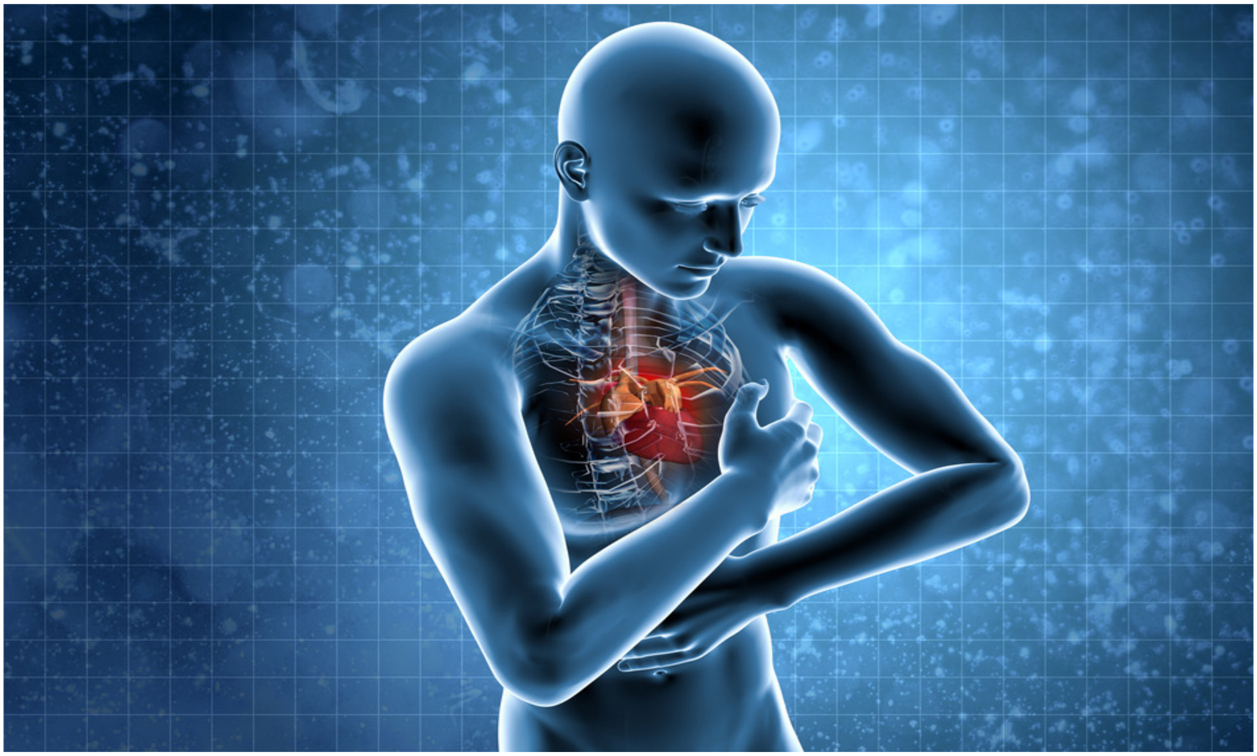
SpringBoard

Predicting Heart Disease

By: Eric Wolf



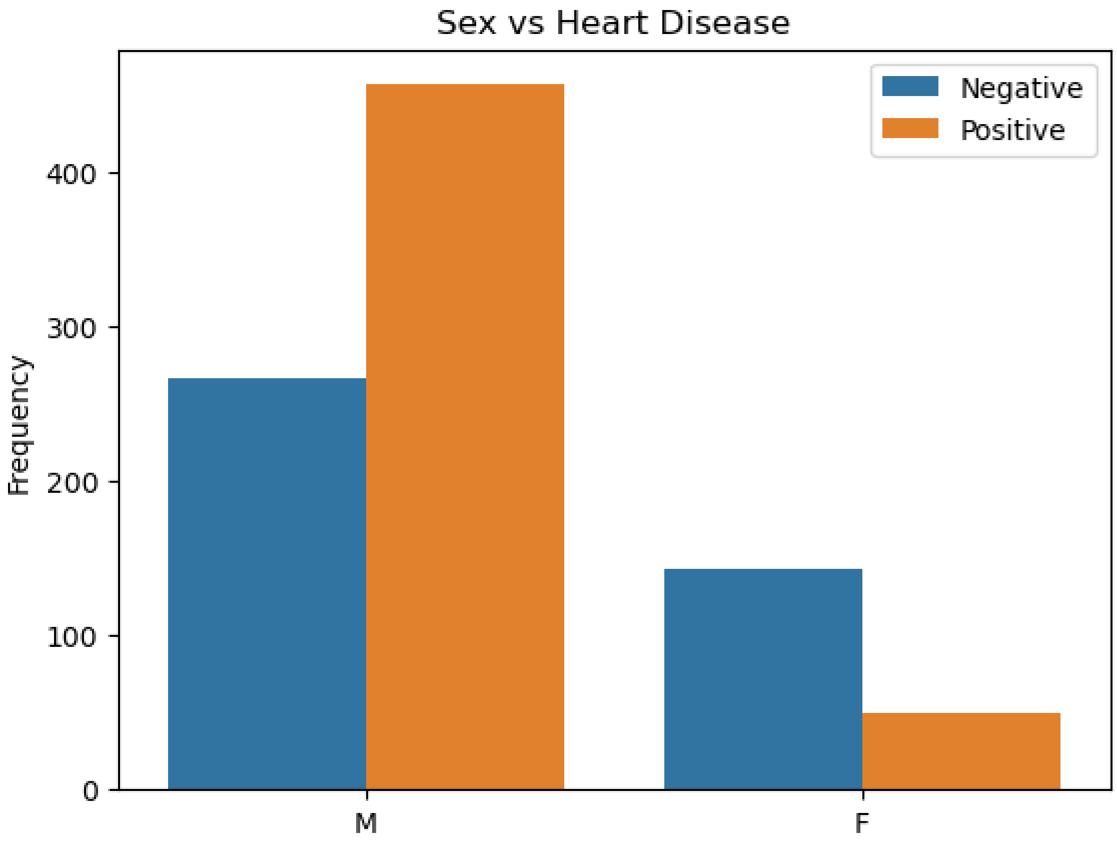
Introduction

Cardiovascular diseases (CVDs) account for 31% of all deaths worldwide with four out of five CVD deaths attributed to heart attack or stroke. Patients who are known to have CVD or who are considered to be at higher risk due to preexisting factors including but not limited to diabetes, hyperlipidaemia, or hypertension would benefit from an early detection system for heart disease. A dataset merging five large scale studies has been created identifying 11 common features associated with CVD to build and train a predictive model for identification of increased risk for heart disease. This model once implemented will be used to identify patients with the probability of being at a higher risk of heart disease based on known/observable information.

Method

Source

The received data was created by a combination of 5 large scale health studies from across the globe. These studies focused on 12 key attributes: Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Exercise Induced Angina, Oldpeak, The Slope of the Peak Exercise ST Segment (ST Slope), and Presence of Heart Disease. The data received represents a good distribution of patient data with 918 cases, ages ranging from 28 to 77, and a heart disease rate of around 55%. There is an exception with the male/female distribution of the raw data being skewed with 79% of the observations being male in addition to a larger number of positive cases for CVD as shown in the figure below.



Data Wrangling | Exploratory Data Analysis

With the raw dataset received and uploaded, following a thorough assessment to confirm the absence of missing or null values, I delved into a more detailed examination of each index. This began with the scrutiny of categorical features such as gender, chest pain type, and resting electrocardiogram (ECG) readings. The intention was to ensure the presentation of data in a meaningful manner and to identify any potentially erroneous or substituted values. After confirming the integrity of these features, I proceeded to segregate the data by gender due to the underlying assumptions that heart disease may present different symptoms between genders and plotted the resulting values against instances of positive heart disease cases. This analysis aimed to determine whether any necessary adjustments to the model may be warranted for further analysis based on gender-related differences.

Upon discovering no statistically significant discrepancies in the categorical features with respect to gender, I transitioned to the exploration of numerical features present within the dataset. During this process, it was observed that several cholesterol values were recorded as zero, a few Oldpeak values fell below the acceptable range, and there was an instance of a patient having a resting blood pressure (BP) value of zero. I took note of these discrepancies and recognized the need for adjustments to ensure proper processing of these categories within the dataset for future analysis.

Having identified these issues, my analysis extended further into the numerical features, mirroring the approach used previously, to ascertain whether any significant gender-related differences existed. With no notable disparities detected, my attention shifted to data augmentation to rectify missing or erroneous values. Notably, a substantial proportion of missing data points pertained to cases involving patients who tested positive for heart disease. In response, I opted to replace these data points with the mean value, while simultaneously excluding the patient with missing values for both resting BP and cholesterol. This process resulted in a refined and structured dataset, organized within a dataframe which facilitates trend analysis and supports the accurate determination of significant factors for predicting heart disease.

With both the categorical and numerical features wrangled and explored for significant gender related differences I began testing individual features for significance with respect to the presence of heart disease utilizing the chi square test of independence for categorical features and T-tests for numerical features. The insights derived from the exploratory data analysis have illuminated compelling associations between the features within the dataset and the presence of heart disease. These associations span a spectrum from weak to robust.

When we gauge the strength of associations among the categorical features using Cramer's V score, a ranking emerges. At the forefront of this ranking (from highest to lowest Cramer's V score) are ST Slope, Chest Pain Type, Exercise Angina, Sex, Fasting BS, and Resting ECG. It's noteworthy that Resting ECG, while not displaying a particularly formidable connection, exhibits a moderate association.

Turning our attention to the numerical features, a ranking based on t-test outcomes underscores their relationships. At the pinnacle of this ranking (from highest to lowest association) lies Oldpeak, Max HR, Age, and Resting BP. A chart showing the outcomes of each T-test can be found below.

T-Test Results

Variable	t	p-value
Age	8.89	3.1608374545678206e-18
RestingBP	3.59	0.0003427865820658374
Cholesterol	-0.34	0.7342297705776781
MaxHR	-13.26	8.044072537846356e-37
Oldpeak	14.24	1.0953740022189476e-41

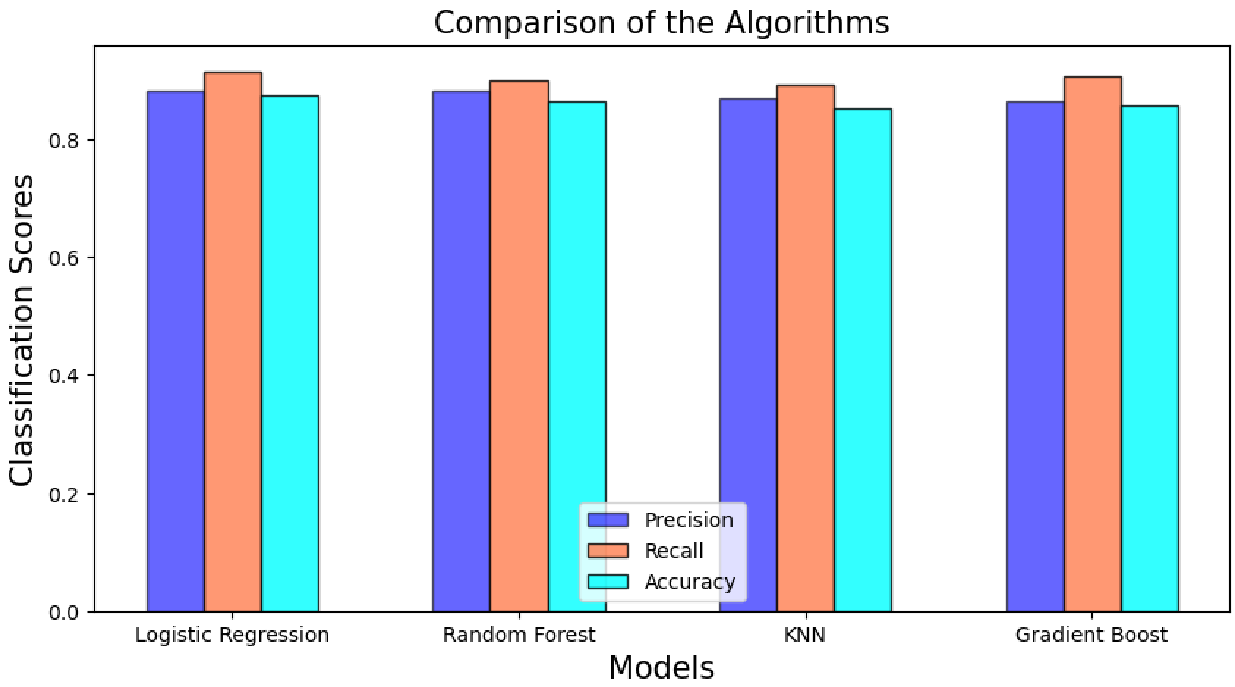
Cholesterol, however, diverges from this trend, demonstrating no statistically significant relationship with heart disease. It's intriguing to observe that Cholesterol, despite this finding, had the highest incidence of missing values within the raw dataset and therefore any association or lack thereof has a possibility of being skewed based upon the interpretation of the missing data. Thus, while the data suggests a lack of discernible connection between Cholesterol and heart disease, the caveat remains that Cholesterol's analysis was potentially hampered by the prevalence of missing values in the original dataset.

Modeling | Results

Commencing the data modeling process required the swift establishment of a foundational baseline model, serving as a reference point for assessing subsequent models. To accomplish this, I initiated the deployment of a dummy classifier utilizing a uniform prediction approach, akin to coin flipping, to generate predictions for the test dataset. This facilitated the evaluation of its performance through the utilization of metrics such as ROC AUC, a comprehensive classification report, and a confusion matrix. This foundational data served as a springboard for the implementation of four carefully selected models, each holding promise in attaining the intended outcomes. The models chosen for this purpose were: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting.

A consistent methodology was followed for each of these models. Beginning with the creation of individual pipelines and subsequent grid or random search cross-validation, adapting based on the specific characteristics of each model and its fitting duration. This comprehensive approach was then complemented by a meticulous assessment of each model's performance, employing the aforementioned scoring methodologies.

Upon conducting a thorough comparison of the attained scores, a clear trend emerged: the Logistic Regression model established itself as the prime contender, boasting the highest scores across several metrics including Precision, Recall, and Accuracy. Given the critical medical context, the prioritization of minimizing false negatives over false positives carried profound significance, thereby elevating the importance of recall in this context. Remarkably, the Logistic Regression model outshone its counterparts across most scoring criteria, a pivotal factor in determining its superiority, and retained the highest recall percentage of 91.43. A comparison of the four models can be found in the figure below.



Future Improvements

In future analyses, it would be advisable to consider the incorporation of supplementary patient data, encompassing potentially critical factors such as height, weight, and the presence of pre-existing heart conditions. These additional attributes have the potential to significantly enhance the predictive capabilities of the model, offering a more comprehensive assessment of an individual's risk for heart failure. By including such information, we can leverage a broader spectrum of patient characteristics to refine our predictions and ultimately increase the model's accuracy and reliability.

Moreover, as we strive for continuous model enhancement, the exploration of more advanced machine learning techniques is warranted. One promising avenue is the implementation of neural networks, which have demonstrated remarkable performance in complex pattern recognition tasks. The introduction of a neural network architecture into the predictive framework may further optimize the scoring metrics, enabling more subtle relationships to be uncovered within the data and achieve even greater predictive metrics.

Furthermore, to bridge the gap between model development and practical application, the creation of a user-friendly interface should be considered. Such an interface would facilitate the seamless deployment of our predictive model in real-world healthcare settings. By developing an intuitive user interface, healthcare professionals can easily input patient data, obtain risk assessments, and make informed clinical decisions based on the model's predictions. This user-centric approach aims to streamline the integration of our predictive model into the healthcare workflow, ultimately benefiting patients and healthcare providers alike.