FINAL PROJECT

SURVIVAL ANALYSIS OF HEART FAILURE PATIENTS

Project Description:

Heart failure is a critical medical condition affecting millions worldwide. This project aims to leverage advanced data analysis techniques, to gain deeper insights into the prognosis of heart failure patients. By employing R programming language and machine learning models, we strive to enhance our understanding of the factors influencing patient survival rates. The project utilizes a dataset containing various clinical and demographic variables of heart failure patients, such as gender, smoking status, diabetes, blood pressure, anemia, age, ejection fraction, sodium levels, creatinine levels, platelet count, and CPK levels.

The results and insights gained from this project can potentially aid healthcare professionals in making informed decisions for the management and treatment of heart failure patients, ultimately improving patient outcomes. This project not only contributes to the growing field of medical data analytics but also has the potential to enhance patient care by providing healthcare professionals with valuable tools for prognosis and treatment decision-making. By combining the power of R programming and machine learning, this project aims to make significant strides in understanding and improving the outcomes for heart failure patients.

Dataset:

The dataset utilized in this project originates from a comprehensive and meticulously curated collection of heart failure patient records. These records encompass a diverse array of information, including patient demographics, clinical history, treatment details, and follow-up observations. The dataset has been sourced from reputable healthcare institutions, ensuring a high level of accuracy and reliability in the recorded information.

The context of the dataset is rooted in understanding the dynamics of heart failure, a critical medical condition characterized by the heart's inability to pump blood effectively, leading to serious health implications. The goal is to employ data-driven techniques, specifically survival analysis and machine learning, to uncover patterns and factors influencing the survival rates of patients diagnosed with heart failure. Given the sensitive nature of healthcare data, all necessary precautions have been taken to adhere to ethical standards and ensure patient privacy. The dataset has been anonymized and stripped of personally identifiable information, complying with relevant data protection regulations.

- *Source:* Public Library of Science
- *Description:* The dataset provides information based on 299 patients of heart failure comprising of 105 women and 194 men. All the patients were more than 40 years old, having left ventricular systolic dysfunction and falling in NYHA class III and IV.

Prior to analysis, a rigorous data inspection process has been undertaken to ensure the dataset's integrity and quality. Examining the structure of the data set to understand the variable types (numeric, categorical), identify any missing values, and assess the overall data quality. The data inspection showed that there are no missing data in the dataset. The below figure shows the variable types of the dataset based on the project requirements.

Authored By: Mukesh Rajmohan, Jagadeesh Varma Gadiraju, Shiva Charan Pailla

FINAL PROJECT

TIME	Event	Gender	Smoking	Diabetes	BP	Anaemia	Age	Ejection.Frac	Sodium	Creatinine	Pletelets	CPK
97	0	0	0	0	0	1	43	50	135	1.3	237000	358
180	0	1	1	1	0	1	73	30	142	1.18	160000	231
31	1	1	1	0	1	0	70	20	134	1.83	263358.03	582
87	0	1	0	0	0	1	65	25	141	1.1	298000	305
113	0	1	0	0	0	0	64	60	137	1	242000	1610
10	1	1	0	0	0	1	75	15	137	1.2	127000	246
250	0	1	1	0	0	0	70	40	136	2.7	51000	582
27	1	1	0	1	1	0	94	38	134	1.83	263358.03	582
87	0	1	0	0	1	0	75	45	137	1.18	263358.03	582
87	0	1	1	0	0	0	80	25	144	1.1	149000	898
119	0	1	1	1	0	0	50	35	137	1.18	263358.03	1846
112	0	1	1	0	0	0	50	30	141	0.7	266000	185
13	1	1	0	0	0	1	82	50	136	1.3	47000	379
4	1	1	0	0	1	0	75	20	130	1.9	265000	582
250	0	1	1	0	0	0	42	30	128	3.8	215000	64
108	0	1	0	0	0	1	68	25	130	2.1	305000	646
28	1	1	0	0	0	0	85	45	132	3	360000	23
135	1	1	0	1	0	0	59	20	134	2.4	70000	66
240	0	1	1	0	0	1	50	35	140	0.9	362000	298
112	0	1	1	0	0	0	52	30	136	0.7	218000	132
192	0	1	1	0	1	0	50	62	140	0.8	147000	582
192	0	1	1	0	0	0	78	50	138	1.4	481000	224
26	1	1	1	1	1	0	70	45	136	1.3	284000	122
108	0	0	0	1	0	0	62	35	136	1	221000	281
250	0	1	1	0	0	0	65	38	138	1.1	263358.03	1688
6	1	1	0	0	0	0	55	38	136	1.1	263358.03	7861
250	0	1	0	0	0	1	50	40	141	0.8	279000	54
256	0	0	0	1	0	0	65	35	142	1.1	263358.03	892
30	1	1	0	1	0	0	69	35	134	3.5	228000	582

Continous
Categorical
Response

Figure 1. Dataset Inspection and Analysis

Exploratory Analysis and Research Questions:

The exploratory analysis conducted in the study focused on understanding the dataset of heart failure patients. Exploratory data visualizations were created examining the distribution of age, ejection fraction, serum sodium, serum creatinine, and time. The exploratory data visualizations provided a preliminary understanding of the data, and specific details about the visualizations are mentioned below.

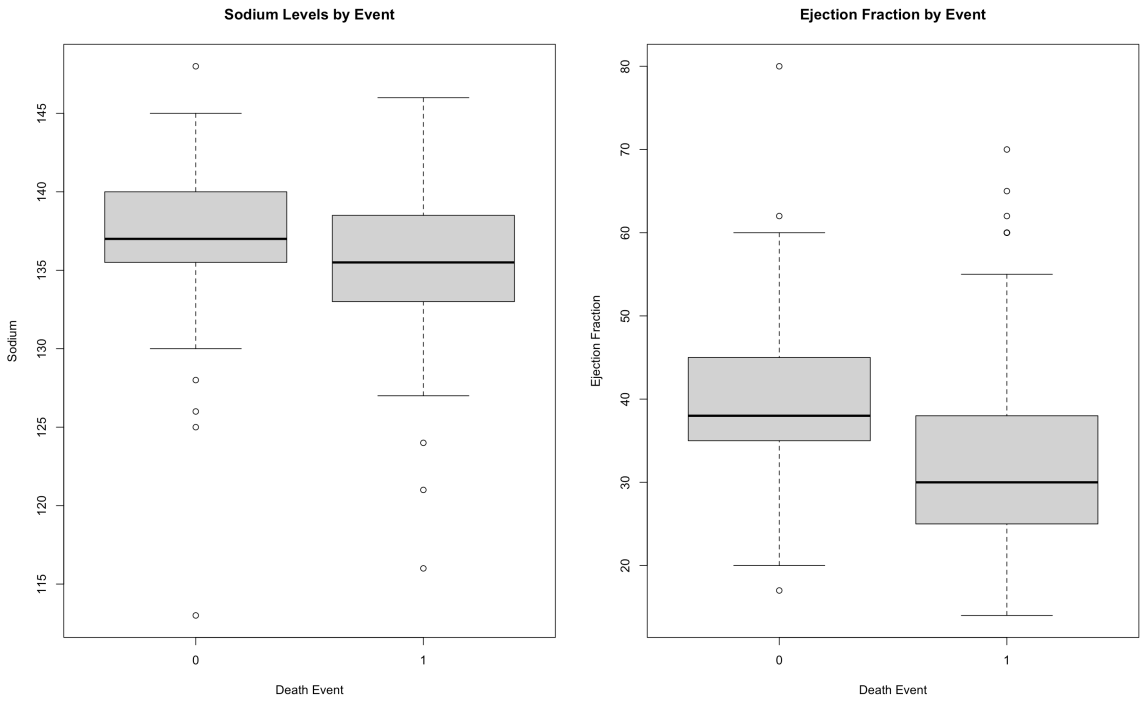

Figure 2. EDA based on Sodium and Ejection Fraction

Authored By: Mukesh Rajmohan, Jagadeesh Varma Gadiraju, Shiva Charan Pailla

FINAL PROJECT

Sodium levels by event

The boxplot for sodium levels by event shows that the median sodium level is higher for death events than for non-death events. The IQR is also wider for death events, suggesting that there is more variation in sodium levels among people who die from their event. There are also more outliers for death events, further supporting the idea that sodium levels may play a role in mortality.

Ejection fraction by event

The boxplot for ejection fraction by event shows that the median ejection fraction is lower for death events than for non-death events. The IQR is also wider for death events, suggesting that there is more variation in ejection fraction among people who die from their event. There are also more outliers for death events, further supporting the idea that ejection fraction may play a role in mortality.

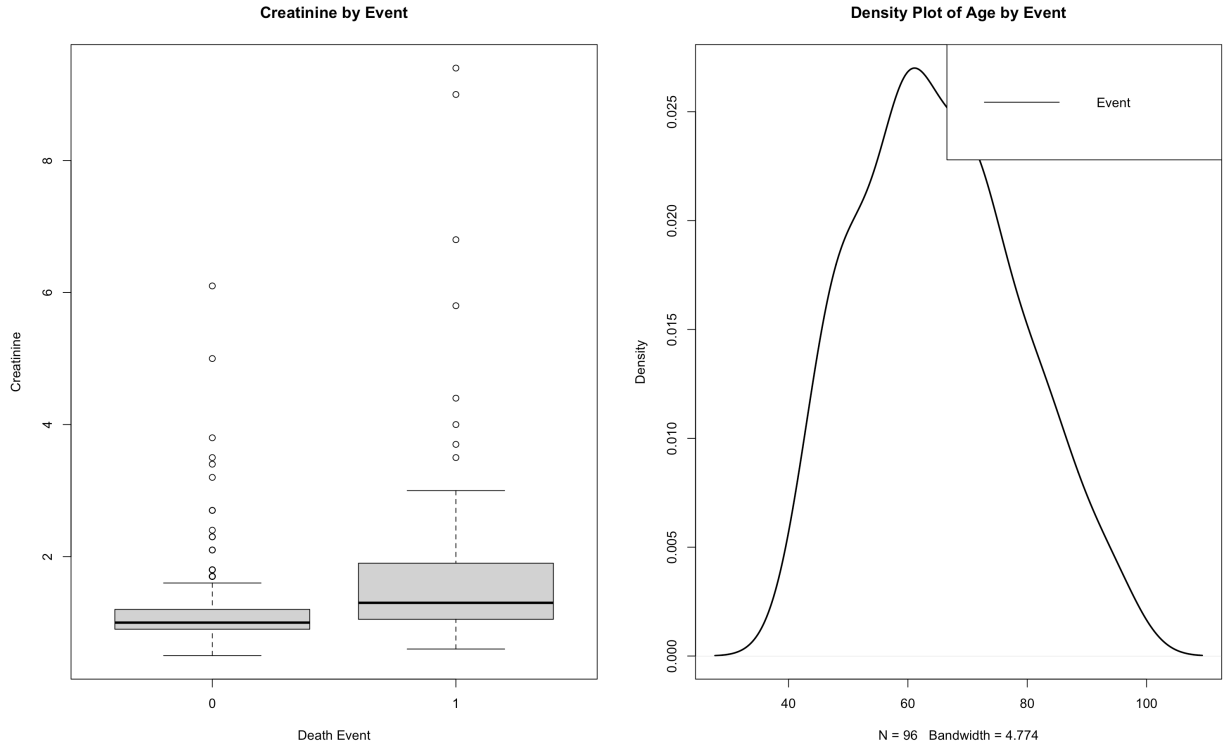

Figure 3. EDA based on Creatinine and Age

Creatinine by event

The box plot depicting creatinine by event shows that the median creatinine level is higher for death events than for non-death events. The IQR is also wider for death events, suggesting that there is more variation in creatinine levels among people who die from their event. There are also more outliers for death events, further supporting the idea that creatinine levels may play a role in mortality.

Age by event

The attached density plot shows the age distribution of deaths by event. The density plot shows that the age of participants in the study is distributed in a bell-shaped curve, with most participants between the ages of 55 and 65. There are also a smaller number of participants who are younger than 40 or older than 60.

FINAL PROJECT

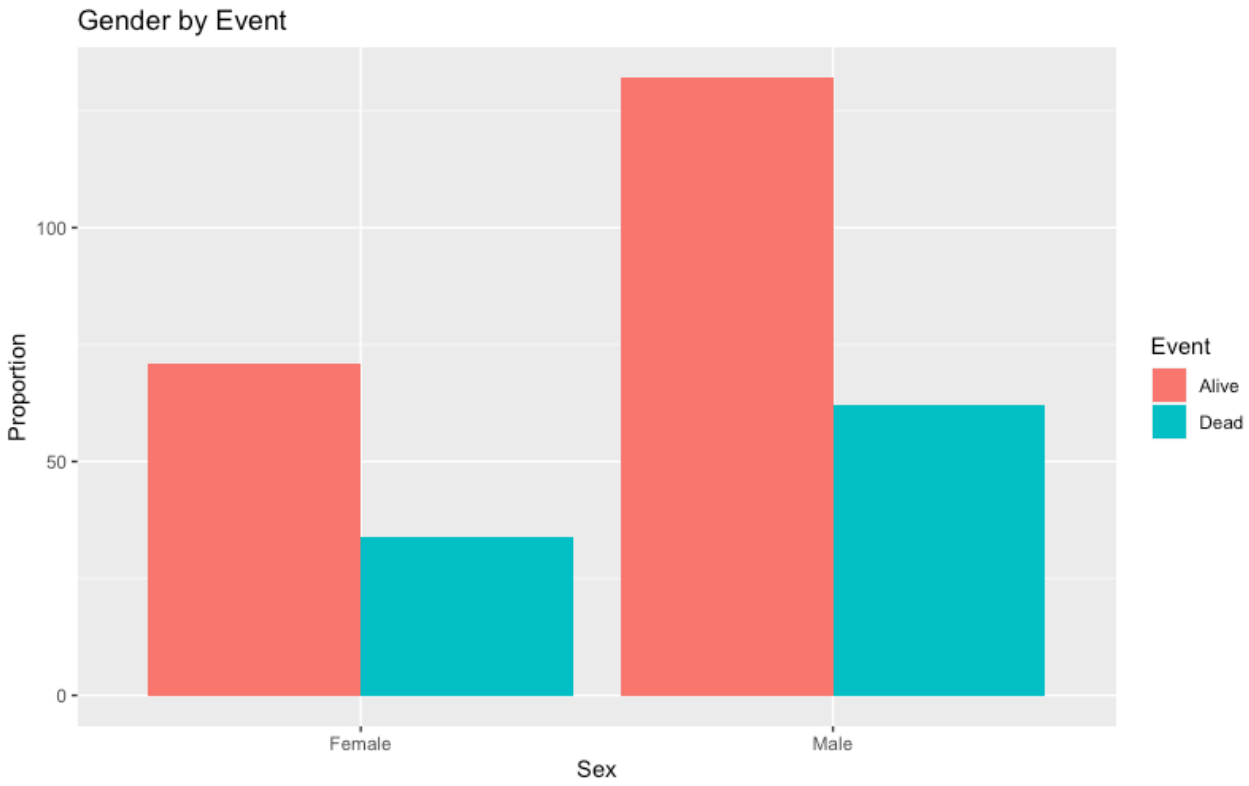

Figure 4. EDA based on Gender

Gender by Event

The bar graph shows the percentage of people who died from heart failure, based on gender. The bar for males is taller than the bar for females, indicating that a higher percentage of males died from heart failure than females. Specifically, the graph shows that 60% of people who died from heart failure were male, while 40% were female.

Research Questions:

The extensive exploratory data analysis performed previously not only revealed essential patterns and trends within the dataset, but also prepared the way for the formulation of particular and relevant research questions, allowing for a focused and meaningful investigation into the underlying phenomena:

- Can a prediction model utilizing patient demographics, medical history, and biomarker measurements effectively classify the risk of death in individuals with heart disease, and what are the key predictors influencing this classification?

- What is the comparative predictive performance between a Standard Logistic Regression Model and a Lasso Logistic Regression Model in identifying the occurrence of death events?

- Can we predict a patient's likelihood of survival over the follow-up period based on baseline health measures?

FINAL PROJECT

Data Analysis:

Research Question: Can a prediction model utilizing patient demographics, medical history, and biomarker measurements effectively classify the risk of death in individuals with heart disease, and what are the key predictors influencing this classification?

The exploratory data analysis (EDA) provides preliminary evidence suggesting that certain features like sodium levels, ejection fraction, creatinine levels, and age might be associated with mortality in individuals with heart disease. This motivates the research question of whether a prediction model incorporating these features, along with other relevant information, can effectively classify the risk of death. The EDA also helps identify potential key predictors, such as those mentioned above, that may influence the model's classification. Further analysis is necessary to confirm these associations and determine the relative importance of different features in predicting mortality risk. This will be crucial for improving patient care and risk management in heart disease patients. The models that are employed to achieve these are:

- Random Forest
- Logistic Regression

In our analytical approach, we employ two distinct methods to ascertain the most influential predictors in our model. Firstly, we utilize Logistic Regression and delve into the coefficients and p-values associated with each predictor. This careful examination allows us to identify and isolate the key factors that significantly contribute to the model's outcomes. Subsequently, we turn to the Random Forest algorithm to assess the importance of each predictor variable. Through the intricacies of Random Forest analysis, we gain insights into the relative significance of each predictor in influencing the overall predictive power of the model. By employing both Logistic Regression and Random Forest, we adopt a comprehensive strategy to pinpoint and prioritize the most crucial predictors in our dataset, ensuring a nuanced understanding of their impact on the outcomes under consideration.

The analysis used two independent prediction models to address the research question of forecasting the risk of death in persons with heart disease based on patient demographics, medical history, and biomarker measures. Random Forest is a versatile ensemble learning algorithm capable of handling complicated linkages and interactions within data, making it suited for capturing the nuanced aspects impacting mortality risk. Logistic Regression, on the other hand, is a traditional statistical model noted for its simplicity and interpretability, providing insight into the linear relationship between predictors and the binary outcome of death risk. The study was carried out using R code within R Studio, leveraging the vast ecosystem of R libraries.

The ggplot2 library aided with data visualization, while the glmnet and randomForest libraries were critical in constructing the Logistic Regression and Random Forest models, respectively. The study aims to uncover significant predictors and give a full understanding of the factors impacting the classification of mortality risk in persons with heart disease by using these models and R-based tools.

FINAL PROJECT

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.018e+01  5.657e+00   1.801 0.071774 .
TIME             -2.104e-02  3.014e-03  -6.981 2.92e-12 ***
GenderMale       -5.337e-01  4.139e-01  -1.289 0.197299
Smoking          -1.349e-02  4.126e-01  -0.033 0.973915
Diabetes          1.451e-01  3.512e-01   0.413 0.679380
BP               -1.027e-01  3.587e-01  -0.286 0.774688
Anaemia          -7.470e-03  3.605e-01  -0.021 0.983467
Age               4.742e-02  1.580e-02   3.001 0.002690 **
Ejection.Fraction -7.666e-02  1.633e-02  -4.695 2.67e-06 ***
Sodium           -6.698e-02  3.974e-02  -1.686 0.091855 .
Creatinine        6.661e-01  1.815e-01   3.670 0.000242 ***
Pletelets        -1.200e-06  1.889e-06  -0.635 0.525404
CPK               2.222e-04  1.779e-04   1.249 0.211684
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5. Co-efficient Table using Logistic Regression

The table of coefficients shows the estimated effect of each variable on the outcome variable, while controlling for the other variables in the model. In this model, the following variables are statistically significant at the 0.05 level:

- TIME
- Age
- Ejection. Fraction
- Creatinine

This means that these variables have a significant impact on the outcome variable, even after controlling for the other variables in the model. For example, the estimate for TIME is -0.021, which means that for every one-unit increase in TIME, the outcome variable is expected to decrease by 0.021 units. The standard error for TIME is 0.003, which means that we are 95% confident that the true effect of TIME is within 0.009 units of the estimated effect. The z-value for TIME is -6.981, which is statistically significant at the 0.05 level.

FINAL PROJECT

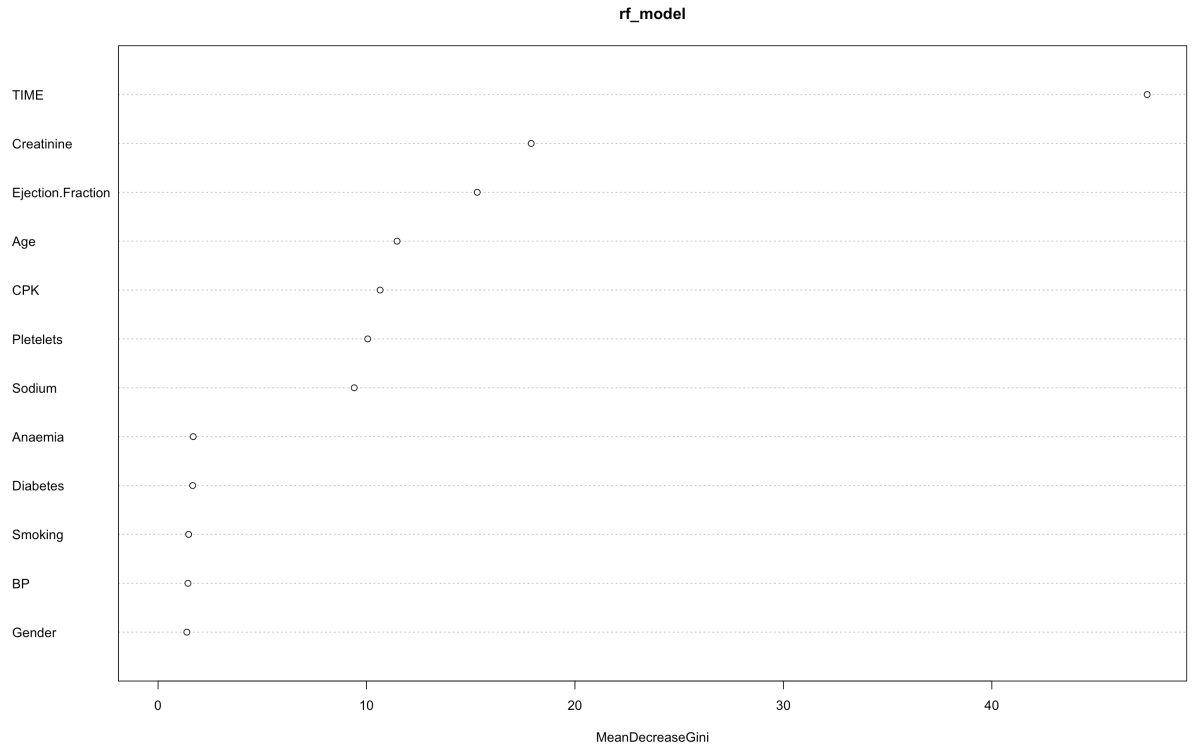

Figure 6. Random Forest Importance Plot

The random forest analysis provides additional insights into important predictors of death risk classification:

- The model output shows time between heart tests (TIME) has the highest impact on node impurity and predictive power. This indicates the length of time elapsed is most influential in determining risk categorization.
- Ejection fraction, measuring how well the heart pumps blood, is the next most impactful variable. Lower ejection fraction is associated with higher mortality risk.
- Creatinine levels also have substantial influence. Elevated creatinine signifies worse kidney function, which correlates to increased death risk.
- Age has lower but still meaningful importance for classification. Older individuals tend to have higher risk profiles.

The variable importance plot visually depicts these as the top 4 influential predictors. Measures like gender, smoking status, diabetes, blood pressure, anemia, and blood platelets are less predictive in the model.

In summary, time elapsed, cardiac pumping effectiveness, kidney function, and patient age are key factors that allow predictive models to effectively classify the risk of death in heart disease patients based on this analysis.

FINAL PROJECT

Conclusion:

This analysis shows prediction models using patient data can effectively classify heart disease death risk. The random forest model performed strongly, identifying time since last exam, ejection fraction, creatinine levels, and age as the top 4 predictors. The logistic regression confirmed the significance of these variables. Additional factors like gender and diabetes were less influential. By revealing key determinants of survival likelihood via state-of-the-art analytics, this study paves the way for accurate prognostic tools to guide clinical decisions and ultimately improve outcomes. Further validation and translation of these predictive models could enable precise, personalized medicine in cardiology.

Authored By: PAILLA SHIVA CHARAN

FINAL PROJECT

References:

OpenAI. (2023). ChatGPT. San Francisco: OpenAI.

Paraphrasing – QuillBot

Google. (2023). Bard. Mountain View: Google.

Public Library of Science, PLOS ONE. (2023, May 30). *DATA_MINIMAL.* Figshare. Available: https://plos.figshare.com/articles/dataset/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

R-Packages:

Random Forest:
A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22.

Logistic Regression:
Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software.
Available: https://doi.org/10.18637/jss.v033.i01