**Prediction of Mortality Due to Heart Disease**

*Background:*

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Chart, scatter chart  Description automatically generated |
| Figures 1 & 2. |

Heart disease is an umbrella term that describes a variety of conditions that can impair cardiac function, diminish quality of life, and ultimately result in death. The Centers for Disease Control and Prevention reports that heart disease is “the leading cause of death for men, women, and people of most racial and ethnic groups in the United States”.1 From 2014 to 2015, heart disease was reported to result in $219 billion dollars in medical expenditures and diminished productivity.2 Of the most common forms of heart disease, myocardial infarctions and coronary artery disease (CAD) have incredibly high incidence with 805,000 myocardial infarctions occurring per year and CAD impacting an estimated 18.2 million Americans.3 Moreover, similar reports are provided worldwide. In Europe it was reported that 41% of deaths are due to either a heart attack or another circulatory disease in 2010 wherein Africa circulatory diseases were the third leading cause of death. [4-5] As this global concern continues to grow, many academic communities have begun looking towards new investigational and statistical techniques to facilitate the diagnosis and betterment of clinical treatments to prevent and treat all branches of heart disease.

*Project Focus & Approach:*

This project aimed to implement Logistic Regression, a fundamental machine learning

algorithm, to predict the mortality of patients as a result of heart disease as derived from clinical metrics. Clinical data was obtained from an online database and data was provided by BMC Medical Informatics and Decision Making in which key clinical metrics associated with heart disease were documented.6 The initial investigation into the data began with an exploratory analysis to identify features that have a significant mean difference between the alive and dead groups. Many features, such as high blood pressure, smoking status, and creatinine phosphokinase level, were reviewed through graphical inspection, but did not display any clear-cut pattern of classification between the two groups of interest. However, visual representation of serum creatinine levels and ejection fraction, with respect to the death event variable, demonstrated a rough distinction (Figure1.). This distinction was slightly improved when age was incorporated into the plot (Figure 2.). Independent two-sample t-tests were then conducted to provide additional evidence as to which metrics had significantly different means between the alive and dead groups. After completion of the exploratory data analysis, a supervised logistic regression algorithm was created.

*Implementation of Logistic Regression:*

The logistic regression algorithm is used for classification analysis and yields predicted values within the range of 0 to 1. This predictive range is a result of the logistic regression utilizing a cost function that employs the sigmoid function (Figure3.), which limits the range of the cost function output to between 0 and 1. Choosing the logistic regression algorithm was the logical choice for the predictive analysis of this data as the outcome of interest to be predicted, death event, was a binary variable. This meant that the implemented algorithm needed to predict a value of either 0 or 1, alive or dead. The algorithm was initialized with a set learning rate (alpha) of 0.001 and a set number of iterations (1000). The cost function within the algorithm was then trained on the training data, employing gradient descent to minimize the cost function (J(θ)). Bias and weighted coefficients of the sigmoid function were initialized at 0 but were adjusted after each iteration in accordance with the closeness of the prediction and the set learning rate. After gradient descent was completed and the model trained, the testing data was fed into the algorithm which used the adjusted (trained) sigmoid weight and bias to yield a probability. If the resulting probability was greater than or equal to 0.5, the predicted value became 1. Conversely, if the observed probability was less than 0.5, the predicted value became 0. The logistic regression algorithm ultimately returns a vector of predicted outcome values which were derived from the data provided to the model. This vector of predicted outcomes was then compared against the testing data, the real clinical outcomes. The accuracy of the model was determined by how many outcomes the algorithm correctly predicted as is reported as a percentage.

*Tools & Modules Utilized:*

This project utilized a multitude of modules and packages to facilitate and accomplish the analysis. Manipulation of data primarily relied on pandas data frame features for the organization and selection of data. Additionally, sklearn’s Bunch feature enabled the passing of organized numpy arrays into the logistic algorithm function. Sklearn was also utilized for the train\_test\_split function which facilitated splitting the heart disease data set into training and testing data. The logistic regression relied solely on python’s numpy module. Data visualization was carried out through the plotnine module which provided ggplot functionality in python. Statistical analysis was carried out utilizing pandas built-in mean function and scipy.stats. Lastly, the os module was utilized to set the working directory and csv was used to import and export csv files. All data analysis and coding were carried out in Visual Studio Code utilizing python 3.8.2.

There were 3 python scripts created for the analysis of this project. Two are supporting scripts which contain the function for the logistic regression and the function for formatting a csv file into the appropriate numpy array to be used in the algorithm. The last script calls the functions from the supporting scripts and carries out pandas data frame manipulation, plot generation, exploratory statistics, and implementation and accuracy check of the logistic regression.

*Results:*

|  |  |  |
| --- | --- | --- |
| **Table 1.** | | |
| **Metric** | **Unadjusted p-value** | **Bonferroni Corrected p-value** |
| Ejection Fraction | 2.45x10^-06 | 1.72x10^-05 |
| Creatinine Phosphokinase | 0.28 | 1.96 |
| Serum Creatinine | 2.19x10^-07 | 1.53x10^-06 |
| Age | 8.92x10^-06 | 6.24x10^-05 |
| Serum Sodium | 6.89x10^-04 | 4.82x10^-03 |
| Platelets | 0.40 | 2.80 |
| Time | 9.12x10^-23 | 6.38x10^-22 |

|  |  |
| --- | --- |
| **Table 2.** | |
| **Logistic Model** | **Predictive Accuracy** |
| Model 1 | 0.81 |
| Model 2 | 0.27 |
| Model 3 | 0.83 |
| Model 4 | 0.85 |

Multiple independent two-sample t-tests were performed. As multiple comparisons were conducted using the same dependent variable, Bonferroni corrected p-values were calculated. Analysis showed that age, ejection fraction, serum creatinine, serum sodium, and time had statistically different means between the alive and dead groups. Creatinine phosphokinase and platelets levels were found not to be statistically significant between the alive and dead groups. The features that were passed into the logistic regression algorithm were determined by the significance found above. Logistic model 1 was found to have a predictive accuracy of 0.81, model 2 was found to have a predictive accuracy of 0.27, model 3 was found to have a predictive accuracy of 0.83, and model 4 was found to have a predictive accuracy of 0.85.

*Discussion:*

|  |  |
| --- | --- |
| **Table 3.** | |
| **Logistic Model** | **Included Features** |
| Model 1 | Ejection Fraction, Serum Creatinine |
| Model 2 | Age, Ejection Fraction, Serum Creatinine |
| Model 3 | Age, Ejection Fraction, Serum Creatinine, Serum Sodium, Time |
| Model 4 | Ejection Fraction, Serum Creatinine, Serum Sodium, Time |

The predicted behavior for the first logistic model followed an expected trajectory as both features shared a statistically significant difference between the alive and dead groupings. From these significance levels it was expected for the model to return a relatively high accuracy rate when utilizing ejection fraction and serum creatinine. However, the drastic reduction in accuracy in model 2 that occurred as a result of incorporating age was unexpected. The addition of age to Figure 1 appeared to accentuate the classification pattern seen in the plot (Figure 2). Additionally, mean age was found to be significantly different between the alive and dead groups. Despite this, the addition of the age feature to logistic regression model 2 impaired the predictive ability of the algorithm. This dampening of predictive power due to the age feature appeared to be offset when the remaining statistically significant continuous variables were incorporated into the model. This led to an increase in predictive accuracy raising model 3 to 0.83. In response to the observed interaction between the logistic regression algorithm and the age feature, model four was created to include all statistically significant features excluding age. This confirmed the researcher’s suspicion of the negative impact of the age feature on the model as predictive accuracy of model 4 had the highest accuracy of all models with a predictive accuracy of 0.85. Based on the results of this study it is apparent that a logistic regression model could be used to accurately predict death due to heart disease within a reasonable degree of error.

*Limitations & Future Projects:*

There are many limitations on the logistic regression model employed in this project. The first limitation on this logistic regression is that this algorithm does not correctly interpret categorical or binary variable features. By implementing the correct handling of binary and categorical variables within the logistic regression function the algorithm will be able to train on a greater number of training data sources. With access to more data features, the model will have more examples to train on which could potentially increase predictive accuracy. The second limitation of this study is that there are only 7 continuous variable features included in this dataset. In the realm of clinical research there are a multitude of clinical metrics that could be associated with heart disease. Another method to improve this logistic regression algorithm would be to increase the number of features and training examples to train the algorithm on. The last limitation is to fully understand why age negatively impacts the predictive accuracy of the logistic regression model and to fully replicate the statistics calculated for this report to ensure overall accuracy.

*References*

1. Centers for Disease Control and Prevention. [Underlying Cause of Death, 1999–2018](http://wonder.cdc.gov/ucd-icd10.html). CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.
2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. [Heart disease and stroke statistics—2019 update: a report from the American Heart Associationexternal icon](https://doi.org/10.1161/cir.0000000000000659). *Circulation.* 2019;139(10):e56–528.
3. Fryar CD, Chen T-C, Li X. [Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K]](https://www.cdc.gov/nchs/data/databriefs/db103.pdf). NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.
4. European Public Health Alliance. 2010 7-February-2011; Available from: http://www.epha.org/a/2352
5. Statistics South Africa. 2008 7-February-2011]; Available from: http://www.statssa.gov.za/publications/P03093/P030932006.pdf
6. Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

*Out-of-Class Learning:*

1. Stanford University – Andrew Ng – Machine Learning (1.1-6.7)
2. WolframMathWorld – Sigmoid Function

***Appendix; Figures:***

|  |
| --- |
| Diagram  Description automatically generated |
| Figure 3. Sigmoid function |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 4. |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 5. |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 6. |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 7. |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 8. |

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 9. |