Robust Cardiovascular Disease Prediction Using Logistic Regression

Snigdha Datta

Colorado State University Global

MIS500: Foundations of Data Analytics

Dr. Isaac Gang

July 05, 2019

Abstract

Cardiovascular disease, commonly known as heart disease, is one of the leading causes of death in the United States as it is worldwide. Early detection of the disease can save thousands of lives and billions of dollars in healthcare costs. A statistical model with the ability to accurately predict heart disease could be of immense help to the patients, their families, the medical community, and the healthcare system. Hospitals and providers collect many patient health metrics during screening and routine lab tests. Such a sample dataset from the University of California, Irvine Machine Learning repository is used in this experiment to develop a robust heart disease prediction model. Sets of Initial Hypotheses are formulated, and the most significant predictor variables are identified using the Wald test. The statistical significance of the proposed model is tested using the Likelihood-Ratio test. Keeping in mind the simplicity, usability, and explainability of results to the medical community, a Logistic Regression model that predicts the heart disease class with a high degree of accuracy is presented in this paper.

Introduction

According to the CDC, the term "heart disease" refers to several heart conditions, the most common of which is coronary artery disease, a major cause of heart attacks. Other related conditions involve the inability of the heart to pump blood efficiently due to malfunction of the heart valves resulting in heart failure. While some people are born with heart disease, it is essential to point out that anyone, including children, can develop heart disease. The condition occurs when plaque builds up in the arteries causing them to narrow over time, reducing blood flow to the heart. Habits such as smoking, eating an unhealthy diet, and not getting enough exercise increase the risk of having heart disease. The recent CDC data suggests that heart disease causes 1 in 4 deaths in the United States, equating to about 655,000 lives and $219 billion each year, including healthcare services, medicine, and loss in productivity. Essentially, a person dies every 36 seconds due to heart disease in the United States (CDC, 2020).

The symptoms of underlying heart disease are not always visible. It may go undetected for a prolonged period causing irreversible damage to the human body. One of the biggest challenges while dealing with heart diseases is the associated medical costs exacerbated by late detection. Since early detection reduces medical costs and saves lives, efforts to promote it are critical and one of the key goals for this paper. Many important metrics collected during screening and routine lab tests can be extracted and utilized to find hidden patterns in the data and lead to early detection of the disease. The emergence of Machine Learning (ML) and Artificial Intelligence (AI) coupled with the availability of quality datasets has presented us with an opportunity to make early detection possible algorithmically. To that end, I surveyed similar efforts in Section 2. I discuss the methodology in detail in Section 3 and present the results with relevant analysis in Section 4. I conclude my work with future direction in Section 5 while providing references in Section 6.

Background

Modeling and predicting cardiovascular disease algorithmically may prove very helpful in the early detection of heart disease, saving thousands of lives and billions of dollars in healthcare costs. Leveraging ML methods have been deemed valuable in this regard by many authors before us, and they applied several such techniques to address the problem. For example, Fredrick David & Belcy (2018) used data mining techniques to build multiple classification models for predicting the heart disease class. The UCI heart disease dataset was used to compare the results of three algorithms- Random Forest (RF), Decision Tree, and Naïve Bayes (NB). At 81% accuracy, the RF result was the best among the three. Leveraging ML methods have been deemed valuable in this regard by many authors before us, and they applied several such techniques to address the problem. For example, Fredrick David & Belcy (2018) used data mining techniques to build multiple classification models for predicting the heart disease class. The prediction results of a two-component concomitant variable Poisson mixture regression model were better than the standard model. However, the best performing model with the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values was the zero-inflated Poisson mixture regression model. Hassani et al. (2020) applied Hybrid Neural Network and Decision Tree using the UCI datasets and achieved 87.3 precision. Singh et al. (2018) used the WEKA Data Mining tool to implement an Artificial Neural Network (ANN) based heart disease prediction system. A Multi-layer Perceptron Neural Network model with backpropagation was trained and tested using a 60-40 train-test split heart disease dataset containing 303 records. The experimental results show that the system could predict heart disease with almost 100% accuracy. Jan et al. (2018) also used the WEKA Data Mining tool to create an intelligent heart disease prediction system. An ensemble model combining Support Vector Machine (SVM), ANN, NB, Logistic Regression, and RF classifiers was created to predict cardiovascular disease recurrence. The UCI Cleveland and Hungarian datasets were used in the analysis. The RF ensembled model achieved the highest accuracy of 98.17%. Ibrahim et al. (2019) leveraged adversarial learning to design a “fair” model to distribute therapies across race and gender groups equitably. Fernandez-Lozano et al. (2018) and Sajeev & Maeder (2019) used a Generalized Linear Model to predict complications in peritoneal dialysis patients. They surveyed PubMed for relevant prediction models finding 229 articles in total. Usman (2018) highlight the inherent problem with feature selection while building a heart disease prediction model. Two slightly different cuckoo-inspired algorithms, the cuckoo search algorithm (CSA) and the cuckoo optimization algorithm (COA), were used for feature selection on multiple datasets. The reduced features were used to build, train, and test four classification algorithms: NB, RF, Multi-layer Perceptron, and SVM. The experimental results show that CSA performed better than COA for minimal feature selection. SVM achieved the highest classification accuracy rate on the Eric and Hungarian, and Stat log datasets. Khateeb & Usman (2017) achieved 80% accuracy with their prediction model using K-NN 14 data attributes.

Methodology

The ML algorithm of choice in this work is Logistic Regression (LR). LR isn’t regression at all. In fact, it is a parametric classification technique, and thus it perfectly suits this use case. Instead of a standard linear function of a straight line, it models using a sigmoid function to shrink the domain of possible predictor values between 0 and 1 compared to a standard linear function that allows values from -∞ to + ∞ (James et al., 2013). For multiple predictors, an LR can be mathematically represented as:

$$p(X) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}; \, where\ X = (x_1, x_2, \ldots, x_n)\ for\ n\ predictors. \quad (1)$$

It can also be represented as a *log-odds* or *logit* function:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n\ ;\ \beta_0 \ldots \beta_n\ are\ the\ parameters\ of\ the\ function. \quad (2)$$

The parameters of the equation are estimated using the Maximum Likelihood function. The goal is to estimate the values for parameters $\widehat{\beta_0}, \widehat{\beta_1} \ldots \widehat{\beta_n}$ such that the $p(X)$ is close to 1 for the positive class and close to 0 for the negative class (James et al., 2013). Mathematically, the Maximum Likelihood function can be represented as:

$$\ell(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i:\, y_i=1} p(x_i) \prod_{i':\, y_{i'}=0} (1 - p(x_{i'})) \quad (3)$$

Unlike complex ANN-based connectionist models, the results of an LR are easier to explain. It also takes a lot less data to train such models. Since the purpose of building such a model is to aid the medical community in the early detection of heart disease, the explainability of results is crucial. To trust the model results, the medical community expects to clearly understand why the model predicted what it predicted.

This experiment will be divided into five distinct phases and follow a standard Data Analytics Lifecycle (EMC Education Services, 2015). The high-level process flow and critical subtasks for each phase are shown in Figure 1. The Identification of data source, programming resources, preliminary data analysis, framing of the correct problem statement, formulating a set of Initial Hypotheses (IHs), and defining the success and failure criteria for the project are performed in the Data Discovery phase. The data cleansing and conditioning, summary, and visual statistical analysis are performed in the Data Preparation phase to get additional insight into the nature of the data. A more detailed analysis of the potential predictor variables using various descriptive statistical analysis techniques is performed in the Model Planning phase. Moreover, the hypotheses testing and model selection are also performed in this phase. In the Model Building phase, the initial dataset is split into a training and a testing dataset. The training dataset is used to train the selected statistical model, and the testing dataset is used to test its ability to predict the outcome class correctly. Finally, all the findings and inferences from the experimental results are communicated.

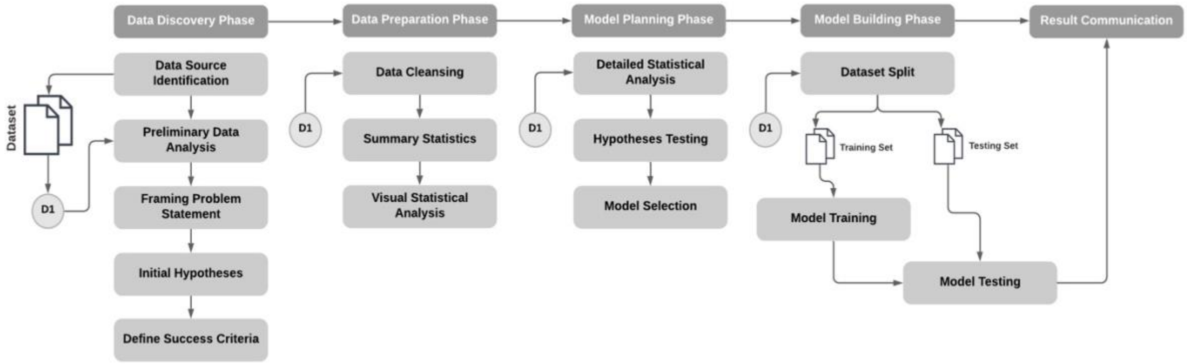


Figure 1: High-Level Process Flow

Analysis and Results

Data Discovery

The Cleveland Heart Disease dataset (Detrano, 1988), from the University of California Irvine Machine Learning Repository's publicly available large datasets, is used for this experiment. The dataset contains sample data collected from VA Medical Center, Long Beach, and Cleveland Clinic Foundation. Though the original dataset contains 76 data elements, the processed dataset which is being used contains only 14 selected and widely used data elements. The variable names, variable data types, and descriptions have been consolidated in a tabular format below.

It is essential to perform a preliminary data analysis to become familiar with the data content, quality, limitations and understand any interdependencies among the data elements. The initial analysis results show that the dataset contains 303 rows and 14 columns. 8 out of the 13 predictor variables are categorical, and 5 are continuous. The outcome variable is categorical.

Table 1: Data element description

Variable Name	Variable Type	Description
Age	Integer	Age in years.
Gender	Integer	Gender (Values: 1 = male; 0 = female).
CP	Integer	Chest pain type (Values: 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic).
Trestbps	Integer	Resting blood pressure (Measured in mm Hg on admission to the hospital).
Chol	Integer	Serum cholesterol measured in mg/dl.
FBS	Integer	Fasting blood sugar > 120 mg/dl (Values: 1 = true; 0 = false).
RestECG	Integer	Resting electrocardiographic results (Values: 0 = normal; 1 = having ST-T; 2 = hypertrophy)
Thalach	Integer	Maximum heart rate achieved.
Exang	Integer	Exercise induced angina (Values: 1 = yes; 0 = no).
Oldpeak	Number	ST depression induced by exercise relative to rest.
Slope	Integer	The slope of the peak exercise ST segment (Values: 1 = up-sloping; 2 = flat; 3 = down-sloping).
CA	Factor	The number of major vessels (Values: 0-3) colored by fluoroscopy.
Thal	Factor	Thallium stress test result (Values: 3 = normal; 6 = fixed defect; 7 = reversible defect.
Num	Integer	The predicted attribute - diagnosis of heart disease (angiographic disease status) (Values: 0 = < 50% diameter narrowing; 1 = > 50% diameter narrowing).

Problem Statement and Initial Hypotheses

Formulating a set of IHs is crucial during the Data Discovery phase of the Data Analytics Lifecycle. Learning about the data sources, their domain, and framing the right problem the analytical system is trying to solve are essential before the hypotheses can be formulated. Understanding the data domain provides the proper context to comprehend the data's characteristics and a meaningful way to interpret it. The proposed analytical system is attempting to predict the heart disease class using one or more predictor variables. The results of this experiment must be able to answer the following questions:

1. Is there a statistically significant relationship between one or more predictor variables and the outcome class?
2. Can an acceptably accurate heart disease prediction model be designed using the selected predictor variables?

Based on the preliminary data analysis, it is likely that more than one significant predictor is contributing to the increased risk of heart diseases. The below IHs are formulated using the preliminary data analysis results:

- Null Hypothesis(H_0): There is no statistically significant relationship between one or more predictor variables and the outcome class.
- Alternate Hypothesis (H_a): There is a statistically significant relationship between one or more predictor variables and the outcome class.

For the purposes of this experiment, an acceptable model is required to achieve at least 80% accuracy in correctly predicting the positive class. The success criteria of this experiment are accordingly set.

Data Preparation

A detailed data exploration, data cleansing, conditioning, and visual statistical analysis on the dataset is performed. The statistical analysis results will help identify the right set of predictor variables for building an accurate statistical model. A working data frame is created using the original dataset. A new categorical variable- *HeartDisease* is created from the original outcome variable- *Num*. The levels: 0,1,2,3, and 4 of the original variable are re-coded to 0 and 1. Meaningful labels are added to the re-coded values- *Healthy* for 0 and *Heart Disease* for 1. As the original outcome variable, *Num* is no longer required, it is dropped from the dataframe. The variables *Gender, CP, FBS*, *RestECG, Exang,* and *Slope* with discrete numerical values are re-coded as categorical and meaningful labels are added. The variables *CA* and *Thal* are already categorical. However, they have missing values in specific categories. The missing values are removed from the dataframe. The dimensions of the cleansed dataset are validated. It is observed that 6 rows having NULL values are successfully removed from the data frame. The cleansed dataset has 14 variables, including 1 outcome variable, *HeartDisease*. 6 variables, which were initially quantitative and discrete, are successfully re-coded as categorical variables. The dataframe now contains 9 categorical variables, including the outcome variable, *HeartDisease*. As expected, 5 variables remain quantitative and continuous.

A detailed visual statistical analysis on all the dataset variables is performed. *Age, Trestbps, Chol, Thalach,* and *Oldpeak* are continuous numerical variables; Boxplot is used for analyzing the critical features of the distribution, such as overall centrality, spread, and skewness. Bar plot, a common visualization technique for qualitative data, is used to visualize the characteristics of *HeartDisease* and the rest of the categorical predictor variables.

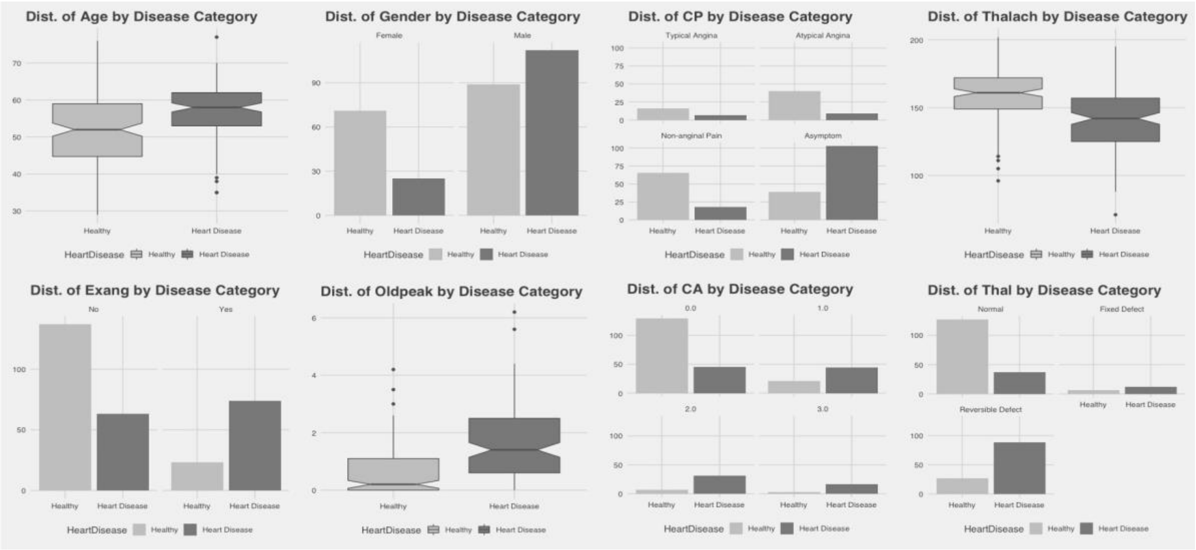


Figure 2: Visualization results for Heart Disease predictors

The dataset contains more data in the healthy category data than in the heart disease category, indicating a slight skewness. Outliers are detected for the variables- *Age, Tresbps, Chol, Thalach,* and *Oldpeak*. The following are the inferences drawn based on data exploration and visual statistical analysis results:

1. The predictor variables, *Tresbps, Chol, FBS*, and *RestECG,* do not seem to have any significant relationship or compelling factor contributing to the increased risk of heart diseases.
2. Further analysis needs to be performed on the remaining predictor variables- *Age, Gender, CP, Thalach, Exang, Oldpeak, Slope, CA,* and *Thal*.

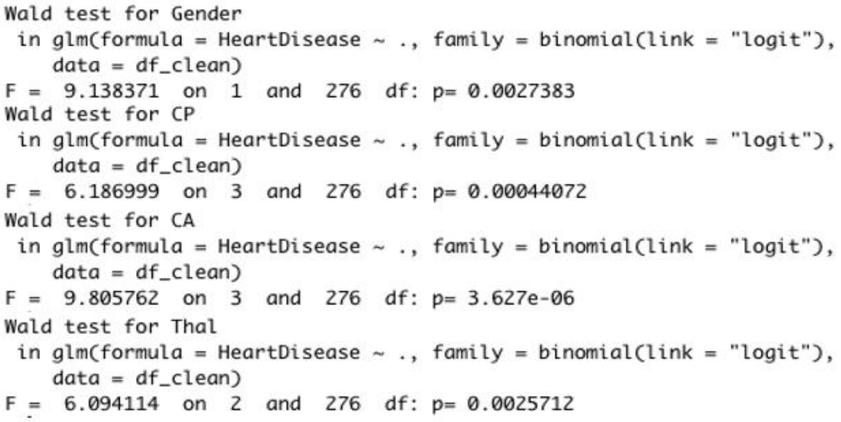
Model Planning

A more detailed statistical analysis on all the potential predictor variables is performed during the Model Planning phase. Understanding the data properties, relationships with other variables, and their significance in influencing the outcome variable class will determine the list of significant predictors. The Wald test for individual predictors and the Likelihood-Ratio (L-R) test for the Logistic Regression model will test the IHs. The goal is to reject the NULL Hypothesis and accept the Alternate Hypothesis. The most significant predictor variables and an appropriate statistical model will be used in the next phase for model building. The full model parameter estimates are generated using the Generalized Linear Model (GLM) in R. Since the outcome variable is categorical and has only two levels, the binomial family function, *logit* is specified. Important statistics, such as the p-values of each predictor's parameter estimates, are recorded and analyzed. Based on the p-values, it can be inferred that the variables, *Age, Gender, CP, Tresbps, Thalach, Slope, CA,* and *Thal* are probable significant variables. The relative importance of individual predictor variables is assessed using the *varimp* function of R.

Finally, the Wald test for predictor variables is performed to evaluate the statistical significance of each coefficient in its ability to influence the outcome variable of the model. Based on the results from variables of importance and the Wald test, only the most significant variables whose p values are <= 0.05 are being considered as predictor variables. These results will also be used during hypothesis testing. The correct statistical testing techniques and modeling methods are determined by the characteristics of the predictor and outcome variables. There are 4 predictor variables, and they are all categorical. The outcome variable, *HeartDisease,* is also categorical. There are two distinct groups of data in the outcome variable, *Healthy*, and *Heart Disease*. So, the outcome variable is also binomial. LR is a suitable modeling method as both the predictor variables and the outcome variable are categorical.

Hypotheses Testing

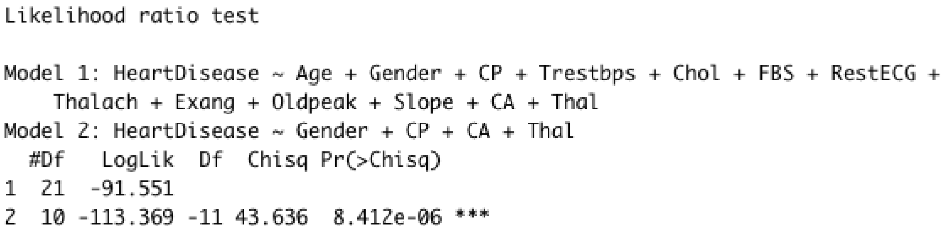
The Wald test results for individual predictors are used to select the most significant predictor variables for the LR model. It is observed that regression coefficients of only the variables- *Gender, CP, CA,* and *Thal* have p-value <= 0.05. All other predictor variables can be removed from the model.



```
Wald test for Gender
 in glm(formula = HeartDisease ~ ., family = binomial(link = "logit"),
    data = df_clean)
F =  9.138371  on  1  and  276  df: p= 0.0027383
Wald test for CP
 in glm(formula = HeartDisease ~ ., family = binomial(link = "logit"),
    data = df_clean)
F =  6.186999  on  3  and  276  df: p= 0.00044072
Wald test for CA
 in glm(formula = HeartDisease ~ ., family = binomial(link = "logit"),
    data = df_clean)
F =  9.805762  on  3  and  276  df: p= 3.627e-06
Wald test for Thal
 in glm(formula = HeartDisease ~ ., family = binomial(link = "logit"),
    data = df_clean)
F =  6.094114  on  2  and  276  df: p= 0.0025712
```

Figure 3: Wald test results for *Gender, CP, CA,* and *Thal*.

L-R test is a suitable hypothesis test for LR models. It compares the likelihood of fit of the proposed model compared to a full model. In R, either the *lrtest* or the Analysis of Variance (ANOVA) function can be used for comparing the two models. The full model represents the NULL Hypothesis, $H_{0,}$ and the proposed model represents the Alternate Hypothesis, H_a. It is necessary to test if the observed difference in model fit is statistically significant. Since the p-value corresponding to the likelihood ratio chi-square statistic of the proposed model is <= 0.05, it can be safely assumed that the proposed model is statistically significant and fits the data better than a full model. The result is used to reject the NULL Hypothesis(H_0) and accept the Alternate Hypothesis (H_a).



```
Likelihood ratio test

Model 1: HeartDisease ~ Age + Gender + CP + Trestbps + Chol + FBS + RestECG +
    Thalach + Exang + Oldpeak + Slope + CA + Thal
Model 2: HeartDisease ~ Gender + CP + CA + Thal
  #Df   LogLik  Df  Chisq Pr(>Chisq)
1  21  -91.551
2  10 -113.369 -11 43.636  8.412e-06 ***
```

Figure 4: LR test comparing the full model reduced model.

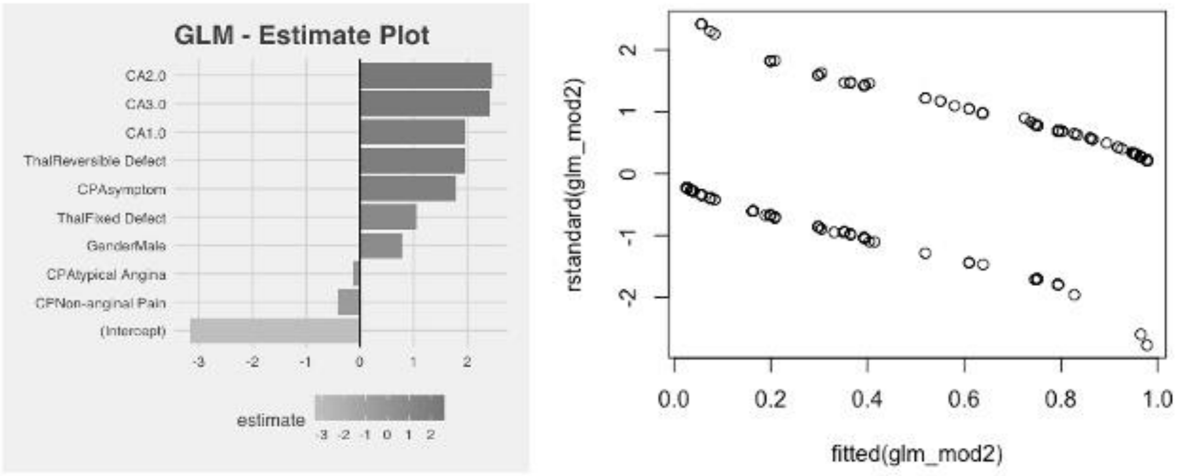


Figure 5: GLM estimate plot and standard vs. fitted model plot

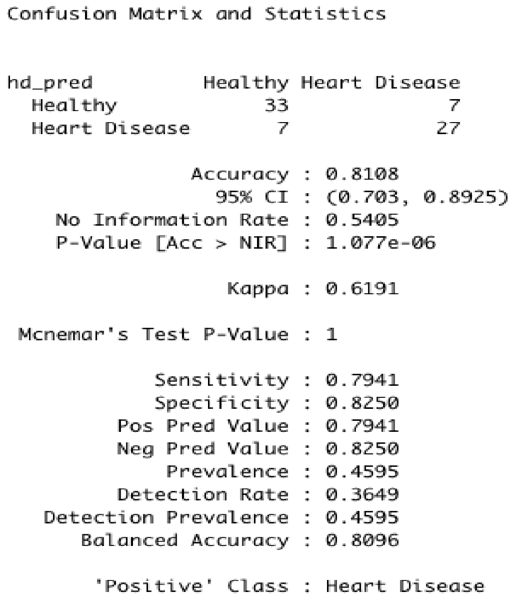
Model Building, Training, and Testing

The working dataset is first split into a training and a testing dataset. The training dataset has 75% of the data (223 rows), and the testing dataset has the rest 25% of the data (74 rows). The training dataset is used to train the proposed LR model. A repeated K-fold cross-validation technique is used to customize the training control parameters, and the training data is partitioned into K equal-sized partitions called "folds." The K-1 folds are used to train the model, while one-fold is used for evaluating the accuracy of the training. For this experimental setup, a 10-fold cross-validation technique is used. The Receiver Operating Characteristic (ROC), Sensitivity, and Specificity values of the trained model are recorded. The testing dataset is used to evaluate the ability of the model to correctly predict the outcome when new, never-seen data is presented to the model. The statistical results such as Accuracy, Kappa, p-value, and the confusion matrix results are recorded as shown in Figure 6.

Results

Based on the Wald test results for individual predictors and the L-R test for the LR model, the Alternate Hypothesis (H_a) was accepted by successfully rejecting the NULL Hypothesis. A Multiple Logistic Regression (MLR) model with 4 predictor variables, *Gender, CP, CA,* and *Thal,* predicted the *HeartDisease* category with 81% accuracy.

The complete R code for the experiment has been uploaded to GitHub and publicly shared. The code can be accessed via the URL- https://github.com/dattasd/analytics-ai-ml.git



```
Confusion Matrix and Statistics

hd_pred         Healthy Heart Disease
  Healthy            33             7
  Heart Disease       7            27

               Accuracy : 0.8108
                 95% CI : (0.703, 0.8925)
    No Information Rate : 0.5405
    P-Value [Acc > NIR] : 1.077e-06

                  Kappa : 0.6191

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7941
            Specificity : 0.8250
         Pos Pred Value : 0.7941
         Neg Pred Value : 0.8250
             Prevalence : 0.4595
         Detection Rate : 0.3649
   Detection Prevalence : 0.4595
      Balanced Accuracy : 0.8096

       'Positive' Class : Heart Disease
```

Figure 6: Confusion Matrix and other statistical results of the final model.

Conclusion

In this paper, I presented a Logistic Regression model that successfully predicts cardiovascular disease with an 81% accuracy, which is comparable to and, in most cases, outperforms models that use similar setup and data splits. To the best of my knowledge, this work is the only one that recommends LR as the solution. Compared to the black-box nature of complex ANN-based models, the results of LR are far more explainable. It also takes a lot less data to train such models. The purpose of this model is to help the medical community make accurate and early heart disease predictions. Though model accuracy is a critical factor, the simplicity, usability, and explainability aspects should not be neglected. Only then can AI and ML-based solutions such as this be trusted, accepted, and widely used. In the future, I plan to investigate hybrid techniques and models with the hope of achieving greater than 90% accuracy on the same dataset and using similar data splits.

References

[1] CDC. (2020, September 8). *Heart disease facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

[2] Davies, T. M. (2016). *The book of R: A first course in programming and statistics*. No Starch Press.

[3] Detrano, R. (1988). *UCI machine learning repository: Heart disease data set*. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] EMC Education Services. (2015). *Data science and big data analytics: Discovering, analyzing, visualizing and presenting data.* John Wiley & Sons.

[5] Fernandez-Lozano, C., Valente, R. A., Díaz, M. F., & Pazos, A. (2018). A generalized linear model for cardiovascular complications prediction in PD patients. *Proceedings of the First International Conference on Data Science, E-learning and Information Systems*. https://doi.org/10.1145/3279996.3280039

[6] Fredrick David, H. B., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *ICTACT Journal on Soft Computing*,*9*(01), 1817-1823. https://doi.org/10.21917/ijsc.2018.0253

[7] Hassani, M. A., Tao, R., Kamyab, M., & Mohammadi, M. H. (2020). An approach of predicting heart disease using a hybrid neural network and decision tree. *Proceedings of the 2020 5th International Conference on Big Data and Computing*. https://doi.org/10.1145/3404687.3404704

[8] Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Global explanations of neural networks. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314230

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer Science & Business Media.

[10] Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology*, *9*, 33-45. https://doi.org/10.2147/rrcc.s172035

[11] Khateeb, N., & Usman, M. (2017). Efficient heart disease prediction system using k-nearest neighbor classification technique. *Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017*. https://doi.org/10.1145/3175684.3175703

[12] Mufudza, C., & Erol, H. (2016). Poisson mixture regression models for heart disease prediction. *Computational and Mathematical Methods in Medicine*, *2016*, 1-10. https://doi.org/10.1155/2016/4083089

[13] Sajeev, S., & Maeder, A. (2019). Cardiovascular risk prediction models. *Proceedings of the Australasian Computer Science Week Multiconference*. https://doi.org/10.1145/3290688.3290725

[14] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*, *13*, 121-124. https://doi.org/10.2147/ijn.s124998

[15] Usman, A. M., Yusof, U. K., & Naim, S. (2018). Cuckoo inspired algorithms for feature selection in heart disease prediction. *International Journal of Advances in Intelligent Informatics*, *4*(2), 95. https://doi.org/10.26555/ijain.v4i2.245