Preventive Pulse

Predictive Analytics for Heart Health

Team Members:
Tyler Gabriele
Brian Lee
Richard Dickson
John Apel Jr

Introduction

Just about everyone who is alive today has been impacted by the devasting toll of heart disease, either directly or indirectly through friends and family members. The vast number of related illnesses that are linked to heart disease is very extensive. Diseases such as Arrhythmias, Cardiomyopathy, heart failure, Coronary Artery Disease, and Pericardium, are just a few to mention according to the Cleveland Clinic. Probably the most important stat that any person could read in America, is heart disease is the leading cause of death. A recent study was conducted by the CDC in 2020 and 2022, and tens of thousands of patients reported all sorts of symptoms, while other potential life factors were also observed. Ultimately, the severity of this illness is not only a massive issue throughout the world, but it is very tricky for medical professionals to detect due to the overlap in symptoms with other illnesses. (As stated in a different recent article by the American Academy of CPR and First Aid)

Literature Review

Going into this project, we had a fundamental belief that machine learning could expose patterns, within this large amount of data, that could allow medical professionals to better discover underlying signs of heart disease with their patients. If steps can be taken in lowering the number of patient fatalities using Machine Learning techniques, we want to exhaust all techniques and resources for achieving this goal. These stakeholders include, but are not limited to, doctors, nurses and other types of physicians that regularly observe factors within the population of people. With our model achieving high accuracy levels with acceptable f1 scores for our selected model classifiers, we can not only determine some of the key variables that our stakeholders need to focus on, but we can also improve our metrics in judging if someone has or doesn't have heart disease, which would lower the number of False Positives and False Negatives. Ultimately, we want to empower medical professionals in detecting early warning signs of heart disease, much sooner than they're accustomed to.

Data

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Our data could be found by accessing the link above. After doing some research and exploring different datasets within Kaggle, we found this dataset that was published by the CDC, which led us to learning about the Behavioral Risk Factor Surveillance System (BRFSS). This data was collected from over 300,000 people surveyed, all of which suffered from symptoms of heart disease but may or may not have had it. These patients reside throughout the US and provided several datapoints for the CDC to measure.

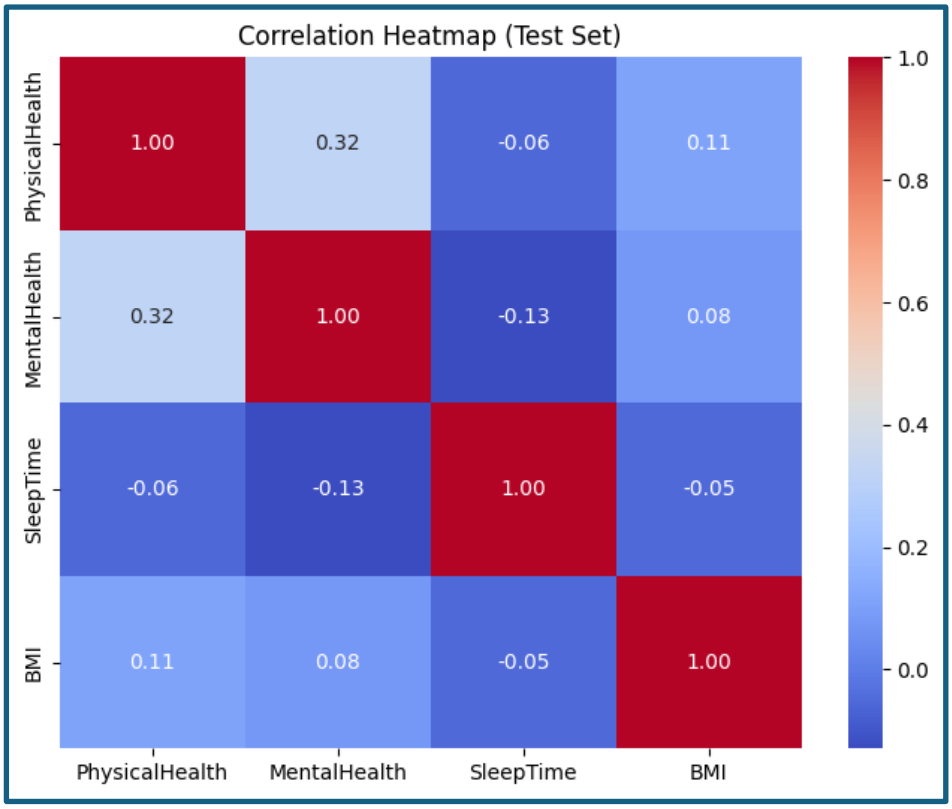
Our initial training dataset had a vast number of columns and rows that we molded into a workable sample dataset from our large population of data (over 440,000 rows and 40 columns). We then separated a portion of data for running a testing dataset, for evaluating the effectiveness, accuracy, and strength of our new model's prediction. After

running a correlation matrix and measuring our variables to our binomial target variable "Has Disease," and by also running PCA for dimensionality reduction, we selected the most relevant variables. We then had a revised dataset of 18 columns with a nice mix of numerical and categorical data. We then created dummy variables to measure some categorical variables (ex. Race, General Health, Age Category & Diabetic History) It's also important to note that we used the pandas library within python for this process of data preprocessing. We also converted our data from SAS to CSV format. Lastly, our dataset was published by Kamil Pytlak, had 503k total views, and 90.8k downloads. Our model also included libraries such as numpy, seaborn, and matplotlib for generating our visualizations.

Methods

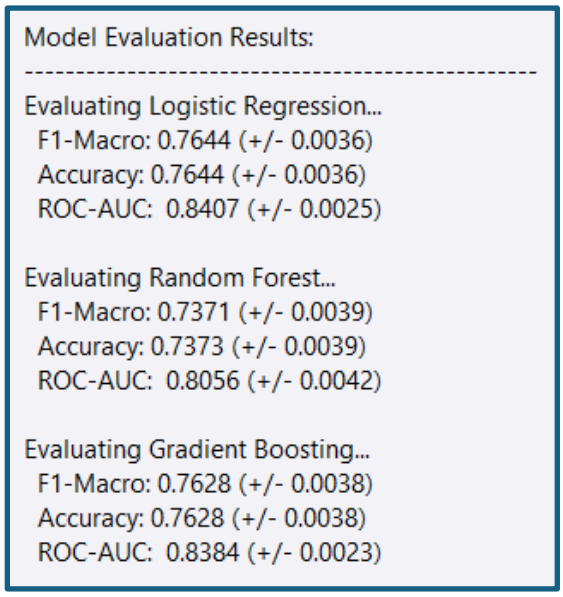
We began with running some basic visualizations between our target variable, "Has Disease" for representing the No's and Yes's for patients with heart disease, and the relationship with other variables with that column. The basic visualizations we created provided us with much needed perspective regarding our data. We used mainly histograms with some scatterplots, to help us examine patterns and identify outliers (along with inconsistencies, and missing values.)

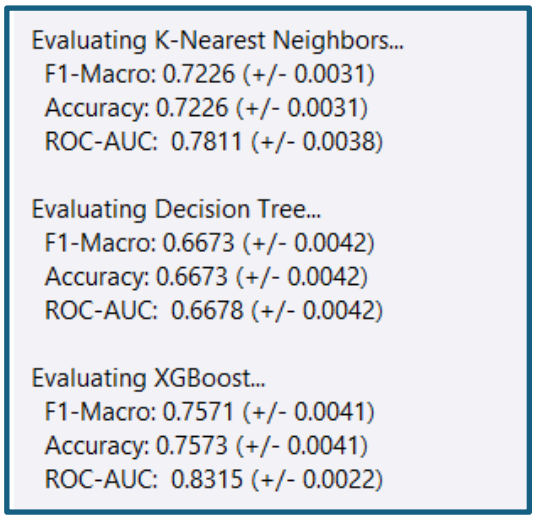
The correlation matrix was also a great resource with identifying patterns and the relationships between variables in our dataset. (Findings of new correlation matrix here) Revisiting our process of creating our dummy variables, we used OneHotEncoder for transforming our data and StandardScaler for normalizing our dataset. Later on, we measured correlation in our Test set to compare how our variables measured, compared to one another.

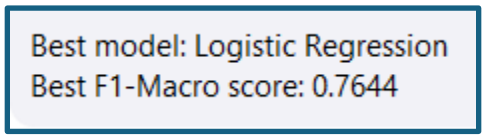


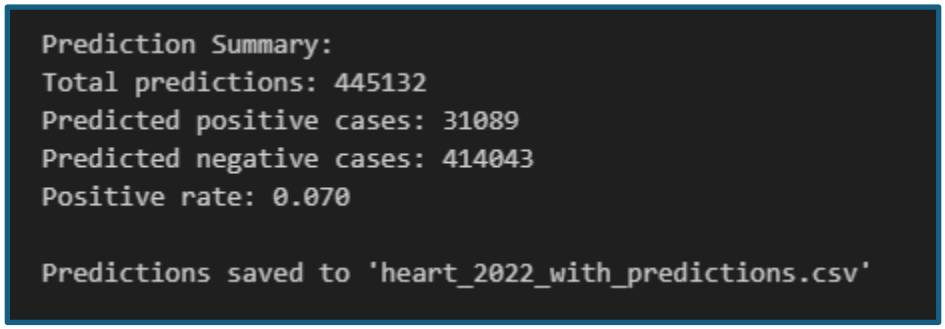
We will also be using multiple classifiers for achieving tasks such as tuning, imputation, encoding, and other techniques for us to adjust our data properly. Despite our data being initially imbalanced and having some initial trouble running our model, we were determined in achieving a high F1 score of over 50% with an acceptable accuracy of 80% or higher. This fundamental goal was rooted in our hope of creating an accurate prediction model that could improve the guessing game that medical professionals are currently stuck with.

Results





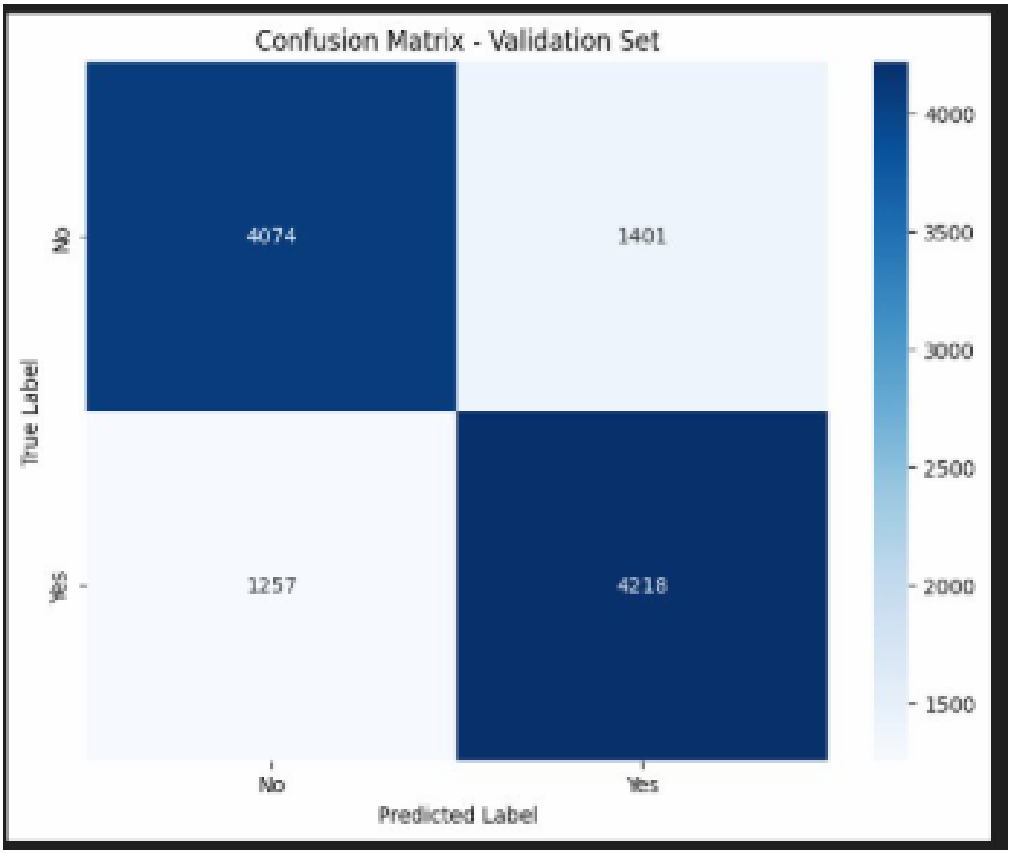




5-Cross Validation was setup as a variable. Then a pipeline was developed to implement SMOTE and each classifier that we were going to use. Then that pipeline was ran through cross-validation score using the 5-fold to get the results of each classifiers accuracy, ROC_AUC, and F-1 score

Discussion

After developing the model, we were able to achieve a high F-1 score and high accuracy of over 0.75 on our top classifier (Logistic Regression.) We were able to use this dataset to predict heart disease in over 31,000 people in our 445,000 person dataset. Medical professionals (stakeholders) would be able to use this model to develop to accurate predict patients that have, or are developing, heart disease. Unfortunately, we aren't able to provide doctors with catching early signs of heart disease, but we are able to provide the biggest contributors in patients that are susceptible in having heart disease.



Limitations

We knew going into this research and model building, that confirmation bias could be a huge factor. This appears to be a central issue when working within health-related datasets. We allowed PCA to be the central determinate in deciding which variables to run with our target variable for this reason, to avoid falling into any traps in assuming one variable was more important than another. So we avoided bringing in personal or societal assumptions about what is considered "healthy" and what has the biggest impact on if someone has or does not have heart disease.

We also ran into issues with the structural mismatch between our training dataset (a 2020 CDC-based survey) and the 2022 dataset. So we used a splitting method from both datasets to prevent us from being pigeon-holed into the surrounding factors of each

separate study. Each dataset originated from the (BRFSS) survey that has been modified from each two-year period. They also had a large overlap in basic features and terminology. We used feature mapping, dropped unmatched columns, and performed value conversions to align the two datasets. Additionally, the 2022 data lacked some of the same target labels, and presented a few new ones, which prevented us from properly evaluating all of the predictions for accuracy post-modeling. As a result, our project emphasizes modeling pipeline, not validation on new labels.

Lastly, the large size of the 2022 dataset, that was over 139 MB, did in fact present computational challenges as we predicted. GitHub's file size restrictions made it tricky to collaborate when we were working on this project separately. Luckily, we were able to file share without too many issues. Our initial sample of data was massively imbalanced, so we also re-sampled our data to include more results. This helped us get a bigger and more accurate picture of what the 2020 survey data actually was reporting. To mitigate this, we plan to down sample the dataset during development and use efficient data loading and processing techniques. This was another issue that we predicted would arise during our proposal period, so we weren't too surprised by this result.

Future Work

Some of the categorical questions are very subjective. Questions such as "Are you generally healthy and "Are you physically healthy" help the surveyor understand the perspective of the patient being surveyed but steers us away from seeing the factual results of if the person is generally healthy or physically healthy. We especially saw this in running our test data as Physical & Mental Health were very rightly skewed. We believe this bias might serve a better purpose for analyzing a different type of study but veer's away from the central point of this study. (Along with its accuracy.)

Also, when measuring the variables to the target variable, "Has Disease," this study can include questions/predictors for future studies. This continued maintenance would accommodate the new features made within our model. From the 2020 set to 2022, several columns were added, or renamed, from variable different names. This process needs to continue to stay up to date in future work, during each re-evaluation.

Citations

Dataset link: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

2020 data: https://www.cdc.gov/brfss/annual_data/annual_2020.html
2022 data: https://www.cdc.gov/brfss/annual_data/annual_2022.html

https://www.onlinecprcertification.net/blog/how-do-doctors-determine-if-someone-has-heart-disease/

https://my.clevelandclinic.org/health/diseases/24129-heart-disease