**Predicting Heart Failure using Machine Learning**

(binary classification problems)

IST 687 – Introduction to Data Science

Adrian Torres

Paige Madison

Frank Williams

**Table of Contents**

Introduction .........................................................................................

Business Questions ..............................................................................

Data Cleanse/Munge/Preparation...........................................................

Descriptive Statistics & Visualizations ....................................................

Data Modeling & Visualizations...............................................................

Machine Learning Algorithms Used and Explanation ................................

Summary................................................................................................

**Introduction**

In our project predicting heart attacks, we leverage advanced data analytics and machine learning techniques to forecast the likelihood of heart failure occurrences among at-risk populations. By analyzing comprehensive datasets encompassing demographic factors, medical histories, and lifestyle choices we aim to develop robust predictive models capable of identifying individuals predisposed to heart failure. Through this approach , we also aim to pave the way for targeted interventions that mitigate their occurrence. This project represents a crucial step towards harnessing the power of predictive analytics to revolutionize cardiovascular care.

**Business Questions**

Can we accurately predict heart attacks?

Which variables are the most important for predicting heart attacks?

Are non-biological ailments, like depression, important for predicting heart attacks?

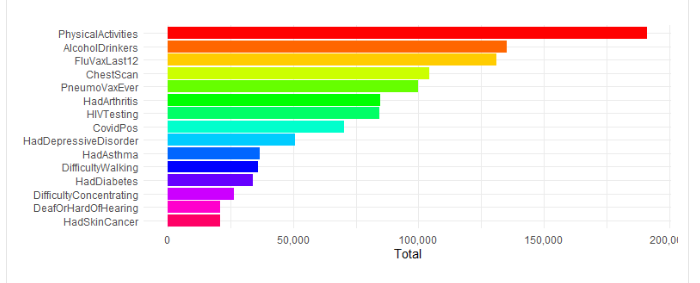
**Data Cleanse/Munge/Preparation**

The dataset covers factors relevant to heart disease, including high blood pressure, high cholesterol, smoking, diabetes status, obesity, physical activity, and alcohol consumption, sourced from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) annual telephone surveys across the United States.

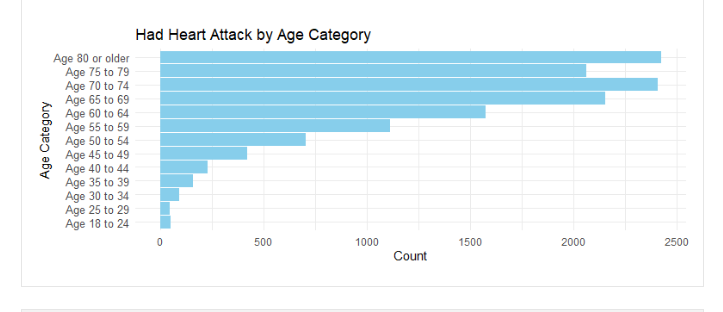
The recent dataset, from 2023, has been refined to include the most relevant variables totaling 40 and is available both with and without missing values, originating from the CDC's BRFSS, which conducts over 400,000 adult interviews yearly, providing a comprehensive overview of health status in the United States.

The dataset can be utilized for various purposes, particularly for applying machine learning techniques such as logistic regression, support vector machines (SVM), and random forest classifiers, with the variable "HadHeartAttack" treated as binary and techniques such as fixing weights or undersampling recommended due to class imbalance, with a logistic regression model already constructed and integrated into an application for heart condition assessment.

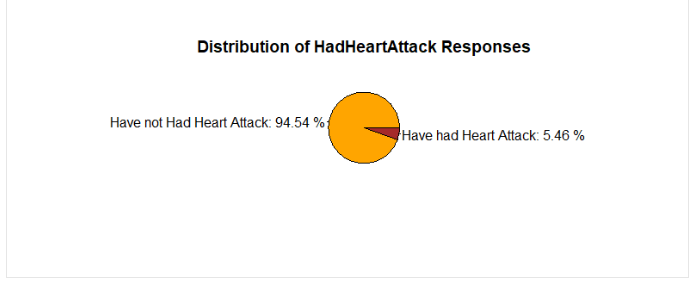
**Descriptive Statistics & Visualizations**



* This visualization provides the top 15 columns that frequently coincide with users who have experienced a heart attack. First, we removed the 'HadHeartAttack' column to then be able to identify these commonly occurring variables with patients who have had heart attacks. This offers preliminary insights into which factors carry the most significant influence in predicting the likelihood of a patient suffering a heart attack



* One of our early questions aimed to determine the predominant age group among patients who experienced heart attacks. As displayed in the visual, we discovered that individuals aged 70-74 and those aged 80 or above constituted the highest proportions of heart attack cases.

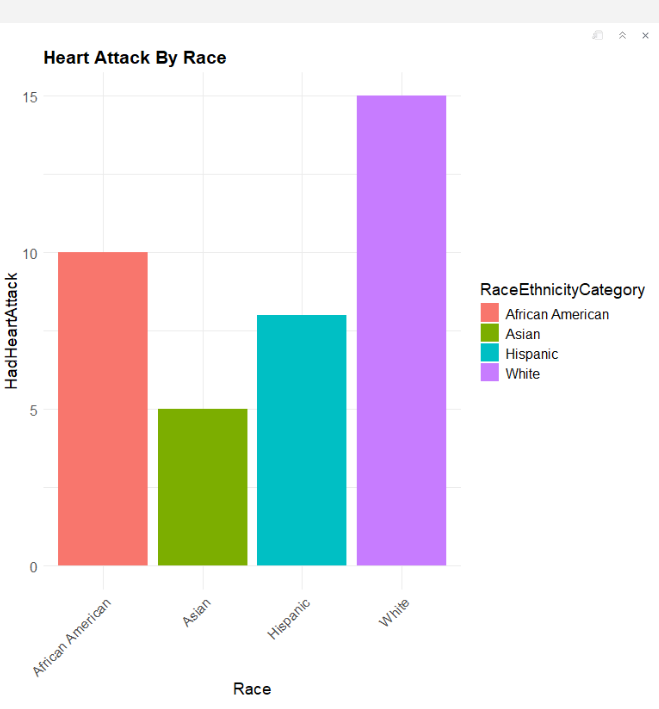


* To gain deeper insights into our dataset, it was important to understand the proportion of patients who had previously encountered heart attacks. As illustrated in the above pie chart, a significant majority, accounting for 94.54% of the patients, had no history of heart attacks, while a smaller percentage, 5.46%, had experienced Heart Attacks.

A graph of a number of blue rectangular bars

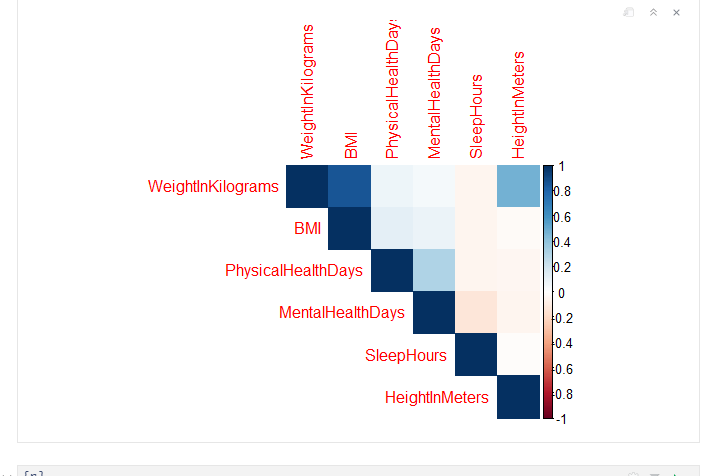
Description automatically generated

We closely examined the 'GeneralHealth' column, shedding light on the overall health condition of each patient in the dataset. We discovered a substantial majority, 83.7% of patients, reported health statuses ranging from Good to Excellent. Conversely, 16.3% of patients indicated their health status as Poor or Fair.



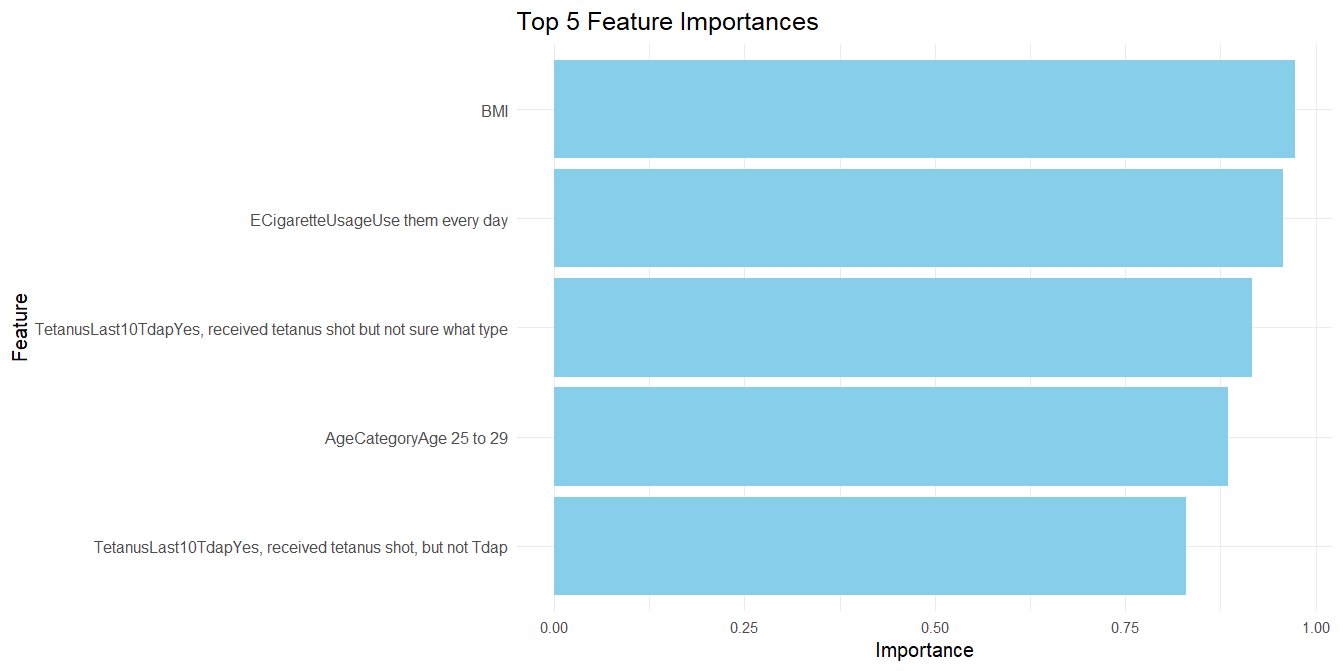
* To understand the impact of race on patients who have experienced a heart attack, we created a graph displaying the distribution of heart attack occurrences among different racial groups. Initially, the graph suggests a higher prevalence among white patients, followed by African Americans. However, upon further analysis using a correlation matrix, we found that race does not exhibit a positive correlation with heart attack occurrence within this dataset.

**Correlation Statistics**

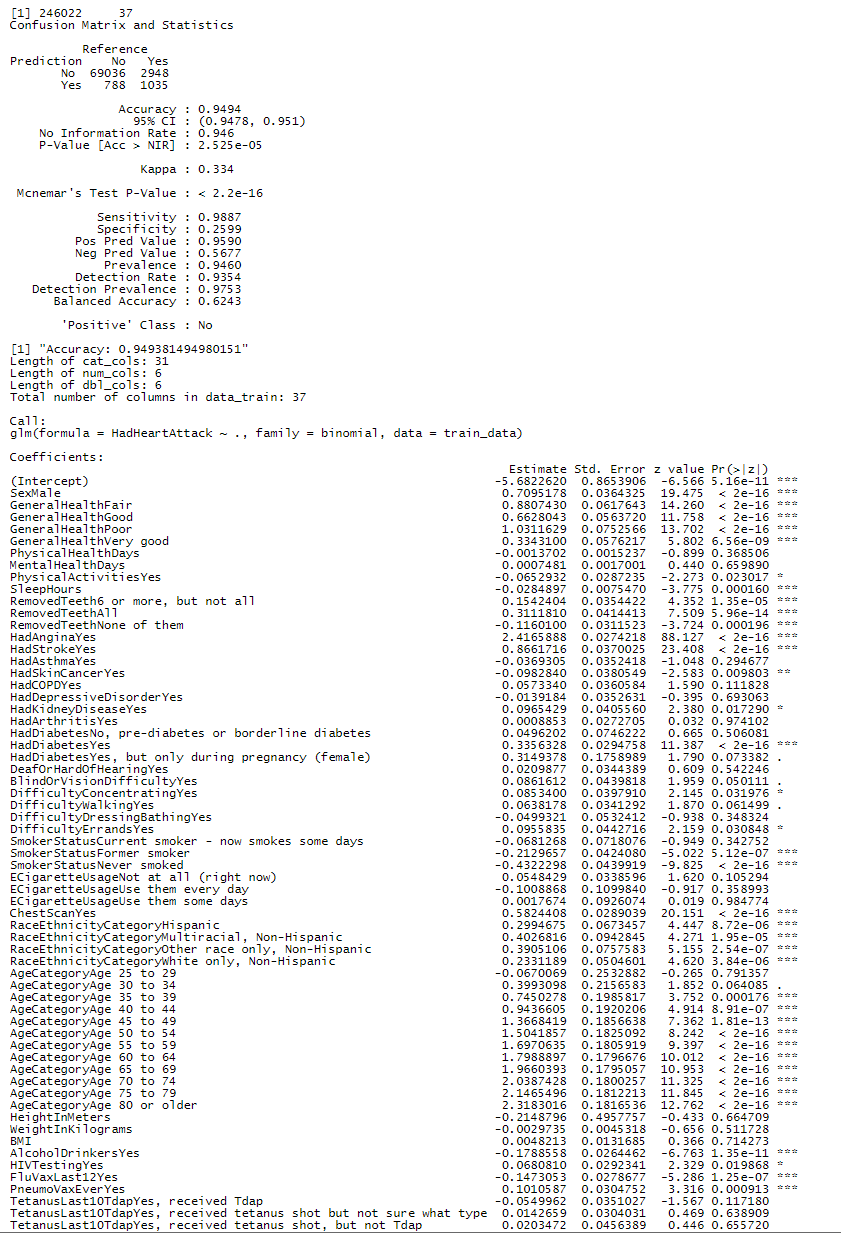


* Utilizing correlation statistics, we identified the key variables significantly associated with the likelihood of patients experiencing heart attacks. Notably, variables such as WeightInKilograms, BMI, and Physical health days exhibited positive correlations, as depicted in the matrix above. This addresses two of the questions we aimed to answer, providing the top three variables that are most correlated with predicting heart attacks. Additionally, we were able to determine that non-biological ailments such as depression do not show any positive correlation in the dataset we examined.

Top 5 Feature Importance



* The following variables exhibited the lowest **p-values**, indicating their significant impact on predicting heart attacks in the undersampled **logistic** **regression** **model**. These variables are ranked as the most significant predictors.



Machine Learning Algorithms Used and Explanation

A screenshot of a computer program

Description automatically generated

First, relevant columns were selected and excluded, and data types of each column were verified. Specifically, the column “State” was excluded as it didn’t appear to be relevant to predicting heart attacks.

A screenshot of a computer program

Description automatically generated

After, a data subset was created, which converted categorical variables to usable factors, and Nas were removed from that subset. Next, the data was split using 70% of the data to train and 30% to test. Then, a logistic regression model was fitted using the training data, which produced the output below:

A screenshot of a computer

Description automatically generated

Figure : Logistic Regression Output

# Referenced Equations

The model output for the logistic regression showed the following:

* There were 69,036 true negatives (TN) or correctly predicted “No” values.
* There were 1,035 true positives (TP) or correctly predicted "Yes" values.
* There were 788 false positives (FP) or incorrectly predicted "Yes" values.
* There were 2,948 false negatives (FN) or incorrectly predicted "No" values.

The **Accuracy metric** was 0.9494 or 94.94%, for our logistic regression model, which meant that it correctly predicted heart attacks 94.94% of the time. The **Balanced Accuracy** metric had a value of 62.43%, which suggested that even with an imbalanced data set, it still managed to score above a score of 0.5, which would be the equivalent of guessing. Additionally, because Balanced Accuracy considers both the sensitivity (i.e., the true positive rate) and the specificity (i.e., the true negative rate), it is a realistic metric for accuracy or the model’s ability to predict correctly.

The **No Information Rate** (**NIR**), or the proportion of predicted values that belong to the majority class, was 94.6%, which meant of the predicted outcomes 94.6% was the majority class (i.e., the “No” class). The **P-Value [Acc>NIR]** has a value of suggests that the model’s No Information Rate was significantly different from the Accuracy value. The p-value that compares both the Accuracy and No Information Rate, assesses if the model’s performance would be significantly better than would be expected by predicting the “No” response or the majority class in most cases. Since the p-value was significantly lower than 0.05, it suggests that the observed difference between the Accuracy and NIR, is not due to chance or randomly guessing, which makes the Accuracy Metric meaningful.

**Cohen’s Kappa value** considers the possibility of agreement between the actual and predicted classes happening by chance and has a scale from -1 to 1. Our model had a Cohen’s Kappa value of 0.334, which suggested fair agreement between the actual and predicted classes.

The **Mcnemar’s Test’s p-value** had a value of which suggested that there were significant differences between the observed discordant pairs (i.e., the instances where the predicted values and actual values are different or the FP +FN) and expected discordant pairs (i.e., the expected complete agreement between the predicted and actual values or the FP + FN), the latter value is considered under the assumption that our model and a hypothetical copy of our model (i.e., the copy of the model where there is no difference between the actual and predicted values) performed equally. In summary, it means that the disagreement between actual and predicted values, is not by chance alone, and there is possibly systematic bias due to the imbalanced dataset.

The **Detection Rate** or the proportion of true positive instances (i.e., the number of “No” values), which were detected by the model, has a rate of 0.9354, which indicated that the model detected 93.54% of positive instances.

The **Prevalence**, or the measure of the positive class in the dataset, or the amounts of “No” in our data set, was 0.946, which indicates that 94.6% of our instances belonged to the positive class. The **Detection Prevalence** is the predicted positive cases by the model and has a value of 0.9753 or 97.53%.

We also ran a decision tree algorithm to see if there was an alternate model that could produce similar results.

A screenshot of a computer

Description automatically generated

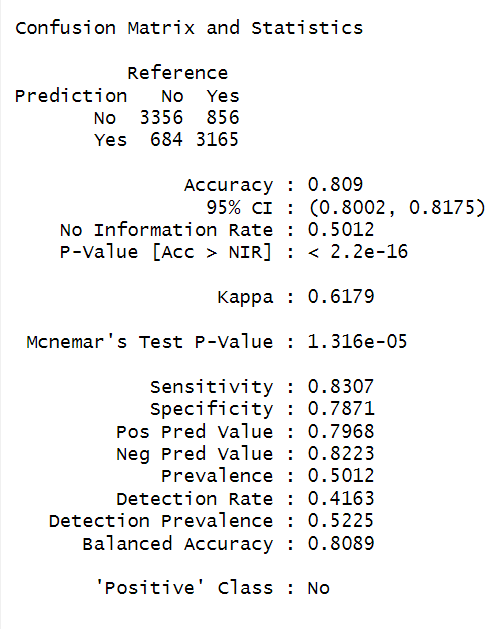
Overall, the model produced a similar output with a slightly lower accuracy score, so we kept the logistic regression model. We also tried a 5-fold cross-validation to assess the performance of the model and got the following output.

A screenshot of a computer screen

Description automatically generated

As can be seen by the accuracy metric, using **k-folds** did not increase the accuracy score, but did increase the time it took to fit the model using the subset version of our data.

Lastly, we tried to **under sample** the “No” response, by identifying the major and minority class, calculating the numbers of samples in each class, **undersampling** the “No” class (i.e., the “No” response), combining the undersampled majority class with the minority class, shuffling and splitting the newly created dataset (i.e., the dataset with the undersampled “No” response). Afterwards, the model was fitted, and its accuracy and confusion matrix were printed. The following is the output from **undersampled “No”** **model**:



Summary

In summary, our project focuses on predicting heart attacks using advanced data analytics and machine learning techniques, aiming to forecast the likelihood of heart failure occurrences among at-risk populations. Through analysis of comprehensive datasets encompassing demographic factors, medical histories, and lifestyle choices, robust predictive models are developed to identify individuals predisposed to heart failure and pave the way for targeted interventions. Key business questions include the accuracy of heart attack prediction, identification of important variables, and the relevance of non-biological ailments like depression. Data preparation involves refining a dataset sourced from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), and applying machine learning techniques such as logistic regression and support vector machines. Descriptive statistics and visualizations provide insights into factors influencing heart attack likelihood, age demographics, patient health statuses, racial disparities, and correlation statistics revealing key variables associated with heart attacks.