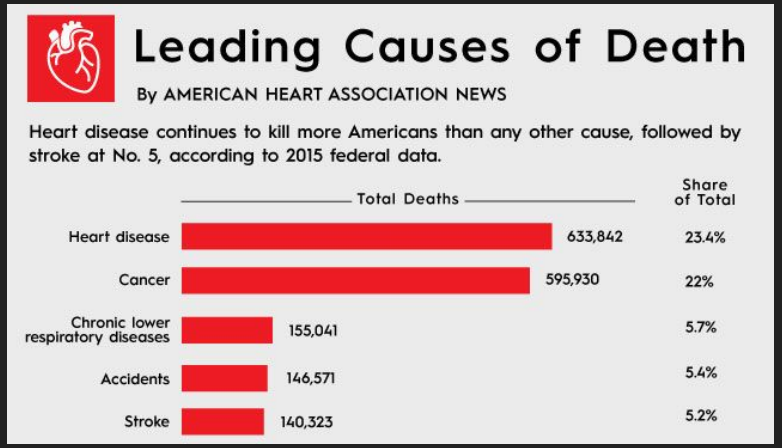
Heart Failure

Michael Dyer

The Problem



Source: Centers for Disease Control and Prevention

- Cardiovascular Disease is the #1 cause of death Globally.
- Heart Disease (a part of CVD) is the leading cause of death in the US.
- Heart Failure is a major form of Heart Disease.

The Solution

- Use a machine learning model to predict if a patient will develop Heart Disease

Why does this matter?

- 31% of worldwide deaths are attributed to CVD's
- Heart Disease costs $108,000,000,000 per year
- Finding a way to pre-treat/manage could save employers money

The Data

About this file

The data contains 918 observations with 12 attributes.

# Age 	
age
	∆ Sex
	sex
		∆ ChestPainType
		chest pain type
			# RestingBP
			resting blood pressure
				# Cholesterol
				serum cholesterol
					# FastingBS
					fasting blood sugar
						∆ RestingECG
						resting electrocardiogram results
							# MaxHR
							maximum heart rate achieved
								✓ ExerciseAngina
								exercise induced angina
									# Oldpeak
									oldpeak = ST
										∆ ST_Slope
										the slope of the peak exercise ST segment
											# HeartDisease
											target

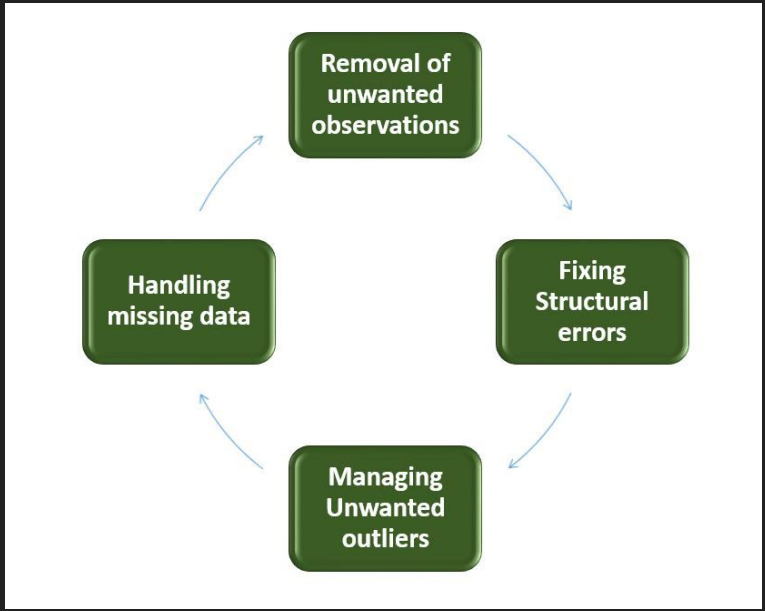
28 – 77	M 79% F 21%	ASY 54% NAP 22% Other (219) 24%	0 – 200	0 – 603	0 – 1	Normal 60% LVH 20% Other (178) 19%	60 – 202	true 371 40% false 547 60%	-2.6 – 6.2	Flat 50% Up 43% Other (63) 7%	0 – 1
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
43	F	TA	100	223	0	Normal	142	N	0	Up	0

Taken from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Data Wrangling

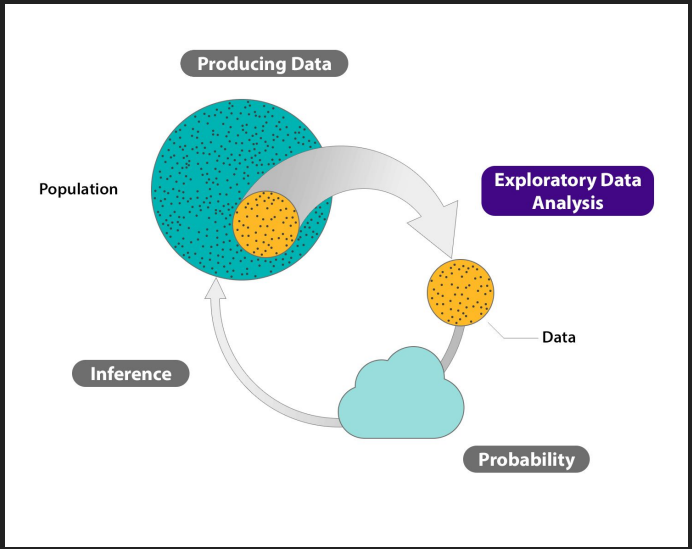
- Original dataset included 918 rows and 12 columns.
  - 7 of the 11 categories were categorical
  - 5 of the 11 categories were numerical
- Converted to Dataframe
- All work done on Jupyter Notebook

Data Cleaning

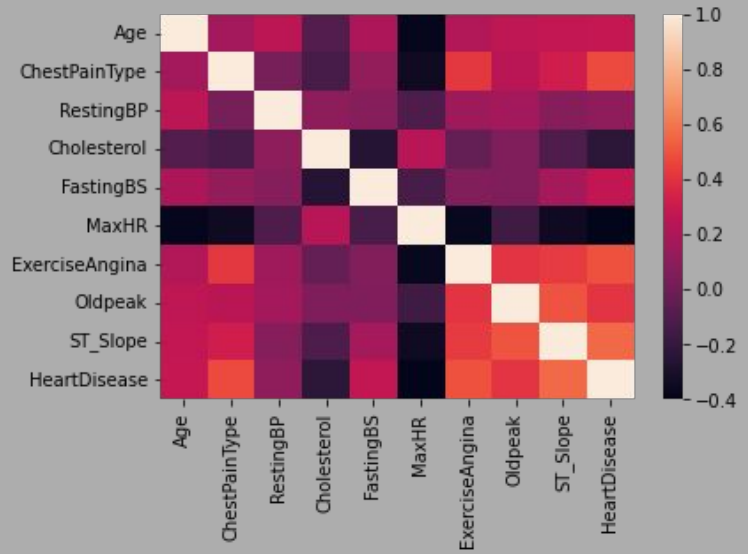


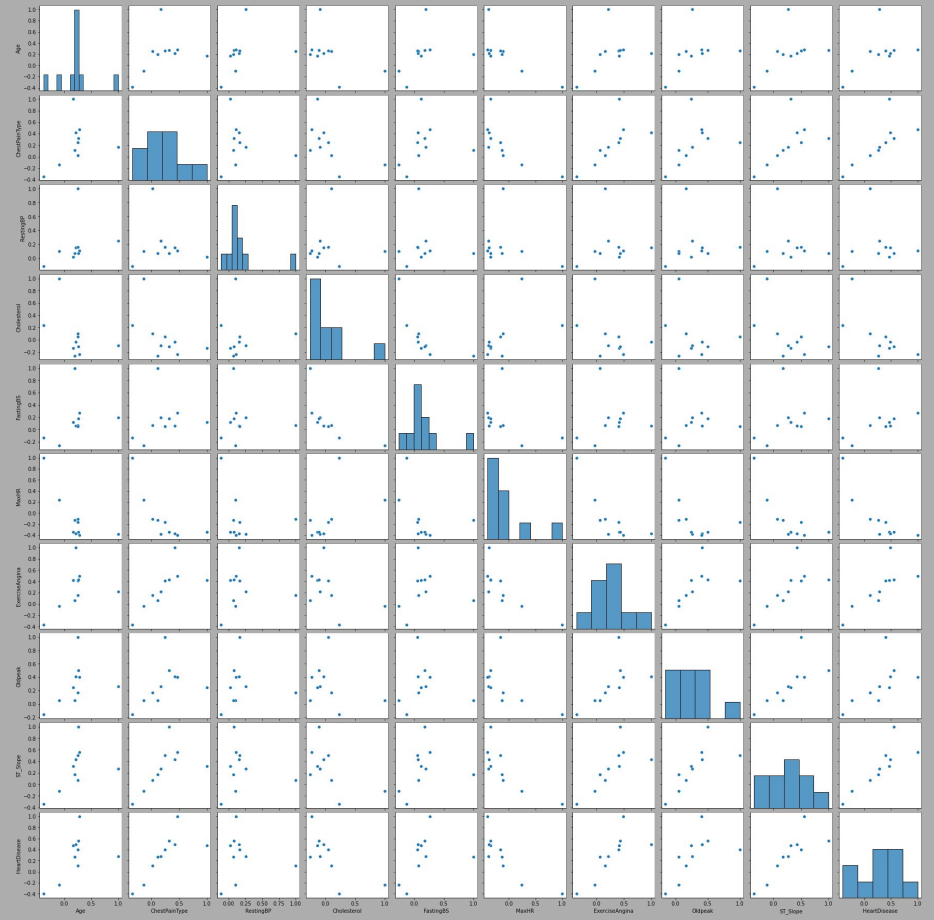
- No data was ‘NaN’
  - 172 Cholesterol values of ‘0’
  - Imputed missing Cholesterol values with Mean
- Performed Outlier Analysis
  - Not many outliers
- Re-formatted columns from strings to integers (‘N’ to ‘0’, ‘Y’ to ‘1’)
- Overall this was a very clean dataset

Exploratory Data Analysis

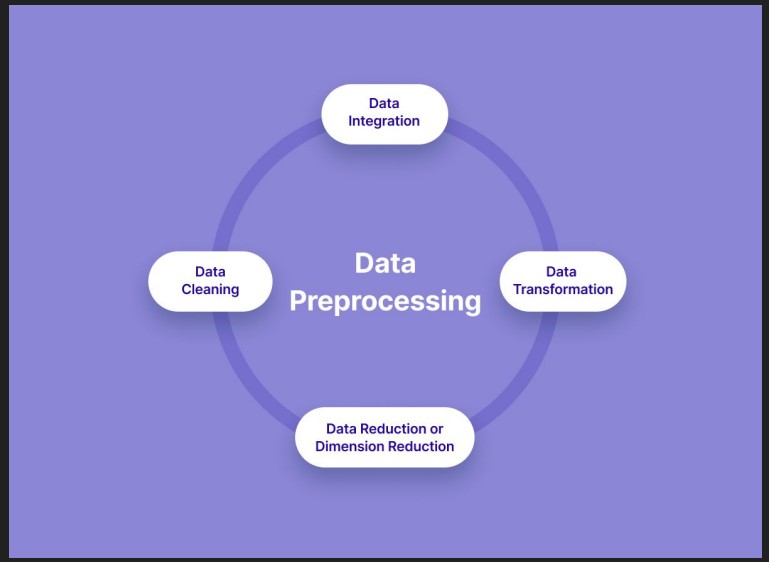


- Data is distributed normally
- Found strongest Correlation of features to Heart Disease
  - ST_Slope, ChestPainType, Exercise Angina
- Chi-Square Test for 2-categorical features
- T-tests for categorical vs. numerical features
- Features showed varying levels of correlation with Heart Disease


Age
ChestPainType
RestingBP
Cholesterol
FastingBS
MaxHR
ExerciseAngina
Oldpeak
ST_Slope
HeartDisease
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Pre-Processing



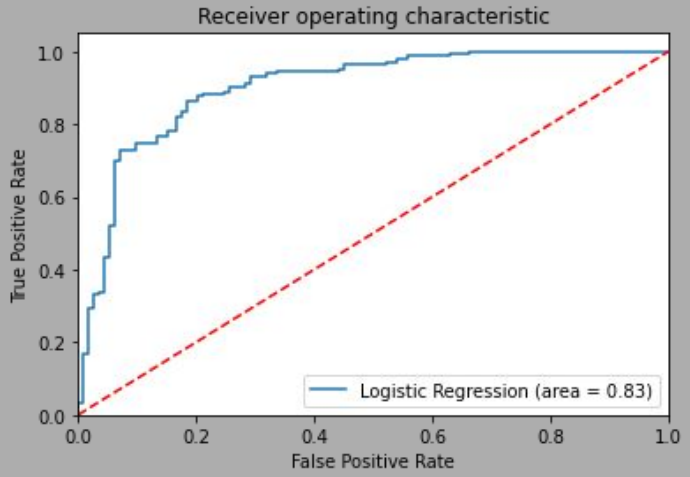
- Scaled Numerical features of dataset
- Created Test/Train splits for model evaluation
- No dimension reduction needed

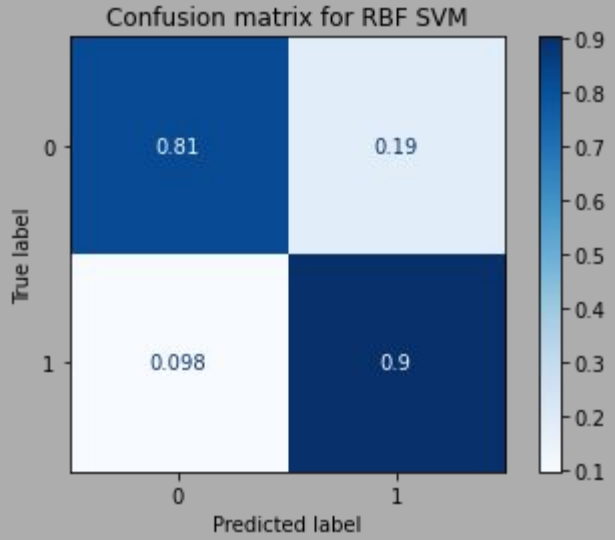
	Age	ChestPainType	RestingBP	Cholesterol	MaxHR	Oldpeak	FastingBS	ExerciseAngina	ST_Slope	HeartDisease	Sex_F	Sex_M	RestingECG_LVH
0	-1.433140	-1.345086	0.410909	0.825070	1.382928	-0.832432	0	0	0	0	0	1	0
1	-0.478484	-0.270422	1.491752	-0.171961	0.754157	0.105664	0	0	1	1	1	0	0
2	-1.751359	-1.345086	-0.129513	0.770188	-1.525138	-0.832432	0	0	0	0	0	1	0
3	-0.584556	0.804242	0.302825	0.139040	-1.132156	0.574711	0	1	1	1	1	0	0
4	0.051881	-0.270422	0.951331	-0.034755	-0.581981	-0.832432	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
913	-0.902775	-2.419749	-1.210356	0.596393	-0.188999	0.293283	0	0	1	1	0	1	0
914	1.536902	0.804242	0.627078	-0.053049	0.164684	2.357094	1	0	1	1	0	1	0
915	0.370100	0.804242	-0.129513	-0.620168	-0.857069	0.293283	0	1	1	1	0	1	0
916	0.370100	-1.345086	-0.129513	0.340275	1.461525	-0.832432	0	0	1	1	1	0	1
917	-1.645286	-0.270422	0.302825	-0.217696	1.422226	-0.832432	0	0	0	0	0	1	0

Scaled Data

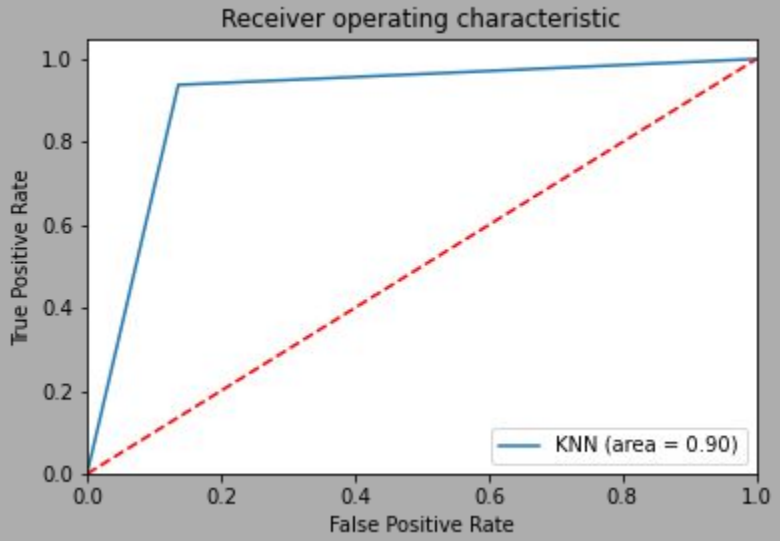
Modeling

- Created several models to find best fit
  - Logistic Regression, SVM, Trees, Random Forest, KNN
- Most performed decently out of the box
  - Logistic Regression had 84% accuracy and 87% recall w/no tuning
- Hyperparameter Tuned KNN model had best performance metrics.
  - 90% accuracy, 94% recall
- Optimized for Recall (False Negatives) since it is Medical problem


Receiver operating characteristic
True Positive Rate
False Positive Rate
Logistic Regression (area = 0.83)
0.0
0.2
0.4
0.6
0.8
1.0


Confusion matrix for RBF SVM
True label
Predicted label
0.81
0.19
0.098
0.9
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Tuning Model



- Hyperparameter Tuned KNN Model
  - n_neighbors=17
  - leaf_size=1
  - p=1

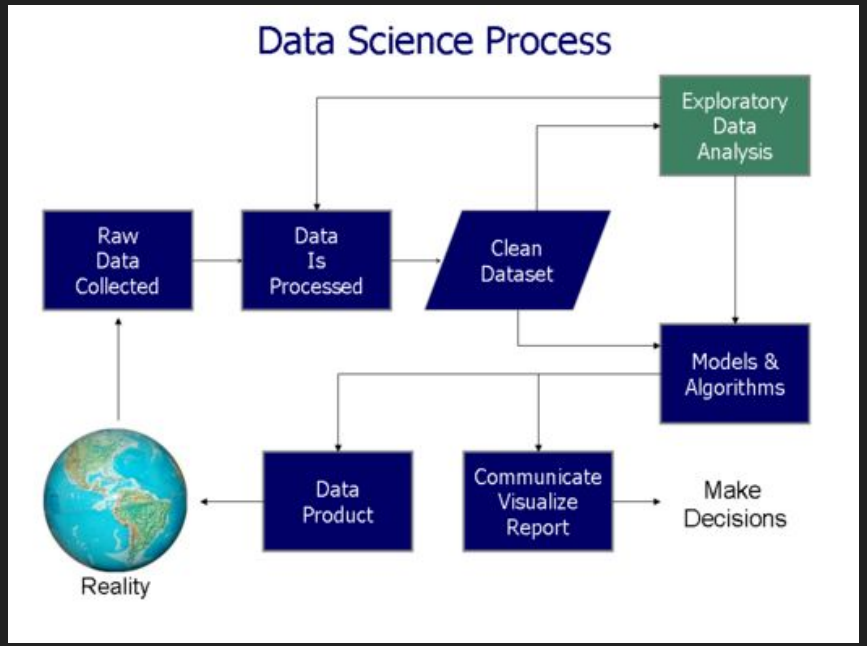
Takeaways

Performance Metrics

- Accuracy: 0.9021739130434783
- Balanced accuracy: 0.9005681818181819
- Precision score: 0.8823529411764706
- Recall score: 0.9375

- Use Tuned KNN Model
- Accuracy - 90%
- Recall - 94%
- Less false negatives = more early detection

Future Research



- Feature Engineering
- Data Collection
- Re-Tuning Model