

**Exploratory Data** 

Analysis

**Process** 

#### **Overview**

- Data is sourced from the CDC as part of the Behavioral Risk Factor Surveillance System and was cleaned by a user on Kaggle.
- Heart Disease is one of the leading causes of death amongst most races in the US.
- About half of all Americans have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.
- Other key indicators: diabetic status, obesity (high BMI), not enough physical activity, or excessive alcohol consumption
- 18 Attributes and 319,795 instances.



### **Project Objective**

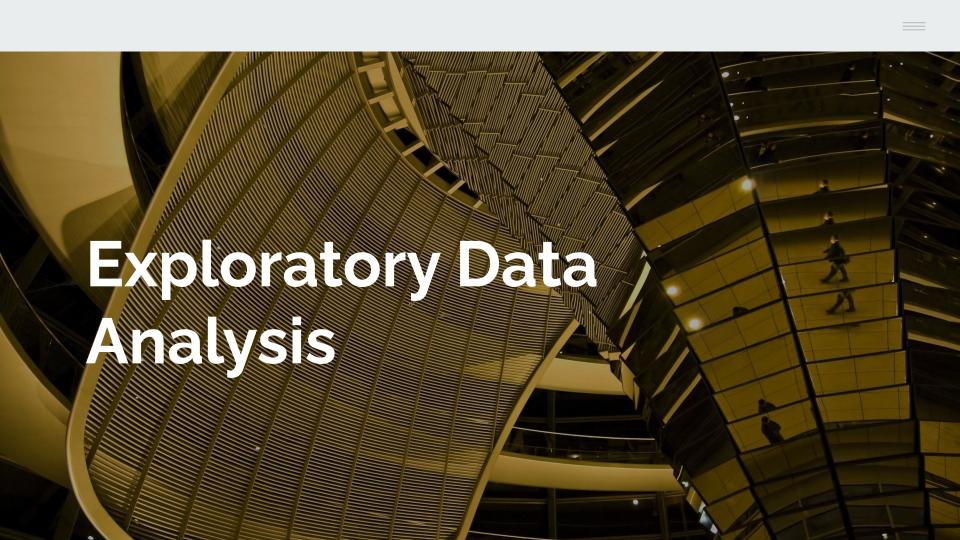
Leverage machine learning methods to detect "patterns" form the data that can predict a patient's condition (presence or absence of heart disease). Adjust weights/undersampling/oversampling and apply classifier models (logistic regression, random forest, Naive Bayes)

#### **Dataset Variables**

- HeartDisease (binary)
- BMI (continuous) Body Mass Index
- Smoking (binary) have you smoked ≥ 100 cigarettes in your life?
- AlcoholDrinking (binary) Yes if male and >14 drinks/week or female and >7 drinks/week
- Stroke (binary) Ever had a stroke
- PhysicalHealth (Continuous) including physical illness and injury, how many days of the past 30 was your physical health not good?
- MentalHealth (Continuous) how many days of the past 30 was your mental health not good?
- DiffWalking (binary) Do you have serious difficulty walking or climbing stairs?
- Sex (binary) male or female
- AgeCategory (category, 13 levels) Age category

#### Dataset Variables (cont'd)

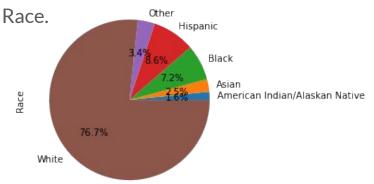
- Race (categorical, 6 levels) Race/ethnicity
- Diabetic (categorical, 4 levels) have they ever been diagnosed with diabetes?
- PhysicalActivity (binary) Adults who reported engaging in physical activity/exercise over the last 30 days outside of their regular job
- GenHealth (categorical, 5 levels) How would you rate their overall health?
- SleepTime (continuous) how much sleep do they get in a 24 hour period?
- Asthma (binary) have they been diagnosed with asthma?
- KidneyDisease (binary) excluding kidney stones, bladder infection, or incontinence, have they ever been diagnosed with kidney disease?
- SkinCancer (binary) have they ever been diagnosed with skin cancer?

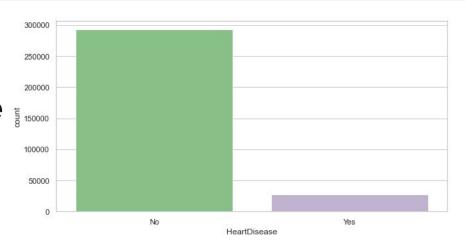


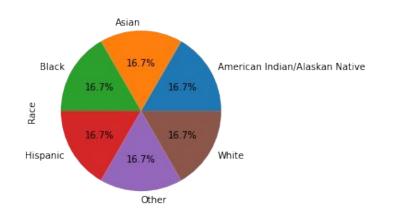
#### **Heart Disease and Race**

The dataset is highly imbalanced with most instances classified as "No" for Heart Disease (292422 negative, 27373 positive)

Overwhelming population of "White" in







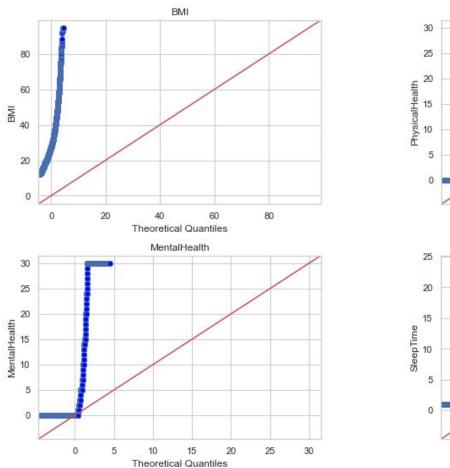
### **Handling Imbalanced Dataset:**

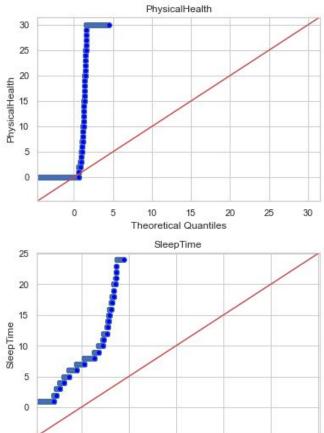
- 292,422 Negative for Heart Disease
- 27,373 Positive for Heart Disease
- Resample on Race

### **Continuous Variables**

	ВМІ	PhysicalHealth	MentalHealth	SleepTime
count	319795.00	319795.00	319795.00	319795.00
mean	28.33	3.37	3.90	7.10
std	6.36	7.95	7.96	1.44
min	12.02	0.00	0.00	1.00
25%	24.03	0.00	0.00	6.00
50%	27.34	0.00	0.00	7.00
75%	31.42	2.00	3.00	8.00
max	94.85	30.00	30.00	24.00

### **Checking for Normality**





Theoretical Quantiles

## **Correlation Heat Map**

HeartDisease	1	0.11	-0.032	0.2	0.2	0.07	0.23	0.035	0.17	-0.1	-0.011	0.041	0.15	0.093	0.052	0.17	0.029	0.0083
Smoking	0.11	1	0.11	0.061	0.12	0.085	0.13	0.065	0.054	-0.097	0.021	0.024	0.035	0.034	0.023	0.12	0.085	-0.03
AlcoholDrinking	-0.032	0.11	1	-0.02	-0.035	0.0042	-0.06	0.037	-0.057	0.017	0.0016	-0.0022	-0.028	-0.0057	-0.039	-0.017	0.051	-0.0051
Stroke	0.2	0.061	-0.02	1	0.17	-0.0031	0.14	-0.004	0.1	-0.079	-0.0093	0.039	0.091	0.048	0.02	0.14	0.046	0.012
DiffWalking	0.2	0.12	-0.035	0.17	1	-0.069	0.24	-0.016	0.21	-0.28	-0.044	0.1	0.15	0.065	0.18	0.43	0.15	-0.022
Sex	0.07	0.085	0.0042	-0.0031	-0.069	1	-0.067	0.019	-0.013	0.048	-0.01	-0.069	-0.0091	0.013	0.027	-0.041	-0.1	-0.016
AgeCategory	0.23	0.13	-0.06	0.14	0.24	-0.067	1	0.16	0.19	-0.12	0.044	-0.058	0.12	0.26	-0.0047	0.11	-0.16	0.1
Race	0.035	0.065	0.037	-0.004	-0.016	0.019	0.16	1	-0.052	0.057	0.05	-0.018	0.0037	0.13	-0.037	-0.00085	-0.014	0.036
Diabetic	0.17	0.054	-0.057	0.1	0.21	-0.013	0.19	-0.052	1	-0.13	-0.011	0.05	0.14	0.033	0.2	0.15	0.033	0.00045
PhysicalActivity	-0.1	-0.097	0.017	-0.079	-0.28	0.048	-0.12	0.057	-0.13	1	0.024	-0.042	-0.082	-0.0013	-0.15	-0.23	-0.096	0.0038
GenHealth	-0.011	0.021	0.0016	-0.0093	-0.044	-0.01	0.044	0.05	-0.011	0.024	1	0.0073	-0.011	0.019	0.036	-0.036	-0.0044	-0.0042
Asthma	0.041	0.024	-0.0022	0.039	0.1	-0.069	-0.058	-0.018	0.05	-0.042	0.0073	1	0.04	-0.0004	0.092	0.12	0.11	-0.048
KidneyDisease	0.15	0.035	-0.028	0.091	0.15	-0.0091	0.12	0.0037	0.14	-0.082	-0.011	0.04	1	0.062	0.051	0.14	0.037	0.0062
SkinCancer	0.093	0.034	-0.0057	0.048	0.065	0.013	0.26	0.13	0.033	-0.0013	0.019	-0.0004	0.062	1	-0.034	0.042	-0.033	0.041
ВМІ	0.052	0.023	-0.039	0.02	0.18	0.027	-0.0047	-0.037	0.2	-0.15	0.036	0.092	0.051	-0.034	1	0.11	0.064	-0.052
PhysicalHealth	0.17	0.12	-0.017	0.14	0.43	-0.041	0.11	-0.00085	0.15	-0.23	-0.036	0.12	0.14	0.042	0.11	1	0.29	-0.061
MentalHealth	0.029	0.085	0.051	0.046	0.15	-0.1	-0.16	-0.014	0.033	-0.096	-0.0044	0.11	0.037	-0.033	0.064	0.29	1	-0.12
SleepTime	0.0083	-0.03	-0.0051	0.012	-0.022	-0.016	0.1	0.036	0.00045	0.0038	-0.0042	-0.048	0.0062	0.041	-0.052	-0.061	-0.12	1
	HeartDisease	Smoking	AcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	Asthma	KidneyDisease	SkinCancer	BMI	PhysicalHealth	MentalHealth	SleepTime

- 1.0

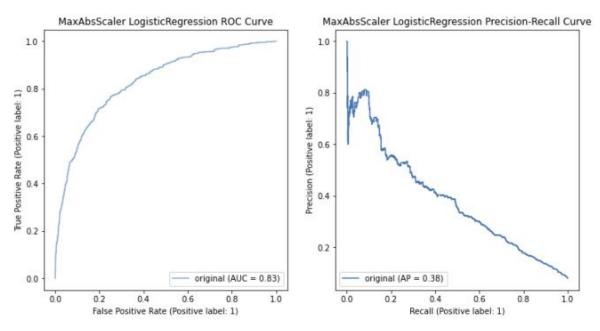
#### **Models Chosen**

- Logistic Regression
  - Simple to run and a good first step and helps select future models
  - Fastest to converge and not computationally intensive to train
- Random Forest
  - Chosen for its proficiency for handling high dimensional datasets
  - Algorithm will always converge with the datapoints
- Naive Bayes
  - Assumes no relation between the variables, and performs well with categorical variables.

## **Hyperparameter Tuning**

- Logistic Regression C values: [0.05, 0.5, 0.9]
- Random Forest Max Leafs: [5, 10, 15]
- Random Forest Number of Estimators: [500, 1000, 2000]
- Naive Bayes alpha values: [0.05, 0.5, 0.9]

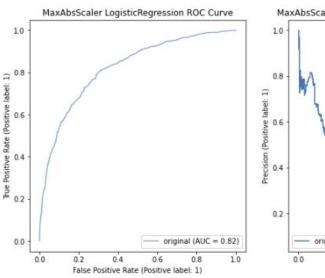
Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Precision	0.5797	0.5678	0.5696	0.5786	0.5653
Recall	0.1712	0.1650	0.1674	0.1734	0.1690
F1	0.2643	0.2557	0.2587	0.2668	0.2603
False Positive Rate	0.0107	0.0107	0.0108	0.0108	0.0112
False Negative Rate	0.0715	0.0717	0.0714	0.0709	0.0717

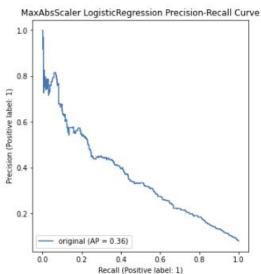


#### Log Transformation

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparam eter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Precision	0.5689	0.5637	0.5720	0.5657	0.5749
Recall	0.1556	0.1586	0.1551	0.1575	0.1562
F1	0.2443	0.2475	0.2441	0.2465	0.2457
False Positive Rate	0.0101	0.0105	0.0099	0.0103	0.0098
False Negative Rate	0.0729	0.0723	0.0722	0.0721	0.0722

#### Log Transformation





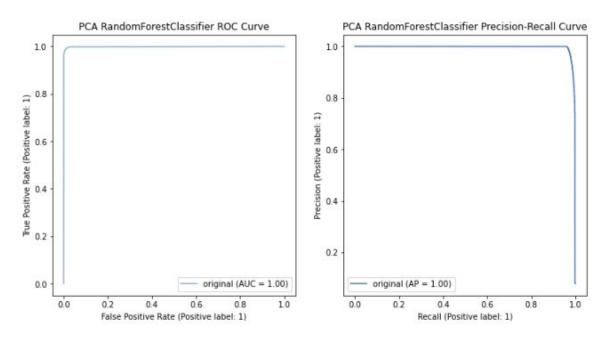
#### Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Threshold	0.5787	0.4881	0.5470	0.5760	0.5823

#### Log Transformation and Feature Scaling

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (C = 0.05)	Hyperparameter Tuning (C = 0.5)	Hyperparameter Tuning (C = 0.9)
Threshold	0.5801	0.5790	0.5216	0.5244	0.5801

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Precision	0.9930	0.9964	1.0	0.9330	0.9521
Recall	0.9645	0.9643	0.0	0.0104	0.0612
F1	0.9785	0.9801	0.0	0.0207	0.1150
False Positive Rate	0.0005	0.0002	0.0	6.4271	0.0002
False Negative Rate	0.0030	0.0030	0.0858	0.0846	0.0805



#### Log Transformation

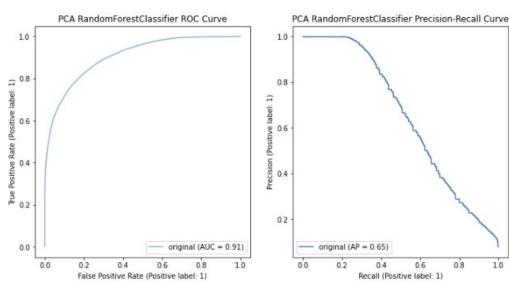
Scaler	MaxAbsScaler		Hyperparameter Tuning (n_estimators = 500 / max_leaf_nodes = 5)	Hyperparameter Tuning (n_estimators = 1000 / max_leaf_nodes = 10)	Hyperparameter Tuning (n_estimators = 2000 / max_leaf_nodes = 15)
Precision	0.8261	0.8339	1.0	0.8413	0.8924
Recall	0.4036	0.4064	0.0	0.0139	0.0372
F1	0.5423	0.5465	0.0	0.0273	0.0715
False Positive Rate	0.0072	0.0069	0.0	0.0002	0.0003
False Negative Rate	0.0511	0.0510	0.0856	0.0838	0.0826

0.6

0.8

1.0

#### Log Transformation



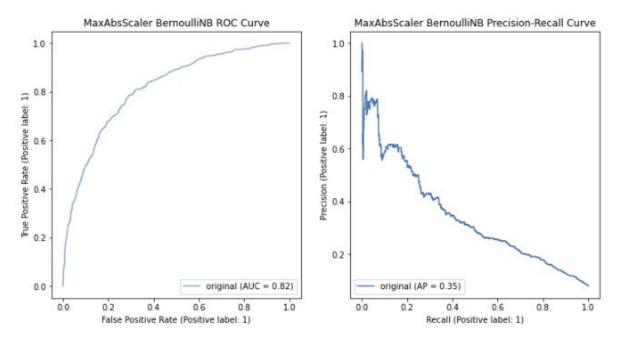
#### Feature Scaling

		PCA	Hyperparameter	Hyperparameter	Hyperparameter
Scaler	MaxAbsScaler	Dimensionality	Tuning (n_estimators = 500 /	Tuning (n_estimators = 1000 /	Tuning (n_estimators = 2000
		Reduction	max_leaf_nodes = 5)	max_leaf_nodes = 10)	/ max_leaf_nodes = 15)
Threshold	0.0400	0.04	0.04	0.04	0.04

#### Log Transformation and Feature Scaling

		PCA	Hyperparameter	Hyperparameter	Hyperparameter
Scaler	MaxAbsScaler	Dimensionality	Tuning (n_estimators = 500 /	Tuning (n_estimators = 1000 /	Tuning (n_estimators = 2000
		Reduction	max_leaf_nodes = 5 <b>)</b>	max_leaf_nodes = 10)	/ max_leaf_nodes = 15)
Threshold	0.05	0.0499	0.0500	0.0499	0.0499

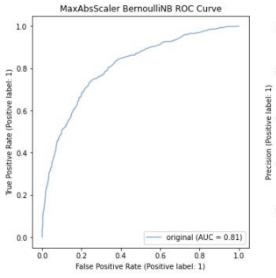
Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparameter Tuning (alpha = 0.05)	Hyperparameter Tuning (alpha = 0.5)	Hyperparameter Tuning (alpha = 0.9)
Precision	0.3217	0.3188	0.3198	0.3208	0.3189
Recall	0.4454	0.4393	0.4458	0.4514	0.4450
F1	0.3736	0.3695	0.3725	0.3750	0.3715
False Positive Rate	0.0806	0.0807	0.0812	0.0816	0.0818
False Negative Rate	0.0476	0.0482	0.0474	0.0468	0.0477

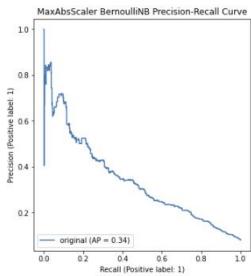


#### Log Transformation

Scaler	MaxAbsScaler	PCA Dimensionality Reduction	Hyperparam eter Tuning (alpha = 0.05)	Tuning (alpha =	Hyperparameter Tuning (alpha = 0.9)
Precision	0.3521	0.3566	0.3582	0.3553	0.3575
Recall	0.3826	0.3899	0.3866	0.3901	0.3894
F1	0.3667	0.3725	0.3718	0.3719	0.3728
False Positive Rate	0.0604	0.0605	0.0594	0.060	0.0602
False Negative Rate	0.0529	0.0524	0.0526	0.0523	0.0525

#### Log Transformation





#### Feature Scaling

		PCA	Hyperparameter	Hyperparameter	Hyperparameter
Scaler	MaxAbsScaler	Dimensionality	Tuning	Tuning	Tuning
		Reduction	(alpha = 0.05)	(alpha = 0.5)	(alpha = 0.9)
Threshold	1.6284	1.6872	1.6291	1.6274	1.6287

#### Log Transformation and Feature Scaling

		PCA	Hyperparameter	Hyperparameter	Hyperparameter
Scaler	MaxAbsScaler	Dimensionality	Tuning	Tuning	Tuning
		Reduction	(alpha = 0.05)	(alpha = 0.5)	(alpha = 0.9)
Threshold	1.9146	1.9136	1.9139	1.9146	1.9134

#### **ANOMALY DETECTION**

- Used algo called Isolation Forest on continuous data
- Wanted to see if there was any unexpected data points within the dataset
- Aims to see how well the algorithm performs with the prediction of anomalies.

#### **ML Models: Isolation Forest**

Scaler	None		
Precision	0.1459		
Recall	0.5436		
F1	0.2301		
False Positive Rate	0.2731		
False Negative Rate	0.0392		

#### Recommendations

- Health-centric data, more specific to the individual
  - HIPAA compliance will be an significant obstacle
- Identifying socio-economic status
  - o People with greater income, will have access to wider resources to better their health
- Specifying location (zip code)
  - Certain cities/states may have better health/Medicaid programs than others.

# Thank you

