Heart Failure Prediction: Final Report

Stat 140XP

Audrey Huang, Tess Wellington, Ethan Shahzad, Derek Diaz

December 2024

Abstract

As the leading cause of death worldwide, heart disease is a critical public health issue. Understanding its risk factors and identifying preventive measures are essential for reducing mortality. In this project, we analyze a heart failure prediction dataset from Kaggle, sourced from Faisalabad Hospital in Pakistan, to determine the most significant predictors of mortality by heart failure. Our approach includes data cleaning, feature selection, and predictive modeling using decision trees.

1 Introduction

Heart disease is the leading cause of death globally, making it crucial to identify its causes and develop preventive measures. Our study utilizes a heart failure prediction dataset from Kaggle, collected from Faisalabad Hospital, Pakistan. Our goal is to determine which factors significantly impact heart failure outcomes.

First, we will utilize data wrangling to clean our dataset for relevant variables. Next, we will use feature selection to select significant variables for predicting mortality from heart failure. Last, we will build a decision tree model to predict future results from our selected features.

2 Data Description

The dataset consists of 368 observations and 60 variables, of which 18 are well-defined and relevant for analysis. The excluded variables lacked descriptions, making them unsuitable for meaningful interpretation. The selected variables encompass both demographic and medical information for each patient. The demographic statistics include age, age group, gender, locality (rural or urban), and marital status. The medical statistics include smoking status, depression status, number of visiting time, chest pain type, diabetic status, cholesterol, fasting blood sugar, resting blood pressure, resting electrocardiographic results, thalach heart rate, platelet count, hemoglobin count, and whether the patient died as a result of heart failure.

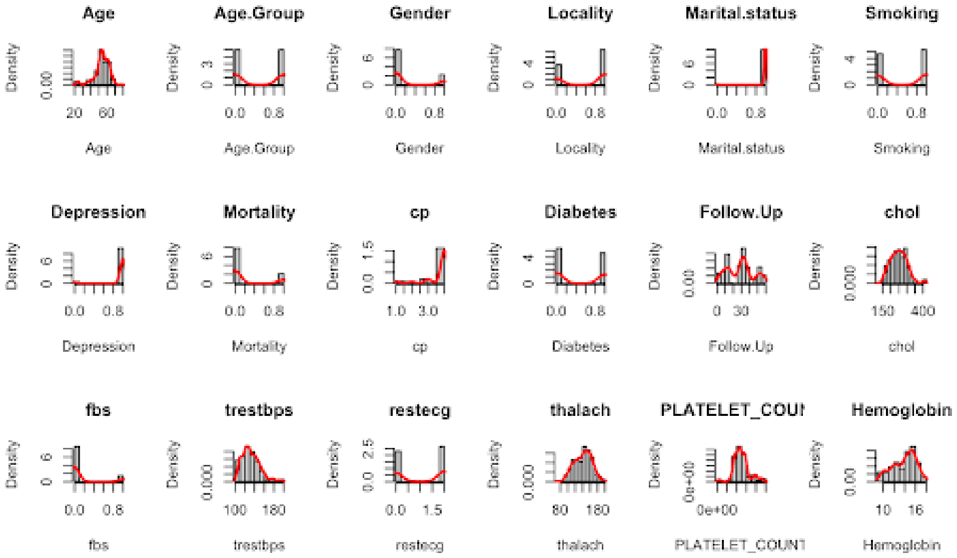

Figure 1: Distribution of Variables

The histogram in Figure 1 provides a graphical summary of variable distributions.Although age group, diabetes, and smoking appear to be evenly distributed, we can see significant differences in gender, locality, marital status, fbs, depression, and mortality. For our continuous variables, most appear to have a normal distribution, with slight skewing for a few variables. However a few variables do not appear normally distributed. Chest pain appears to be significantly left skewed while follow.up and restecg appear to be multimodal and bimodal respectively.

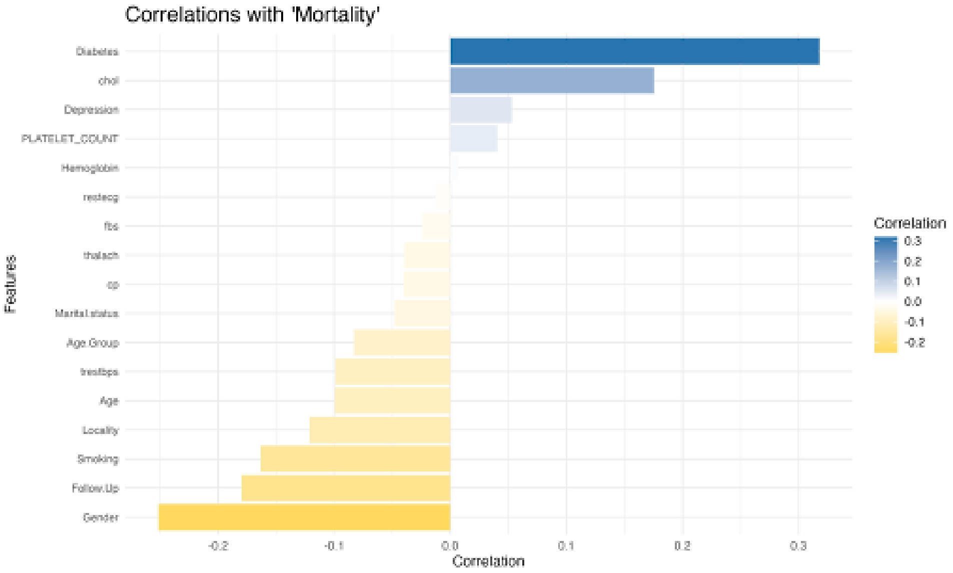

Figure 2: Correlation of Variables with Mortality

Figure 2 ranks variables by correlation with mortality. As they are ranked, we can see that diabetes and cholesterol are among the top two variables with the highest correlation. We can also see that gender, follow up, and smoking rank are the bottom 3 variables, with no correlation to mortality.

3 Feature Selection

After exploring the correlation of each variable, we moved on to feature selection to answer our research question. To determine the most significant predictors of heart failure, we used mortality as a response variable and treated all other known variables as predictors. We explored stepwise regression processes using both AIC and BIC as criteria.

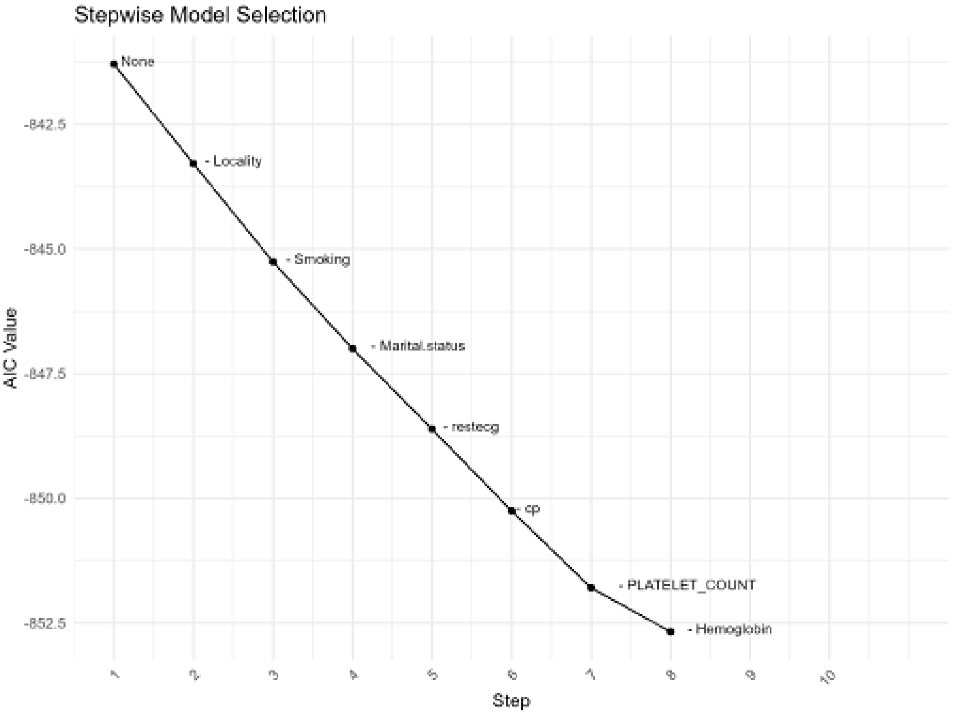

Figure 3: Backward Stepwise Model Selection using AIC

Figure 3 illustrates the process of backward stepwise regression with AIC, starting at a full model with all predictors and iteratively removing predictors with the largest p-value. Predictors were removed until the model achieved the lowest Akaike Information Criterion (AIC) which measures the fit of the model. We chose backward stepwise regression because we had a reasonably small amount of starting predictors and wanted the benefit of looking at the effects of all predictors simultaneously.

$$\begin{aligned}\text{Mortality} \sim\ & \text{Age} + \text{ Age Group} + \text{ Gender} + \text{ Depression} + \text{ Diabetes} + \text{ Follow-Up Visits}\\ & + \text{ Cholesterol} + \text{ Fasting Blood Sugar} + \text{ Resting Blood Pressure} + \text{ Maximum Heart Rat}\end{aligned}$$

Age, age group, gender, depression, diabetes, visiting time, cholesterol, fasting blood sugar, resting blood pressure, and maximum heart rate achieved were found to be the most significant predictors. Note that locality, smoking, marital status, restecg, cp, platelet count, and hemoglobin were determined to be insignificant predictors, which is a consistent finding with the correlations shown in Graph 2. In addition, backwards regression using BIC as the criteria agreed with the above model as the best fit. With substantial evidence from our correlation graph and multiple methods of feature selection, we committed to these variables as our most significant predictors of mortality due to heart failure. However, as discussed below, we decided to remove one predictor due to redundancy.

4 Predictive Modeling

After selecting our significant variables, we move to building a model for prediction. Due to the imbalance of data, where there are significantly more survivors, we chose to use a decision tree for predictive modeling. As the variables age and age group are redundant, we use 9 variables as our X variable. We chose a 70-30 split of training and testing data to train a decision tree via the boosting method.

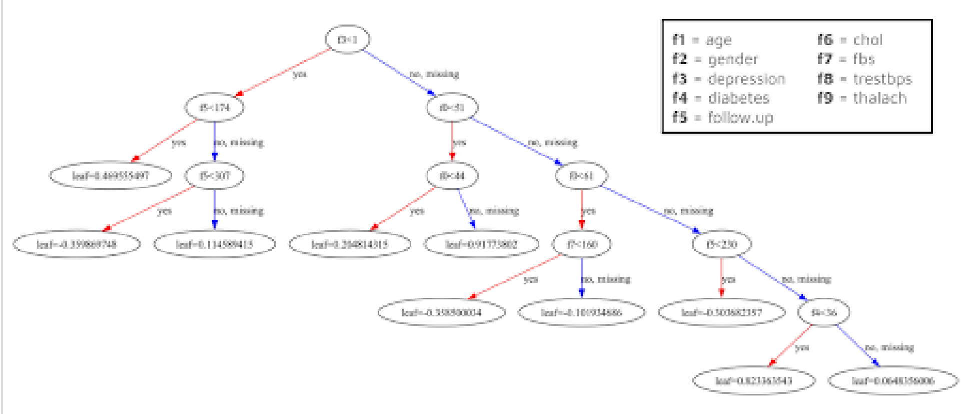

Figure 4: XGBoost Classification Tree for Heart Disease Prediction

The model achieved 96. 4% precision, demonstrating the strong predictive power of our model in determining the risk of mortality.

5 Discussion and Conclusion

Our study successfully identified key risk factors that influence heart failure mortality and built a highly accurate predictive model. Age, gender, depression, diabetes, visiting time, cholesterol, fasting blood sugar, resting blood pressure, and maximum heart rate achieved were all found to be strong predictors used to create an accurate model. These predictors showed strong predictive power with the response variable, aiding in the creation of an efficient model.

We experienced one limitation with our chosen dataset, which was a plethora of vague variables with no given definitions. We determined that moving forward in our analysis with undefined variables would lead to less meaningful interpretations and conclusions. So, we intentionally decided to only include meaningful, well-defined variables to use as predictors. However, this limited our model's predictive ability.

Although we achieved high accuracy with our model, the decision tree approach for modeling comes with limitations. Due to the dataset set having more survivors, the model will have a weaker prediction of the minority class (mortality). Another issue with decision trees is that the method tends to overfit the model, typically with smaller datasets. However, this may not affect our predictive model because our dataset is neither small nor large.

Although it was uncertain how significant these predictors were, we were able to observe their substantial predictive power with the response variable using the nine predictors. In order to have a better understanding of the behavioral and environmental variables influencing the mortality rate of heart disease patients in Pakistan, future research may focus on determining why these nine predictors are significant in predicting risk of mortality by heart failure.

6 References

- Khan, Asghar Ali. *Mortality Rate of Heart Patients in Pakistan Hospital.* Kaggle, 2021. `https://www.kaggle.com/datasets/asgharalikhan/mortality-rate-heart-patient-data`