Predicting Heart Disease

Prem Panchal

Data Science Capstone Project

The Problem

- Heart disease major issue
- Possible ways to detect it faster?
- How would a person know they need to get checked up?

Steakholders

- Hospitals
- Medical Clinics
- Company owners where employees are at risk for heart related disease

Dataset

Combination of 5 heart data sets, combined in 2021

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations

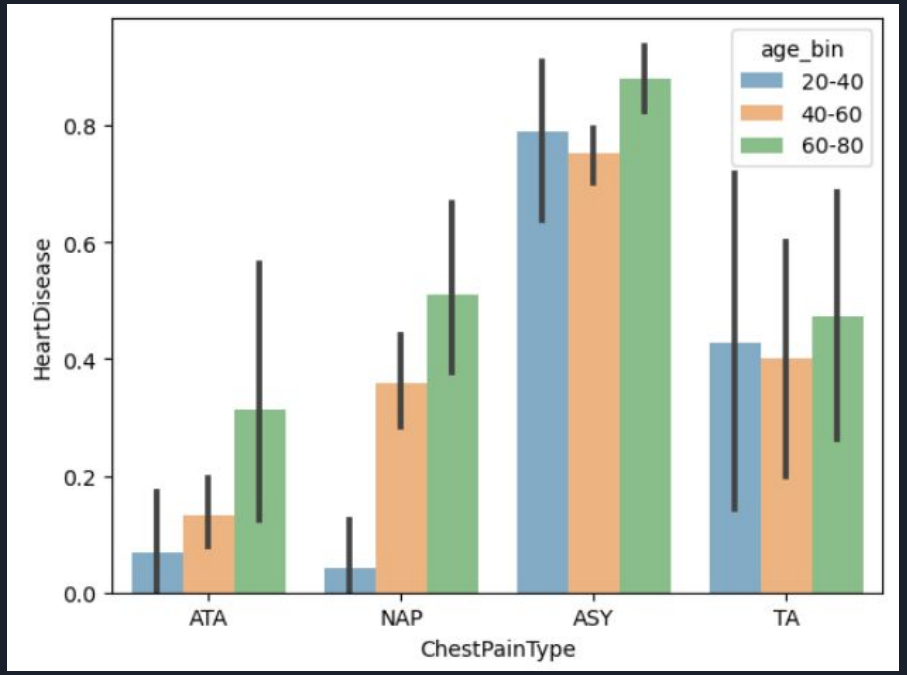
Final dataset: 918 observations

Possible Symptoms

- Age
- Heart Pain Types
- Cholesterol
- Hereditary causes

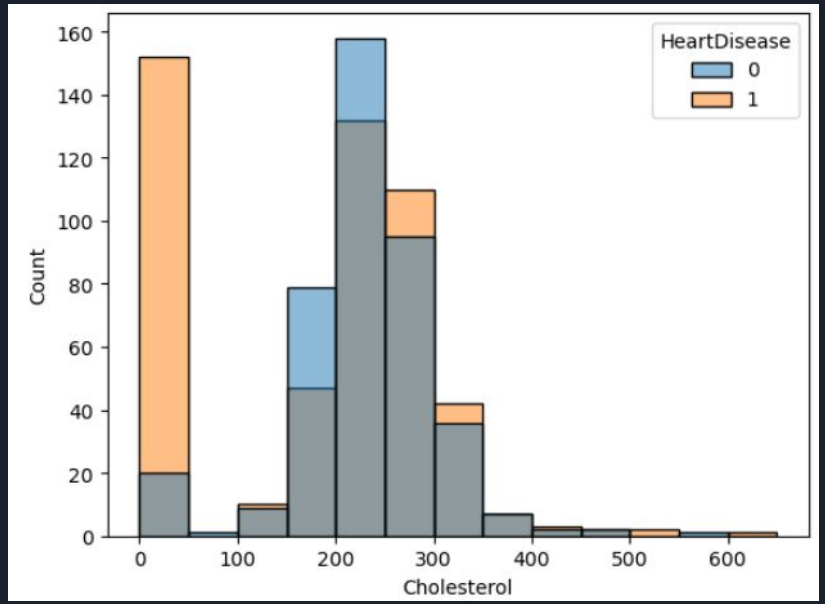
Chest Pain Types

- ATA - Atypical Angina
- NAP - Non-Anginal Pain
- ASY - Asymptomatic
- TA - Typical Angina



Cholesterol

- No Direct Correlation
- Many cases with heart disease actually have a cholesterol of 0
- High cholesterol can be caused by both hereditary issues or dietary issues
- Added new feature into Data set to accommodate hereditary cases

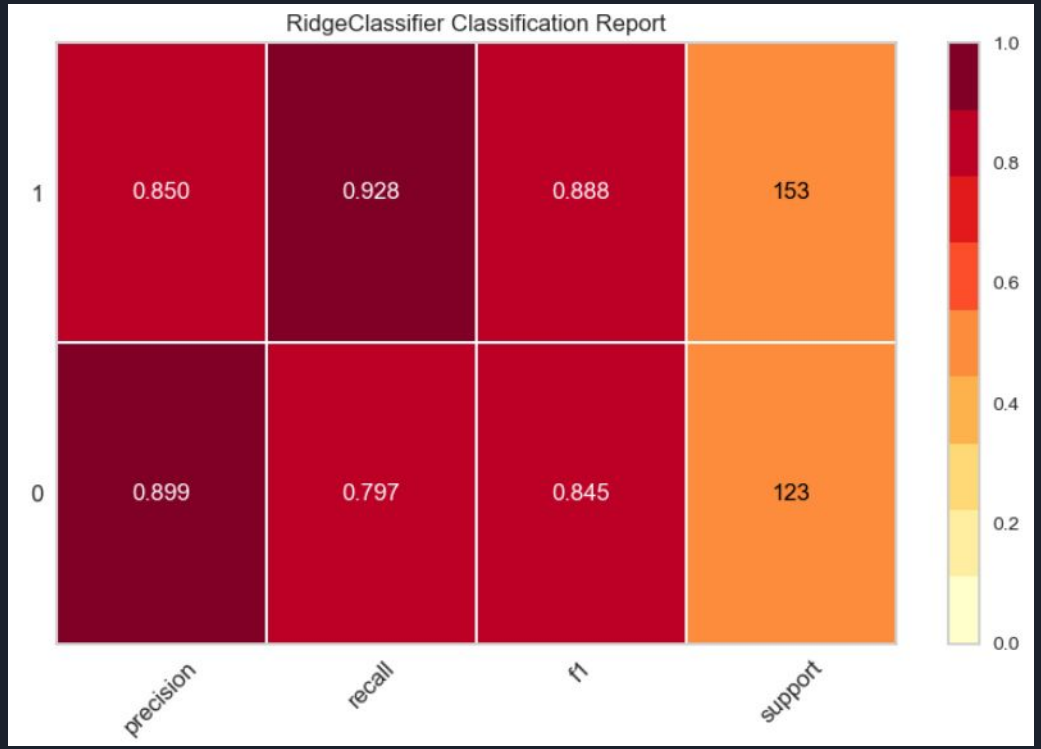


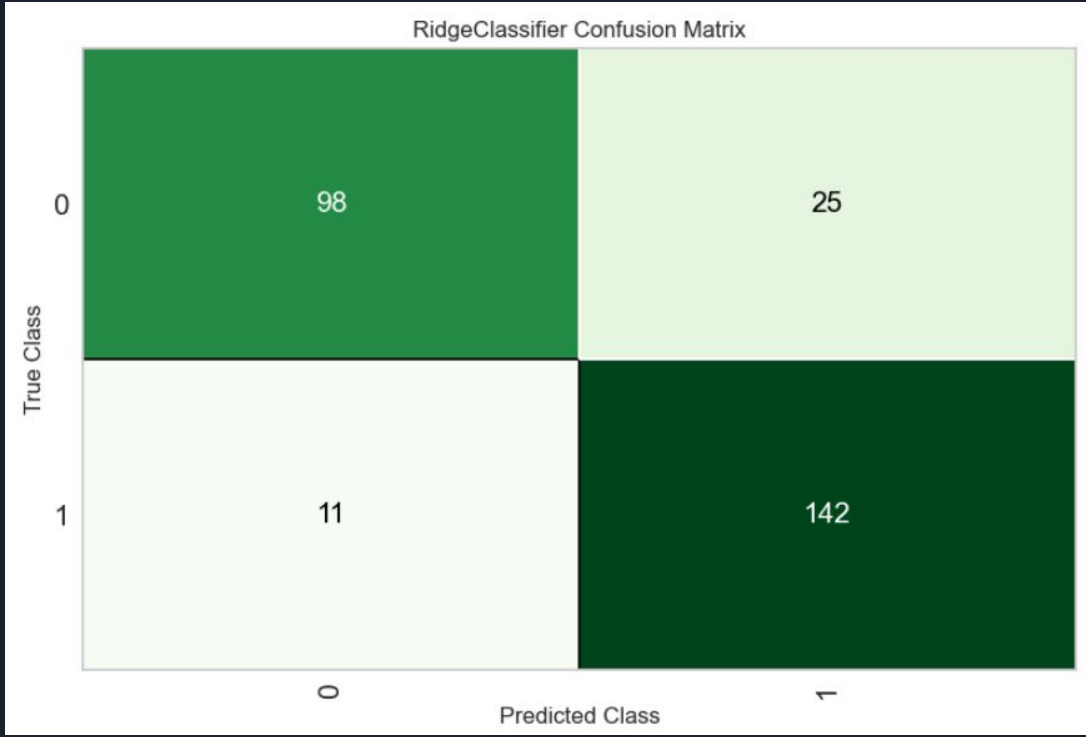
Model Comparisons

- Supervised learning
- Binary Classification
- Fairly even classification distribution

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.8723	0.0000	0.8929	0.8794	0.8851	0.7412	0.7437	0.0630
lr	Logistic Regression	0.8707	0.9245	0.8958	0.8754	0.8842	0.7377	0.7412	0.4350
lda	Linear Discriminant Analysis	0.8692	0.9264	0.8929	0.8745	0.8826	0.7347	0.7375	0.0740
catboost	CatBoost Classifier	0.8691	0.9254	0.8985	0.8724	0.8839	0.7338	0.7377	0.1330
rf	Random Forest Classifier	0.8644	0.9264	0.8985	0.8650	0.8798	0.7243	0.7289	0.0880
gbc	Gradient Boosting Classifier	0.8597	0.9239	0.8901	0.8643	0.8754	0.7148	0.7197	0.0800
et	Extra Trees Classifier	0.8582	0.9150	0.8926	0.8583	0.8739	0.7117	0.7160	0.0950
ada	Ada Boost Classifier	0.8535	0.9096	0.8587	0.8762	0.8660	0.7041	0.7072	0.0850
lightgbm	Light Gradient Boosting Machine	0.8363	0.9110	0.8645	0.8461	0.8540	0.6677	0.6707	0.1590
nb	Naive Bayes	0.8210	0.9162	0.7492	0.9128	0.8197	0.6459	0.6611	0.0710
dt	Decision Tree Classifier	0.7851	0.7827	0.8027	0.8096	0.8040	0.5656	0.5703	0.0700
knn	K Neighbors Classifier	0.6776	0.7259	0.7182	0.7112	0.7130	0.3439	0.3453	0.2210
svm	SVM - Linear Kernel	0.5902	0.0000	0.6883	0.7167	0.6040	0.1525	0.1913	0.0650
dummy	Dummy Classifier	0.5530	0.5000	1.0000	0.5530	0.7121	0.0000	0.0000	0.0680
qda	Quadratic Discriminant Analysis	0.4470	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0700

Model Performance




RidgeClassifier Confusion Matrix
0
1
True Class
98
25
11
142
0
1
Predicted Class

Model Improvements

- Obtain more data
  - Possibly through synthetic medical data
- Look into other features that could make the data more accurate

Conclusion

- Ridge Classifier most accurate model
- While highly accurate only serves as a baseline
  - Ideally available to both patients and doctors to determine more extensive testing
- Adding more data to the model iteratively can help doctors more quickly determine if testing needs to be done

Thank you!

Prem Panchal
Email: prem0217@gmail.com
Project: https://github.com/prem-0217/CapstoneTwo