Prem Panchal

Final Report

Heart Disease Detection

Problem Statement

Heart Disease is a major problem all over the world and is often discovered too late. There are many possible symptoms that can overlap with other diseases and it can easily be overlooked if there is not some past occurrence to have it checked immediately. To this end, is there a simpler way to detect the disease early on so that proper treatment can be done as soon as possible.

With my dataset of just under 1000 patient data, I have created a tool that can take into account a variety of attributes amongst patients and determine if a patient is likely to have heart disease at an average precision of about .875. This model can be used by both doctors and patients alike to determine if closer examination of the heart needs to take place.

Data Wrangling

The raw dataset was actually incredibly clean already. The dataset is patient data with very few instances of missing data, as extensive testing is generally done to conclude if a patient has heart disease. With no null values in the data, the next step was to check for any anomalies with the data; this was most notable as the minimum values for resting blood pressure and cholesterol were 0. As these values seemed odd at the time, I decided to keep them in the dataset for future insights, and ended with a dataset shape of 918 by 12.

Exploratory Data Analysis

One of the first key things I looked into during EDA was the different forms of chest pain types to understand how they could affect the model, and an estimate of how many cases of heart disease are found in each type.

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

As shown in figure 1 below, it was common for asymptomatic patients to actually be found with heart disease, and that there is a lot of data confirming that chest pain is not a reasonable way to confirm if heart disease is present in the patient.

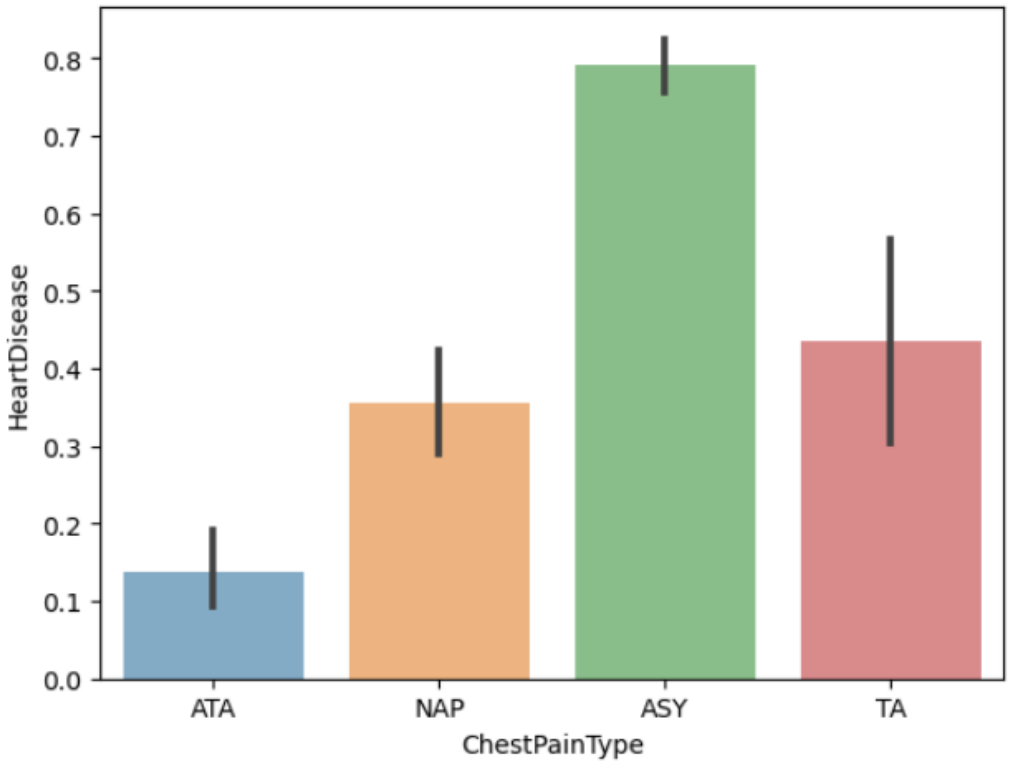

Figure 1

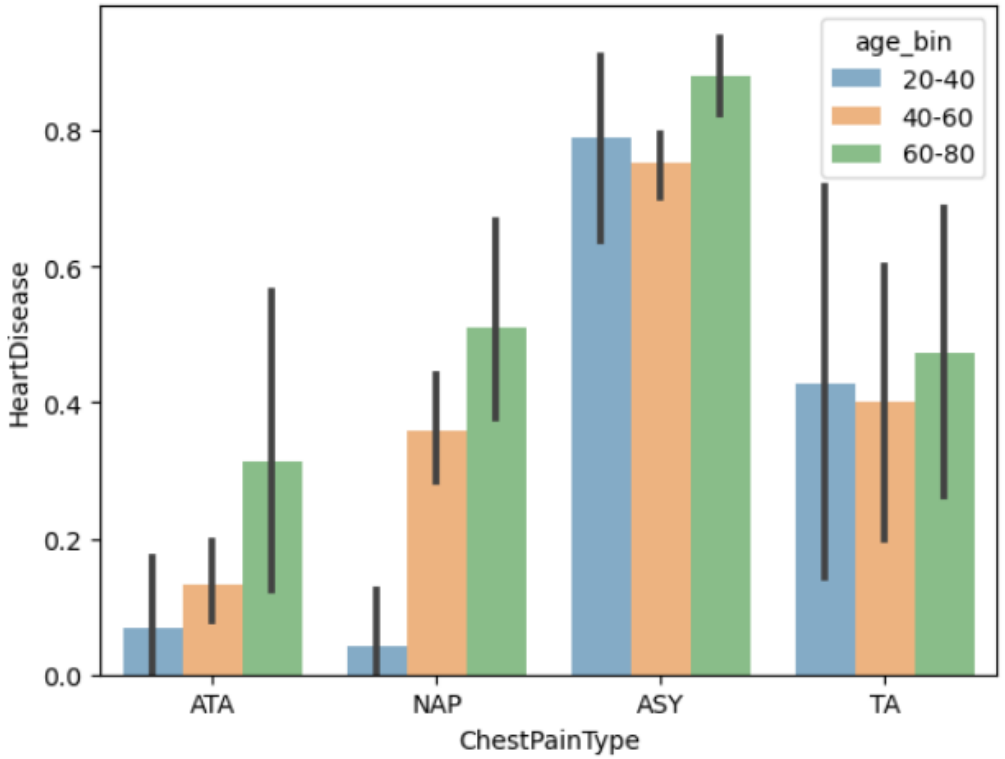

Figure 2

After comparing only chest pain types, I began to look towards age and seeing how that could affect the distribution. The data only had patient data between the ages of 20-80, so after diving the ages into groups of 20 years, I created figure 2; figure 2 above shows the expected general trend that older age patients are more likely to have heart disease.

The last major feature with a lot of anomalies that I could see easily was cholesterol. With very little medical background myself, I had gone in with the mindset that there would be some noticeable trend between high cholesterol and heart disease. After creating figure 3 below, there were 2 major issues I had found with my assumptions; first being that there seems to be no direct correlation between cholesterol and heart disease, and second being that more than half the instances of patients with heart disease actually are marked with 0 cholesterol.

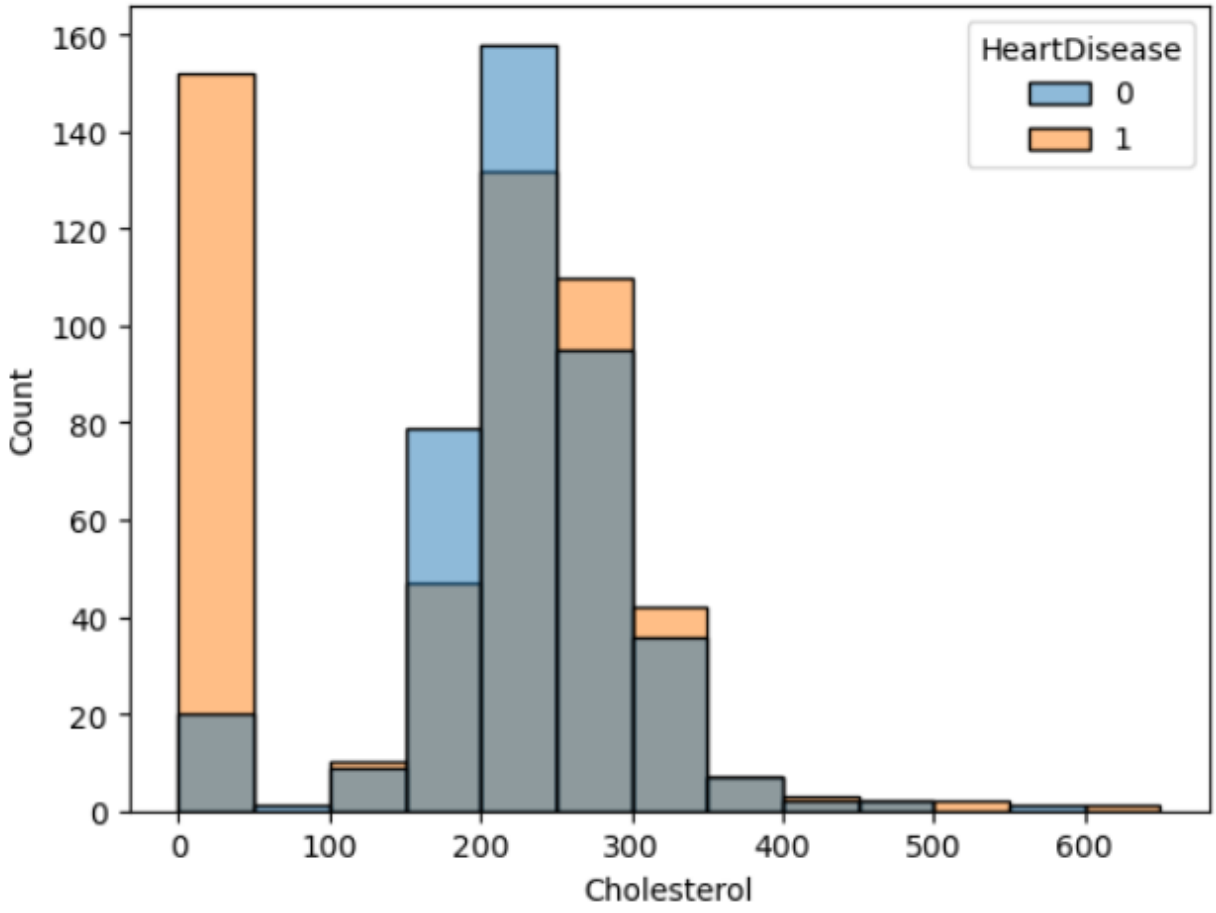

Figure 3

‘There are actually many outliers for the cholesterol data in both directions, but I had chosen to not remove any data as cases of high or low cholesterol can be found in the real world. That being said, there are instances where heart disease or cholesterol values can be hereditary; these instances can skew the model and need to be feature engineered as such. Using the mean for both resting blood pressure and cholesterols, a hereditary feature was appended to the data set such that it would be marked as a hereditary heart disease if their bp and cholesterol were below the means.

Model Selection

Using Pycaret, I tested a wide variety of models as shown below in figure 4. Using accuracy as the main metric, I found a ridge classifier to be the best model given the data distribution as it also had comparable recall and precision scores.

	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

Figure 4

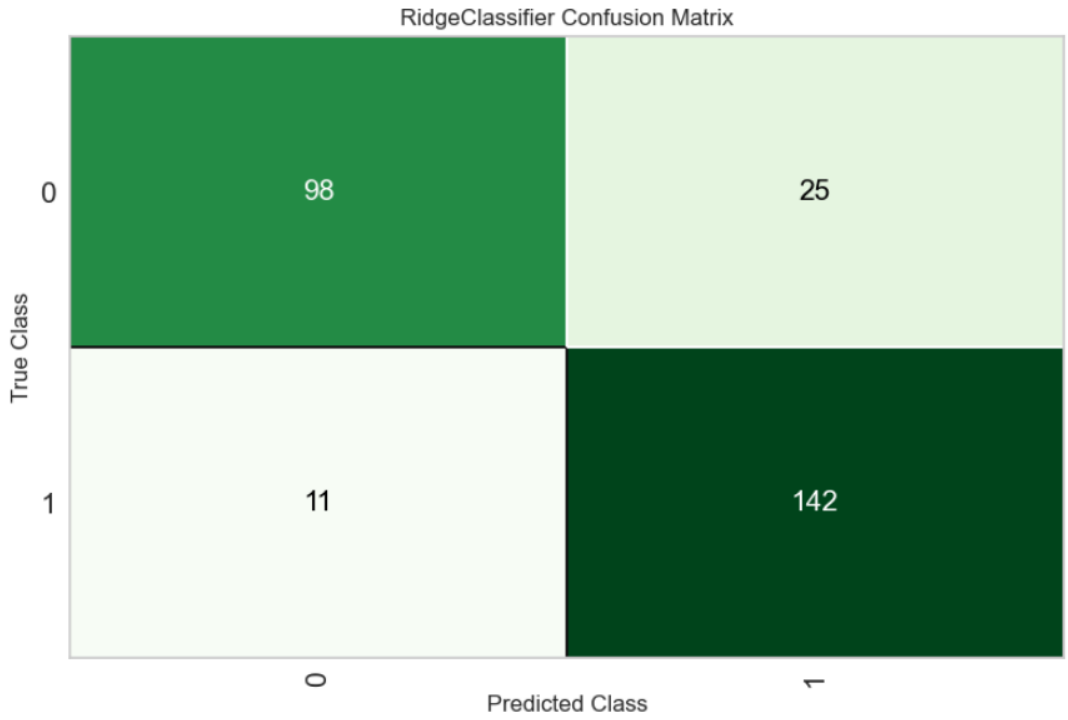

Figure 5

After making the conclusion to stick with the ridge classifier, I then proceeded to use cross validation on the classifier and found that the standard k=1 happened to be the best fit for the model for this context. The confusion matrix above in figure 5 shows the models performance in depth.

Takeaways

With the given data, the ridge classifier is the best model to be used to determine if someone has heart disease. With how major heart disease is on a person's life however, it should definitely be taken as just a guideline and proper procedures should always be taken to confirm if the patient truly has heart disease; ideally both patients and doctors could have access to this model and input the patient data to see if any extra testing needs to be done.

The model can also be used in an iterative sense. After a patient is tested for heart disease, this data can be input in the model and stored so that any new trends with patient data and having heart disease can be seen; this can then be used by doctors to recommend patients if they should get checked up for heart disease in the future.

Future Research

While this model is a solid baseline for detecting heart disease, there is always room for more data. To make the model more accurate, I would first look more into heart disease data and see what other features could be added to the model. Another way to improve the model would be look into a way to get more patient data; this is much harder as there are many legal complications with it, but possibly going through synthetic medical data could help improve the model as well.