Heart Disease

Matt Rosmarin, Mohammed Mahmood Al Zakwani, Nate Swan, Owen Tatlonghari, Michael Antonucci

The Problem

- Heart disease is a broad term for a range of conditions that affects how the heart functions
- Mainly involves problems with the hearts structure, rhythm or blood vessels
- Leading cause of death globally
- 18 million people die from heart disease per year, 695k in US
- Personal issue for most people
- Develops silently over times with no early symptoms
- A lot of different factors contribute to heart disease (high blood pressure, smoking, obesity)

Dataset Overview

Source: CDC 2020 Behavioral Risk Factor Surveillance System

Overview

- Total Observations: 319,795 rows
- Trimmed it down to 50,000
- Features: 17 predictor variables
- Target Variable: Heart Disease
    - Binary Classification: Yes or No

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
2	No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
3	S	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
4	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
5	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
6	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
7	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
8	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
9	No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No	Good	9	Yes	No	No
10	No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, borderline diabetes	No	Fair	5	No	Yes	No
11	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No
12	Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No
13	No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes	Very good	5	No	No	No
14	No	28.37	Yes	No	No	0	0	Yes	Male	75-79	White	Yes	Yes	Very good	8	No	No	No
15	No	28.15	No	No	No	7	0	Yes	Female	80 or older	White	No	No	Good	7	No	No	No
16	No	29.29	Yes	No	No	0	30	Yes	Female	60-64	White	No	No	Good	5	No	No	No
17	No	29.18	No	No	No	1	0	No	Female	50-54	White	No	Yes	Very good	6	No	No	No
18	No	26.26	No	No	No	5	2	No	Female	70-74	White	No	No	Very good	10	No	No	No
19	No	22.59	Yes	No	No	0	30	Yes	Male	70-74	White	No, borderline diabetes	Yes	Good	8	No	No	No
20	No	29.86	Yes	No	No	0	0	Yes	Female	75-79	Black	Yes	No	Fair	5	No	Yes	No
21	No	18.13	No	No	No	0	0	No	Male	80 or older	White	No	Yes	Excellent	8	No	No	Yes
22	No	21.16	No	No	No	0	0	No	Female	80 or older	Black	No, borderline diabetes	No	Good	8	No	No	No
23	No	28.9	No	No	No	2	5	No	Female	70-74	White	Yes	No	Very good	7	No	No	No
24	No	26.17	Yes	No	No	0	15	No	Female	45-49	White	No	Yes	Very good	6	No	No	No
25	No	25.82	Yes	No	No	0	30	No	Male	80 or older	White	Yes	Yes	Fair	8	No	No	No
26	No	25.75	No	No	No	0	0	No	Female	80 or older	White	No	Yes	Very good	6	No	No	Yes
27	No	29.18	Yes	No	No	30	30	Yes	Female	60-64	White	No	No	Poor	6	Yes	No	No
28	No	34.34	Yes	No	No	21	8	Yes	Female	65-69	White	No	Yes	Fair	9	No	No	No
29	No	31.66	Yes	No	No	5	0	No	Male	60-64	White	No	Yes	Very good	5	No	No	No
30	No	24.89	No	No	No	1	0	No	Female	55-59	White	No	Yes	Very good	7	No	No	No
31	No	36.58	No	No	No	0	0	No	Female	60-64	White	Yes	No	Good	5	No	No	Yes
32	No	25.84	Yes	No	No	5	0	No	Male	70-74	Black	No	Yes	Good	8	No	No	No
33	No	30.67	No	No	No	4	4	Yes	Female	80 or older	White	No	Yes	Fair	8	Yes	No	No
34	No	45.35	No	No	No	30	0	Yes	Male	70-74	White	Yes	No	Good	8	No	No	No
35	No	19.02	Yes	No	No	0	5	No	Female	60-64	White	No	Yes	Very good	9	No	No	No
36	No	38.97	No	No	No	0	0	Yes	Female	70-74	Black	No	No	Good	6	No	No	No

Data Cleaning & Preprocessing

Feature Engineering Steps:
Created **ComorbidityCount** by summing: Stroke, Diabetic, Asthma, KidneyDisease, SkinCancer

Created **UnhealthyDays** by adding: PhysicalHealth + MentalHealth (capped at 30 days)

Created **RiskBehavior** column: 1 if Smoking or Alcohol Drinking is "Yes"

Categorized **SleepTime** into new SleepCategory:
Very Short (<6 hrs)
Short (6–6.9 hrs)
Normal (7–8.9 hrs)
Long (9+ hrs)

Data Cleaning & Preprocessing

- Converted categorical variables with "Yes"/"No" values into binary (0 = No, 1 = Yes)

- Used pd.get_dummies() to convert multi-category columns like:

  - AgeCategory (e.g., "55-59")
  - Race
  - GenHealth
  - SleepCategory

- No missing values found — dataset was already cleaned

- Target variable defined: HeartDisease (0 = No, 1 = Yes)

-Exported cleaned dataset and split into train/test using pickle files for consistent use across all models

Exploratory Data Analysis (EDA)

- Only 8.7% of the people have heart disease → highly imbalanced

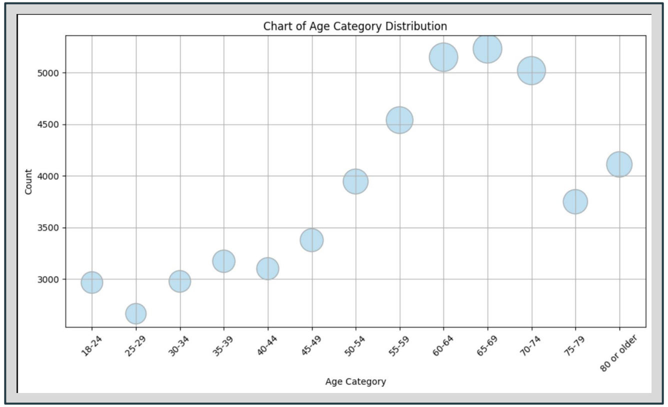
- Age is a strong predictor — older groups have higher disease rates

- Poor general health is closely linked to heart disease

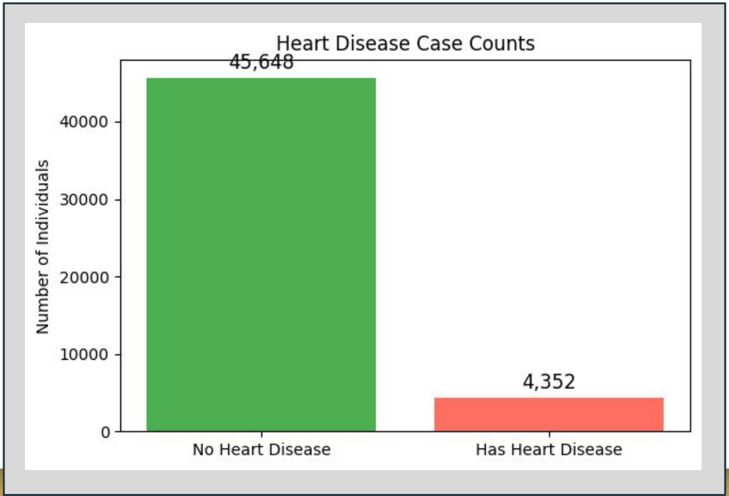
- People with multiple comorbidities are more likely to be affected

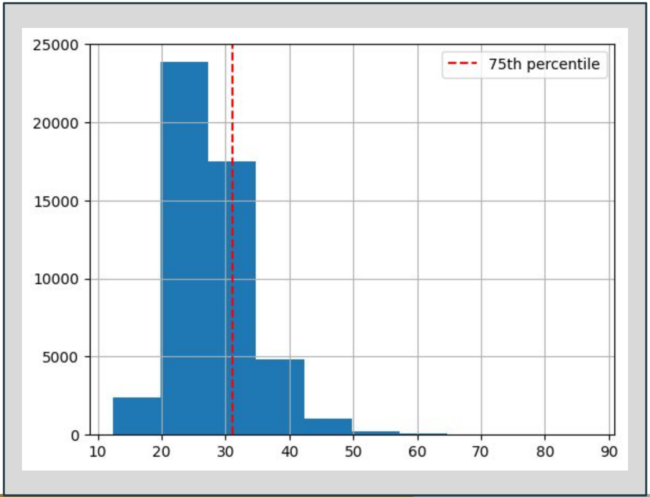
- Physical activity and sleep patterns also show relevant trends

Exploratory Data Analysis (EDA)



	BMI	PhysicalHealth	MentalHealth	SleepTime
count	50000.000000	50000.000000	50000.000000	50000.00000
mean	27.971388	3.539560	3.984260	7.12938
std	6.239799	8.094921	7.979439	1.49613
min	12.400000	0.000000	0.000000	1.00000
25%	23.710000	0.000000	0.000000	6.00000
50%	26.960000	0.000000	0.000000	7.00000
75%	31.010000	2.000000	4.000000	8.00000
max	87.050000	30.000000	30.000000	24.00000





Modeling Approach

Split data into training and testing sets using train_test_split:

→ Ensures models are tested on unseen data

Used consistent X_train/X_test/y_train/y_test across all models by loading from .pickle files

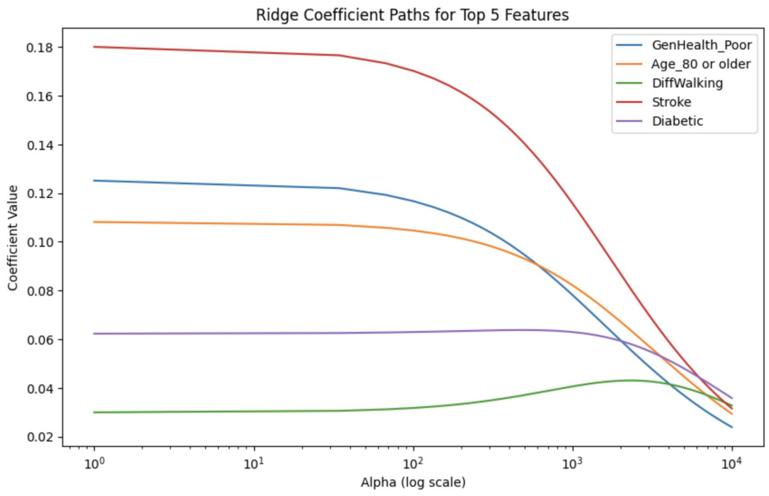
Built and tested a variety of models, each with a unique purpose

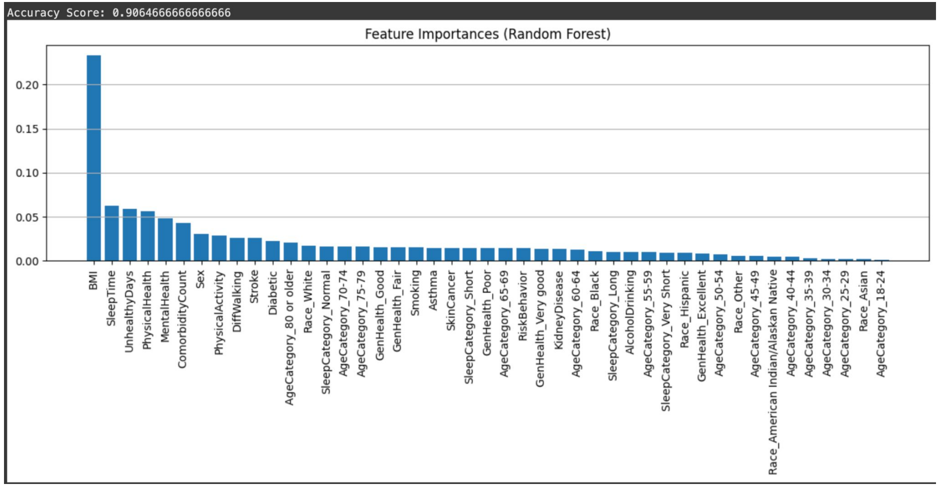
Models were evaluated using:

→ Accuracy, Precision, Recall, F1-score and more

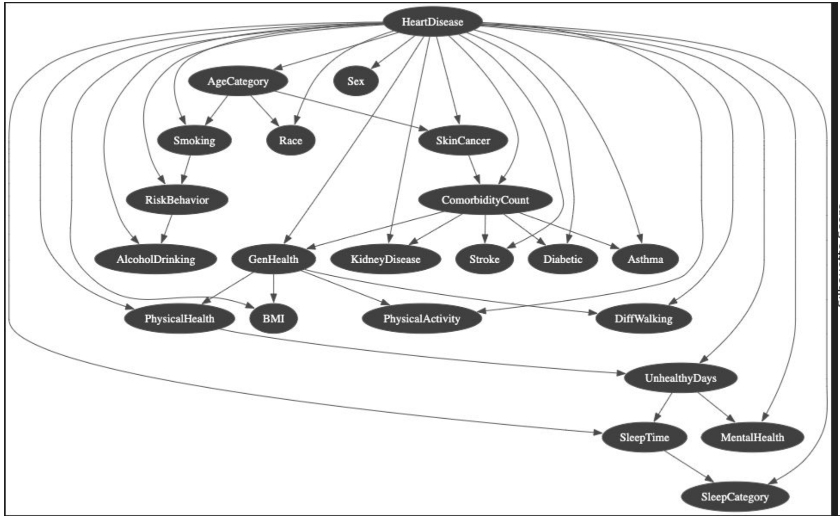
→ Confusion Matrix and more for visual insights to model predictions and more

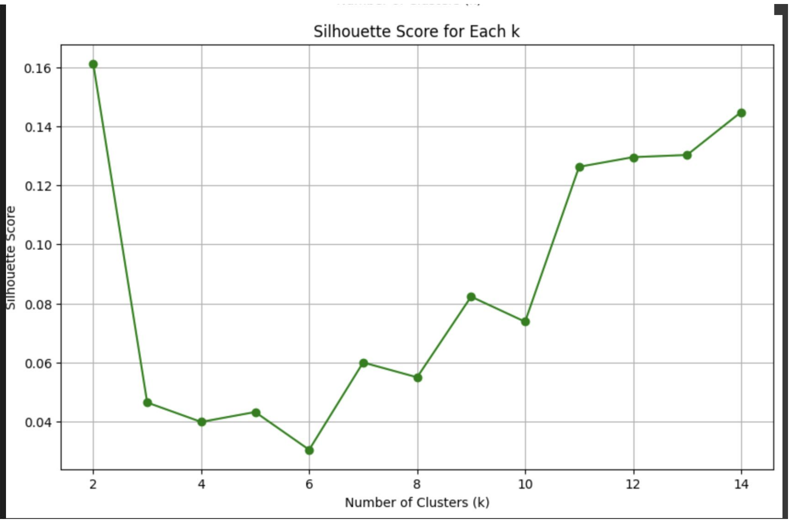
Our Models


Ridge Coefficient Paths for Top 5 Features
GenHealth_Poor
Age_80 or older
DiffWalking
Stroke
Diabetic
Coefficient Value
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
10^0
10^1
10^2
10^3
10^4
Alpha (log scale)


Accuracy Score: 0.9064666666666666
Feature Importances (Random Forest)
0.20
0.15
0.10
0.05
0.00
BMI
SleepTime
UnhealthyDays
PhysicalHealth
MentalHealth
ComorbidityCount
Sex
PhysicalActivity
DiffWalking
Stroke
Diabetic
AgeCategory_80 or older
Race_White
SleepCategory_Normal
AgeCategory_70-74
AgeCategory_75-79
GenHealth_Good
GenHealth_Fair
Smoking
Asthma
SkinCancer
SleepCategory_Short
GenHealth_Poor
AgeCategory_65-69
RiskBehavior
GenHealth_Very good
KidneyDisease
AgeCategory_60-64
Race_Black
SleepCategory_Long
AlcoholDrinking
AgeCategory_55-59
SleepCategory_Very Short
Race_Hispanic
GenHealth_Excellent
AgeCategory_50-54
Race_Other
AgeCategory_45-49
Race_American Indian/Alaskan Native
AgeCategory_40-44
AgeCategory_35-39
AgeCategory_30-34
AgeCategory_25-29
Race_Asian
AgeCategory_18-24

Our Models


HeartDisease
AgeCategory
Sex
Smoking
Race
SkinCancer
RiskBehavior
ComorbidityCount
AlcoholDrinking
GenHealth
KidneyDisease
Stroke
Diabetic
Asthma
PhysicalHealth
BMI
PhysicalActivity
DiffWalking
UnhealthyDays
SleepTime
MentalHealth
SleepCategory


Silhouette Score for Each k
Silhouette Score
0.16
0.14
0.12
0.10
0.08
0.06
0.04
2
4
6
8
10
12
14
Number of Clusters (k)

Model 1: Boosted Tree

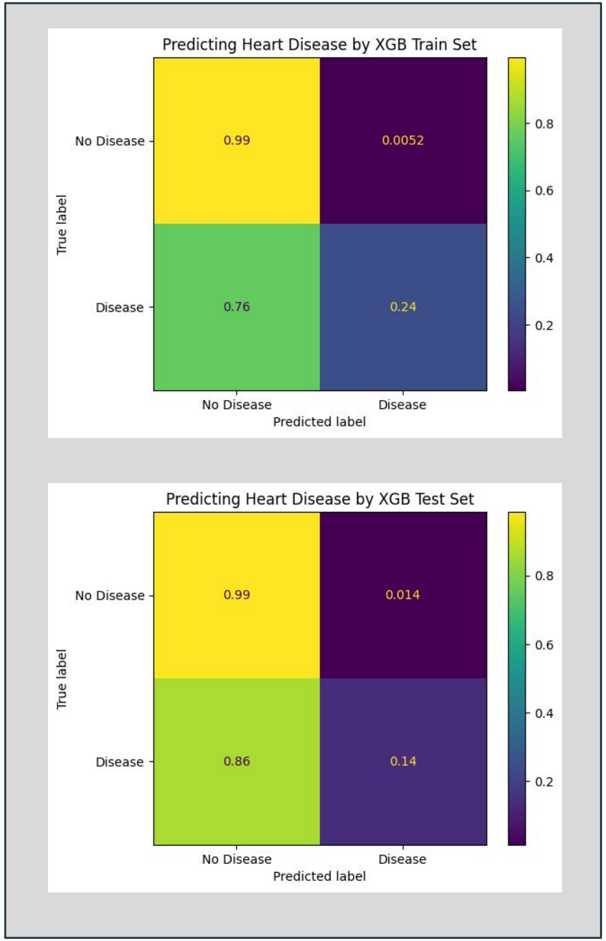
XGBoost Classifier

Base Score: (0.8543, 0.7976)

Tuning

- ➔ K-Fold Cross Validation: 5 Folds
- ➔ Parameters:
    - ◆ Max Depth
    - ◆ Number of estimators
    - ◆ Eta
    - ◆ Min Child Weight
    - ◆ Scale Position Weight
- ➔ Randomized Grid Search: 50 Iterations

Tuned Score: (0.9293 0.9119)



Model 1: Boosted Tree

Accuracy: 0.91

Precision: 0.48

Recall: 0.14

F1 Score: 0.22

ROC AUC: 0.82

The model has high **accuracy** (0.91) but low **recall** (0.14), meaning it misses most heart disease cases. **Precision** (0.48) is also low, leading to false positives. The **F1 score** (0.22) reflects poor balance between precision and recall, while the **ROC AUC** (0.82) shows good overall class differentiation. Improving recall is crucial to avoid missing heart disease cases.

Probable Cause: Dataset Class Imbalance between NoHeartDisease/HeartDisease respectively,(45,648/4,352)

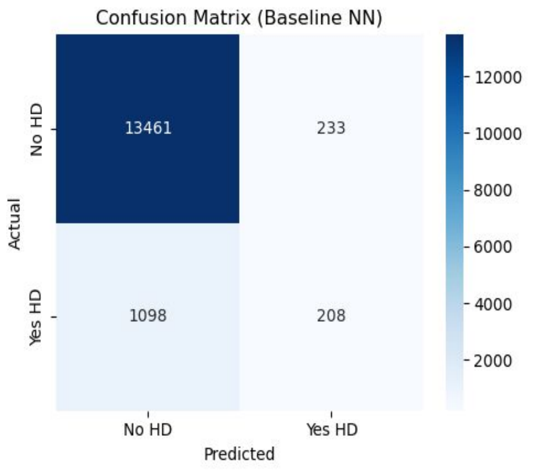
Model 2: Neural Network

- Trained for 30 epochs to predict heart disease
- Dataset was highly imbalanced (only 9% had disease)
- Without fixing this, the model would just predict "no disease" for everyone
- We used class weighting to make the model care more about disease cases
- This helped improve the model’s ability to detect real heart disease

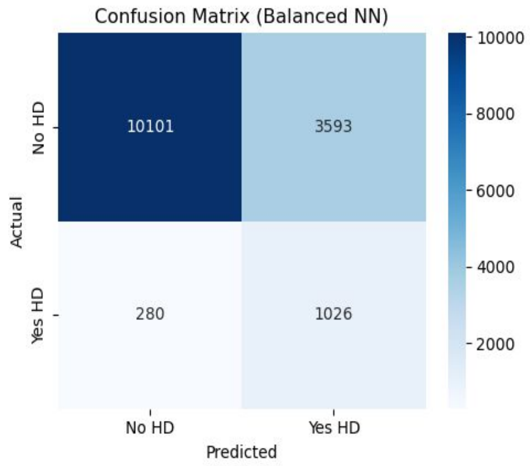
Model 2: Neural Network

Impact of Class Weighting

Metric	Before Weights	After Weights
Accuracy	0.92	0.76
Recall (1)	0.15	0.79
F1-Score (1)	0.23	0.35



Accuracy dropped, but recall and F1-score improved, meaning the model became much better at detecting actual heart disease cases.



Model Comparison

Metric	Boosted Tree	Neural Network (Weighted)
Accuracy	0.91	0.76
Precision	0.14	0.79
Recall	0.22	0.35

Both models started off predicting “No Disease” most of the time due to class imbalance.

Boosted Tree stayed that way — giving high accuracy but missing most actual cases (low recall).

Neural Network originally had similar behavior, but after adding class weights, it improved a lot in catching real disease cases.

Conclusion: Class weighting made the Neural Network better for detecting rare but important outcomes, while Boosted Tree stayed better at general prediction accuracy.

Final Reflections & Takeaways

Key Lessons Learned:

- Accuracy alone can be misleading, especially with imbalanced datasets.
- Class weighting helped the neural network focus more on rare cases like heart disease.
- If we had more time, we would refine our preprocessing and address data imbalance better.
- Tracking features like **BMI** and **SleepTime** may help individuals reduce heart disease risk.

Final Thoughts:

- Key predictors: **ComorbidityCount**, **AgeCategory**, **GenHealth**, and **BMI**.
- Boosted Trees provided high accuracy and clear feature importance.
- The Neural Network, after class weighting, improved at identifying true positive heart disease cases.
- Using different models gave us a well-rounded understanding and showed that combining approaches supports stronger real-world decisions.