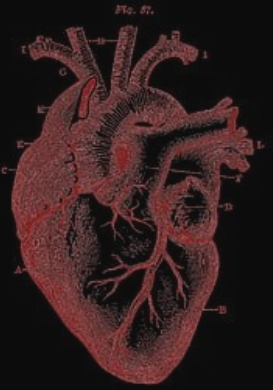
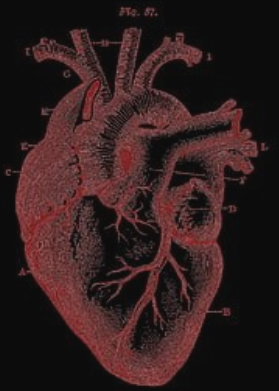
Heart Disease Prediction

Improving patient outcomes with early detection of disease using the 1988 Cleveland Clinic heart disease dataset.



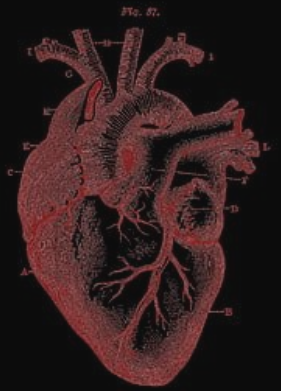
Quinn Meier

Why?



- Ischemic heart disease - leading cause of death globally (16% in 2020)
- Cardiovascular disease more broadly (31.8 % in 2017)
- Early detection and treatment improves patient outcomes

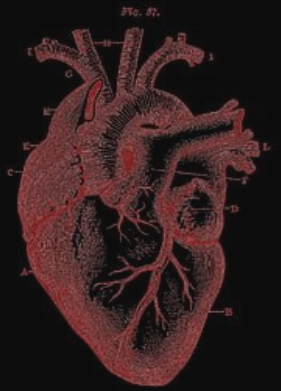
Data



- Cleveland Clinic, 1988
- Sourced from UC Irvine Machine Learning Repository
- 303 patients, 76 anonymized attributes, 14 used in analysis

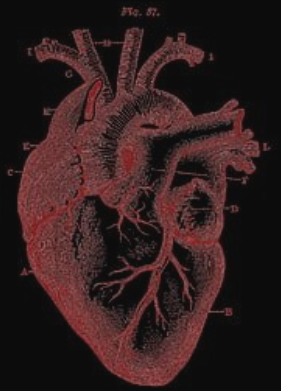
Cleveland Dataset

Method



- Binary classification
- Focus on model interpretability and performance due to medical context

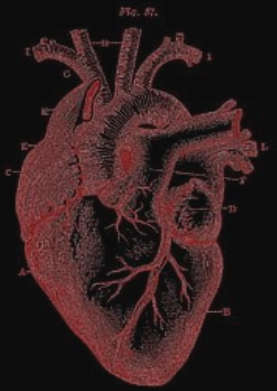
Data Cleaning

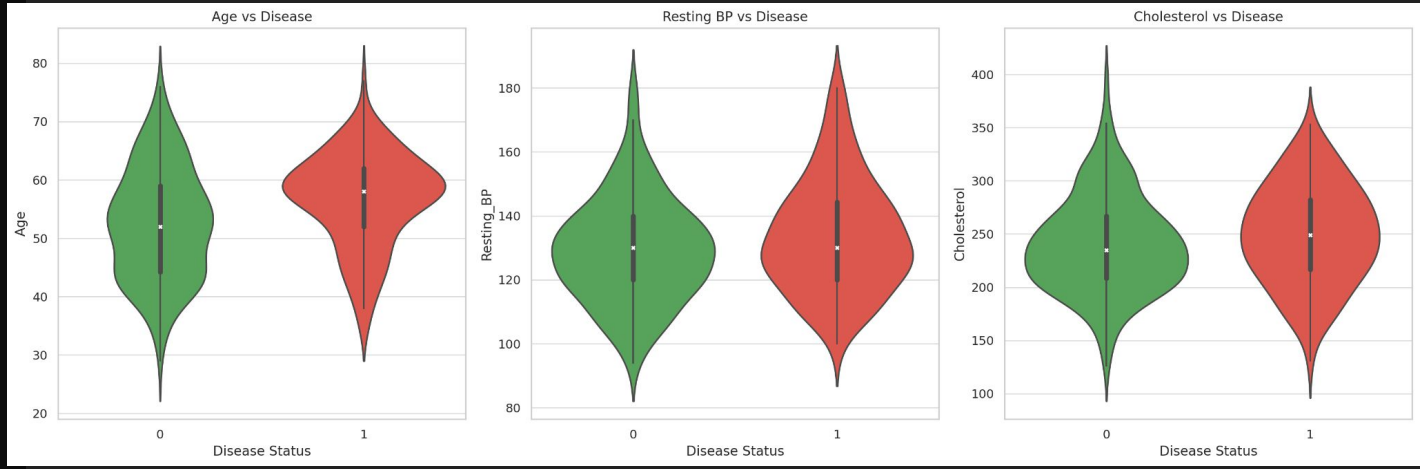


- 4 rows removed for missing categorical variables
- 9 rows removed for outliers (z-score > 3) in blood pressure and cholesterol.
- Final shape: 290 x 14

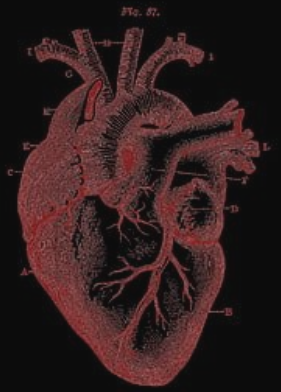
EDA

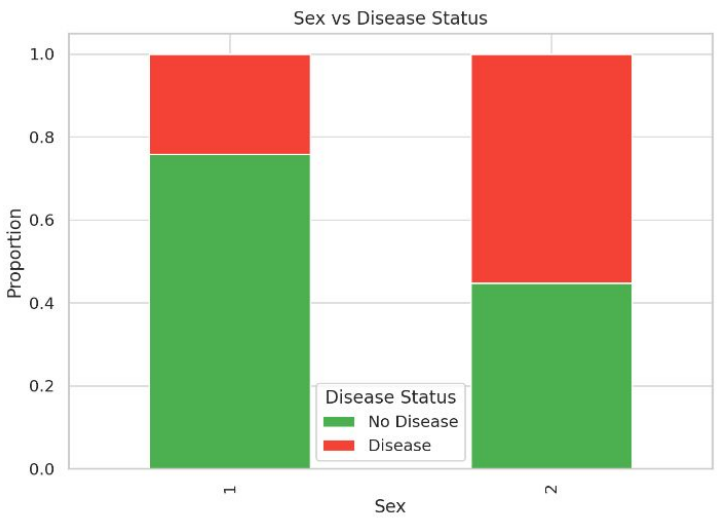
- Many expected trends observed - reversible defect, older patients, sex
- Some unexpected - blood pressure, chest pain type - asymptomatic, cholesterol

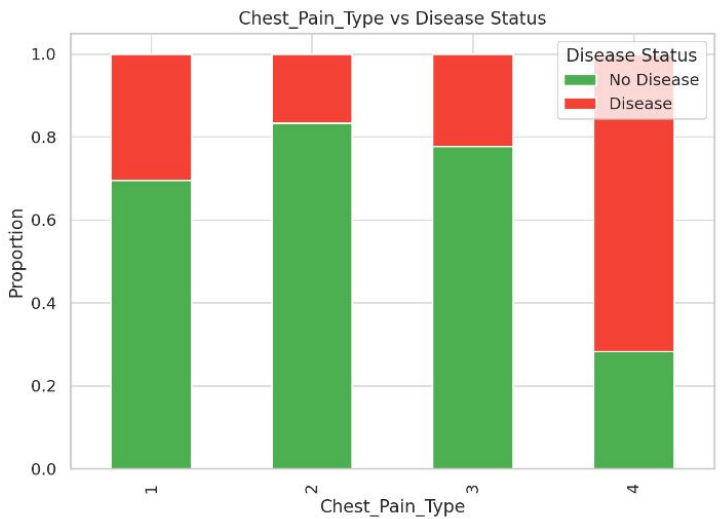


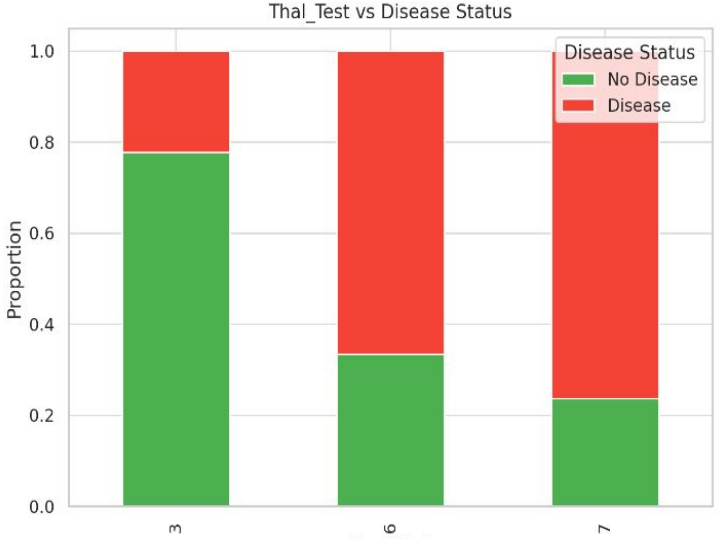


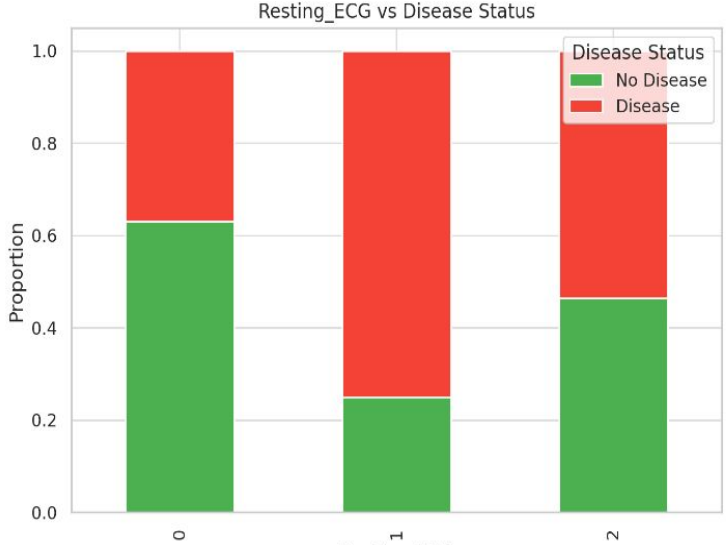
EDA



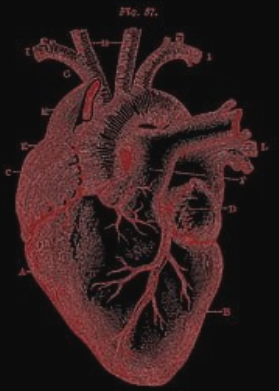

Sex vs Disease Status
Proportion
Sex
Disease Status
No Disease
Disease


Chest_Pain_Type vs Disease Status
Proportion
Chest_Pain_Type
Disease Status
No Disease
Disease


Thal_Test vs Disease Status
Proportion
Disease Status
No Disease
Disease

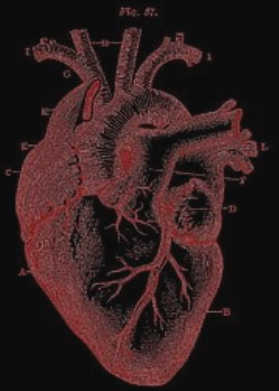

Resting_ECG vs Disease Status
Proportion
Disease Status
No Disease
Disease

Modeling



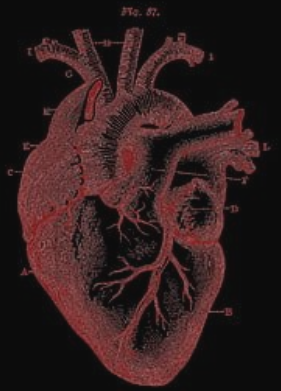
- 60/40 train-test split, stratified y
- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forest
- Gradient Boosting

Accuracy & Recall



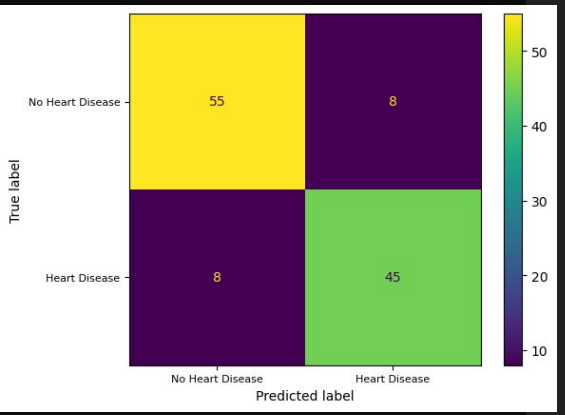
- LR: 85%, 86%
- SVM: 84%, 80%
- DT: 78%, 78%
- RF: 86%, 86%
- GB: 86%, 88%

Metrics



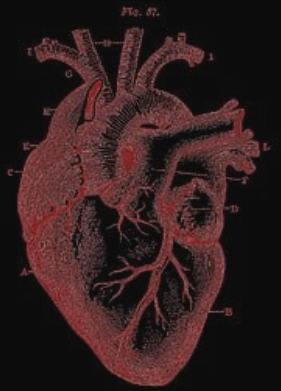
models	Logistic Regression	SVM	Decision Tree	Random Forest	Gradient Boosting
hyperparameters					
random_state	123	123	123	123	123
training sample size	60%	60%	60%	60%	60%
solver / kernel / criterion	liblinear	linear	gini	n/a	n/a
C	0.1	0.1	n/a	n/a	n/a
max_depth	n/a	n/a	3	3	2
max_features	n/a	n/a	17	17	17
min_samples_split	n/a	n/a	3	3	3
min_samples_leaf	n/a	n/a	2	1	1
learning_rate	n/a	n/a	n/a	n/a	0.1
subsample	n/a	n/a	n/a	n/a	0.8
n_estimators	n/a	n/a	n/a	2500	20
performance					
accuracy	0.85	0.84	0.78	0.86	0.86
recall	0.86	0.80	0.78	0.86	0.88
ROC	0.92	0.92	0.83	0.93	0.92

Conclusion



- Logistic regression best all-around model - interpretability, simplicity, fewest false negatives
- Hyperparameter tuning, data wrangling
- Feature importances
  - Asymptomatic chest pain
  - Nuclear stress test - coronary stenosis
  - ST depression - ECG reading possibly indicating blocked arteries

Future Work



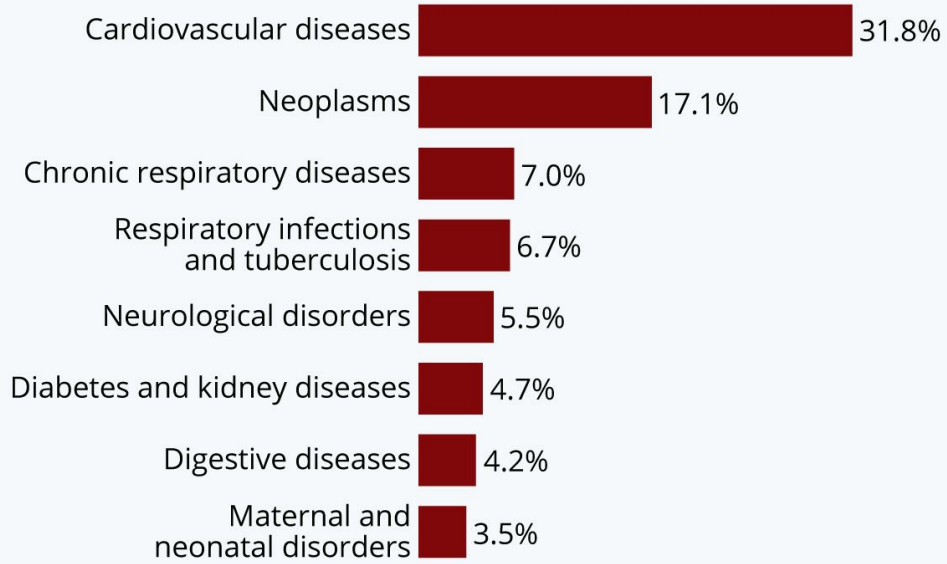
- Stratified along demographic features such as race, sex, etc.
- Combined with time-series analysis to predict cardiovascular events
- Applied to other diseases and conditions

Questions?

Backup

Top Global Causes of Death

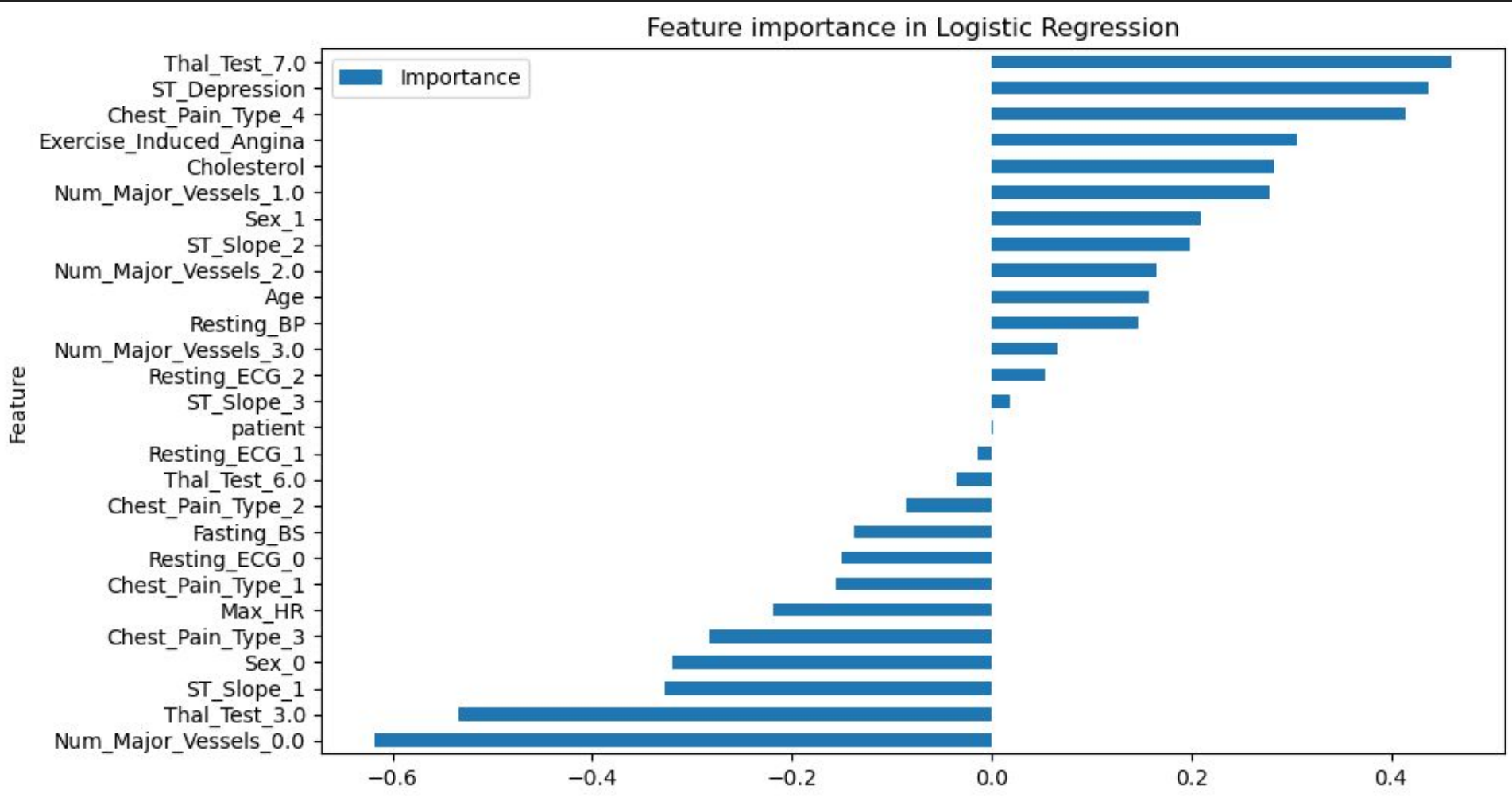
Share of all global deaths in 2017, by most common causes



Source: World Economic Forum / Institute for Health Metrics and Evaluation

features

- age: patient age in years
- sex: patient sex (1 = male, 0 = female)
- cp: Type of chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
- trestbps: patient resting blood pressure (mmHg measured at intake into hospital)
- chol: patient cholesterol level (mg/dl)
- fbs: patient fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- restecg: patient resting electrocardiograph measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: patient maximum heart rate achieved
- exang: exercise-induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
- slope: the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
- ca: number of major vessels (0-3)
- thal: thallium tracer test result, coronary stenosis detected when healthy heart cells uptake tracer is called reversible defect, while no uptake due to prior heart attack is fixed defect. (3 = normal; 6 = fixed defect; 7 = reversible defect)
- target: presence of heart disease (0 = not present, 1 = present.)


Feature importance in Logistic Regression
Importance
Feature
Thal_Test_7.0
ST_Depression
Chest_Pain_Type_4
Exercise_Induced_Angina
Cholesterol
Num_Major_Vessels_1.0
Sex_1
ST_Slope_2
Num_Major_Vessels_2.0
Age
Resting_BP
Num_Major_Vessels_3.0
Resting_ECG_2
ST_Slope_3
patient
Resting_ECG_1
Thal_Test_6.0
Chest_Pain_Type_2
Fasting_BS
Resting_ECG_0
Chest_Pain_Type_1
Max_HR
Chest_Pain_Type_3
Sex_0
ST_Slope_1
Thal_Test_3.0
Num_Major_Vessels_0.0
−0.6
−0.4
−0.2
0.0
0.2
0.4