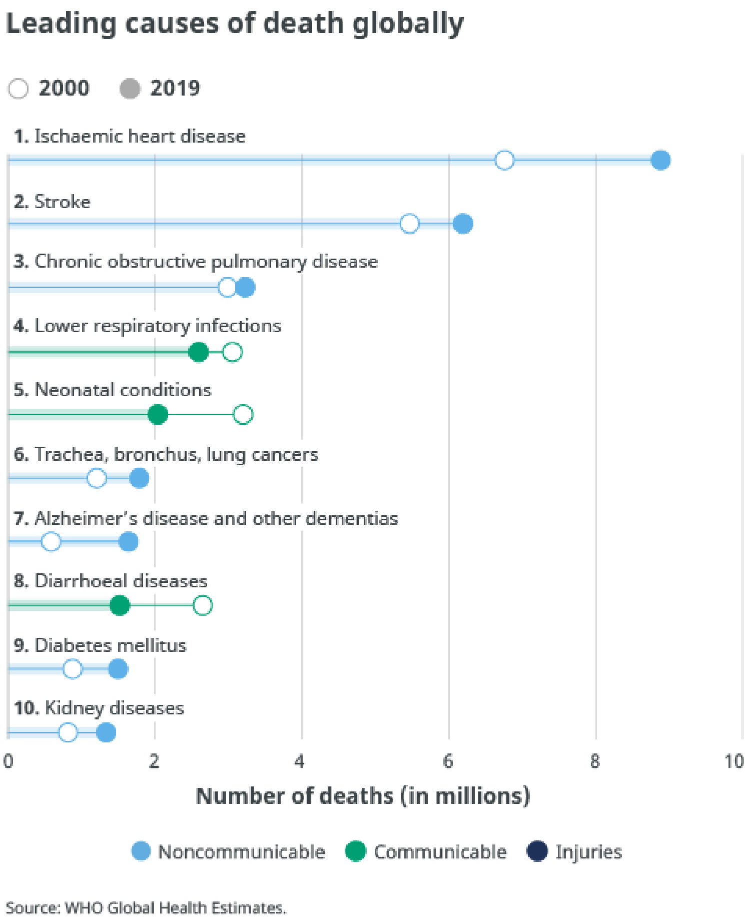
Final Report:
Heart Failure Prediction Analysis

Problem Statement:

Heart Failure leads to over 600,000 deaths in America alone. The chart below from the World Health Organization (WHO) shows that heart failure was the most significant cause of death in 2000 and 2019, only growing its lead even further in 2019.



Source: WHO Global Health Estimates.

Fig 1. Breakdown of # of deaths by cause

Being able to predict heart failure can save millions of lives around the world. I want to dig into this issue not only to uncover findings of what leads to heart failure but also to help medical professionals detect signs of heart failure before it becomes a serious problem. A study[2] done by Mirella Fry and her team found that living with heart failure causes disruptions in every day lives of patients. Common symptoms of heart failure include breathlessness, fatigue, irregular heart rate, and impaired thinking, among others; directly impacting the sufferers' quality of life.

I am going to analyze the Heart Failure Prediction dataset to find features that have a strong indication of an individual suffering from Heart Failure.

Data:

The raw dataset[5] was combined from 5 different studies over 11 of the most common features that lead to heart failure. I downloaded this data from Kaggle; it had 918 rows and 12 columns, 1 of which is whether the individual had heart failure or not. The data set was mostly clean but I needed to modify some things to get it ready for further analysis.

Features in the dataset:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

I looked into all the integer and float features using the describe function of pandas to make sure the min, max, and count numbers were accurate. There were several individuals whose cholesterol levels were not tracked properly (tracked as ‘0’), so I replaced the zeroes with the median value of the rest of the ‘Cholesterol’ column. This way I didn’t lose valuable information from such a small data set.

I then used the pandas qcut function to break down 4 quantitative columns into 4 quartiles. The columns I broke down into quartiles are Age, Max Heart Rate, Cholesterol, and Resting Blood pressure for each individual.

The final shape of my data set is 917 rows and 16 columns.

Exploratory Data Analysis:

In this step of the process, I wanted to understand which of the features in the data set most likely indicated the likelihood of an individual having heart failure. I looked at the correlation between each feature and the target parameter.

For starters, I looked into how the data was split for people with and without heart failure. It is a fairly even split with 55.3% of the individuals in the dataset having heart failure and 44.7% without heart failure.

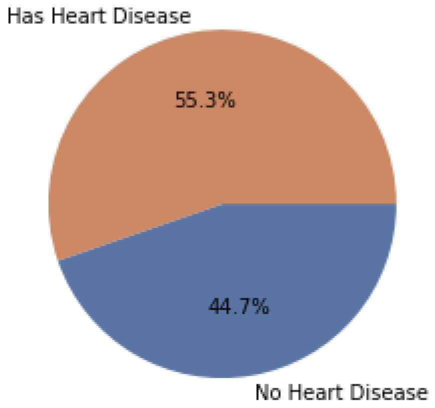


Fig 2. % of participants with and without Heart Disease

I noticed individuals in the second quartile for Cholesterol levels had the highest percentage of people that had heart failure compared to all the other quartiles, as seen in Fig 3 below. According to a study done by Korean National Health Insurance (source below), higher cholesterol levels lead to a higher likelihood of getting heart disease, so it was surprising to see that the data didn't show the same trend. But, I realized that these data could be skewed because I replaced all '0' values in the feature with the median cholesterol level in the initial data wrangling section. This doesn't immediately disprove what is seen in the data because the same method was performed whether or not someone has heart failure but I want to be wary of this throughout the analysis.

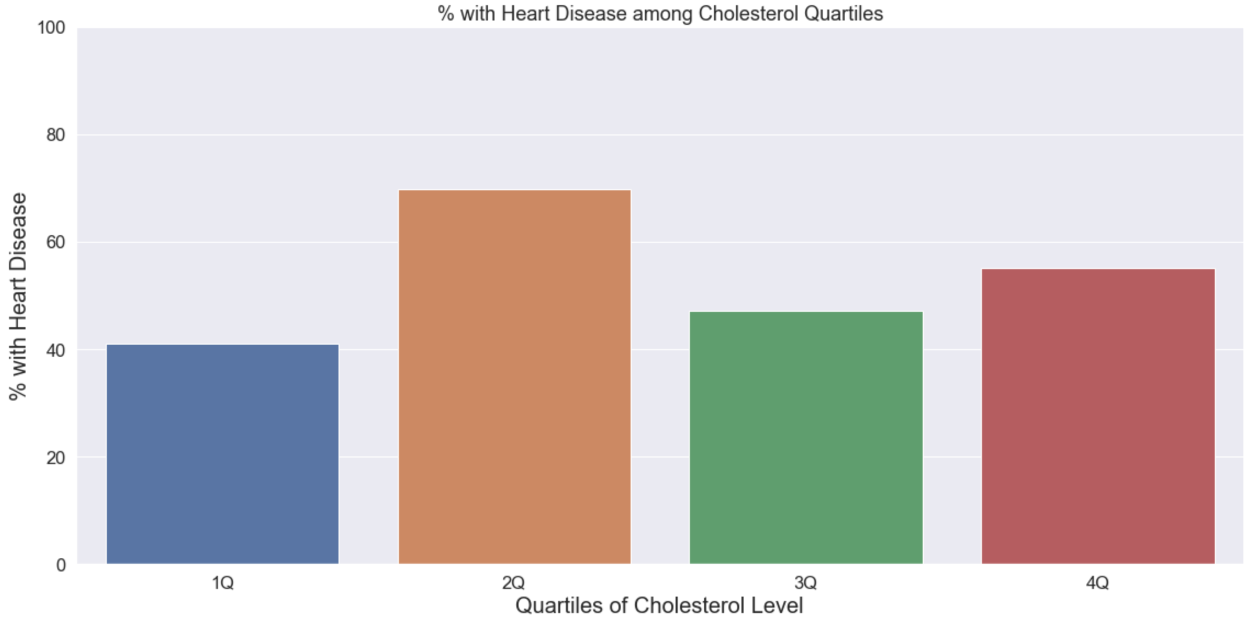


Fig 3. % of participants with Heart Disease by their Cholesterol level's Quartile

I then looked at whether the Max Heart Rate of an individual was an indicator of heart failure. Fig 4 shows that the higher the max heart rate of an individual, the lower the chance of them having heart failure. This goes against the popular belief that a lower heart rate is better for your cardiovascular functions. A study[4] from the Journal of Cardiology (JOC) states that lowering heart rate had a large contribution to improving the survival rate of individuals with heart disease. I believe both findings could be accurate because the JOC study is tracking resting HR while the dataset I am using tracks Max HR. This could be due to the heart functioning properly when it needs to, so perhaps Max HR and resting HR are not so comparable.

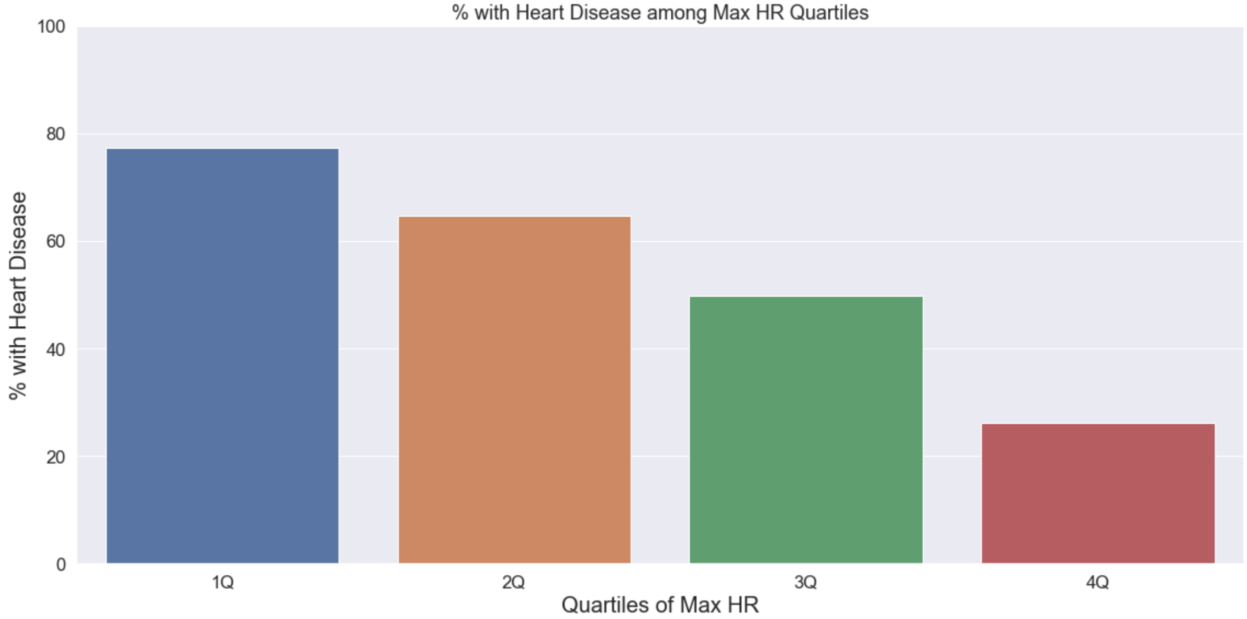


Fig 4. % of participants with Heart Disease by their Max Heart Rate Quartile

The ST segment/heart rate slope tracks an individual's response during exercise stress testing. It is considered the most accurate ECG criterion for diagnosing heart failure. Fig 5 below shows that if an individual has an ST slope they are more likely to have heart disease. An individual with an ‘Up’ ST slope has a significantly lower likelihood of heart failure than an individual with a ‘Down’ or ‘Flat’ ST slope. The ST slope has been the strongest indicator of heart failure among the features so far.

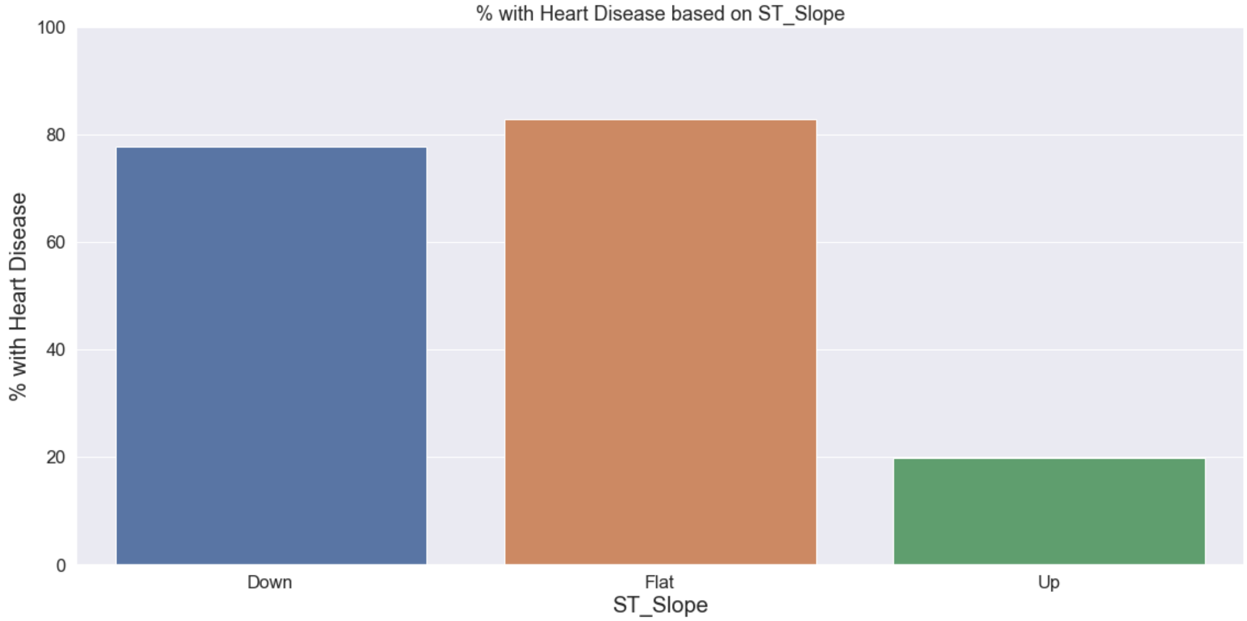


Fig 5. % of participants with Heart Disease by their RestingECG reading

Another strong indicator of heart failure is the type of chest pain an individual experiences. As seen in Fig 6 below, if a person has Asymptotic chest pain (aka silent Heart attack) they are significantly more likely to have heart disease than the other types of chest pain. Chest pain as a whole is not a strong indicator of heart failure but asymptomatic chest pain is.

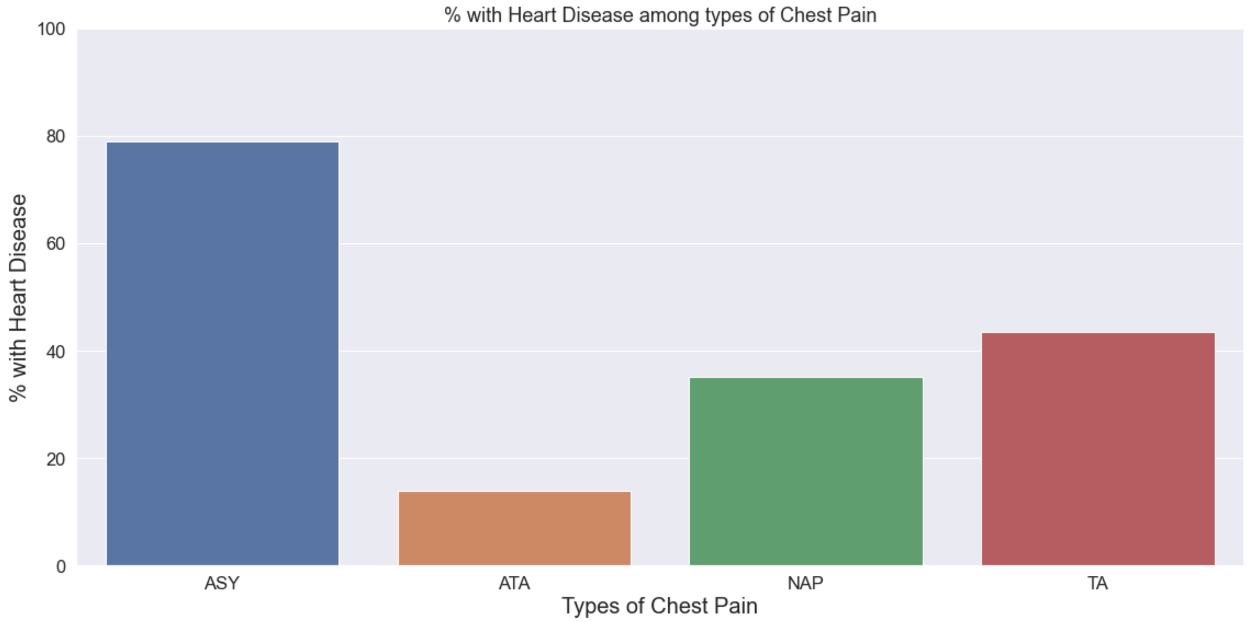
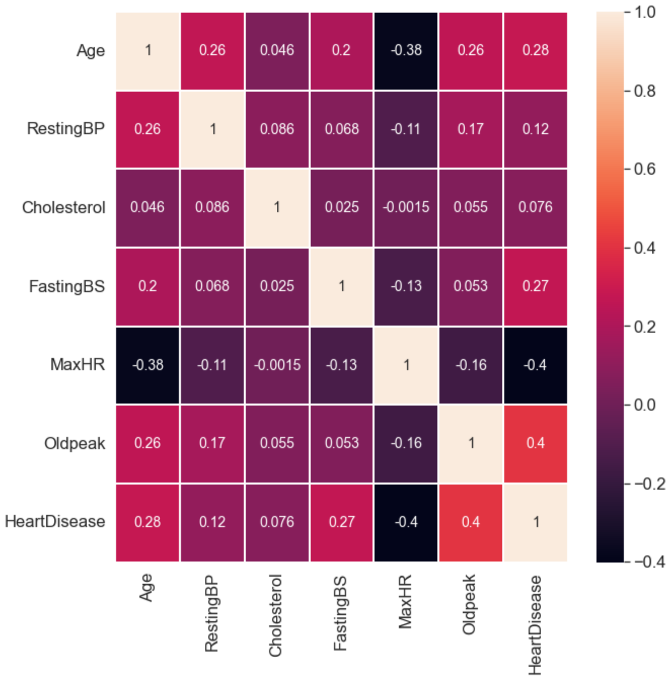


Fig 6. % of participants with Heart Disease by the type of Chest Pain experienced

I created a correlation matrix to see if there are any confounding variables within the dataset. I found that Heart Disease and an individual's Oldpeak level have the highest correlation but it is not significant. There is also a negative correlation between Max Heart Rate and Heart Disease which we noticed in Fig 4. Max Heart Rate and Age also have a similar negative correlation but it is not significant. Since there are no significant correlations between features, it is safe to assume there will not be multicollinearity issues.



	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1	0.26	0.046	0.2	-0.38	0.26	0.28
RestingBP	0.26	1	0.086	0.068	-0.11	0.17	0.12
Cholesterol	0.046	0.086	1	0.025	-0.0015	0.055	0.076
FastingBS	0.2	0.068	0.025	1	-0.13	0.053	0.27
MaxHR	-0.38	-0.11	-0.0015	-0.13	1	-0.16	-0.4
Oldpeak	0.26	0.17	0.055	0.053	-0.16	1	0.4
HeartDisease	0.28	0.12	0.076	0.27	-0.4	0.4	1

Fig 7. Correlation Matrix for all the quantitative features in the dataset

Finally, I wanted to find the most important features that indicate an individual having Heart Failure so I used Random Forest Classifier to create a feature importance list, Fig 8. This was as expected, upward-facing ST slope was the strongest indicator of an individual not having heart disease. Asymptotic chest pain is second on the feature importance list as expected based on Fig 6.

I didn't expect to see the Oldpeak feature so high on the feature importance list, but it certainly makes sense. Oldpeak tracks an individual's ST slope by exercise relative to their rest. Logically, this makes sense because an individual's ST slope significantly changing from exercise to rest can indicate an issue.

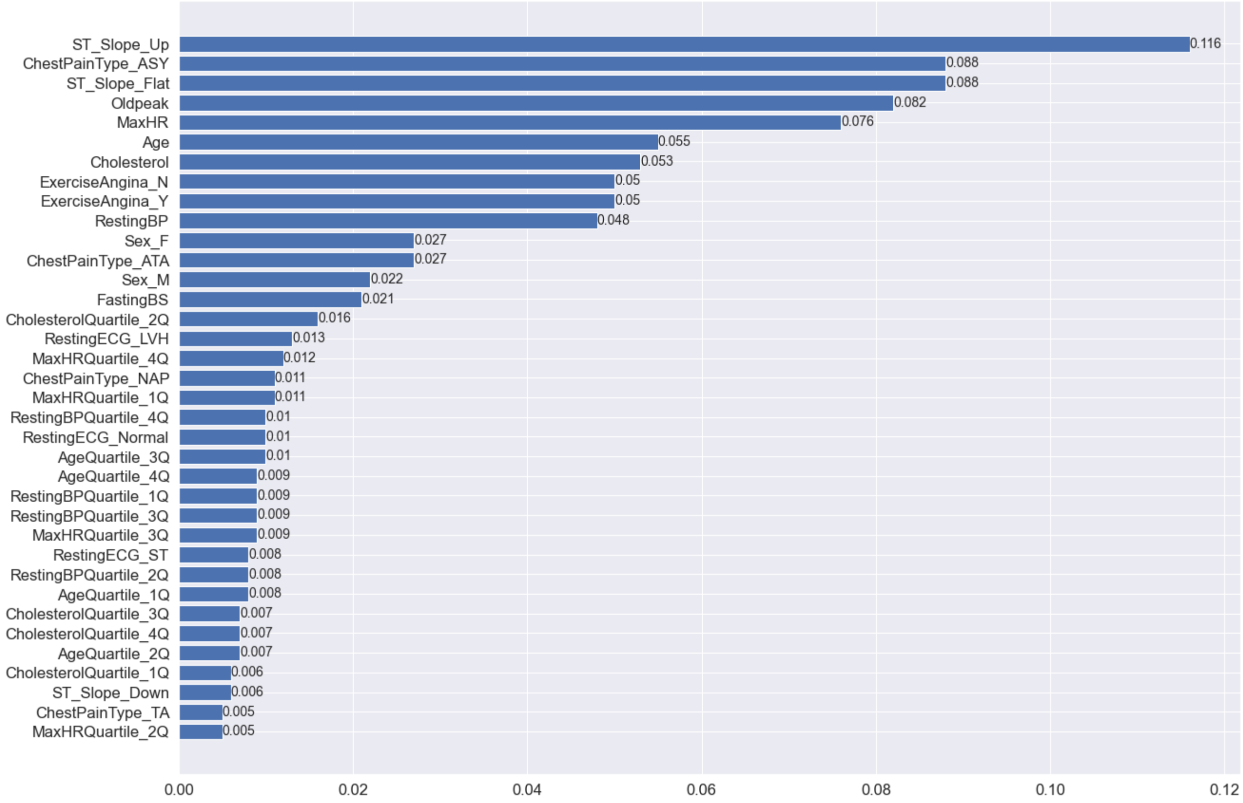


Fig 8. Feature importance chart for all the features in the dataset

Modeling/Takeaways:

I ran 3 separate ML models to predict Heart Failure: Logistic Regression, Random Forest Classifier, and K-Nearest Neighbors. For this business case, recall needs to be prioritized since failing to classify someone who is at risk of heart disease has potentially deadly repercussions. In contrast, predicting someone as being at risk who isn't doesn't have the same level of danger associated with it.

For each model, I did a grid search for all the hyperparameters; this was done to find the parameters that led to the highest accuracy score. I then used Recursive Feature Elimination for Logistic Regression and Random Forest Classifier to find the optimal number of features to include in the data for each model (Unable to use this for KNN due to restrictions in sklearn). Highlighted in Table 1 below.

	Logistic Regression	Random Forest Classifier	K-Nearest Neighbors
Best Parameters	C=1, max_iter=50, penalty='l1', and solver='liblinear'	max features = log2 and max_depth = 5	N_neighbors = 7
Optimal Feature Set	11	5	All
ROC-AUC Score	0.933	0.907	0.901

Table 1. 3 ML models with optimal parameters, feature set, and ROC-AUC score for this use-case

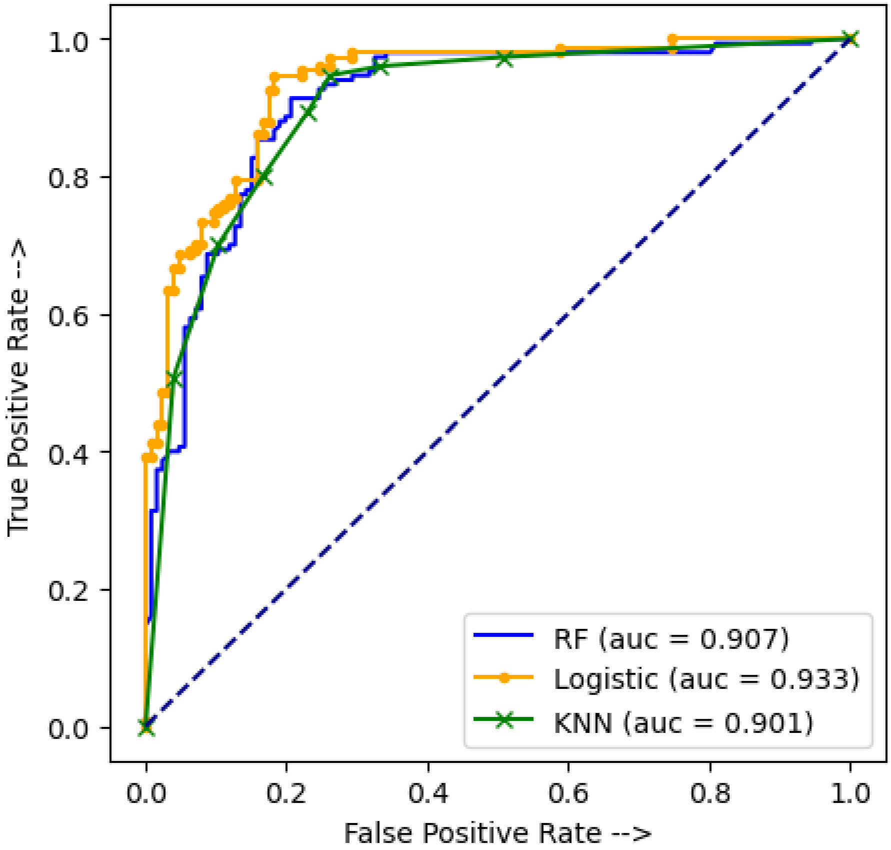


Fig 9. ROC-AUC Curve for all 3 models

The Logistic Regression model performed the best out of the 3 using the ROC-AUC curve. Fig 10 below shows the precision recall curve for all 3 models with optimal thresholds.

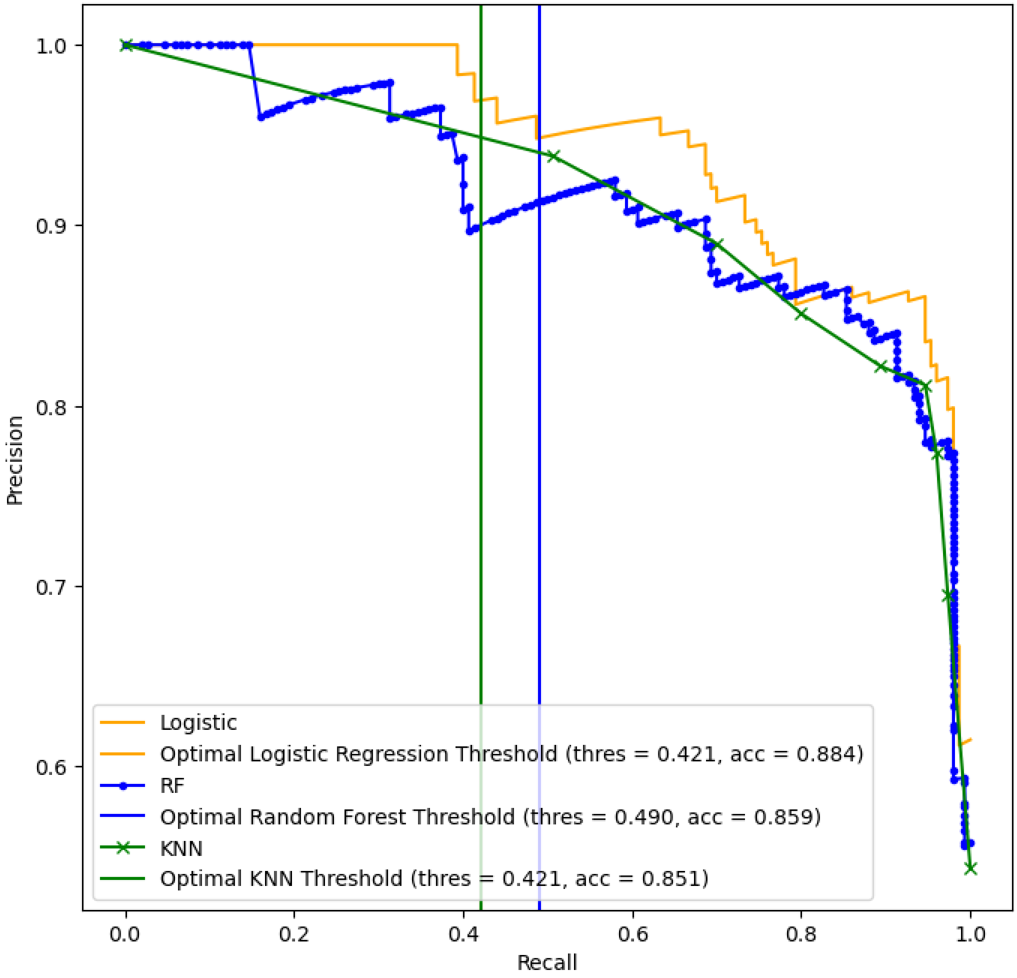


Fig 10. Precision-Recall curve for all 3 models, Vertical lines are optimal thresholds for each model (Logistic Regression and KNN have the same optimal threshold of 0.42)

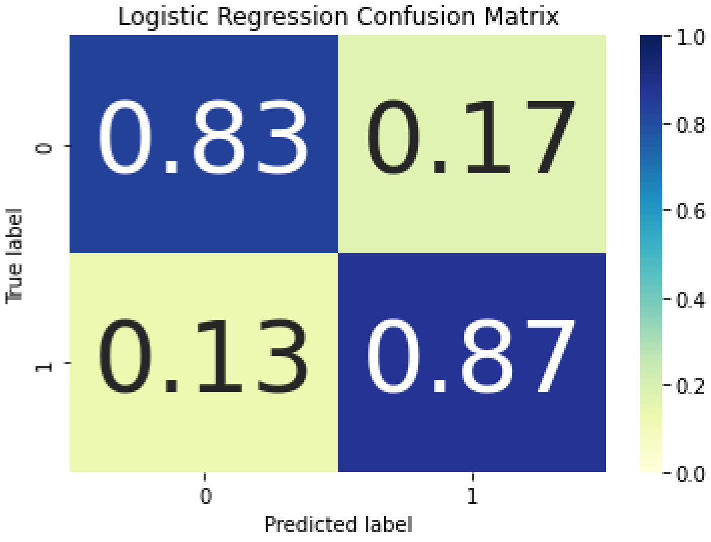


Fig 11. Logistic Regression Confusion Matrix

	Precision	Recall	F1-score	Support
0	0.85	0.83	0.84	126
1	0.86	0.87	0.87	150
Accuracy			0.86	276
Macro Avg	0.85	0.85	0.85	276
Micro Avg	0.85	0.86	0.85	276

Fig 12. Classification Report for the Logistic Regression Model

Taking a look at the confusion matrix for all 3, the Logistic Regression model accurately predicted someone having Heart Disease 87% of the time, while the Random Forest predicted 90% of the time, and KNN only 89% of the time. But, the Logistic Regression model has more AUC than the other models leading me to believe that it is better at predicting 0 as 0 and 1 as 1 for this classification problem.

The Logistic Regression model is the best ML model for this problem because it has an accuracy at 0.86, while the other 2 models are close they don't have as high of AUC as the Logistic Regression model. Logistic Regression is clearly ahead of KNN and Random Forest after evaluating the model metrics.

The Logistic Regression model is the best ML model for this use case. Using the 11 features this data set accounted for, a Logistic Regression model is the most accurate at predicting an individual's chance of having heart failure.

The most important features for predicting heart failure are the type of chest pain an individual experiences, the direction of their ST slope, Max heart rate, and their Oldpeak. These features were kept in the data set for all 3 models and had the biggest impact on prediction. Doctors should keep the focus more on these features when testing individuals for heart disease as they are the most impactful indicators.

Future Study:

There was a lot of important information that was uncovered during this analysis but there is still more to learn. I would like to look further into why a higher max HR led to a lower likelihood of having heart failure when studies show that a lower resting HR is better for not developing heart disease. It would be interesting to learn the relationship between max and resting HR and how they impact the likelihood of heart failure.

Sources:

1. WHO CHART - https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. Living with Heart Failure Study - https://bmcprimcare.biomedcentral.com/articles/10.1186/s12875-016-0537-5
3. Cholesterol Study - https://www.ahajournals.org/doi/10.1161/JAHA.118.008819
4. HR Study - https://www.journal-of-cardiology.com/article/S0914-5087(12)00125-6/fulltext
5. Dataset - https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction