Choose Your Own Capstone Project: Predicting Survival in Patients with Heart Failure HarvardX PH125.9x Data Science Capstone

Adam J. E. Blanchard

May 25, 2021

Contents

1 Overview 2
1.1 Introduction 2
1.2 Aim of the Project 3
1.3 Dataset 3
2 Data Wrangling and Inspection 5
3 Exploratory Data Analysis 6
4 Traditional Inferential Analysis 13
5 Modeling Approaches 17
5.1 Data Preparation 18
5.2 Model Training and Tuning 18
6 Results on Validation Dataset 22
7 Discussion 23
7.1 Limitations 23
7.2 Future Directions 24
8 References 25
9 Appendix A - Supplemental Figures 26
10 Appendix B - Environment 34

1 Overview

Cardiovascular diseases are the leading cause of death worldwide. They also result in hundreds of billions of dollars in direct and indirect costs. This project aimed to use machine learning models to predict patient mortality in a sample of 299 patients with heart failure. The dataset was collected from two hospitals in Pakistan in 2015 and made available to the public. This project describes the exploration of the dataset using visualization and summary statistics, inferential analyses of the relationships between the features and patient death, and the development of various machine learning models. Several variables were found to be associated with patient death depending on the analysis. Five variables were found to be related to patient death in the bivariate analyses and three remained in the multivariate logistic regression: age, ejection fraction, and serum creatinine. The machine learning models were able to predict death using all the features in the dataset as well as these three selected features. The performance of the predictive models was evaluated based on several metrics using cross-validation in the training dataset, as well as in a partitioned test dataset.

1.1 Introduction

Cardiovascular diseases (CVDs) are a group of serious health disorders that impact the heart and blood vessels, including coronary heart disease, valvular heart disease, heart attack, heart failure, cerebrovascular diseases (strokes), and other conditions (Chicco & Jurman, 2020; Heart and Stroke Foundation Canada, 2020a). Overall, cardiovascular diseases are the leading cause of death worldwide, resulting in the deaths of approximately 17 million people each year and representing around 31% of global deaths per year (World Health Organization, 2017).

In the United States, CVDs are the leading cause of death for men, women, and most ethnic groups with around 100 people dying every hour (Centers for Disease Control and Prevention, 2020a). CVDs cost the United States approximately $219 billion a year in health care, medication, and lost productivity. In Canada, CVDs are the second leading cause of death with around 12 people dying each hour, and it is the costliest disease economically with an estimated direct and indirect cost of $21.2 billion (Heart and Stroke Foundation, 2020a; Public Health Agency of Canada, 2017). There are several well-known behavioural risk factors for developing CVDs, including tobacco use, unhealthy diet, and excessive alcohol use, as well as other risk factors including diabetes, hypertension, and obesity (Public Health Agency of Canada, 2017; World Health Organization, 2017).

Heart failure is a condition in which the muscles of the heart fail to adequately pump blood (Ahmad et al., 2017; Centers for Disease Control and Prevention, 2020b; Heart and Stroke Foundation, 2020b). Heart failure typically develops after the heart has been damaged or weakened; the muscles are then incapable of adequately pumping blood around the body. As well, the heart may not accommodate the flow of blood from the lungs to heart. As a result of these issues, the lungs and other parts of the body (e.g., ankles) may begin to back up with fluid. As a result, the individual may experience tiredness and difficulty breathing, which can ultimately lead to excessive fluid in the lungs (acute pulmonary edema). Heart failure is an extremely serious condition that can worsen if left untreated. There is no cure for heart failure, but several treatment and preventative measures are available. The most common causes are damage to the heart from a heart attack (myocardial infarction) and high blood pressure (hypertension).

Heart failure affects many people. Approximately 6.2 million in the United States and 600,000 in Canada are living with heart failure (Centers for Disease Control and Prevention, 2020b; Heart and Stroke Foundation, 2020b). As heart failure is potential life threatening, early diagnosis and treatment are imperative. Consequently, the ability to accurately forecast the development and course of this disorder, and other CVDs, is imperative. Although many risk factors and causes for heart failure has been identified, there is still much work to be done in this area. In particular, researchers and clinicians have yet to achieve high accuracy in predicting survival in patients with heart failure (Ahmad et al., 2017; Chicco & Jurman, 2020).

Machine learning in this context can be an efficient tool for developing predictive models and identifying useful features or risk factors for death from heart failure (Ahmad et al., 2017; Chicco & Jurman, 2020). That is, machine learning is a useful tool for predicting and identifying the most useful predictor variables. Machine learning has much potential in this context, as health records provide a plethora of potential useful information for uncovering relationships between various medical, physiological, and behavioural characteristics (and are often stored electronically). Various authors have attempted to use machine learning in this context; however, these reports generally yielded modest levels of accuracy at best and failed to achieve a consensus on the most important risk factors (Chicco & Jurman, 2020).

1.2 Aim of the Project

This report was prepared as the final project for the HarvardX PH125.9x Data Science Capstone course; the final course required for the Professional Certificate in Data Science. The specific requirements of this projects instructed students to:

- Apply machine learning techniques (i.e., at least two different models or algorithms must be used, with at least one being more advanced than linear or logistic regression)
- Use a publicly available dataset that must be automatically downloaded with the code
- Provide a written report documenting the analysis and presenting findings with supporting statistics and figures, as well as all the code used to prepare the report

This project aimed to examine the relationship between a number of potential predictor variables for death/survival in patients with cardiovascular disease, and then use machine learning to develop predictive models using a publicly available dataset. The dataset consisted of medical information for patients in heart failure collected in 2015 at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, Pakistan (Ahmad et al., 2017). Specifically, this project (a) begins with a brief description of the variables, (b) details the data wrangling processes used (i.e., downloading, cleaning, and checking the data), (c) explores the dataset using summary statistics, visualization, and traditional inferential statistics, and (d) describes the training of several machine learning algorithms and examines their performance in a test dataset

1.3 Dataset

In this report, I analyzed a publicly available dataset that was released by Ahmad and colleagues (2017). The dataset contains the records of 299 patients with heart failure from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, Pakistan from April to December 2015. All patients were diagnosed with left ventricular systolic dysfunction and prior heart failure. The dataset contains 299 rows each representing a patient and 13 columns each representing a feature of the patients (see Table 1).

Table 1: Variable Description, Measurement, and Range

Feature	**Description**	**Measurement**	**Range**
Age	Age of patients in years	Numeric - years	40 - 95
Anaemia	Decrease in red blood cells	Boolean	0, 1
Creatinine Phosphokinase	Level of CPK in the blood	Numeric - mcg/L	23 - 7,861
Diabetes	Presence of diabetes	Boolean	0, 1
Ejection Fraction	Percentage of blood leaving heart	Boolean	14 - 80
High Blood Pressure	Presence of hypertension	Numeric - percentage	0, 1
Platelets	Level of platelets in the blood	Numeric - kp/mL	25.01 - 850.00
Serum Creatinine	Level of creatinine in the blood	Numeric - mg/dL	0.50 - 9.40
Serum Sodium	Level of sodium in the blood	Numeric - mEq/L	114 - 148
Sex	Biological sex - man or woman	Binary	0, 1
Smoking	Presence of smoking	Boolean	0, 1
Time	Number of days to follow-up	Numeric - days	4 - 285
Death Event	Death of patient during follow-up	Boolean	0, 1

Adapted from Chicco & Jurman (2020)

mcg/L = micrograms per liter

kp/mL = kiloplatelets/microliter

mEq/L = milliequivalents per litre

As seen, the dataset contains 13 variables covering a range of clinical and behavioural features. Seven of the variables are continuous, while the remaining six are binary. Notably, the outcome or target is binary or dichotomous, representing the survival (death event = 0) or death (death event = 1) of the patients during the follow-up period. For additional details about the data, please see the original publication (Ahmad et al., 2017).

Notably, it is unclear exactly how the time variable (i.e., length of follow-up period in days) was measure across patients, and how the follow-up period ended if the patient died. As seen below, the time variable is strongly correlated with death in the dataset. This may be an artifact of the measurement of this variable, without more specific indication of how this variable was measured (see Limitations).

Several authors have previously analyzed this dataset using a variety of different techniques. For instance, Ahmad and colleagues (2017) used survival analysis to predict death and identify the most important features in the dataset. Their analysis found that age, ejection fraction, serum creatinine, serum sodium, anemia, and high blood pressure were statistically related to the likelihood of death. As well, Zahid and colleagues (2019) used a similar survival analysis approach separated by gender to produce separate predictive models. The final models both included seven variables (four shared across gender and three unique to each gender). Finally, Chicco and Jurman (2020) employed a combination of regression and machine learning techniques. They compared the results of 10 different machine learning strategies using all the variables in the dataset. Using feature ranking, they also identified serum creatinine and ejection fraction as the two most important variables, and created separated models using just these variables.

2 Data Wrangling and Inspection

The complete code to download the data is available in the supplemental code file. The included code will download the data and create two datasets. Due to known limitations in converting logical (i.e., Boolean) variables to factors in R, the included code creates two datasets that store the binary variables as logical variables and factors, respectively. The code also adds labels to the factor variables. The age variable was also converted to an integer variable, as only two values were not in integer format. Inspection of the data reveals no missing data for any variables and/or patients.

Table 2 presents the first ten rows of the data. As seen in the table, each row represents a single patient, and each column represents a different feature or variable. This is consistent with the description of the data having 299 patients (rows) and 13 features (columns). An examination of each of the variables in the dataset also reveals that each variable appropriately corresponds to the type in the description provided in Table 1.

Table 2: Examination of the Heart Data Structure

Age	**Anaemia**	**Creatinine Phosphokinase**	**Diabetes**	**Ejection Fraction**	**High Blood Pressure**	**Platelets**	**Serum Creatinine**	**Serum Sodium**	**Sex**	**Smoking**	**Time**	**Death Event**
75	No	582	No	20	Yes	265000	1.9	130	Male	No	4	Yes
55	No	7861	No	38	No	263358	1.1	136	Male	No	6	Yes
65	No	146	No	20	No	162000	1.3	129	Male	Yes	7	Yes
50	Yes	111	No	20	No	210000	1.9	137	Male	No	7	Yes
65	Yes	160	Yes	20	No	327000	2.7	116	Female	No	8	Yes
90	Yes	47	No	40	Yes	204000	2.1	132	Male	Yes	8	Yes
75	Yes	246	No	15	No	127000	1.2	137	Male	No	10	Yes
60	Yes	315	Yes	60	No	454000	1.1	131	Male	Yes	10	Yes
65	No	157	No	65	No	263358	1.5	138	Female	No	10	Yes
80	Yes	123	No	35	Yes	388000	9.4	133	Male	Yes	10	Yes

Overall, the data was extremely clean and user friendly in its original format. The conditions of the dataset required very little in terms of data wrangling (i.e., no need to gather, spread, separate, unite, or join). As mentioned, the data was also stored as a single case along each row with features or variables represented by columns. This structure was appropriate for the current project. As evident in Tables 1 and 2, the dataset contains no string or date variables. As there were no string variables included, there was no need to detect, locate, extract, or replace these variables. Similarly, the lack of date variables meant no need to process these variables into more useable formats. Notably, some of the variable names have been changed slightly since the original publication reporting on the dataset (Ahmad et al., 2017; Chicco & Jurman, 2020). These changes were already present in the downloaded version of the data.

3 Exploratory Data Analysis

Figure 1 presents a series of simple bar graphs presenting the raw frequency and percentage of each of the dichotomous variables in the dataset. Of note the graph in the upper left of the figure presents the frequency of the outcome (i.e., death event). From this plot, we see that more patients survived than not. Across the entire dataset, 203 patients survived (67.89%) and 96 patients died (32.11%) during the follow-up period. As well, the dataset contained 194 men (64.9%) and 105 women (35.1%). Notably, none of the dichotomous predictor variables display an exceptionally low base rate of occurrence. The outcome (i.e., patient mortality) and patient smoking have the lowest rates of occurrence at around 32% each.

Figure 1: Frequency Distributions of the Dichotomous Variables

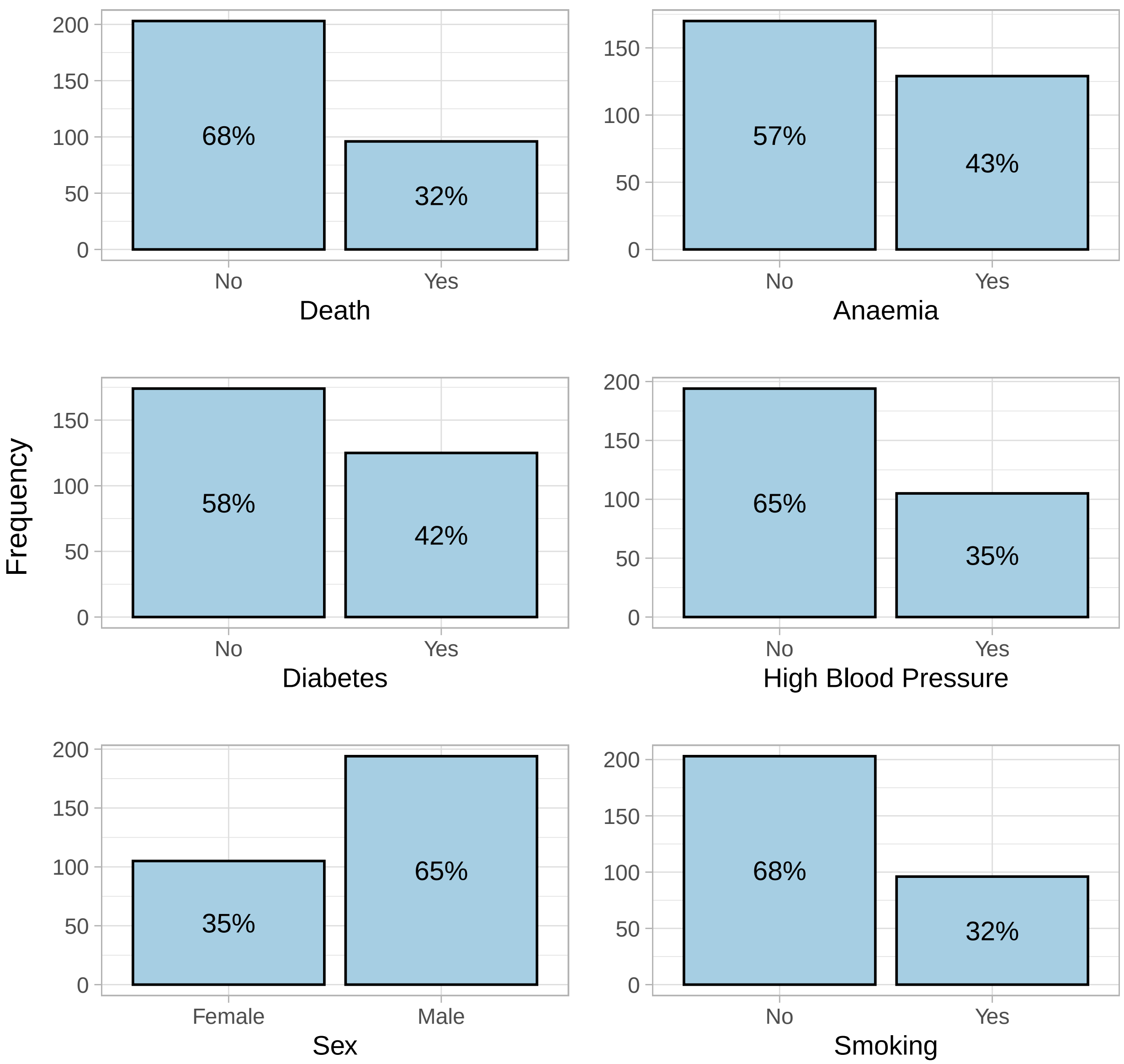
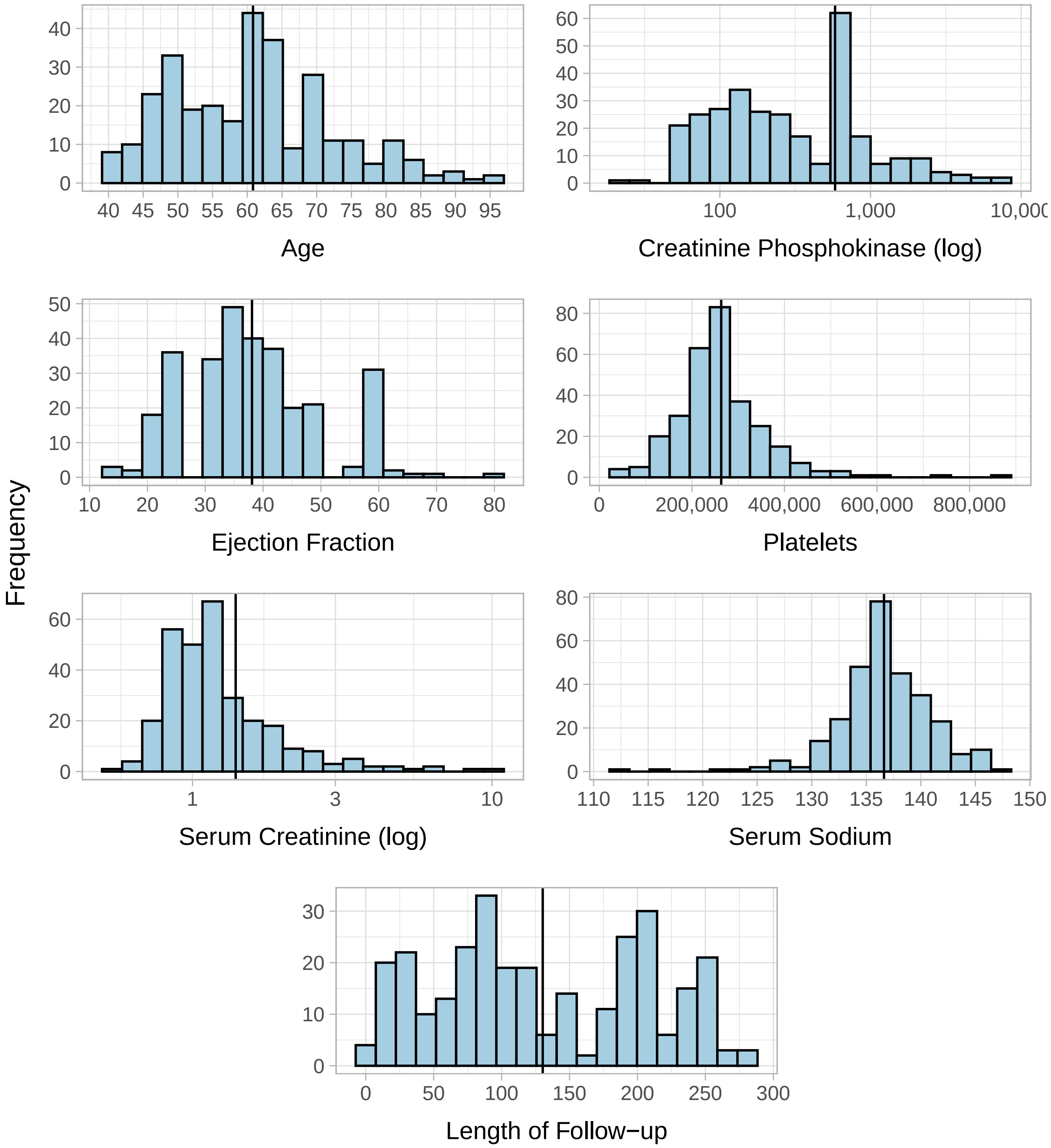


Figure 2 presents a series of frequency distributions for all the continuous variables in the dataset with a vertical line indicating the mean value. As seen in this figure, there is considerable variability across the continuous variables. Two of the distributions appear unimodal and relatively symmetric (i.e., platelets and serum sodium), although they still present with some degree of asymmetry with a single long tail. Notably, creatinine phosphokinase and serum creatinine are presented in log base

10 scale to ease interpretation. These distributions are positively skewed with many observations clustered at the lower end of the distribution and a long tail in the upper end. As well, three of the distributions (i.e., age, ejection fraction, and length of follow-up) have numerous peaks; that is, they appear multimodal. Due to the asymmetric and multipeaked nature of some of these distributions, it may not be appropriate to rely on statistical procedures that are based on the normal distribution. That is, the assumption that these variables have normal distributions in the population or multivariate normal distributions may be violated and caution should be used in employing procedures with these assumptions.

Figure 2: Frequency Distributions of the Continuous Variables



Next, the variables were examined separated by patient death. Figure 3 displays a series of stacked bar graphs of all the dichotomous variables separated by death event. The axis presents the proportion of cases while the text displays the raw frequency and death is represented by colour. Based on these visualizations, it seems that the rates of these variables were relatively consistent between patients who survived and those who did not. The largest differences are evident for high blood pressure (around 8% higher likelihood of death in those with high blood pressure) and anaemia (around 6% higher likelihood of death in those with anaemia). In contrast, there appears to be less than a percent difference in mortality rates across the other three variables (i.e., diabetes, sex, and smoking).

Figure 3: Bar Graphs of the Dichotomous Variables by Patient Death

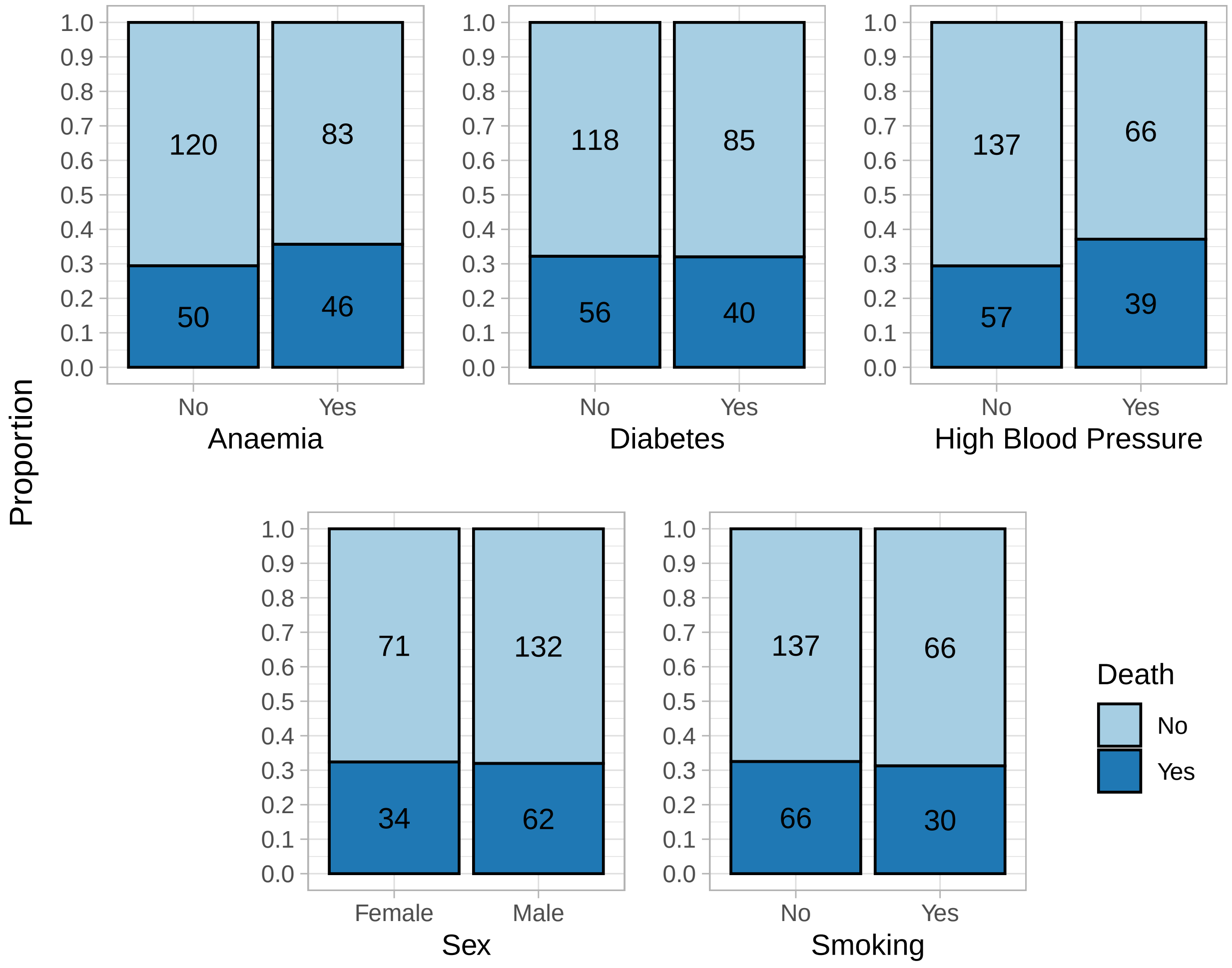


Table 3 presents descriptive statistics for the dichotomous variables. Specifically, the table presents the frequency and percentages for each of the variables overall and separated by death event. The table confirms the visualizations seen in Figure 3 with the rates of each feature remaining relatively consistent between patients who survived and those who did not. In particular, the mortality rates are quite consistent across diabetes, sex, and smoking.

Table 3: Descriptive Statistics for the Dichotomous Variables: Overall and by Death Event

		Death Event		
Characteristic	**Overall, N = 299**	**No, N = 203**	**Yes, N = 96**	**p-value**
Anaemia				0.25
No	170.00 (56.86%)	120.00 (59.11%)	50.00 (52.08%)	
Yes	129.00 (43.14%)	83.00 (40.89%)	46.00 (47.92%)	
High Blood Pressure				0.17
No	194.00 (64.88%)	137.00 (67.49%)	57.00 (59.38%)	
Yes	105.00 (35.12%)	66.00 (32.51%)	39.00 (40.62%)	
Diabetes				0.97
No	174.00 (58.19%)	118.00 (58.13%)	56.00 (58.33%)	
Yes	125.00 (41.81%)	85.00 (41.87%)	40.00 (41.67%)	
Sex				0.94
Female	105.00 (35.12%)	71.00 (34.98%)	34.00 (35.42%)	
Male	194.00 (64.88%)	132.00 (65.02%)	62.00 (64.58%)	
Smoking				0.83
No	203.00 (67.89%)	137.00 (67.49%)	66.00 (68.75%)	
Yes	96.00 (32.11%)	66.00 (32.51%)	30.00 (31.25%)	

[1] n (%)

[2] Pearson's Chi-squared test

Figure 4 displays a series of boxplots of all the continuous variables separated by death event. Several of these variables appear to differ between patients who survived and those who did not. Based on the boxplots, it appears that patients who died were generally older, with lower ejection fraction, higher serum creatinine, and shorter follow-up periods. Serum sodium may also differ across death event, but the pattern is less apparent. In contrast, creatinine phosphokinase and platelet levels do not appear to differ across death event.

Figure 4: Boxplots of the Continuous Variables by Patient Death

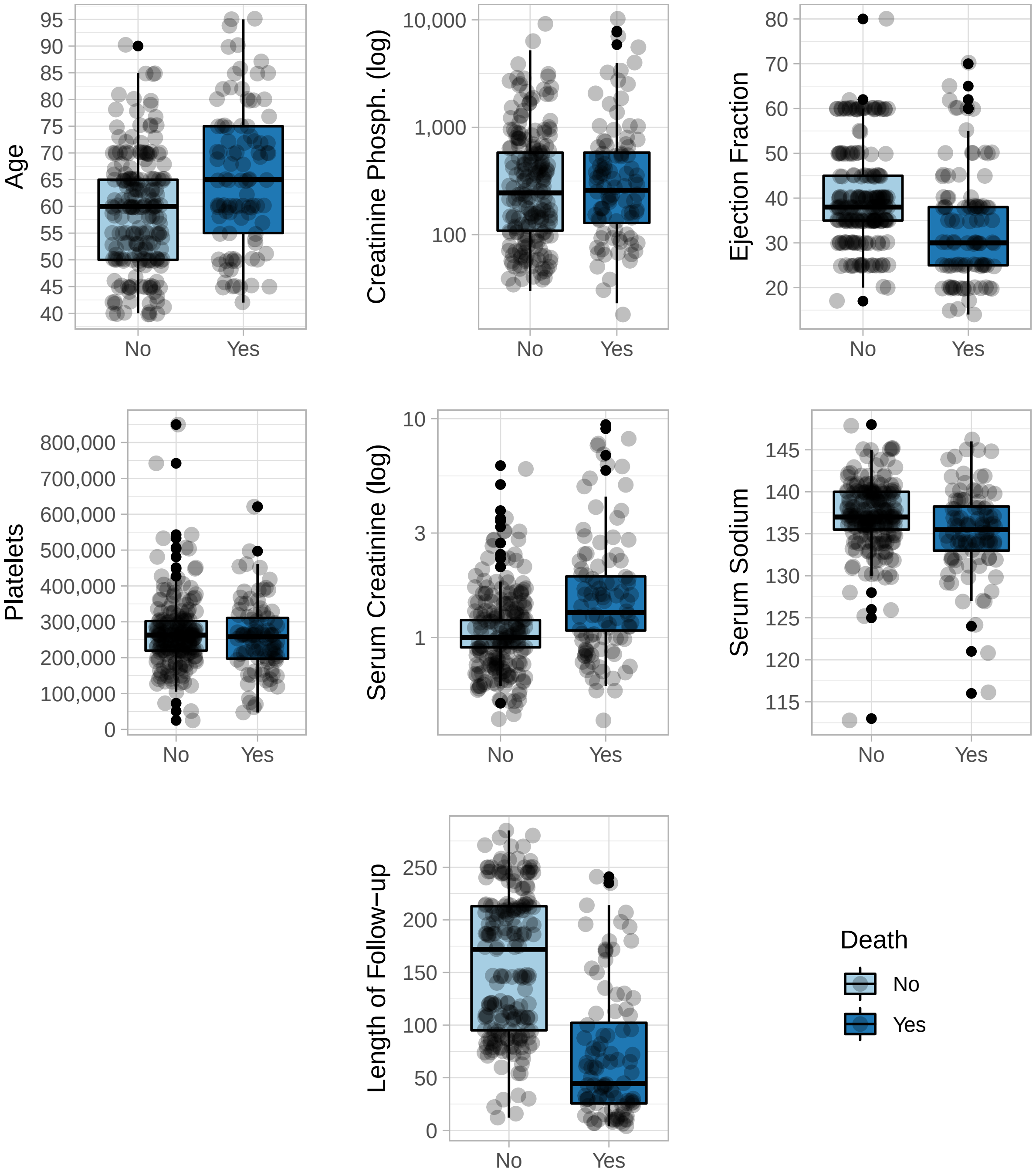


Table 4 presents descriptive statistics for the continuous variables. Specifically, the table presents summary statistics of the continuous variables over the entire dataset and separated by patient mortality. Evident from this table is the positive skew seen in creatinine phosphokinase and serum creatinine, as they demonstrate considerable divergence between the mean and median values both overall and across death event. Of note, the time variable (i.e., days to follow-up) shows inconsistent differences between the mean and median values across the groups. Consistent with the plots seen in Figure 4, this table highlights the fact that many of these features varied across patients based on their survival during the follow-up period.

Table 4: Descriptive Statistics for the Continuous Variables: Overall and by Death Event

		Death Event		
Characteristic	**Overall, N = 299**	**No, N = 203**	**Yes, N = 96**	**p-value**
Age				<0.001
Mean (SD)	60.83 (11.89)	58.76 (10.64)	65.21 (13.22)	
Median (IQR)	60.00 (51.00, 70.00)	60.00 (50.00, 65.00)	65.00 (55.00, 75.00)	
Creatinine Phosphokinase				0.68
Mean (SD)	581.84 (970.29)	540.05 (753.80)	670.20 (1,316.58)	
Median (IQR)	250.00 (116.50, 582.00)	245.00 (109.00, 582.00)	259.00 (128.75, 582.00)	
Ejection Fraction				<0.001
Mean (SD)	38.08 (11.83)	40.27 (10.86)	33.47 (12.53)	
Median (IQR)	38.00 (30.00, 45.00)	38.00 (35.00, 45.00)	30.00 (25.00, 38.00)	
Platelets				0.43
Mean (SD)	263,358.03 (97,804.24)	266,657.49 (97,531.20)	256,381.04 (98,525.68)	
Median (IQR)	262,000.00 (212,500.00, 303,500.00)	263,000.00 (219,500.00, 302,000.00)	258,500.00 (197,500.00, 311,000.00)	
Serum Creatinine				<0.001
Mean (SD)	1.39 (1.03)	1.18 (0.65)	1.84 (1.47)	
Median (IQR)	1.10 (0.90, 1.40)	1.00 (0.90, 1.20)	1.30 (1.08, 1.90)	
Serum Sodium				<0.001
Mean (SD)	136.63 (4.41)	137.22 (3.98)	135.38 (5.00)	
Median (IQR)	137.00 (134.00, 140.00)	137.00 (135.50, 140.00)	135.50 (133.00, 138.25)	
Length of Follow-up				<0.001
Mean (SD)	130.26 (77.61)	158.34 (67.74)	70.89 (62.38)	
Median (IQR)	115.00 (73.00, 203.00)	172.00 (95.00, 213.00)	44.50 (25.50, 102.25)	

[1] Mean (SD) or Median (IQR)

[2] Wilcoxon rank sum test

Figures A1 through A5 present visualizations of the continuous variables across each of the five dichotomous features grouped by death event (see Appendix A for supplemental graphs). Each of the figures focused on a different dichotomous feature, anaemia in Figure A1, diabetes in Figure A2, high blood pressure in Figure A3, sex in Figure A4, and smoking in Figure A5. For each of the dichotomous features represented on the vertical axes, the figure presents a series of visualizations for each of the continuous variables represented on the horizontal axes separated by death event represented by colour and shape. The graphs also include the mean values represented by the larger black shapes.

Inspection of these figures highlights the impact of each of the continuous features on the likelihood of death across the dichotomous features. Overall, these figures reveal no obvious interactions:

- With respect to anaemia (Figure A1), the effects of the continuous variables appear to be consistent between those with and those without anaemia. Only creatinine phosphokinase varies slightly in mean values across those with and without anaemia.
- Looking at diabetes (Figure A2), the continuous variables appear to display similar patterns across those with and those without diabetes. Minor differences are seen in some variables; slightly larger differences in age and ejection fraction are present for those without diabetes.
- Examining Figure A3, the patterns in the continuous variables with respect to patient death appear to be rather consistent across those with and without high blood pressure. Slightly larger mean differences are apparent in creatinine phosphokinase and ejection fraction for those with diabetes.
- The same consistency in the impacts of the continuous features is seen across sex in Figure A4. The effect of age and ejection fraction appear slightly larger in men than women, whereas platelet levels may have greater effect in women than men.
- Finally, the variables do not display any obvious variability in effects across smoking (Figure A5). From these graphs, age and ejection fraction appear to be more discriminative in smokers than non-smokers.

Additionally, Figures A6 and A7 present visualizations of the continuous features across each other grouped by death event (see Appendix A). Figure A6 presents days to follow-up on the vertical axes with various other continuous features along the horizontal axes separated by death event represented by colour and shape. The graph also includes ellipses around the data separated by patient death. It is apparent from Figure A6 that overall patients are more likely to die earlier in the follow-up period (i.e., fewer days in the follow-up period), which is consistent with what was observed in Figure 4 and Table 4. Also consistent with the previous figures, the likelihood of death appears to be higher in older patients, as well as patients with lower ejection fraction and higher serum creatinine levels.

Figure A7 presents the bivariate distributions of select continuous variables (i.e., age, ejection fraction, serum creatinine, and serum sodium) separated by death event represented by colour and shape. The graph also includes ellipses around the data separated by patient death. Once again, this figure reveals the influence of age, ejection fraction, and serum creatinine on the likelihood of death. Overall, based on these visualizations, no strong, clear, or identifiable interactions seem to be present.

4 Traditional Inferential Analysis

Several inferential procedures were used to analyze the data. Specifically, Mann-Whitney U tests, Pearson's chi-squared tests, and correlations were used to investigate the relationship between each of the features and patient death. Each of these tests was used based on the type of feature being investigated (i.e., dichotomous, or continuous) to determine the effect of that feature on the likelihood of death. In addition, multivariate logistic regression was used to determine the relationship between all the features and patient death.

The relationships between the dichotomous features and death event were first examined using Pearson's Chi-squared test for independence. This procedure tests the likelihood that an observed difference in the frequencies across the categories is due to chance (Howell, 2013; Privitera, 2018). That is, this test examines the extent to which the frequencies of two categorical variables are dependent upon each other and, thus, related to each other. Table 3 presents the results of the Chi-squared tests along with the descriptive statistics for all the dichotomous features (see above). As seen, none of the dichotomous predictors influenced the likelihood of patient death according to these analyses (i.e., all p-values in Table 3 are non-significant). In other words, the likelihood of dying does not vary across levels of these dichotomous features.

Mann-Whitney U tests (also known as the Wilcoxon rank sum tests) were used for the continuous variables. This test examines whether the distributions of each of the feature are the same across levels of the outcome (Howell, 2013; Privitera, 2018). That is, the procedure tests whether the distribution of each feature in those that survived and those that died are from the same population. Table 4 presents the results of the Mann-Whitney U tests in the final column (see above). According to these analyses, five of the continuous variables are related to patient death: age, ejection fraction, serum creatinine, serum sodium, and length of follow-up. This is consistent with the visualizations in Figure 4.

In addition, correlations between all the variables were examined. Correlation coefficients describe the strength and direction of a given type of relationship between two variables (Howell, 2013; Privitera, 2018). Linear relationships were investigated in this instance. Correlations were examined independently for each variable using Pearson's correlations, point-biserial correlations, and phi correlations as appropriate depending on the types of variables (ie., two continuous, one continuous and one dichotomous, or two dichotomous, respectively).

Figure 5 presents the correlations between all the variables in the dataset. This figure displays the correlation coefficients represented by size and colour. The significance level of each of the correlations is also indicated in the figure (i.e., * = $p < .05$, ** $< .01$, *** = $p < .001$). The results of these analyses are consistent with the data presented in Tables 3 and 4. The same five features are related to patient death: age, ejection fraction, serum creatinine, serum sodium, and length of follow-up. In particular, patients are more likely to die if they are older, and have higher serum creatinine, lower ejection fraction, and lower serum sodium. As well, time or length of follow-up was correlated with patient death; however, it is unclear whether this is due to the measurement of this variable or a legitimate effect (see Limitations).

Several other notable relationships are evident in Figure 5. In addition to being correlated with patient death, age is also correlated with serum creatinine and length of follow-up (see Figure A6 for bivariate plots). Ejection fraction is correlated with serum sodium and sex (see Figure A7 and A4 for bivariate plots, respectively). Serum creatinine is also correlated with serum sodium and length of follow-up (see Figure A6 and A7, respectively).

Focusing on the prediction of patient mortality, Table 5 displays the correlation coefficients and significance level for each of the features in descending order of strength. The strongest relationship is evident for length of follow-up ($r = -.527$). However, due to the reasons mentioned, this variable may be unreliable. Of the remaining features that are related to patient death, the strongest relationships only account for an estimated 8.7% (serum creatinine) and 7.2 % (ejection fraction) of the variability in patient mortality.

Figure 5: Correlation Matrix

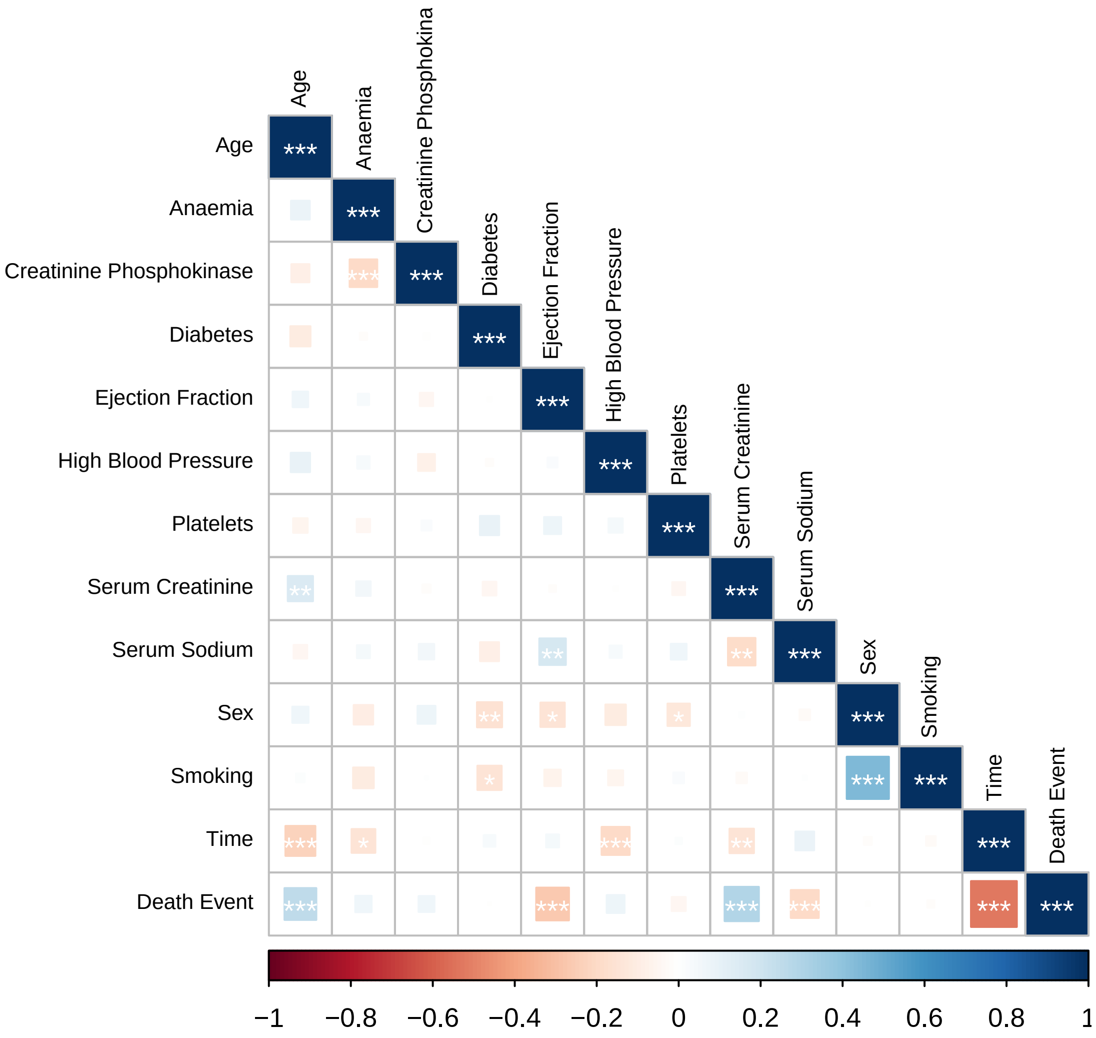


Table 5: Correlations with Death Event

	r squared	r	p-value
Time	0.278	-0.527	0.000
Serum Creatinine	0.087	0.294	0.000
Ejection Fraction	0.072	-0.269	0.000
Age	0.064	0.254	0.000
Serum Sodium	0.038	-0.195	0.001
High Blood Pressure	0.006	0.079	0.171
Anaemia	0.004	0.066	0.253
Creatinine Phosphokinase	0.004	0.063	0.280
Platelets	0.002	-0.049	0.397
Smoking	0.000	-0.013	0.828
Sex	0.000	-0.004	0.941
Diabetes	0.000	-0.002	0.973

Finally, logistic regression was used to determine the relationship between all the features and patient mortality. Logistic regression is a regression modeling technique for fitting a function to data in which the outcome is dichotomous or binary (Howell, 2013). Two logistic regression analyses were performed. First, all the variables were included in the logistic regression model. Then, due to the unknown and possibly unreliable measurement of the time variable, it was removed and the analysis was performed again.

The results of both models' overall goodness of fit are presented in Table 6. This table presents three common pseudo-R-squared coefficients that assess the overall effect of the model. In all instances larger values correspond to better model fit. As seen, the regression model including all the predictors yielded better overall goodness of fit compared to the model without this variable. The time variable appears to be strongly related to patient death, as the second analysis produced quite lower values.

Table 6: Logistic Regression - Overall Model Fit

	All Features	**No Time Feature**
McFadden	0.415	0.205
Cox and Snell (ML)	0.406	0.236
Nagelkerke (Cragg and Uhler)	0.568	0.323

In addition, Table 7 and 8 present the logistic regression results for the individual feature coefficients. Table 7 presents the coefficients for the model including all features, and Table 8 presents the coefficients for the model that does not include the time variable as a predictor. The bolded rows in these tables correspond with the features found to be independently contributing to the regression models.

The results of the logistic regression models are largely consistent with the previous analyses. For instance, with all features included in the regression, four variables are related to patient mortality: age, ejection fraction, serum creatinine, and time. Note that serum sodium, which was correlated with patient death in the bivariate analyses, was no longer independently contributing to the regression model.

When time is not included in the model as seen in Table 8, only three features are related to patient death: age, ejection fraction, and serum creatinine. Note that in this instance, when time is removed, the associations between creatinine phosphokinase and sodium serum with patient death increase (although still slightly below reaching statistical significance).

Table 7: Logistic Regression Coefficients - All Features

	Estimate	**Std. Error**	**Z**	**p**
(Intercept)	10.191	5.656	1.802	0.072
Age	**0.047**	**0.016**	**2.995**	**0.003**
Anaemia	-0.007	0.360	-0.019	0.985
Creatinine Phosphokinase	0.000	0.000	1.248	0.212
Diabetes	0.145	0.351	0.414	0.679
Ejection Fraction	**-0.077**	**0.016**	**-4.694**	**0.000**
High Blood Pressure	-0.103	0.359	-0.286	0.775
Platelets	0.000	0.000	-0.634	0.526
Serum Creatinine	**0.666**	**0.181**	**3.670**	**0.000**
Serum Sodium	-0.067	0.040	-1.686	0.092
Sex	-0.533	0.414	-1.288	0.198
Smoking	-0.014	0.413	-0.033	0.973
Time	**-0.021**	**0.003**	**-6.981**	**0.000**

Table 8: Logistic Regression Coefficients - No Time Feature

	Estimate	**Std. Error**	**Z**	**p**
(Intercept)	1.498	0.794	1.886	0.060
Age	**0.009**	**0.002**	**4.296**	**0.000**
Anaemia	0.055	0.050	1.095	0.275
Creatinine Phosphokinase	0.000	0.000	1.905	0.058
Diabetes	0.017	0.050	0.332	0.740
Ejection Fraction	**-0.011**	**0.002**	**-5.032**	**0.000**
High Blood Pressure	0.068	0.051	1.328	0.185
Platelets	0.000	0.000	-0.285	0.776
Serum Creatinine	**0.106**	**0.024**	**4.408**	**0.000**
Serum Sodium	-0.011	0.006	-1.914	0.057
Sex	-0.063	0.058	-1.071	0.285
Smoking	0.013	0.058	0.229	0.819

Based on traditional descriptive and inferential analyses, many of the variables in the dataset have potential for predicting patient mortality using machine learning models. The time variable, representing the length of the follow-up period, presented the strongest association with patient death, yet the measurement of this variable presents problems for inclusion in predictive models. Resultingly, age, ejection fraction, and serum creatinine presented with the strongest associations with patient mortality. Serum sodium was only associated with patient death in the bivariate analyses, but not the logistic regression analyses.

5 Modeling Approaches

The following sections concern the binary prediction of patient death during the follow-up period using machine learning algorithms. In attempts to predict patient death, six different machine learning models were employed: boosted generalized linear model (GLM), k-nearest neighbours (KNN), generalized additive model using loess (gamLoess), conditional inference random forest (cForest), quantile random forest (Random Forest or rf), and classification and regression tree (rpart).

Each of these approaches is suited for this classification task with their respective advantages and disadvantages (Irizarry, 2019). For instance, KNN models involve grouping the data into strata or neighbourhoods that are thought to be constant and averaging across these neighbourhoods. The number of neighbourhoods is a tuning parameter that was determined using cross-validation (discussed below). The two random forest approaches (cForest and RF) involve constructing many decision trees and then taking the mode or average of these trees for prediction. Random forest techniques also randomly determine which features to include in each decision tree. The default tuning parameters were used from the caret package for all the models (except the GLM models which does not have any tuning parameters).

Based on the exploratory data analysis and inferential analysis, several strategies were employed during the machine learning stage. Overall, it appears that age, ejection fraction, and serum creatinine have the greatest potential for predicting patient death. Although, the length of the follow-up period was also strongly correlated with patient death, it was excluded from these models (see Limitations). There was some evidence that additional variables may be helpful in the predictive models, and many machine learning algorithms benefit from including a larger number of features. As such, the first set of predictive models included all the features (except time), while the second set included only a select number of features. The features for the second set of models were selected based on the bivariate and multivariate relationships presented above (see Inferential Analysis), as well as feature ranking in the first set of predictive models with all features.

In this instance, the task is that of classification, as the outcome is categorical and specifically dichotomous (Irizarry, 2019). As such, many of the machine learning algorithms will provide a decision rule of the form – if $f_1(x_1,x_2,\ldots,x_p) > C$, predict the positive category, otherwise predict the negative category, with C equal to a specific cut-off. As the outcome is dichotomous, our prediction will be either right or wrong. This has implications for how we measure the accuracy of the predictive models. In this project, accuracy was assessed using several metrics. The simplest metric is the proportion of cases that are correctly classified, known as overall accuracy. Table 9 shows the common metrics available to judge accuracy in this context.

Table 9: Model Accuracy Metrics

Feature	**Description**
Accuracy	Proportion of true positives and true negatives over all instances
Kappa	Measure of agreement accounting for random chance*
Sensitivity	Proportion of true positives over actual positives
Specificity	Proportion of true negatives over actual negatives
PPV	Proportion of true positives over predicted positives*
NPV	Proportion of true negatives over predicted negatives*
Precision	Proportion of true positives over predicted positives
Recall	Proportion of true positives over actual positives
F1	Harmonic average of precision and recall*
Prevalence	Proportion of actual positives over total
Detection Rate	Proportion of true positives over total
Detection Prevalence	Proportion of predicted positives over total
Balanced Accuracy	(sensitivity + specificity)/2

Adapted from Irizarry (2019) and the caret package description.

* These metrics are calculated using more complex definitions in the caret package.

5.1 Data Preparation

First, the data was randomly split into training (80%) and test (20%) datasets. An 80/20 split was used due to the relatively small size of the dataset (n = 299). With this split, the test dataset will have greater potential of providing a stable estimate of the accuracy of the models. The size of the dataset presents a limitation as any splitting results in proportionality smaller datasets that may not correspond with the overall patterns amongst the variables (see Limitations). After splitting the dataset, the training dataset included 238 patients with 162 (68.1%) survived and 76 (31.9%) deceased, while the test dataset contained 61 patients with 41 (67.2%) survived and 20 (32.8%) deceased.

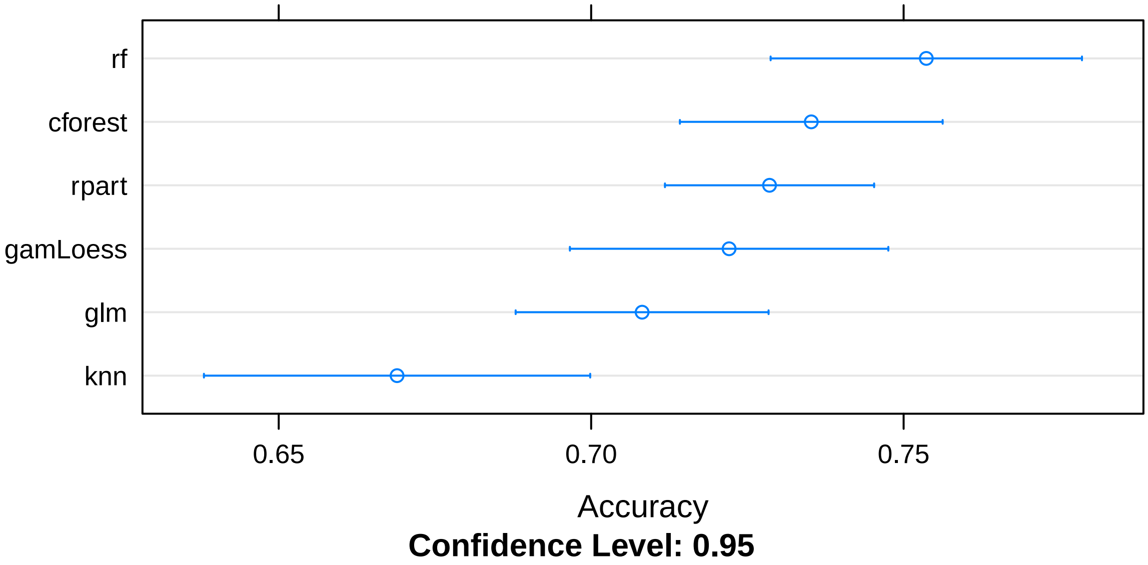
Cross-validation was used during the training phase to maximize the use of the small sample size and to tune the hyper-parameters for the machine learning models. As the dataset is relatively small, computation time was not a major concern in increasing the number of folds in the cross-validation. Larger numbers of folds also provide better estimate of the overall data; thus, 100-fold cross validation was used selecting 90% of the training dataset. Reminder, the code used to partition the dataset and set the cross-validation parameters, as well as all code for this project, is available in the supplemental code file.

Notably, the training dataset will be used for developing several predictive models and tuning the model parameters. The various models will be compared first on their performance in the training dataset with cross-validation. That is, each of the models will be examined based on their performance across the cross-validation in the training dataset, specifically focusing on overall accuracy and kappa values. Subsequently, the best models will be evaluated using the test dataset (i.e., the test dataset will only be used to evaluate the final models). The performance of the models in the test dataset will include all the metrics described in Table 9; however, balanced accuracy and F1 will be used as the primary metrics for comparison across models in the test phase.

5.2 Model Training and Tuning

After splitting the dataset, all the machine learning models were run using all the features (except time) as predictors. Figure 6 presents the overall accuracy of all six models using cross-validation in the training dataset. Overall, the random forest techniques (RF and cForest) yielded the largest overall accuracy, whereas KNN yielded the lowest. Most of the models achieved similar accuracy between 0.71 and 0.75, except the KNN model. The KNN model also showed more variability compared to the other models, with rpart producing the lowest variability in accuracy with cross-validation.

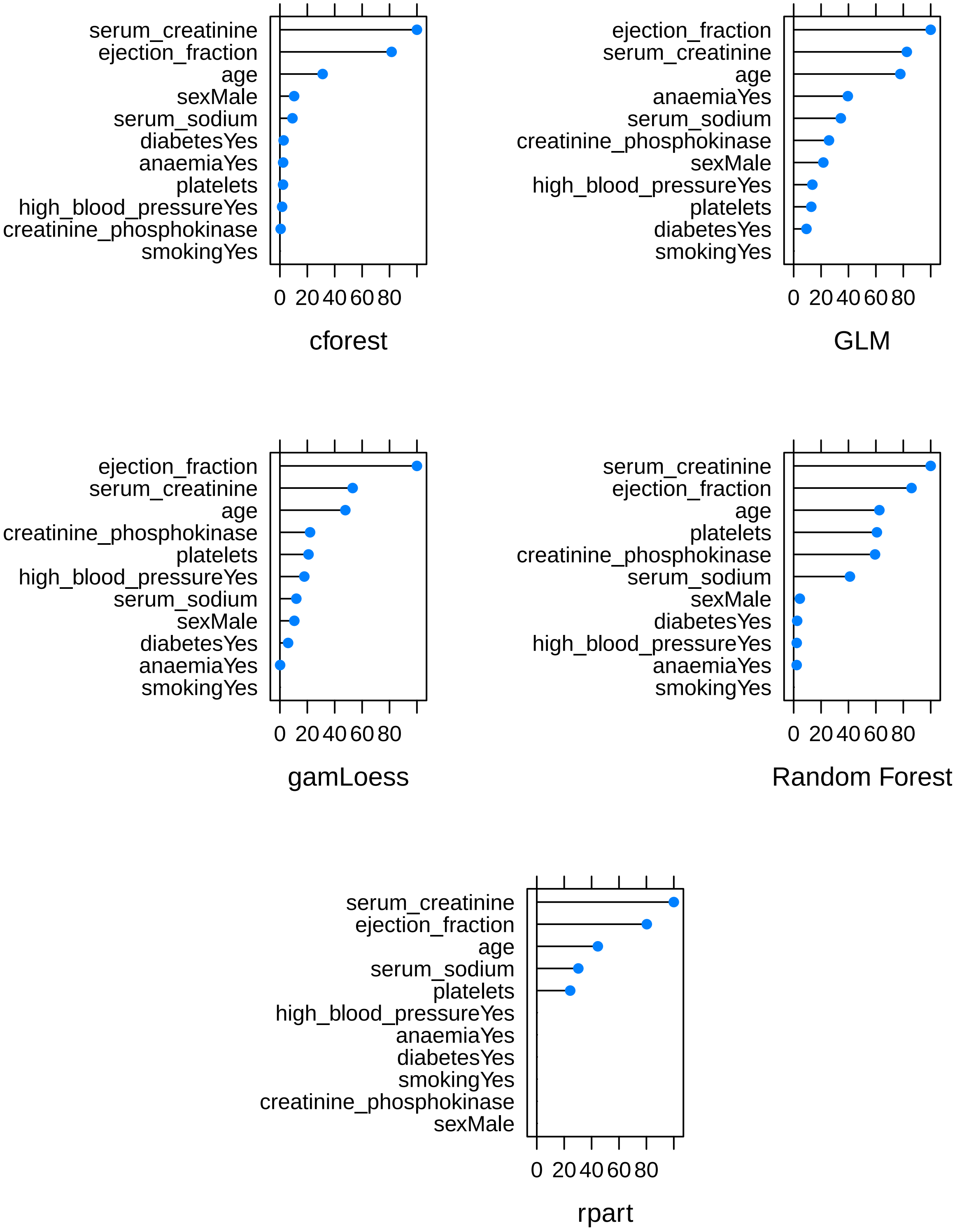
Figure 6: Accuracy in Cross−Validation − All Features



Four of the machine learning models included tuning parameters. Figure A8 presents the results of the tuning parameters for relevant models using cross-validation (see Appendix). The default tuning options provided for each machine learning model were used.

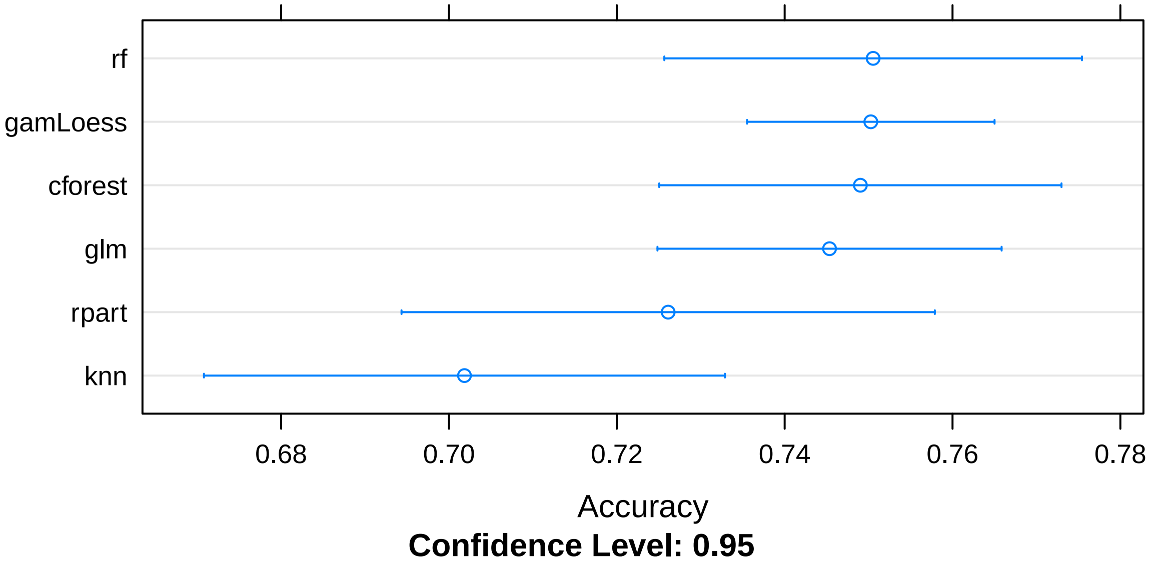
Figure 7 displays the results of the feature ranking analysis provided by several of the machine learning models. These plots indicate the relative importance of each feature in the predictive models. The plots are largely consistent with the patterns identified earlier based on the visualizations and inferential analyses. Across all the models, serum creatinine and ejection fraction were listed as the most important features in the model. That is, all models have these two features as the first and second highest ranking (i.e., serum creatinine ranked first in three models and ejection fraction ranked first in the other two). Age is consistently ranked as the third most important feature in the models. These are the variables that were previously identified as having the strongest relationships with patient mortality.

Figure 7: Variable Importance across Models



Based on the feature ranking presented in Figure 7 and the previous exploratory and inferential analyses, a second set of predictive models was trained using only age, ejection fraction, and serum creatinine as predictors. Figure 8 presents the overall accuracy of the six models with only these three features. Generally, the models yielded higher accuracy using only these three predictors. The two random forest models (cforest and RF), the linear model (GLM), and the additive model (gamLoess) produced the largest accuracies around 0.75. KNN again produced the lowest accuracy in the cross-validation.

Figure 8: Accuracy in Cross−Validation − Select Features



Additionally, Table 10 displays the same results of the machine learning models including both accuracy and kappa values along with standard deviations. Looking at the table, it is more evident that four of the models improved by only including the three features (cForest, GLM, KNN, and Loess). The accuracy and kappa for these models improved with only three features. In contrast, the accuracy was worse but the kappa value better for the random forest model with only three predictors. Finally, the rpart model yielded better accuracy and kappa values when all variables were included in the model.

Table 10: Accuracy in Cross-Validation

	All Features				**Select Features**			
Model	**Accuracy**	**(SD)**	**Kappa**	**(SD)**	**Accuracy**	**(SD)**	**Kappa**	**(SD)**
cForest	0.735	0.029	0.370	0.075	0.749	0.033	0.425	0.083
GLM	0.708	0.028	0.307	0.081	0.745	0.029	0.374	0.086
KNN	0.669	0.037	0.154	0.080	0.702	0.038	0.305	0.101
Loess	0.722	0.036	0.340	0.093	0.750	0.021	0.407	0.067
RF	0.754	0.025	0.428	0.075	0.751	0.035	0.440	0.071
rpart	0.729	0.023	0.374	0.073	0.726	0.043	0.370	0.111

From these results, the generalized linear model (GLM) and k-nearest neighbours (KNN) model were dropped from the analyses. The other four machine learning models (i.e., cForest, Loess, RF, and rpart) were selected based on their performance across both the full feature and selected features models. Therefore, eight final models (i.e., four machine learning techniques with all and selected features) were chosen to be compared in the test dataset. Part of this decision was aimed at allowing for comparisons across models with multiple features, which can increase the model's robustness or generalizability to new data, to those with only select features, which may be more accurate due to the reduction in noise from the other variables.

6 Results on Validation Dataset

Table 11 presents the results of the predictive models in the test dataset. As a reminder, this dataset was partitioned from the original dataset prior to training and tuning the machine learning algorithms; it is only being used to evaluate the final models. The models performed similar in the test dataset compared to the cross-validation results. That is, comparing Table 11, which shows the results in the test dataset, with Table 10, which shows the results using cross-validation in the training dataset, the accuracies were in the same general range (around 0.70 to 0.75).

Table 11: Model Results in the Test Dataset

	All Features				**Select Features**			
	cForest	**Loess**	**RF**	**rpart**	**cForest**	**Loess**	**RF**	**rpart**
Sensitivity	0.600	0.550	0.600	0.400	0.700	0.500	0.450	0.750
Specificity	0.854	0.951	0.829	0.902	0.854	0.927	0.902	0.829
Pos Pred Value	0.667	0.846	0.632	0.667	0.700	0.769	0.692	0.682
Neg Pred Value	0.814	0.813	0.810	0.755	0.854	0.792	0.771	0.872
Precision	0.667	0.846	0.632	0.667	0.700	0.769	0.692	0.682
Recall	0.600	0.550	0.600	0.400	0.700	0.500	0.450	0.750
F1	0.632	0.667	0.615	0.500	0.700	0.606	0.545	0.714
Prevalence	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328
Detection Rate	0.197	0.180	0.197	0.131	0.230	0.164	0.148	0.246
Detection Prevalence	0.295	0.213	0.311	0.197	0.328	0.213	0.213	0.361
Balanced Accuracy	0.727	0.751	0.715	0.651	0.777	0.713	0.676	0.790

Comparing across models that included all features and those including the three selected features, the cForest and rpart models performed better with the selected variables across nearly all metrics. In particular, these models increased their sensitivity and negative predictive values with only three features. In contrast, the Loess model yielded generally better results across all metrics using all features, whereas the random forest model varied depending on the metric being considered.

With respect to balanced accuracy, the best models were rpart with select features, cForest with select features, and Loess with all features. These models also had the three highest F1 scores. In fact, the rpart model using selected features yielded the highest balanced accuracy and F1, as well as sensitivity, negative predictive value, recall, detection rate, and detection prevalence. As well, the Loess model with all features yielded the highest specificity, positive predictive value, and precision, although it had the third highest balanced accuracy and tied for second highest F1.

Notably, the rpart model with select variables that performed best in the test dataset did not perform as well in the cross-validation. Many other models outperformed this rpart model in Table 10 based on accuracy and kappa values; however, it provided the top results in the test datatset (Table 11).

7 Discussion

The aim of this project was to predict patient mortality from behavioural, biological, and physiological features using a publicly available dataset. The dataset consisted of records from 299 patients with heart failure who attended one of two healthcare facilities in Pakistan during 2015. The supplemental file includes all the code used for this project, including the code used to download, inspect, and clean the dataset.

Exploratory data analysis revealed a number of features that may be of particular use in predicting patient death. Based on visualizations, it appeared that age, ejection fraction, serum creatinine, serum sodium, and length of the follow-up were the best features to use for predicting patient death. Inferential analyses including Mann-Whitney U test, Pearson's Chi-squared test, correlations, and logistic regression confirmed the observations made based on the data visualizations. Due to the measurement of the length of the follow-up period (i.e., time variable), it was excluded from subsequent analyses. As a result, three features remained as the most viable individual predictors of patient mortality: age, ejection fraction, and serum creatinine.

Next, six machine learning models were run with all the features included in the models and with only the three most important features included in the models. The dataset was split into a training dataset (80%) and a test dataset (20%). Cross-validation was used in the training dataset to tune the model parameters and assess the potential of each model. The various models were examined first based on their performance in the cross-validation in the training dataset, and subsequently, the final models were compared based on their performance in the test dataset. A variety of metrics described in Table 9 were used to compare the predictive utility of the models. The final models used in the test dataset included a generalized additive model using loess (gamLoess), conditional inference random forest (cForest), quantile random forest (Random Forest or rf), and classification and regression tree (rpart). Overall, the rpart model using only three selected features yielded the most accurate predictions in the test dataset.

The results of this project are generally consistent with the previous published studies that have examined the same dataset. For instance, Ahmad and colleagues (2017) found that age, ejection fraction, serum creatinine, serum sodium, anaemia, and high blood pressure were impacting the likelihood of patient death. Four of these six variables were also identified in the current project as statistically related to patient death. These authors also found that diabetes, sex, and smoking were not related to patient death, which was also seen in the current project. In addition, Chicco and Jurman (2020) used a variety of inferential and machine learning techniques to examine this dataset. They identified ejection fraction and serum creatinine as the most important variables for predicting patient death, largely consistent with the variables identified in the current project. Using 100-fold cross-validation and examining the mean performance across the 100 executions, they found that a random forest model using only the two identified features can be more accurate than including all the features. This is again consistent with the results found here, as some models performed better including only selected variables with high feature ranking. Overall, these results confirm that it is possible to achieve an encouraging level of accuracy in predicting patient death from available features in patients with heart failure.

7.1 Limitations

Several limitations are noteworthy. First, the size of the dataset (n = 299) was a drawback of the current study. The relatively small size of the dataset raised additional challenges in splitting the data into training and test datasets. Splitting the data when the number of observations is low can be problematic as the test dataset is proportionality smaller, which may impact the representativeness and generalizability of the test dataset. A larger sample size would have also provided more reliable results in terms of sample estimates and accuracy metrics (i.e., larger sample sizes provide more accurate estimates of the underlying parameters). Relatedly, it would have been beneficial to have a separate dataset (i.e., not split from the original) to use after developing the models to provide a better test of their accuracy.

Second, the measurement of certain variables posed some limitations. As mentioned above, the time variable representing the number of days in the follow-up period did not have any detailed explanations as to its measurement. The length of the follow-up period was also strongly correlated with patient death. Accordingly, it may be the case that the number of days surviving with heart failure is predictive of the length of time until death, which would be an informative and useful predictor, or it may be the case that the length of the follow-up period correlated with death due to the way this variable was measured in this dataset. As a result of the unreliability of this measurement, it was excluded from the machine learning models. As well, several variables were recorded in the dataset as binary or dichotomous variables, yet they were based on continuous measurements that may have performed differently in predictive models (i.e., anaemia and high blood pressure).

7.2 Future Directions

Machine learning is a useful tool that could be used to improve upon current practices for identifying patients at greatest risk for illness, disorder progression, and death. With respect to the current project, the predictive models should be evaluated in entirely new datasets to continue to understand their predictive accuracy. However, this would require finding another dataset that has the same features and outcome with the same operationalizations and measurement, or the collection of a new validation dataset. Ideally, a much larger dataset could be found to investigate these features further.

Moving forward, several steps could be taken to improve upon the methods and results obtained here. Future work would benefit from gathering additional variables for inclusion in the machine learning models. Generally, health records include a wealth of information. Researchers should consider the information that is typically available to health care professionals. Potential features should be selected based on prior empirical research, theoretical association with the outcome, and ease of access or availability. The interpretability of the features and models should also be considered with respect to implementation and utility of future models. Machine learning models should be further investigated for other major cardiovascular diseases as well as other types of serious illnesses. Predictive models should investigate disease diagnosis, prognosis, and death.

8 References

- Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PLoS ONE*, *12*(7), e0181001. https://doi.org/10.1371/journal.pone.0181001
- Centers for Disease Control and Prevention. (2020a, September 8). *Heart disease facts.* https://www.cdc.gov/heartdisease/facts.htm
- Centers for Disease Control and Prevention. (2020b, September 8). *Heart failure.* https://www.cdc.gov/heartdisease/heart_failure.htm
- Chicco, D., & Jurman, G. (2020). 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). https://doi.org/10.1186/s12911-020-1023-5
- Heart and Stroke Foundation Canada. (2020a). *Conditions.* https://www.heartandstroke.ca/heart-disease/conditions?gclid=Cj0KCQjwna2FBhDPARIsACAEc_V-S913xvyPrMMrQyz_8_3HBuWayLmUXD8Oln7LwPrE4lvc3d3Xm3IaAvaLEALw_wcB&gclsrc=aw.ds
- Heart and Stroke Foundation Canada. (2020b). *Heart failure.* https://www.heartandstroke.ca/heart-disease/conditions/heart-failure
- Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Wadsworth Cengage Learning.
- Irizarry, R. A. (2019, October 24). *Introduction to data science.* https://leanpub.com/datasciencebook
- Privitera, G. J. (2018). *Statistics for the behavioral sciences* (3rd ed.). Sage.
- Public Health Agency of Canada. (2017, February 10). *Heart disease in Canada.* Government of Canada. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html
- World Health Organization. (2017, May 17). *Cardiovascular diseases (CVDs).* https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
- Zahid, F. M., Ramzan, S., Faisal, S., & Hussain, I. (2019). Gender based survival prediction models for heart failure patients: A case study in Pakistan. *PLoS ONE*, *14*(2), e0210602. https://doi.org/10.1371/journal.pone.0210602

9 Appendix A - Supplemental Figures

Figure A1: Continuous Variables by Anaemia and Patient Death

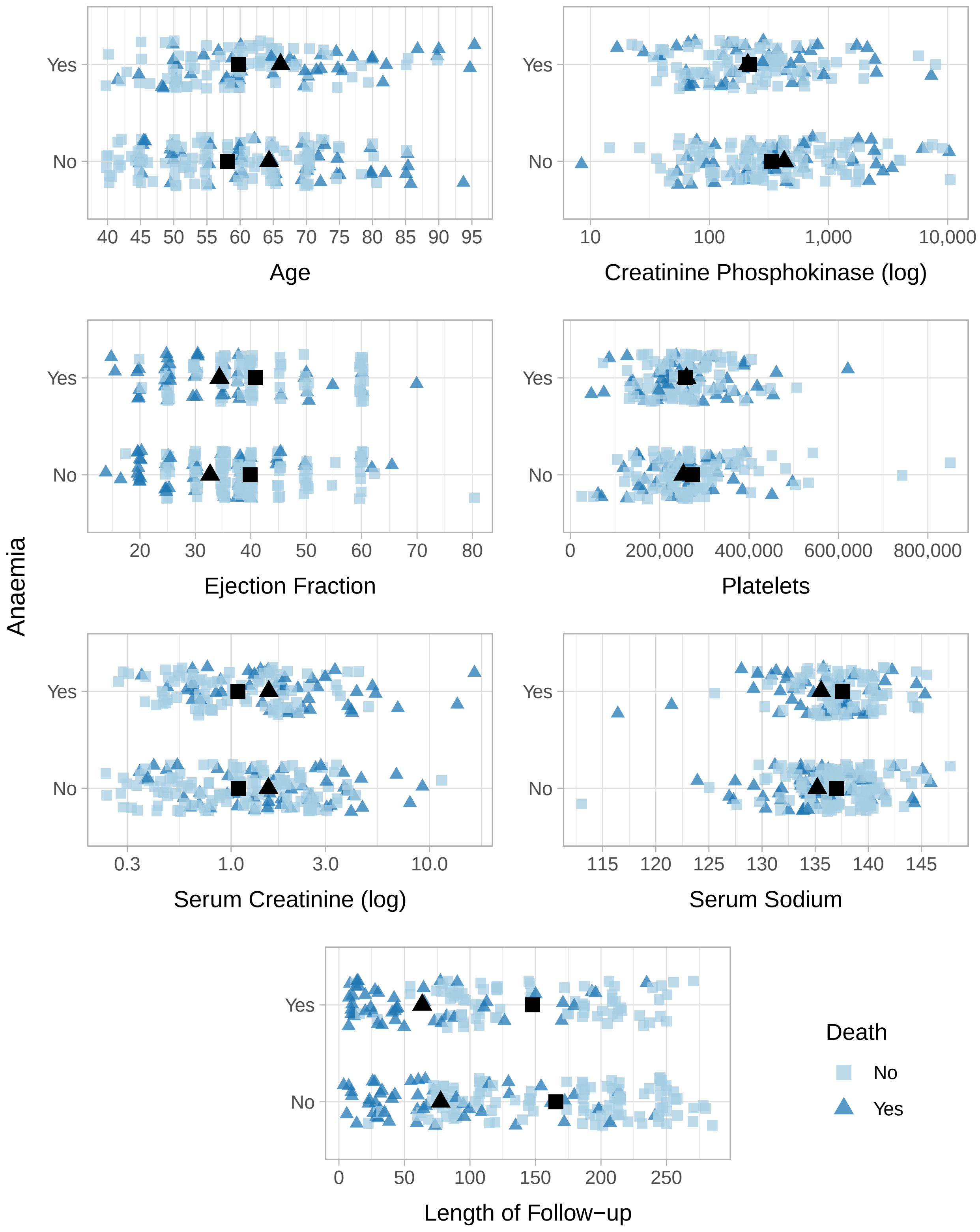


Figure A2: Continuous Variables by Diabetes and Patient Death

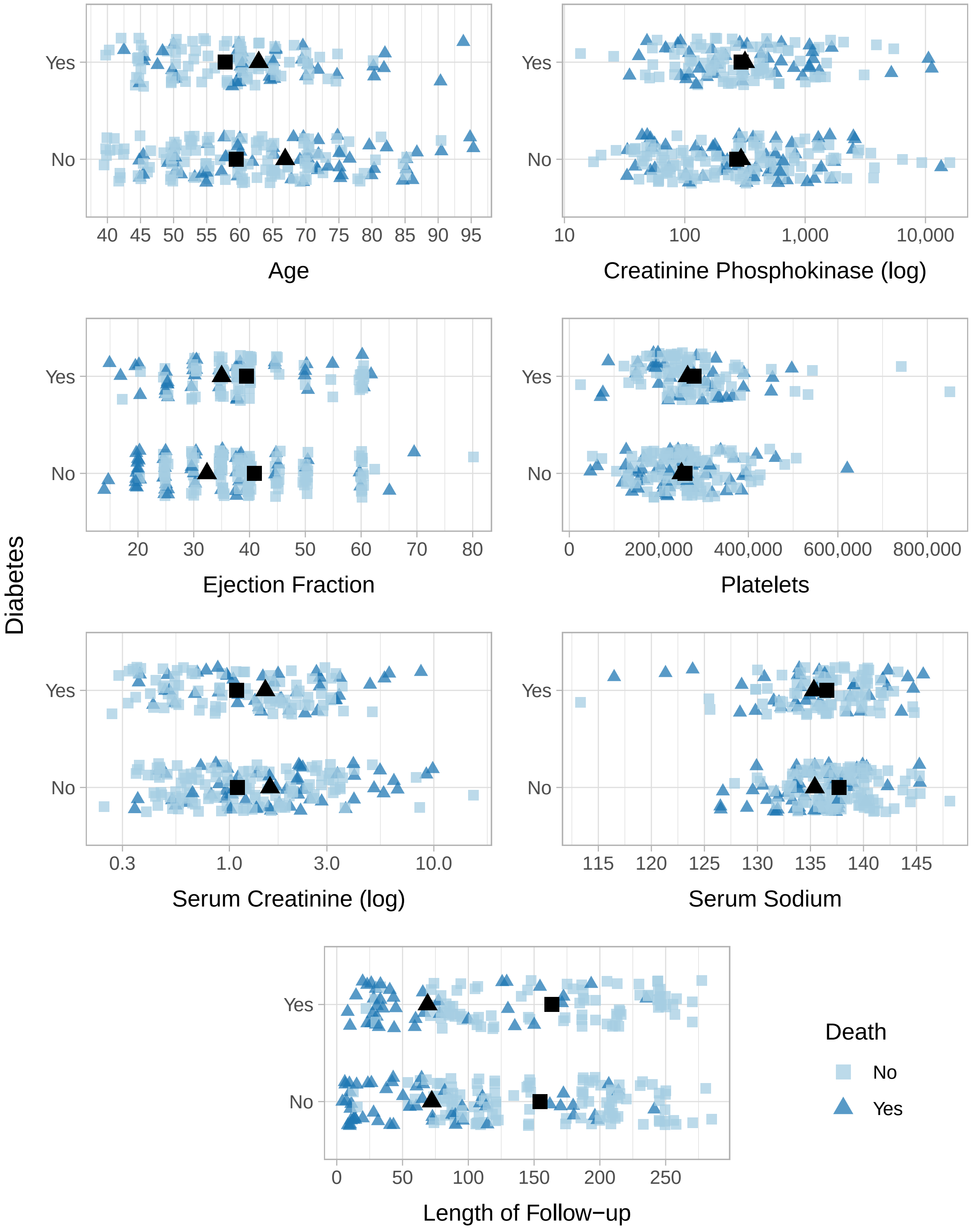


Figure A3: Continuous Variables by High Blood Pressure and Patient Death

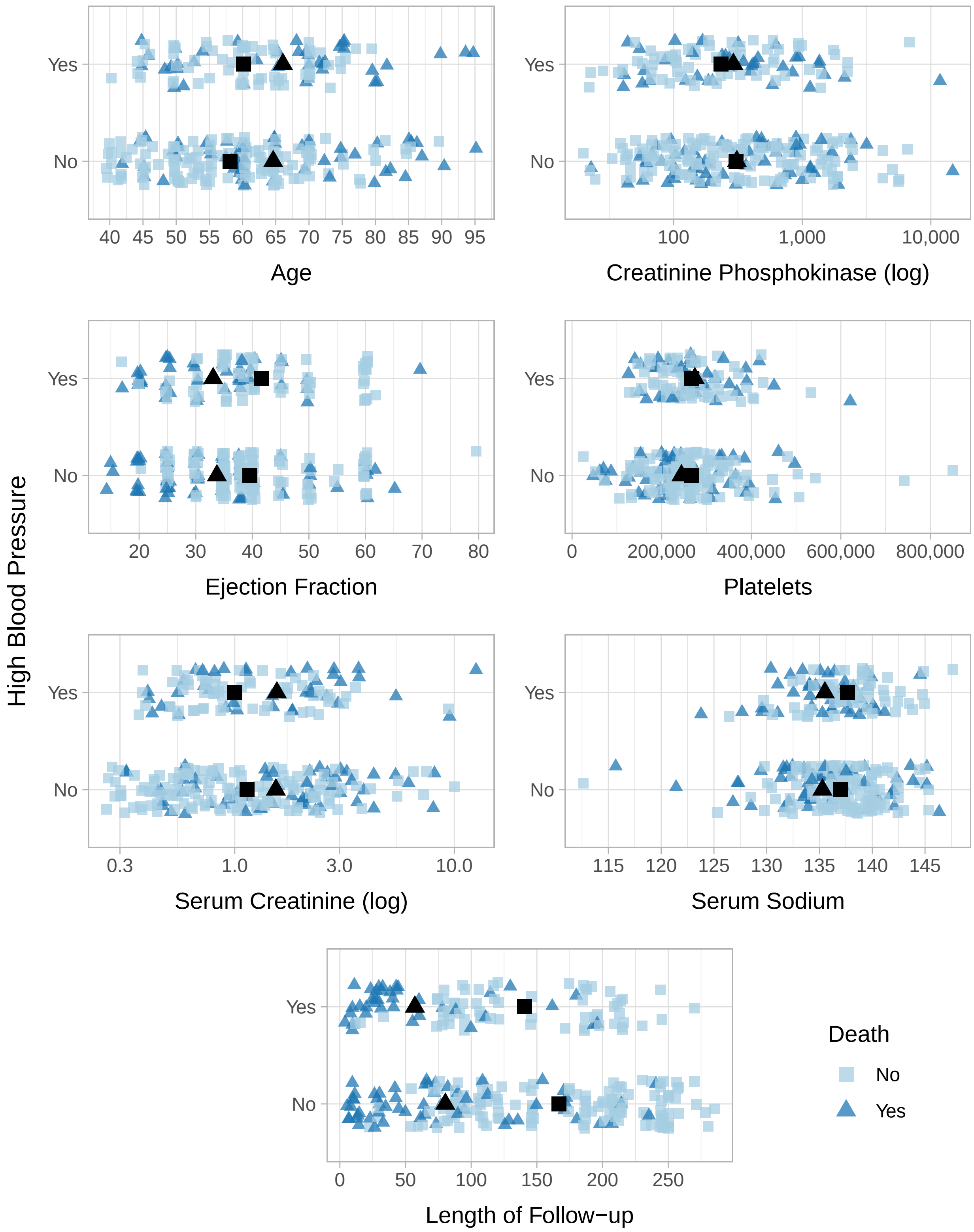


Figure A4: Continuous Variables by Sex and Patient Death

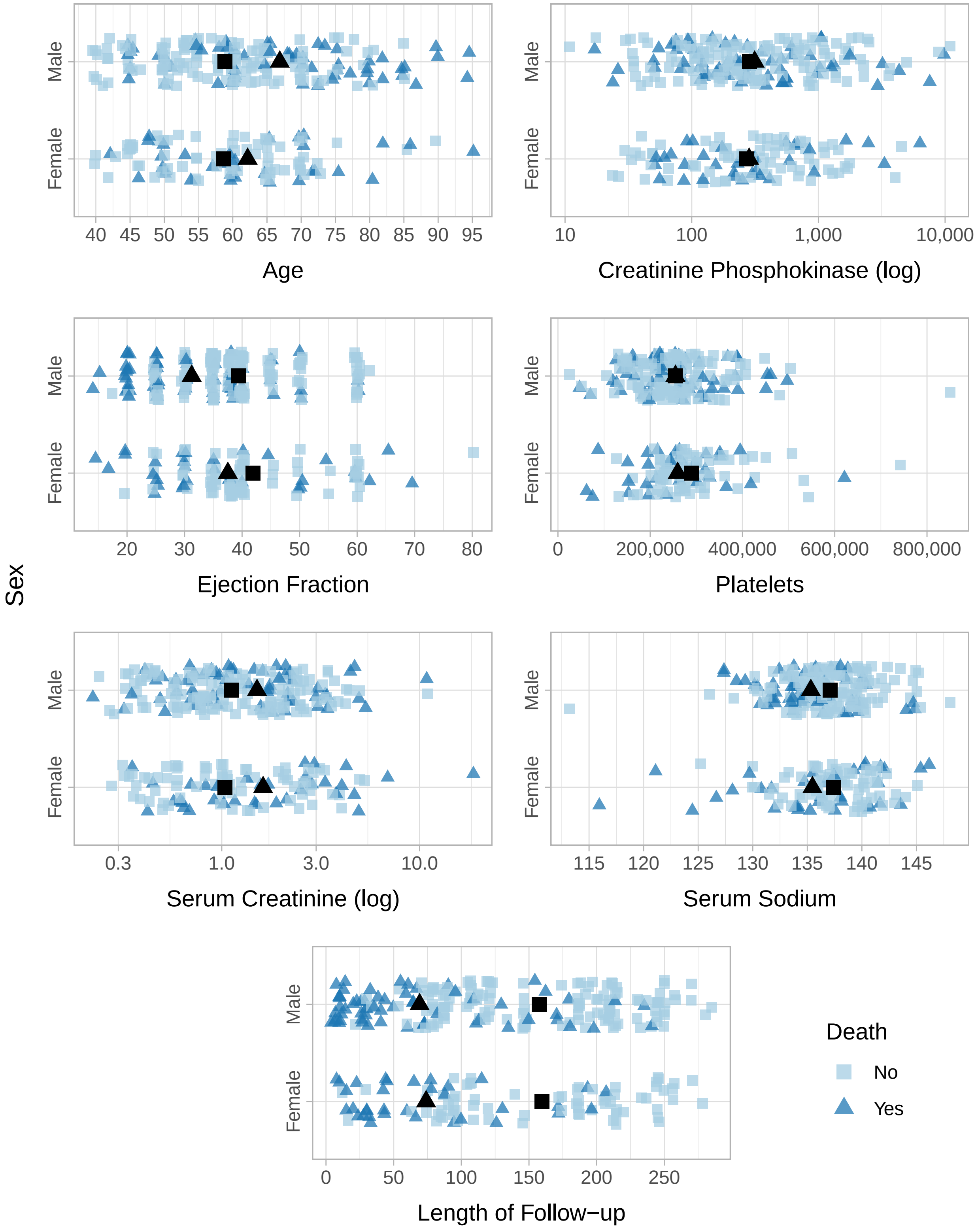


Figure A5: Continuous Variables by Smoking and Patient Death

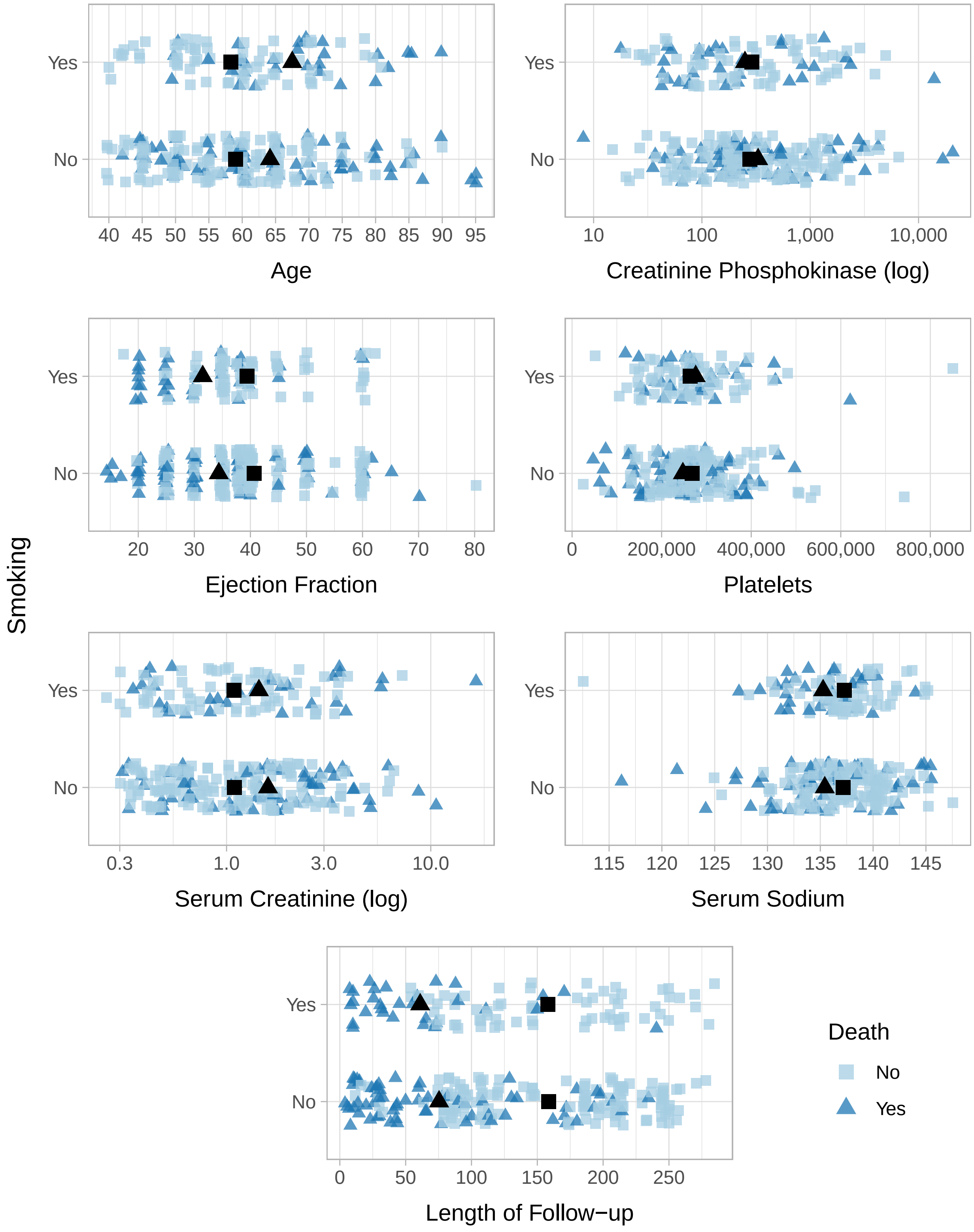


Figure A6: Continuous Variables by Length of Follow−up and Patient Death

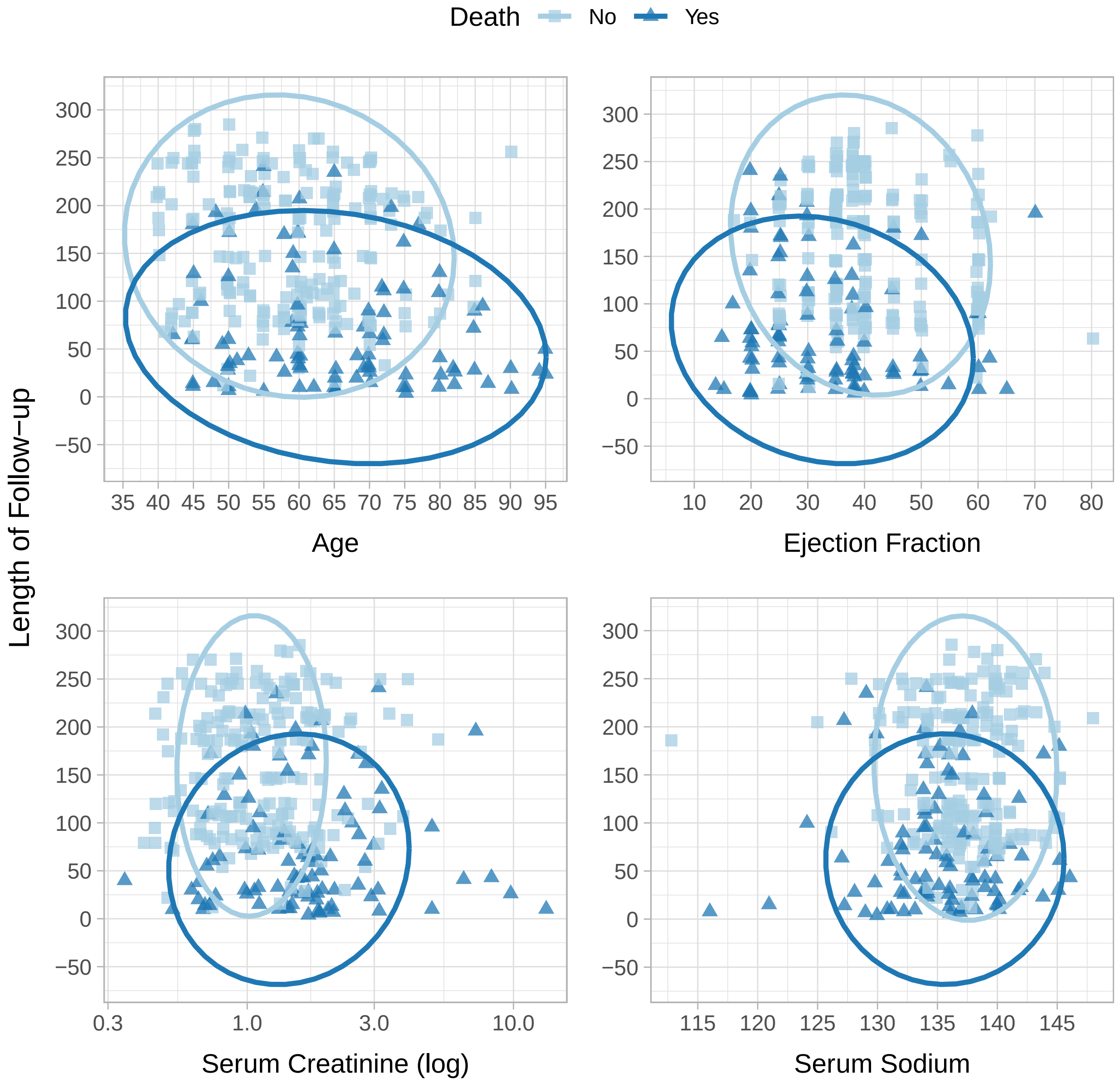


Figure A7: Bivariate Continuous Variables by Patient Death

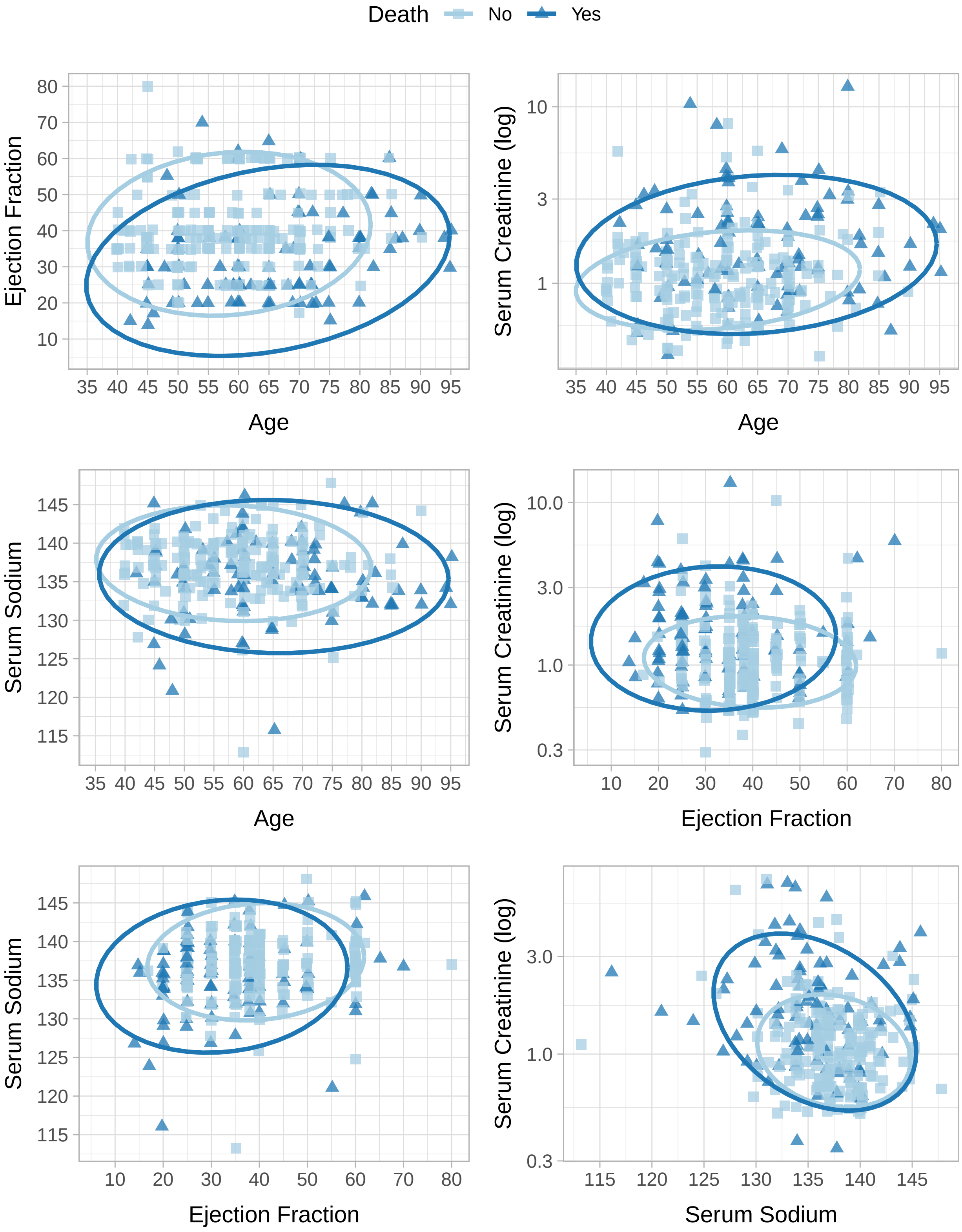
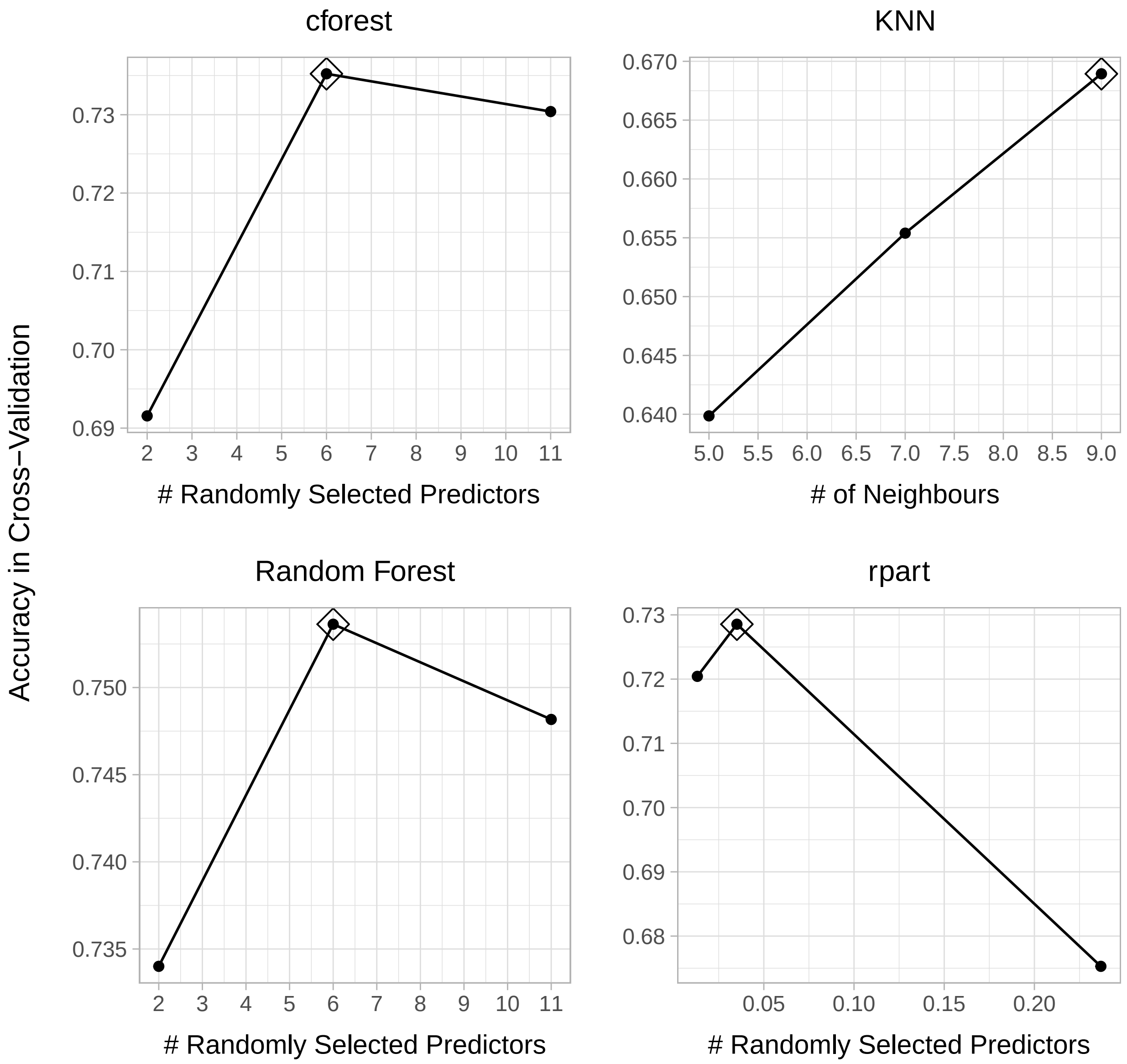


Figure A8: Tuning Parameters across Models − All Features



10 Appendix B - Environment

This project and code were completed using the following specifications:

```
##                _
## platform       x86_64-apple-darwin17.0
## arch           x86_64
## os             darwin17.0
## system         x86_64, darwin17.0
## status
## major          4
## minor          0.3
## year           2020
## month          10
## day            10
## svn rev        79318
## language       R
## version.string R version 4.0.3 (2020-10-10)
## nickname       Bunny-Wunnies Freak Out
```