Predicting the risk of Heart Failure using Generalized Linear Modelling

Rijo Kuruvilla (14171011)

Abstract

Heart Failure (HF) affects over 23 million people worldwide and occurs most commonly in elderly people [1]. Heart Failure (HF) is a chronic progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen [2]. It was considered an emerging epidemic and assessing the factors that causes it became an essential task for the medical community. The purpose of this report is to analyze the dataset, extract the best features from it, and predict the risk of heart failure using supervised classification in conjunction with generalized linear modelling. The Coronary artery disease dataset contains the patient's information, heart (artery) characteristics, and health history. The techniques used to predict the risk of heart failure include exploratory data analysis (EDA), data preprocessing, and feature selection, as well as generalized linear modelling with a variety of link functions. The findings of the study showed that generalized linear modelling with the logit link function produced the best results in determining the factors that increase the risk of heart failure in patients. The analysis concludes that the data at hand is excellent in predicting patients' chances of developing heart failure and yield excellent accuracy results. To identify additional underlying factors that might contribute to heart failure, this study can be further investigated using more complex machine learning models.

Introduction

Heart failure is a clinical syndrome caused by structural and functional defects in the myocardium that cause blood ejection to be impaired [3]. It is most frequently caused by coronary artery disease, hypertension, and diabetes. Obesity, smoking, a family history of heart failure, and specific medical conditions are considered additional risk factors[4]. HF can severely decrease the functional capacity of the patient and increase mortality risk; therefore, it is crucial to effectively treat the disease and to improve the patient's life expectancy. "Across the globe, 17-45% of patients admitted to a hospital with heart failure die within one year of admission and the majority die within five years of admission"[5](Ponikowski et al., 2014). HF also results in a financial burden, not only to the patient but also to the hospital and the country's economy. "The overall economic cost of Heart Failure in 2012 was estimated at $108 billion per annum"[6] (Cook et al., 2014). This value continues to increase with the passing of time, rapidly expanding and industrialized global population. A vast proportion of the world has minimal regard for their health and fitness. A balanced diet coupled with a healthy lifestyle is essential for maintaining good health levels, thereby preventing life-threatening diseases like heart failure. Scientists and government officials all over the world continue to extensively collect, study and analyze features and conditions that can lead to HF. Similarly, this study aims to conduct EDA, extract essential features, and utilize generalized linear models to detect and to predict the risk of HF in patients.

Data

The dataset, 'DataClean-fullage.csv,'(coronary artery disease) has a sample size of 6611 observations and 53 variables (character = 7, numeric = 11, binary integers = 35) and is a subset of the original dataset ('HDHI Admissions Data.csv'). This data was gathered as part of an observational study over a two-year period (1 April 2017 to 31 March 2019) at Hero DMC Heart Institute, Unit of Dayanand Medical College, and Hospital in the district of Ludhiana, which is a state in Punjab, India. Both datasets were retrieved from Kaggle [7][8]. Because no specific sampling procedure is mentioned, it is assumed that the data was collected using simple random sampling. On further examining the data (conducting EDA), the dataset had no missing values as well as no duplicate values. The original ‘HDHI Admissions Data.csv’ was also examined and it consisted of 15,757 observations and 56 variables. This indicates that certain pre-processing steps were conducted on the original dataset. The variables, Cardiogenic Shock (Patients in shock due to cardiac reasons), Shock(Systolic blood pressure < 90 mmHg and when the cause for shock was any reason other than cardiac), Month of the year, Duration of stay, Admission number, Date of admission, Date of Discharge and Housing location were removed from the data. Also, the number of observations was reduced from 15,757 to 6611 with no explanation from the author. No justification for the variable’s elimination was provided, but it is safe to assume that they were irrelevant to the researcher who conducted the original study and that this assumption is also implemented in this study. For the number of reduced observations, it can be assumed it was done to eliminate the duplicated values to validate the assumption of independence between observations, as some supervised machine learning techniques require observations to be independent of each other. This, however, is not a recommended form of practice, as the loss of the data can affect the end results. In contrast to the high imbalance (28.95% - No heart Failure, 71.05% - Heart Failure) in the original dataset (HDHI Admissions Data), the response variable (heart failure) is fairly balanced in the pre-processed dataset (DataClean-fullage) with 52.22% of patients not having it and 47.78% having it (see Appendix B.3 for Frequency distribution plot). Creating a balanced frequency of the response variable might also be a reason for the reduced number of observations in the pre-processed data set. Before conducting the analysis, the variables "sno" (Serial Number of each patient) and "count" were eliminated from the dataset because they had no bearing on the prediction of heart failure. There was no justification or information provided for the variable "count," which consistently had a value of 1 for all observations. The final dataset on which the analysis was conducted consisted of 6611 observations and 51 variables. There are no known interventions, treatments, or grouping techniques that were applied to the data.

- **Methods**

The analysis for this study was carried out with R programming in RStudio studio – version ‘2022.02.3+492' (release name – Prairie Trillium). As mentioned in the ‘Data’ section, EDA was conducted to check for null values and duplicate values in the dataset and none were found. Further examination of the numerical predictors revealed a high correlation between 'Urea' and 'Creatinine'. Before proceeding with the analysis, the variable 'Urea' was kept, and the variable 'Creatinine' was removed from the dataset because it was proven that high blood urea levels are associated with an increased risk of heart failure [9]. Since the response variable is an ungrouped binary data (0,1), Generalized Linear Model (GLM) is used to conduct the analysis for detecting and predicting the risk of heart failure in patients. By allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of each measurement's variance to be a function of its predicted value, GLM generalizes linear regression[10].

The components that define the GLMs for ungrouped binary data are as follows:

i. Random Component: $\mathbf{yi \sim Bin(1, \pi i)}$, $\mathbf{E(yi) = \pi i}$, which means that ‘y’(response variable) follows a Binomial distribution and πi is the probability of having/not having heart failure.

ii. Linear Predictors: $\sum_{j=1}^{P} \boldsymbol{\beta_j x_j}$.

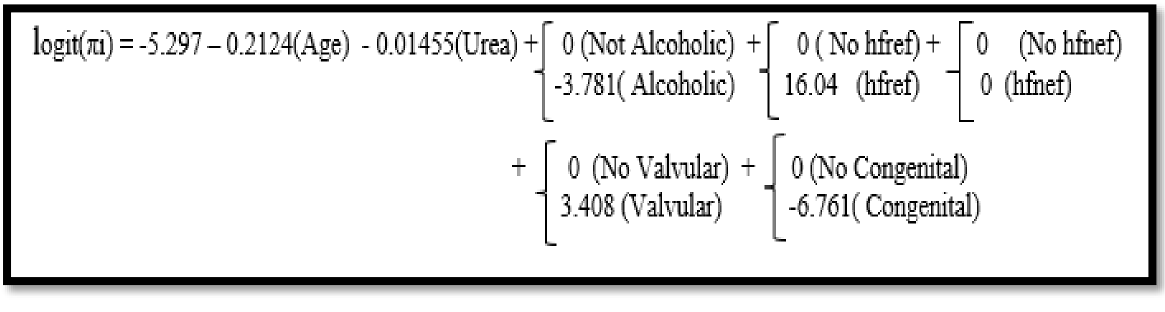
iii. Link Function g(.): The link function connects the linear predictor to the parameter in the distribution. Various Link functions such as logit, probit, and complementary log-log (clog-log) were used to model the data. The logit link function gave the best model results with an Akaike

Information Criteria (AIC) score of 181.02. The lower the AIC score, the better the model performance as AIC penalizes the models with more parameters. The logit link function is also a more feasible choice due to its ease of interpreting the model results. The logit link function for this model is denoted as **$g(\pi i) = \log(\pi i/1-\pi i)$**. After model selection, feature extraction was conducted by using logistic regression to extract the most important factors resulting in HF. The stepwise regression using backward selection was also deployed to extract the best features and to double-check the results of the forward selection. The dataset was split into training data (80%) and test data (20%) before applying logistic regression. Training data is used for feature analysis and model building**,** and test data is used to predict and determine model accuracy. The logistic regression assumptions were then verified**.** The assumption that the response variable is a binary (0,1) categorical variable was verified by examining the structure of the response variable**.** Variance inflation factor **(**VIF**)** was used to test the assumption of little or no multicollinearity between variables**.** VIF values above 10 indicate high collinearity, and VIFs between 5 and 10 indicate the need for further investigation. The VIF values for all important features were between 1.28 and 3. 8, indicating no collinearity between the variables.

The maximum likelihood parameter estimation method with the Newton Raphson iteration is used to estimate the model parameters in Logistic Regression (GLMs with logit link function is called Logistic Regression)[11](Febrianti et al., 2021) (see Appendix A for all mathematical equations of GLM). When fitting a binomial GLM with a logit link function, the sign (+ve or -ve) of the β(estimate) value indicates the direction of the association. $\beta > 0$ indicates a positive association between the predictor and the probability of the event (HF) to occur, while a $\beta < 0$ suggests a decrease in the probability. Potential problems such as checking for complete and quasi-complete separation as well as checking the effect of interaction among predictors were also performed before performing predictive analysis. The model results are discussed in the next section.

- **Results and Discussion**

The likelihood ratio test was carried out between the original model (model with all features) and the nested model (model with significant features) to evaluate the Deviance between the two models because Pearson statistics is not useful for evaluating the performance of logistic regression. Both models fit the data equally well, with a deviance value of -1.9761 and a p-value (0.1598) which is greater than the assumed level of significance (0.05) proved that the Null Hypothesis is true, and therefore the nested (less complex) model was utilized for building the final model. Age, alcohol use, urea levels, hfref (heart failure with reduced ejection fraction), hfnef (heart failure with normal ejection fraction), valvular heart disease, and congenital heart disease were found to be the significant features (see appendix C for all raw codes). The summary results of logistic regression were assessed, and the mathematical equation for the model is shown in FIGURE 1.

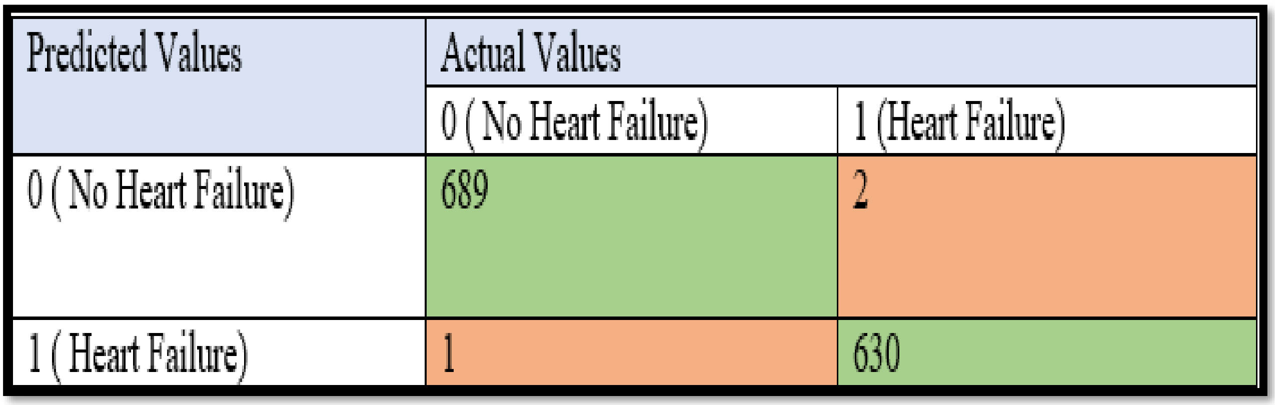


$$\mathrm{logit}(\pi i) = -5.297 - 0.2124(\text{Age}) - 0.01455(\text{Urea}) + \left[\begin{array}{l} 0\ (\text{Not Alcoholic}) \\ -3.781(\text{ Alcoholic}) \end{array}\right. + \left[\begin{array}{l} 0\ (\text{ No hfref}) \\ 16.04\ \ (\text{hfref}) \end{array}\right. + \left[\begin{array}{l} 0\ \ \ (\text{No hfnef}) \\ 0\ \ (\text{hfnef}) \end{array}\right.$$

$$+ \left[\begin{array}{l} 0\ \ (\text{No Valvular}) \\ 3.408\ (\text{Valvular}) \end{array}\right. + \left[\begin{array}{l} 0\ (\text{No Congenital}) \\ -6.761(\text{ Congenital}) \end{array}\right.$$

FIGURE 1: Mathematical Notation of Logistic Regression

On further analyzing the output, ‘hfnef1’ has a high standard error value, indicating quasi-complete separation. However, in this case, the 'do nothing' strategy was used, as the likelihood of other predictors are still valid [12]. As mentioned in the Methods section, the sign of β values, determine the association between the predictor and the likelihood of the event taking place. For quantitative variables, if the value of the predictor (x) variable increases by one unit, the log-odds of the response variable increase linearly by β . For categorical variables, dummy variables are used, by considering one of the levels of the predictor as the reference level. The response variable is then analyzed by how much it increases/decreases from the reference level. The reference level, in this case, is 0 (Not having a heart disease). For numerical variables ‘age’ and ‘urea’, if the age and the urea level of the patient decrease by 1 unit, the log odds of the event (HF) decreases linearly by 0.2124 and 0.01455. respectively. For the categorical variables, ‘alcohol’ and ‘congenital’, the odds of having HF (1), is $e^{-3\cdot781}$ and $e^{-6.761}$ exponentially lesser when a person is a non-alcoholic and does not have congenital heart disease (x = 0), while for ‘href’ and ‘valvular’, the odds of having HF(1), is $e^{+16.04}$ and $e^{+3\cdot408}$ times exponentially greater if the patient has ‘href’ and ‘valvular’ disease (x =1) respectively. The effect of interaction, between alcohol and urea, hfref and hfnef as well as valvular and congenital were studied. The interaction elements were found to have no impact on heart failure (the p-values were found to be not different from zero) and hence were not included in the analysis. A classification table and an Area under the Receiver Operating Characteristic Curve (AUC-ROC) were used to assess the predictive ability for the risk of HF. The outcomes are depicted in FIGURES 2 and 3, respectively. Logistic Regression predicted HF with an accuracy score of 99.77%. It was also replicated ten times using a for-loop, and the accuracy means, and standard deviation means were calculated across ten test sets, yielding 99.86% and 0.000946 respectively. The rows of the classification table (confusion matrix) represent what the algorithm predicted, while the columns represent the actual results. Logistic Regression correctly classifies 689 out of 700 cases with no HF (0) while misclassifying only 2 out of 632 classes with an HF (1). The model's recall (ability to correctly detect) or sensitivity score is 99.86%, while the precision (ability to correctly predict) score for HF detection is 99.68. 99.71% is the True Positive Rate (Specificity) rating. Like the confusion matrix, the AUC-ROC curve also produced an accuracy score of 99.90%, indicating a near-perfect model performance in determining HF.



Predicted Values	Actual Values	
	0 (No Heart Failure)	1 (Heart Failure)
0 (No Heart Failure)	689	2
1 (Heart Failure)	1	630

FIGURE 2: Classification Table

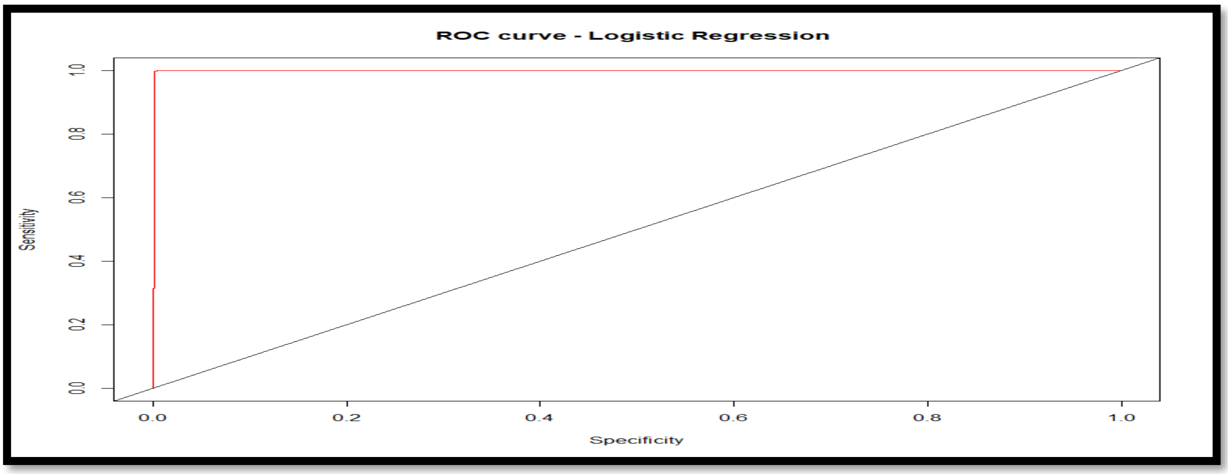


FIGURE 3: ROC curve

- <u>**Conclusion**</u>

A person's previous health history, body fluid content, as well as dietary habits, can be considered essential factors that lead to HF. It is crucial that a person maintains a healthy lifestyle as well as gets regular health checkups to minimize the risk of developing heart failure. Since today's generation does not place a great emphasis on their health, coupled with a lack of self-awareness, has seen a rise in HF cases. This report utilizes the health data ('DataClean-fullage.csv') and applies feature selection techniques (using both forward and backward methods) along with logistic regression to predict the risk of HF in patients. The procedure began with an EDA on the dataset, followed by the selection of the best link function and the extraction of significant features from a set of features to improve the precision and detection of HF in patients. The significant features which were identified were age, alcohol, urea content, hfref (heart failure with reduced ejection fraction, hfnef (heart failure with normal ejection fraction), valvular heart disease, and congenital heart disease. Logistic regression was then applied to these significant features to predict the model performance in detecting HF, with the model displaying excellent accuracy of 99.77%. The AUC-ROC also displayed similar accuracy of 99.90%.

The results derived from the classification techniques will aid researchers in assessing critical factors that cause HF in patients. Advanced machine learning techniques can also be utilized to gain further insights into this analysis and compare model performance. One of the constraints of this study is that this data comprises records from a single hospital in Punjab, which is a state in the country of India, more data can be collected from different states or different countries, to conduct further research and make inferences regarding HF.

References

1. Roger, V. L. (2013, June 19). Epidemiology of Heart Failure [Review of Epidemiology of Heart Failure]. Lippincott Williams & Wilkins. https://doi.org/10.1161/CIRCRESAHA.113.300268

2. National Heart, Lung, and Blood Institute. (2022, March 24). *Heart Failure What Is Heart Failure? | NHLBI, NIH.* Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/heart-failure
3. Inamdar, A. A., & Inamdar, A. C. (2016). Heart Failure: Diagnosis, Management and Utilization. *Journal of clinical medicine*, *5*(7), 62. https://doi.org/10.3390/jcm5070062

4. Better Health Channel. (2012). *Heart disease - risk factors.* Vic.gov.au. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart-disease-risk-factors

5. Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., Jaarsma, T., Krum, H., Rastogi, V., Rohde, L. E., Samal, U. C., Shimokawa, H., Budi Siswanto, B., Sliwa, K., & Filippatos, G. (2014). Heart failure: preventing disease and death worldwide. *ESC Heart Failure*, 1(1), 4–25. https://doi.org/10.1002/ehf2.12005

6. Cook, C., Cole, G., Asaria, P., Jabbour, R., & Francis, D. P. (2014). The annual global economic burden of heart failure. *International Journal of Cardiology*, 171(3), 368–376. https://doi.org/10.1016/j.ijcard.2013.12.028
7. *Coronary Artery Disease Analysis & Prediction.* (n.d.). Kaggle.com. https://www.kaggle.com/code/homelysmile/coronary-artery-disease-analysis prediction/data

8. *Hospital Admissions Data*. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/ashishsahani/hospital-admissions-data?select=HDHI+Admission+data.csv

9. Jujo, K., Minami, Y., Haruki, S., Matsue, Y., Shimazaki, K., Kadowaki, H., Ishida, I., Kambayashi, K., Arashi, H., Sekiguchi, H., & Hagiwara, N. (2017). Persistent high blood urea nitrogen level is associated with increased risk of cardiovascular events in patients with acute heart failure. *ESC Heart Failure*, 4(4), 545–553. https://doi.org/10.1002/ehf2.12188

10. Wikipedia Contributors. (2019, July 22). Generalized linear model. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Generalized_linear_model

11. Febrianti, R., Widyaningsih, Y., & Soemartojo, S. (2021). The parameter estimation of logistic regression with maximum likelihood method and score function modification. *Journal of Physics: Conference Series*, 1725, 012014. https://doi.org/10.1088/1742-6596/1725/1/012014
12. *FAQ What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them? (n.d.).* Stats.oarc.ucla.edu. Retrieved March 1, 2023, from https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/#:~:text=In%20the%20case%20of%20complete

APPENDIX

Appendix A: Mathematical Equations for GLM with logit link function

A.1 Equation 1: Likelihood Function of a binomial distribution

$$\mathbf{L(\pi, n; y) = \prod [\, \pi{}^\wedge(niyi) * (1\text{-}\pi){}^\wedge(ni\text{-}niyi)\,]{}^\wedge(ni/n)}$$

where:
n = number of observations
yi = variable in consideration
πi = probability of the event happening

A.2 Equation 2: Score function for GLM

$$\mathbf{\partial l / \partial \beta j = \sum_{j=1}^{P} \beta_j x_j \,[\, Xi(yi - \mu i) / a(\varphi) * \partial \eta i / \partial \beta j\,]}$$

and for a binomial distribution:

$$\mathbf{\partial l / \partial \beta j = \sum_{i=1}^{p} ni(yi - \pi i) * xij}$$

where:
a(φ) = dispersion parameter for the probability distribution
∂l / ∂βj = partial derivative of the linear predictor with respect to βj

A.3 Equation 3: Results of the GLM with logit link function

The equation below summarizes the effect of each predictor on the effect of Heart Failure.

$$\text{logit}(\pi i) = -5.297 - 0.2124(\text{Age}) - 0.01455(\text{Urea}) + \left\{\begin{array}{l} 0\ (\text{Not Alcoholic}) \\ -3.781(\text{ Alcoholic}) \end{array}\right. + \left\{\begin{array}{l} 0\ (\text{ No hfref}) \\ 16.04\ \ (\text{hfref}) \end{array}\right. + \left\{\begin{array}{l} 0\ \ \ (\text{No hfnef}) \\ 0\ \ (\text{hfnef}) \end{array}\right.$$

$$+ \left\{\begin{array}{l} 0\ \ (\text{No Valvular}) \\ 3.408\ (\text{Valvular}) \end{array}\right. + \left\{\begin{array}{l} 0\ (\text{No Congenital}) \\ -6.761(\text{ Congenital}) \end{array}\right.$$

<u>Appendix B: EDA- Tables and Images</u>

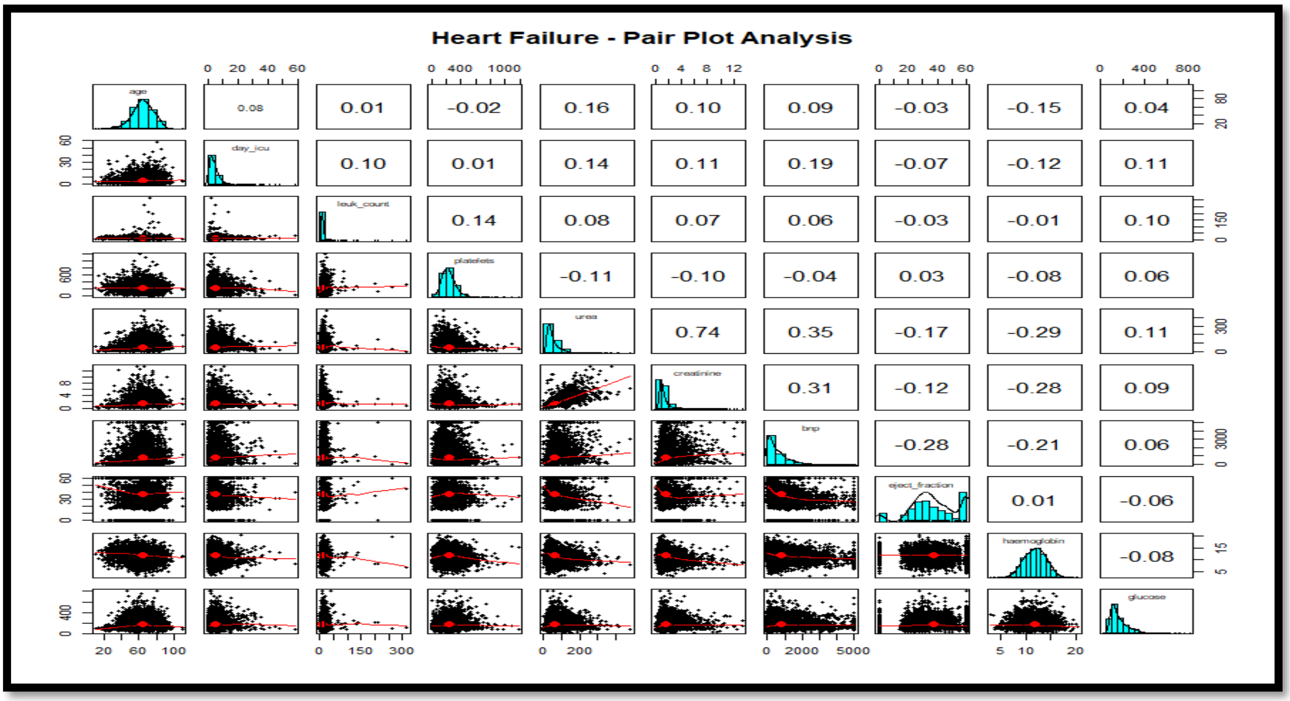
The following images were created to visualize the results reported in this study:

<u>B.1: Data Summary Table of significant predictors</u>

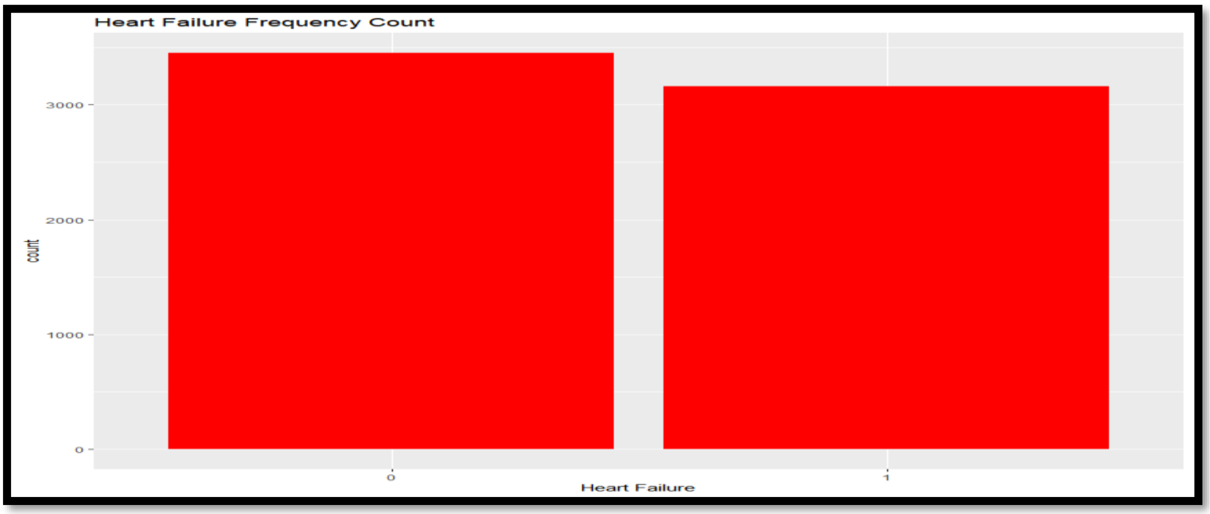
Index	Variable Name	Variable Type	Variable Levels	Variable Description
1	age	Numeric	-	Age of Patients
2	alcohol	Categorical	0,1	If a patient consumes alcohol or not.
3	Urea	Numeric	-	Urea content in the body
4	hfref	Categorical	0,1	If a patient had Heart Failure with Reduced Ejection Fraction or not?
5	Hfnef	Categorical	0,1	If a patient had Heart Failure with Normal Ejection Fraction or not.
6	Valvular	Categorical	0,1	If a patient had Valvular Heart Disease or not.
7	Congenital	Categorical	0,1	If a patient had congenital heart disease or not.

Note: Zero indicates 'No' and One indicates 'Yes'

<u>B.2: Pair Plot to display the association between the numerical variable</u>



B.3: Response Variable (Heart Failure) Frequency Distribution Plot



Appendix C: Raw Code and Summary Outputs

Note: This is only a snapshot of the code. View RMarkdown file for entire code.

C.1 : Data loading and EDA

```
heart_failure = read.csv("DataClean-fullage.csv") # Load the data
str(heart_failure) # structure of the data
summary(heart_failure)  # summary statistics of the data
View(heart_failure)         # View the data
is.na(heart_failure)         # Check for missing values # Comments: No missing values found
names(duplicated(heart_failure))  # Check for duplicated values # Comments: No duplicated values found
```

C.2: GLM with logistic regression on whole features and nested features

```
# Using all variables
glm_hf_1 = glm(heart_failure ~ . , data = heart_failure , family = binomial(link = "logit"))
summary(glm_hf_1)  # AIC: 181.02

# Using significant predictors
glm_hf_1_2 = glm(heart_failure ~  alcohol + hfref + hfnef + valvular  + congenital , data = heart_failure , family = binomial(link = "logit"))
summary(glm_hf_1_2)  # AIC: 101.28
```

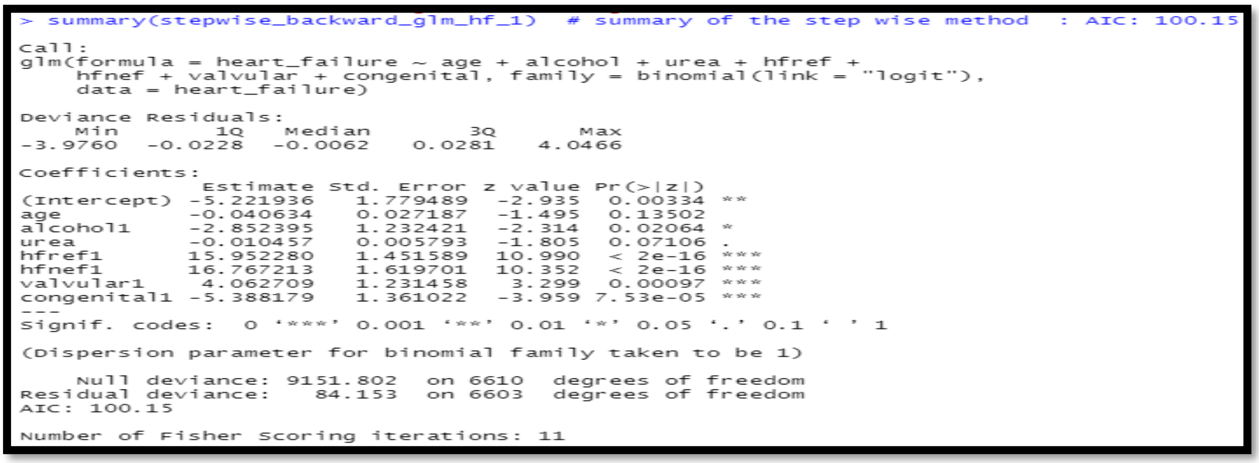
C.3: Likelihood Ratio

```
models_11_12_anova = anova(glm_hf_1_1,glm_hf_1_2, test = "Chisq")

# Since the Deviance result is -1.9761, and there is only a marginal difference in the Residual
```

Deviance, it is concluded that the models are essentially same and therefore the lesser complex model was used for analysis.

C.4: Stepwise Regression (Using backward method) Code and Output

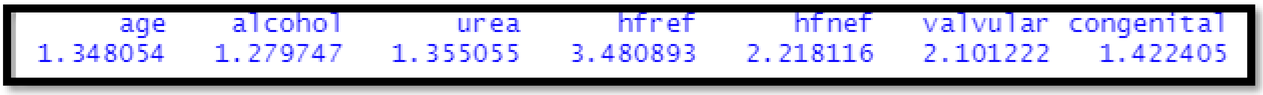
stepwise_backward_glm_hf_1 = step(glm_hf_1, direction = "backward")



```
> summary(stepwise_backward_glm_hf_1)  # summary of the step wise method  : AIC: 100.15

Call:
glm(formula = heart_failure ~ age + alcohol + urea + hfref + 
    hfnef + valvular + congenital, family = binomial(link = "logit"), 
    data = heart_failure)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.9760  -0.0228  -0.0062   0.0281   4.0466  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.221936   1.779489  -2.935  0.00334 ** 
age         -0.040634   0.027187  -1.495  0.13502    
alcohol1    -2.852395   1.232421  -2.314  0.02064 *  
urea        -0.010457   0.005793  -1.805  0.07106 .  
hfref1      15.952280   1.451589  10.990  < 2e-16 ***
hfnef1      16.767213   1.619701  10.352  < 2e-16 ***
valvular1    4.062709   1.231458   3.299  0.00097 ***
congenital1 -5.388179   1.361022  -3.959 7.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9151.802  on 6610  degrees of freedom
Residual deviance:   84.153  on 6603  degrees of freedom
AIC: 100.15

Number of Fisher Scoring iterations: 11
```

C.5: Variance Inflation Factor results

The following image displays the VIF results of each variable.



```
      age   alcohol      urea     hfref     hfnef  valvular congenital 
 1.348054  1.279747  1.355055  3.480893  2.218116  2.101222  1.422405 
```

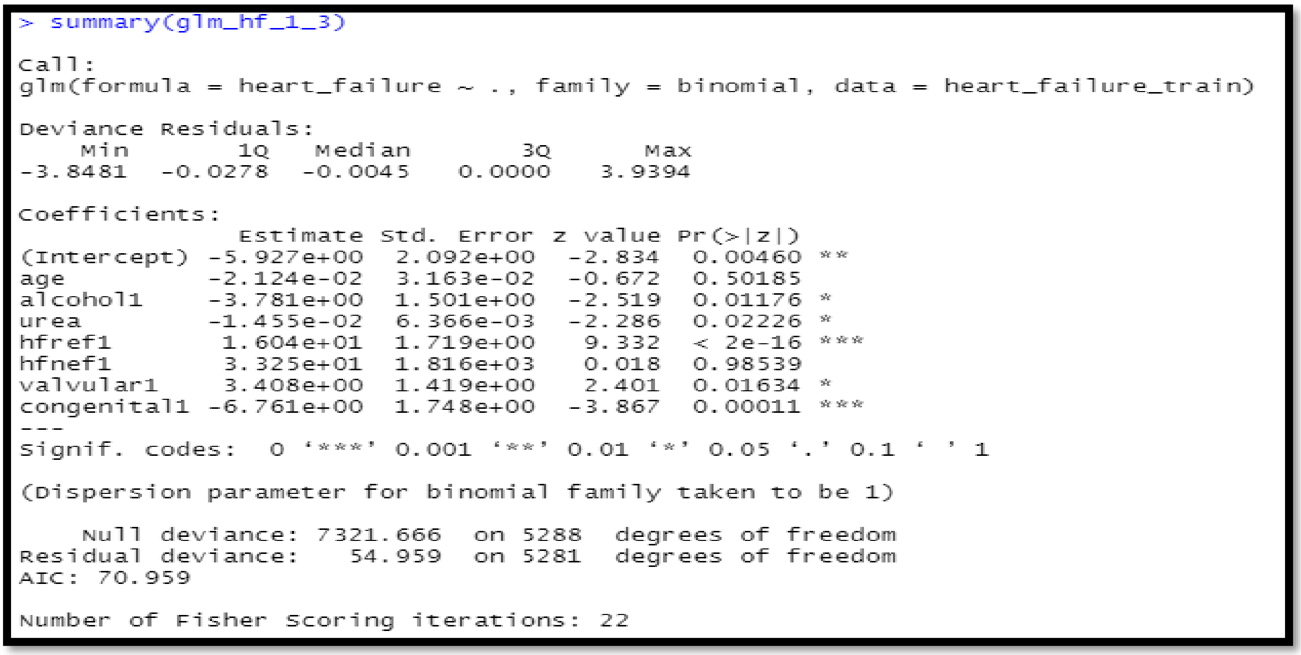
# Variance inflation factor **(***VIF***)** *was used to assess the assumption of little or no multicollinearity between variables***.** *VIF values above 10 indicate high collinearity, and VIFs between 5 and 10 indicate the need for further investigation. The VIF values for all important features were between 1.28 and 3. 8, indicating no collinearity between the variables.*

C.6: Data partitioning and Logistic Regression on Partitioned Data

heart_failure_new = heart_failure[,c(1,7,15,23,24,25,26,35)] *# Create a new subset of the data with selected features # Age , Urea, Alcohol, Hfref, Hfnef, Valvular , Congenital*

set.seed(126) # Set the seed for model consistency
sample = sample.split(heart_failure_new$heart_failure , SplitRatio = .80) *# Split the data*
heart_failure_train = subset(heart_failure_new, sample == TRUE)   *# Training data*
heart_failure_test = subset(heart_failure_new, sample == FALSE)   *# Test data*
dim(heart_failure_train)   *# Dimensions of train data :5289 x 8*
dim(heart_failure_test)   *# Dimensions of test data : 1322 x 8*

glm_hf_1_3 = glm(heart_failure ~ . , data = heart_failure_train , family = binomial) *# Logistic Regression on Training Data*
summary(glm_hf_1_3)   *# Model Summary*

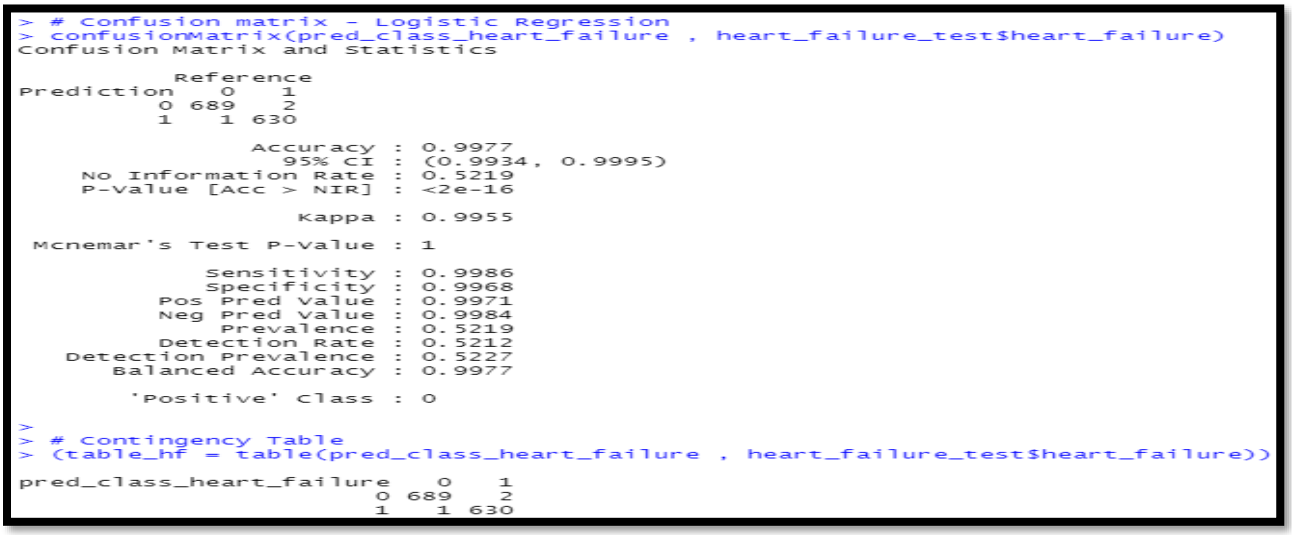


```
> summary(glm_hf_1_3)

Call:
glm(formula = heart_failure ~ ., family = binomial, data = heart_failure_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.8481  -0.0278  -0.0045   0.0000   3.9394  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.927e+00  2.092e+00  -2.834  0.00460 ** 
age         -2.124e-02  3.163e-02  -0.672  0.50185    
alcohol1    -3.781e+00  1.501e+00  -2.519  0.01176 *  
urea        -1.455e-02  6.366e-03  -2.286  0.02226 *  
hfref1       1.604e+01  1.719e+00   9.332  < 2e-16 ***
hfnef1       3.325e+01  1.816e+03   0.018  0.98539    
valvular1    3.408e+00  1.419e+00   2.401  0.01634 *  
congenital1 -6.761e+00  1.748e+00  -3.867  0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7321.666  on 5288  degrees of freedom
Residual deviance:   54.959  on 5281  degrees of freedom
AIC: 70.959

Number of Fisher Scoring iterations: 22
```

Comments: On analyzing the output hfnef1 has a higher Standard Error, indicating Quasi-Complete Separation. However, in this case, the 'do nothing' strategy was used, as the likelihood of other predictors was still valid.

C.7: Confusion Matrix Outputs

This image gives the detailed confusion matrix output with the classification table and accuracy results.



```
> # Confusion matrix - Logistic Regression
> confusionMatrix(pred_class_heart_failure , heart_failure_test$heart_failure)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 689   2
         1   1 630
                                          
               Accuracy : 0.9977          
                 95% CI : (0.9934, 0.9995)
    No Information Rate : 0.5219          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.9955          
                                          
 Mcnemar's Test P-Value : 1               
                                          
            Sensitivity : 0.9986          
            Specificity : 0.9968          
         Pos Pred Value : 0.9971          
         Neg Pred Value : 0.9984          
             Prevalence : 0.5219          
         Detection Rate : 0.5212          
   Detection Prevalence : 0.5227          
      Balanced Accuracy : 0.9977          
                                          
       'Positive' Class : 0               
                                          
> 
> # Contingency Table
> (table_hf = table(pred_class_heart_failure , heart_failure_test$heart_failure))
                        
pred_class_heart_failure   0   1
                       0 689   2
                       1   1 630
```

C.7: Interaction Effects Code

glm_hf_1_4 = glm(heart_failure ~ (alcohol * urea) + age + hfref + hfnef + valvular + congenital, data = heart_failure_train , family = binomial) *# Interaction between Alcohol + Urea*

glm_hf_1_5 = glm(heart_failure ~ alcohol + urea + age + (hfref * hfnef) + valvular + congenital, data = heart_failure_train , family = binomial) *# Interaction between hfref and hfnef*

```
glm_hf_1_6 = glm(heart_failure ~  alcohol + urea + age + hfref + hfnef + (valvular * congenital),
data = heart_failure_train , family = binomial)  # Interaction between valvular and congenital
```