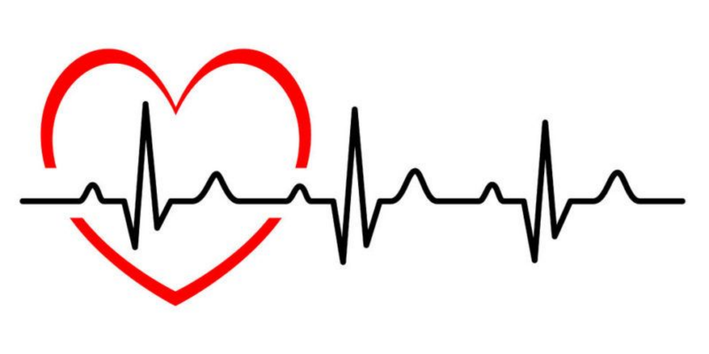
Heart Disease

Codi Steinborn



The Problem

According to the CDC about 610,000 people die of heart disease in the United States every year.

That’s approx. 1 in 4 deaths.

The Solution

Early detection and knowing what signs to look for can help with prevention and reduce the risks of heart related deaths.

Data

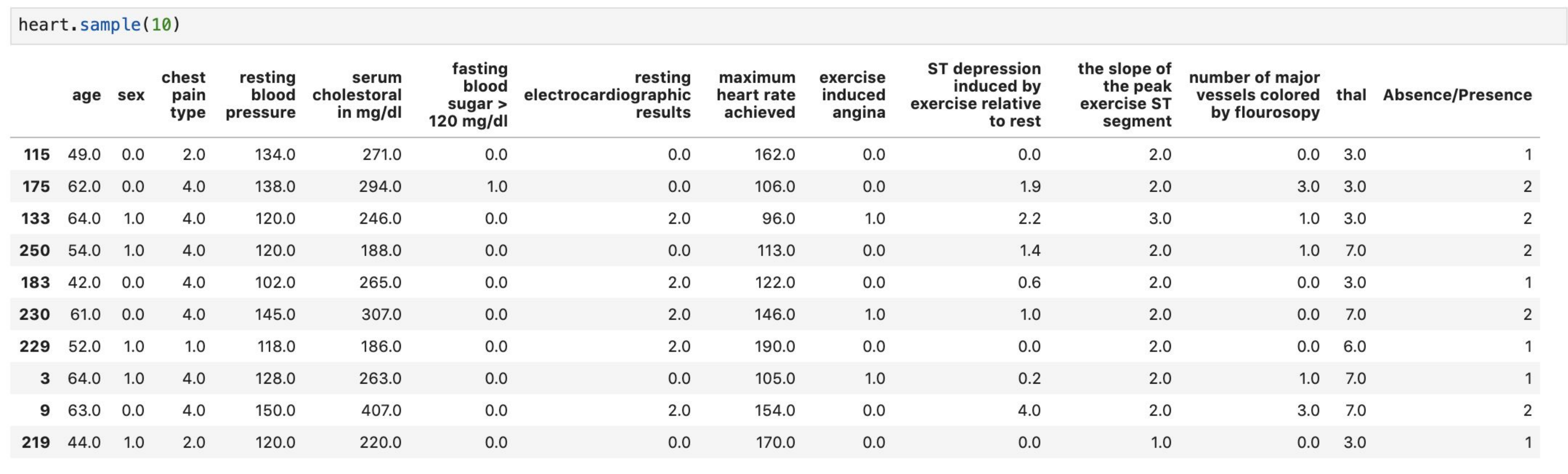
This study will investigate a data set published by the UCI Machine Learning Repository about Heart Disease

Target Variable: Disease_Presence

Feature Variables:

1. Age
2. Sex
3. Chest Pain Type (categorized by values 1-4)
4. Resting Blood Pressure
5. Serum Cholesterol in mg/dl
6. Fasting blood sugar > 120 mg/dl
7. Resting electrocardiographic results
8. Maximum Heart Rate achieved
9. Exercise Induced Angina
10. Oldpeak (ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
12. Number of major vessels colored by flourosopy
13. Thal (categorized by values 3, 6, or 7)

Data cont.



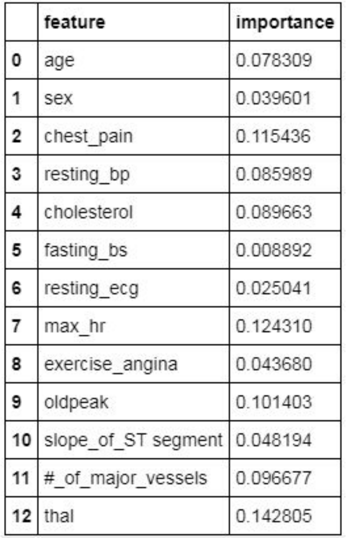
```
heart.sample(10)
```

	age	sex	chest pain type	resting blood pressure	serum cholestoral in mg/dl	fasting blood sugar > 120 mg/dl	resting electrocardiographic results	maximum heart rate achieved	exercise induced angina	ST depression induced by exercise relative to rest	the slope of the peak exercise ST segment	number of major vessels colored by flourosopy	thal	Absence/Presence
115	49.0	0.0	2.0	134.0	271.0	0.0	0.0	162.0	0.0	0.0	2.0	0.0	3.0	1
175	62.0	0.0	4.0	138.0	294.0	1.0	0.0	106.0	0.0	1.9	2.0	3.0	3.0	2
133	64.0	1.0	4.0	120.0	246.0	0.0	2.0	96.0	1.0	2.2	3.0	1.0	3.0	2
250	54.0	1.0	4.0	120.0	188.0	0.0	0.0	113.0	0.0	1.4	2.0	1.0	7.0	2
183	42.0	0.0	4.0	102.0	265.0	0.0	2.0	122.0	0.0	0.6	2.0	0.0	3.0	1
230	61.0	0.0	4.0	145.0	307.0	0.0	2.0	146.0	1.0	1.0	2.0	0.0	7.0	2
229	52.0	1.0	1.0	118.0	186.0	0.0	2.0	190.0	0.0	0.0	2.0	0.0	6.0	1
3	64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
9	63.0	0.0	4.0	150.0	407.0	0.0	2.0	154.0	0.0	4.0	2.0	3.0	7.0	2
219	44.0	1.0	2.0	120.0	220.0	0.0	0.0	170.0	0.0	0.0	1.0	0.0	3.0	1

Model

1. Found a Random Forest Classifier model yielded best results
    a. 98%-100% on Training set
    b. 83%-88% on Test set

2. Most important features
    a. Maximum heart rate achieved
    b. Thal
    c. Chest pain type



	feature	importance
0	age	0.078309
1	sex	0.039601
2	chest_pain	0.115436
3	resting_bp	0.085989
4	cholesterol	0.089663
5	fasting_bs	0.008892
6	resting_ecg	0.025041
7	max_hr	0.124310
8	exercise_angina	0.043680
9	oldpeak	0.101403
10	slope_of_ST segment	0.048194
11	#_of_major_vessels	0.096677
12	thal	0.142805

Evaluation

- Null Model accuracy: ~55%
- My Model’s accuracy: ~85.5%
- Low Bias, Higher Variance - Overfitting

Future Improvements

1. Collect data on more feature variables
    a. family history
    b. smoking habits
2. Collect data from more subjects
3. Regularization

Sources

- https://www.cdc.gov/heartdisease/facts.htm
- http://archive.ics.uci.edu/ml/datasets/statlog+(heart)