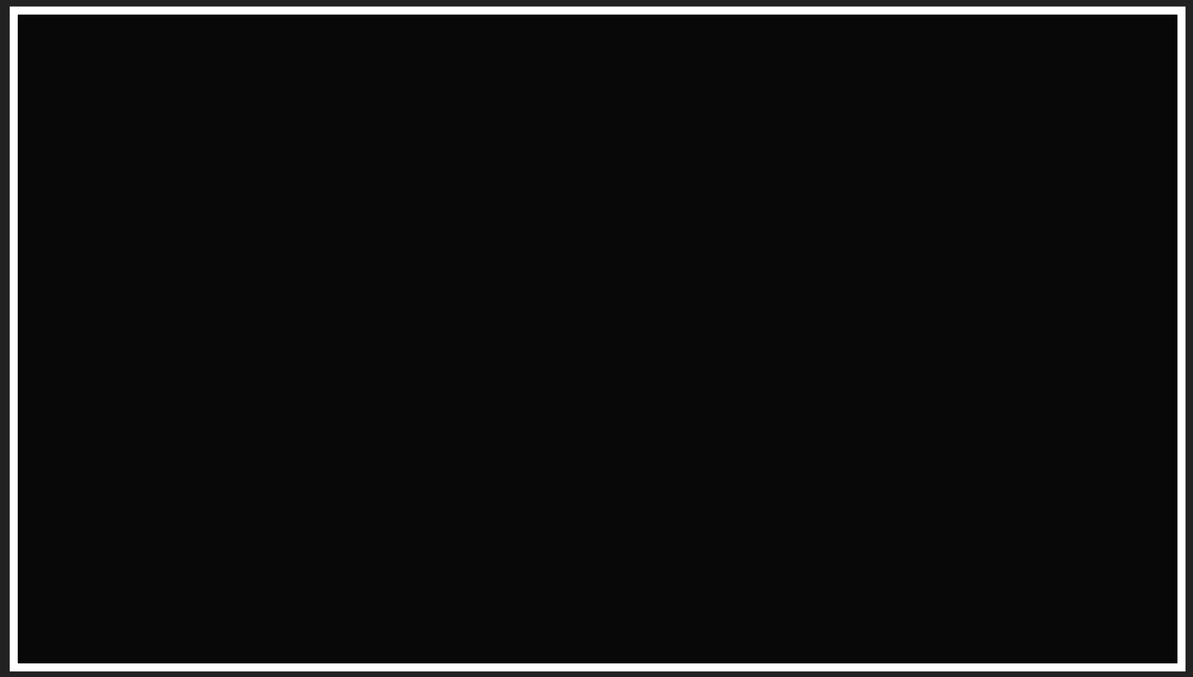


GIF by NHLBI #OurHearts

CLASSIFICATION PROJECT

PREDICTING HEART DISEASE

Sandra Paredes

INTRODUCTION

- **Motivation:** Kaiser Permanente, an HMO, wants to identify patients at high risk for heart disease who would benefit from a heart health program.[1]

- **Research Question:** How might we predict which patients are at high risk of heart disease?

- **Impact Hypothesis:** Reduce the number of patients who develop heart disease (arterial plaque or heart attack).

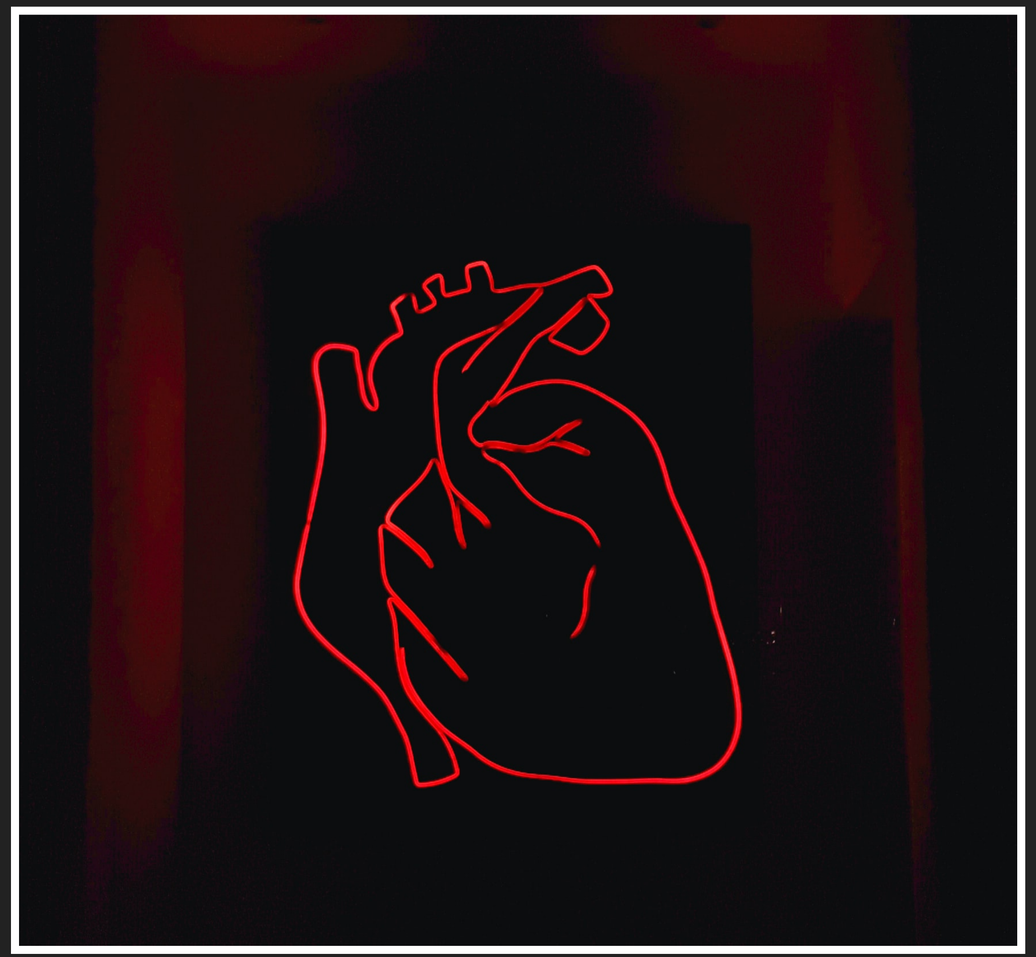


Photo by Alexandru Acea on Unsplash

METHODOLOGY

- Dataset
  - Kaggle, Indicators of Heart Disease [2]
  - Excerpted from CDC BRFSS, 2020 [3]
  - Survey of American adults
  - n= 319,795
- Target
  - Heart Disease
    - coronary heart disease
    - myocardial infarction

Photo by Giulia Bertelli on Unsplash

RESULTS

Workflow:

- Mapped values
- Recall, ROC AUC

Feature engineering:

- Question groups [4]
- Risk factors [5]

Class imbalance handling:

- Precision/Recall Curve
- Decision threshold = 0.05

Figure 1. Correlation Matrix of Select Original Features

	y_heart_disease	demg_age	behavior_tobacco	health_physical	health_mobility	disease_diabetes	disease_kidney	disease_stroke
y_heart_disease	1.000000	0.232325	0.107764	0.170721	0.201258	0.118281	0.145197	0.196835
demg_age	0.232325	1.000000	0.130384	0.110789	0.242552	0.154070	0.122697	0.137280
behavior_tobacco	0.107764	0.130384	1.000000	0.115352	0.120074	0.038993	0.034920	0.061226
health_physical	0.170721	0.110789	0.115352	1.000000	0.428373	0.111644	0.142197	0.137014
health_mobility	0.201258	0.242552	0.120074	0.428373	1.000000	0.152876	0.153064	0.174143
disease_diabetes	0.118281	0.154070	0.038993	0.111644	0.152876	1.000000	0.095186	0.072476
disease_kidney	0.145197	0.122697	0.034920	0.142197	0.153064	0.095186	1.000000	0.091167
disease_stroke	0.196835	0.137280	0.061226	0.137014	0.174143	0.072476	0.091167	1.000000

RESULTS

- Logistic Regression + GridSearchCV:
  - Recall = 0.522244
  - ROC AUC = 0.841228
- Scoring for each model:
  - X_validate_scaled
  - v_validate

Figure 2. Top 5 Best Performing Models

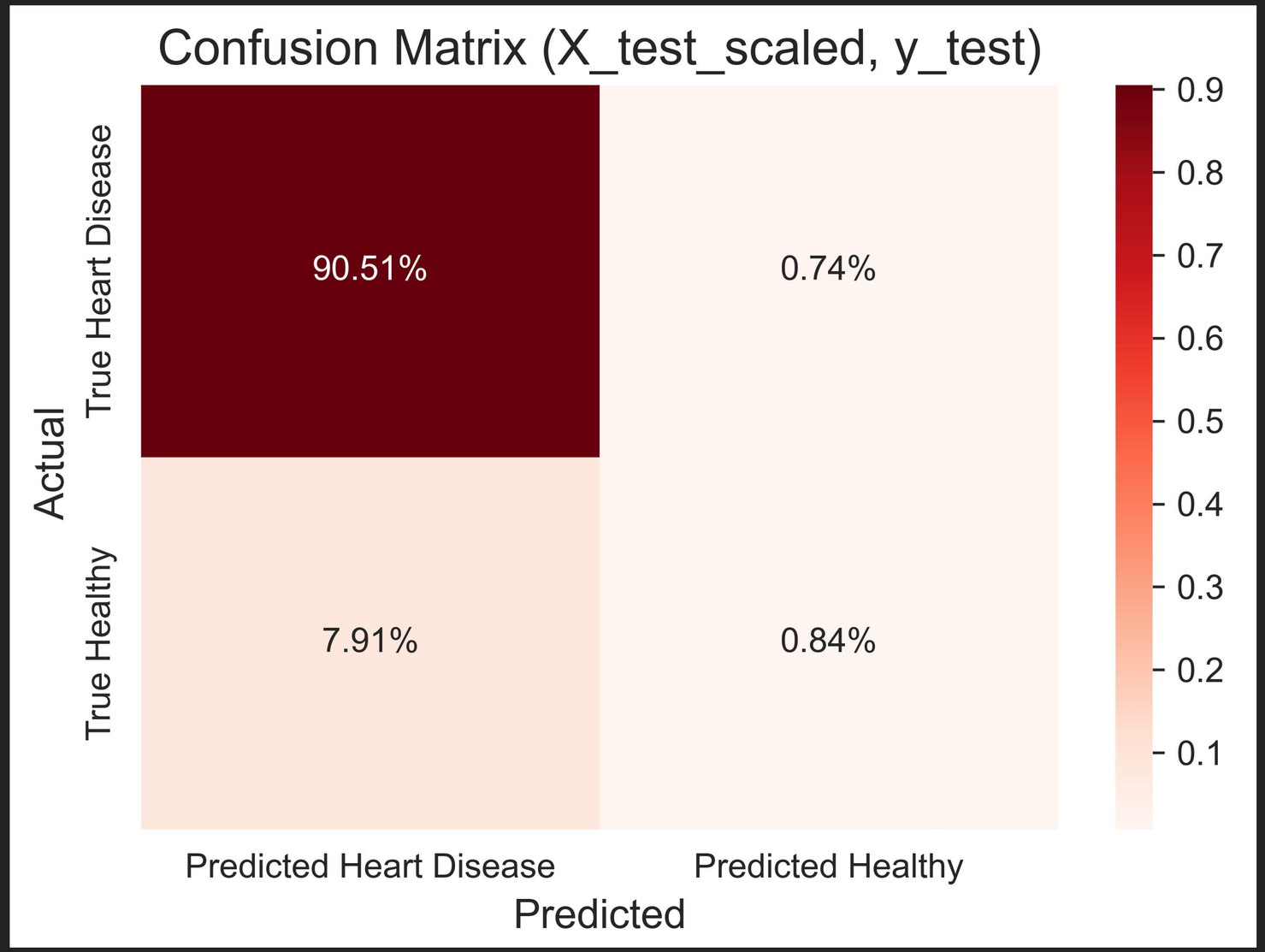
	Model	Variable	Recall	ROC AUC
16	Logistic regression GridSearchCV	log_reg_gridcv	0.522244	0.841228
2	Logistic regression	log_reg	0.519505	0.841142
14	Logistic regression question groups + risk fac...	log_reg_grp_risk	0.466844	0.799779
12	Logistic regression group features	log_reg_grp	0.389535	0.797162
7	Bernoulli NB	bern	0.365447	0.762406

RESULTS

Test Model

- Model predicts heart disease
  - 90.51% correct
- Minimal false negatives
  - 0.74% missed

Figure 4.

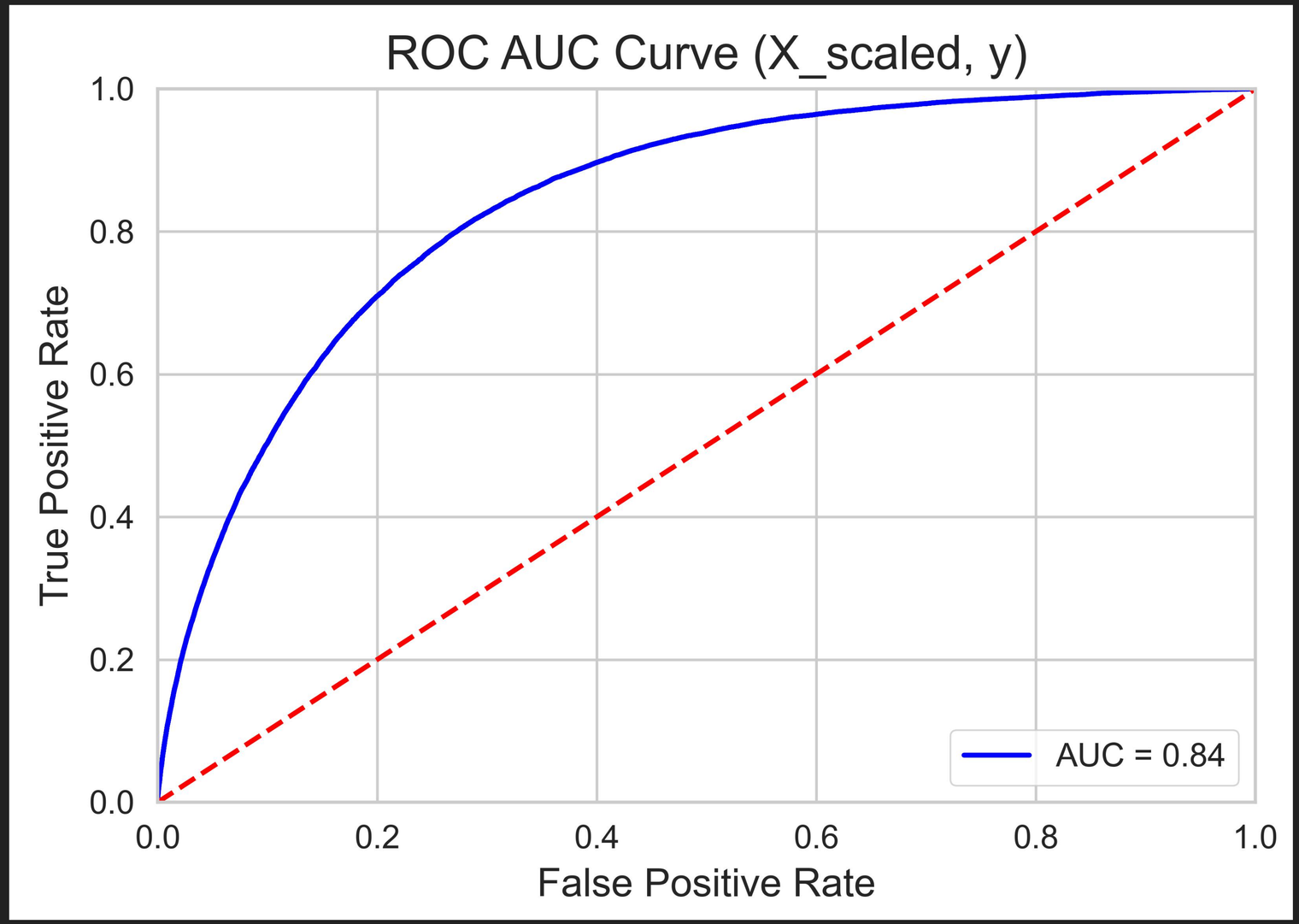


RESULTS

Model Performance

- Model performs well
  - AUC = 0.84

Figure 5.



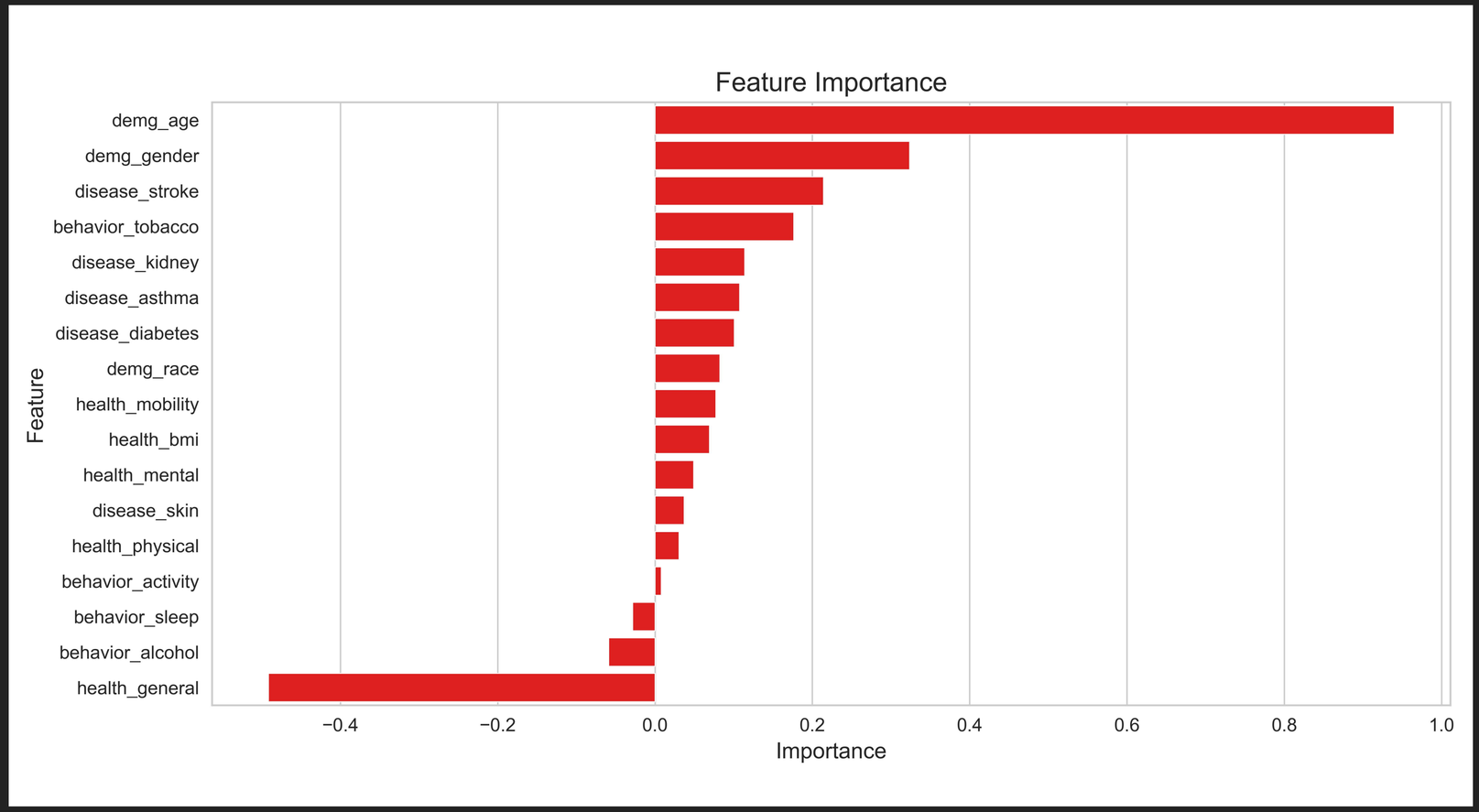
RESULTS

Feature Importance

- Age
- Gender
- Stroke
- Tobacco use
- Kidney disease

- Correlation matrix
  - Identified 4 out 5
  - Missed gender

Figure 6.



RESULTS

Predictions

- 1: 60, man, smoker, no chronic disease
  - No
  - Probability: [[0.65946415 0.34053585]]

- 2: 80, female, stroke, smoker, kidney disease
  - Yes
  - Probability: [[0.38661415 0.61338585]]

- 3: 85 year old, man, not smoker, kidney disease
  - Yes
  - Probability: [[0.41463387 0.58536613]]

Photo by Dan Senior on Unsplash

Photo by Danie Franco on Unsplash

Photo by Aziz Acharki on Unsplash

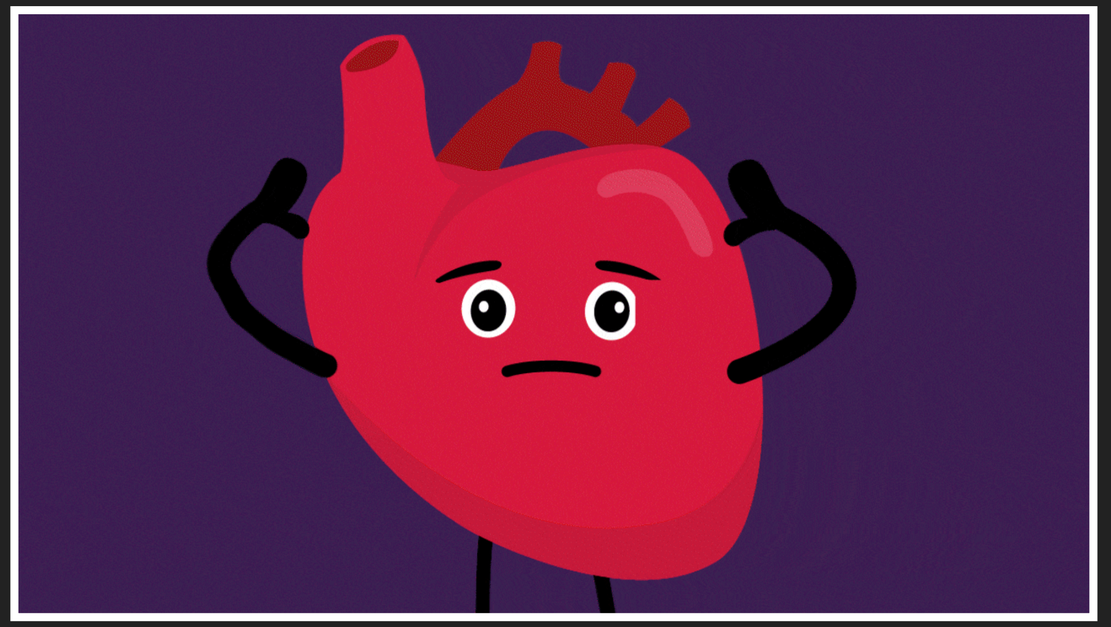
CONCLUSIONS

Insights

- Age most important to predict heart disease.
- One or more chronic diseases raises risk of heart disease.
- Tobacco use also high risk factor.

Recommendations

- Follow up with older patients for pilot program.
- Then patients with multiple chronic diseases.
- Offer smoking cessation programs.



GIF by NHLBI #OurHearts

FUTURE WORK

- Logistic regression model with top 5 features of importance
- Reattempt StackingClassifer()
- Tune model with GridSearchCV parameters and threshold or class_weights.
- Include more survey responses from BRFSS related to lifestyle and behavior.

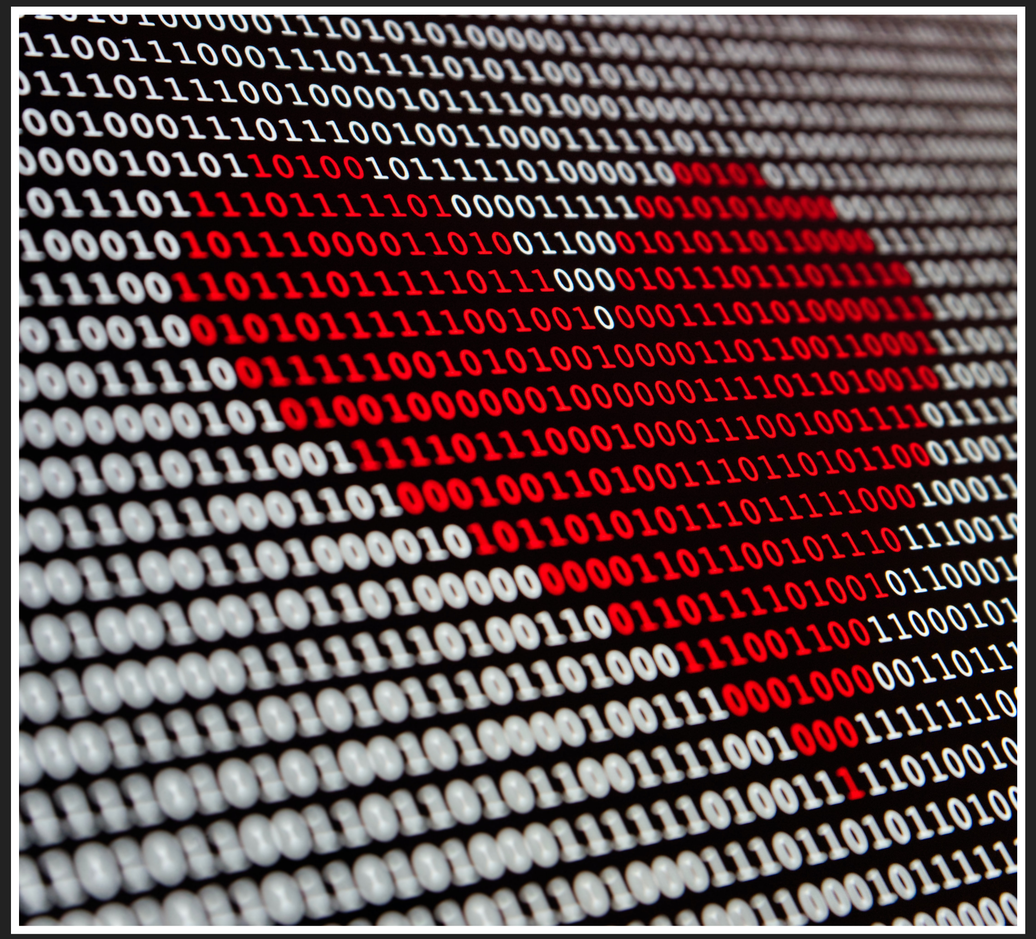


Photo by Alexander Sinn on Unsplash

APPENDIX

- Summary, data, and slides are available at github.com/slp22/classification-project

Photo by Kelly Sikkema on Unsplash

APPENDIX: SOURCES

1. National Heart, Lung, and Blood Institute, Coronary Heart Disease: https://www.nhlbi.nih.gov/health/coronary-heart-disease/causes
2. Kaggle Personal Key Indicators of Heart Disease: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
3. CDC Behavioral Risk Factor Surveillance System: https://www.cdc.gov/brfss/index.html
4. Data Dictionary: https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf
5. CDC Heart Disease Facts: https://www.cdc.gov/heartdisease/facts.htm

APPENDIX: DATA DICTIONARY

Target
- y_heart_disease: Y/N | coronary heart disease (CHD) or myocardial infarction (MI)

Health Behaviors
- behavior_activity: Num (0-30) | # days did physical activity/ exercise other than regular job
- behavior_alcohol: Y/N | heavy drinker, defined as men: 14+/wk, women: 7+/wk (includes beer, wine, malt beverage, liquor)
- behavior_sleep: Num (0-24) | # hours of sleep in a 24-hour period, on average
- behavior_tobacco: Y/N | smoked at least 100 cigarettes in your life

Demographics
- demg_age: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+
- demg_gender: male/female
- demg_race: White, Black, Asian, American Indian/Alaskan Native, Hispanic, Other race

Health Measures
- health_bmi: Num | Body Mass Index (BMI)
- health_physical: Num (0-30) | # days physical health not good, includes physical illness and injury
- health_mental: Num (0-30) | # days mental health not good, includes stress, depression, and problems with emotions
- health_general: Excellent, Very Good, Fair, Poor | Would you say that in general your health is...
- health_mobility: Y/N | serious difficulty walking or climbing stairs

Chronic Disease
- disease_asthma: Y/N
- disease_diabetes: Y/N/Y pregnancy/N borderline
- disease_kidney: Y/N | kidney disease, excludes kidney stones, bladder infection or incontinence
- disease_skin: Y/N | skin cancer
- disease_stroke: Y/N

APPENDIX: HIGH RISK FACTORS

Figure 1. Correlation Matrix of Select Original Features

	y_heart_disease	demg_age	behavior_tobacco	health_physical	health_mobility	disease_diabetes	disease_kidney	disease_stroke
y_heart_disease	1.000000	0.232325	0.107764	0.170721	0.201258	0.118281	0.145197	0.196835

Figure 6. Feature Importance