Predicting Mortality by Heart Failure

Willliam Ofosu Agyapong

5/6/2021

1 Introduction

According to the World Health Organization, cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Of these deaths, 85% are due to heart attack and stroke. Heart failure is a common event caused by cardiovascular diseases.

The early detection of people with cardiovascular diseases or who are at high cardiovascular risk due to the presence of one or more risk factors is paramount to reducing deaths arising from heart failures. As a result, predictive models become indispensable.

1.1 Objective

The dataset explored in this project contains 12 features that can be used to predict mortality by heart failure. The goal of this project, therefore, is to **identify an appropriate classification model that can accurately predict the mortality of patients with heart failure**.

Various classification models, as can be found in the Predictive Model section of this report, will be explored with the hope of coming up with a model that has high prediction accuracy.

1.2 Data Description

The data used in this report were obtained from Kaggle. As reported by the authors of the data, Chicco and Jurman (2020), the data come from medical records of patients having heart failure that were collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad, from April to December, 2015. The dataset consists of 299 observations for 13 different variables.There are no missing values. This dataset is for the prediction of heart failure based on multiple attributes such as diabetes, sex, smoking, high blood pressure, among others. The **Death event** variable is the response or target variable. A list of all variables and relevant information for each variable is included in Table 1. Though the time feature seems to be an important predictor, I feel there is not much information about this variable for me to include it in the modeling process so it was not used in the search for the best model.

Table 1: Variables in the heart failure data set

Variable Name	Description	Measurement Unit	Range/Level
Age	Age of the patient	Years	[40 ,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Binary	0: No, 1: Yes
High blood pressure (HBP)	If a patient has HBP	Binary	0: No, 1: Yes
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L	[23 ,..., 7861]
Diabetes	If the patient has diabetes	Binary	0: No, 1: Yes
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14 ,..., 80]
Sex	Woman or man	Binary	0: Woman, 1: Man
Platelets	Platelets in the blood	kiloplatelets/mL	[25100 ,..., 850000]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.5 ,..., 9.4]
Serum sodium	Level of sodium in the blood	mEq/L	[113 ,..., 148]
Smoking	If the patient smokes	Binary	0: No, 1: Yes
Time	Follow-up period	Days	[4 ,..., 285]
Death event	If the patient died during the follow-up period	Binary	0: No, 1: Yes

Note: *mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre*

Table 2: First and last four observations in the data

	age	anaemia	CPK	diabetes	ejection_frac	HBP	platelets	serum_creatinine	serum_sodium	sex	smoking	time	death_event
1	75	0	582	0	20	1	265000	1.9	130	1	0	4	1
2	55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
3	65	0	146	0	20	0	162000	1.3	129	1	1	7	1
4	50	1	111	0	20	0	210000	1.9	137	1	0	7	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
296	55	0	1820	0	38	0	270000	1.2	139	0	0	271	0
297	45	0	2060	1	60	0	742000	0.8	138	0	0	278	0
298	45	0	2413	0	38	0	140000	1.4	140	1	1	280	0
299	50	0	196	0	45	0	395000	1.6	136	1	1	285	0

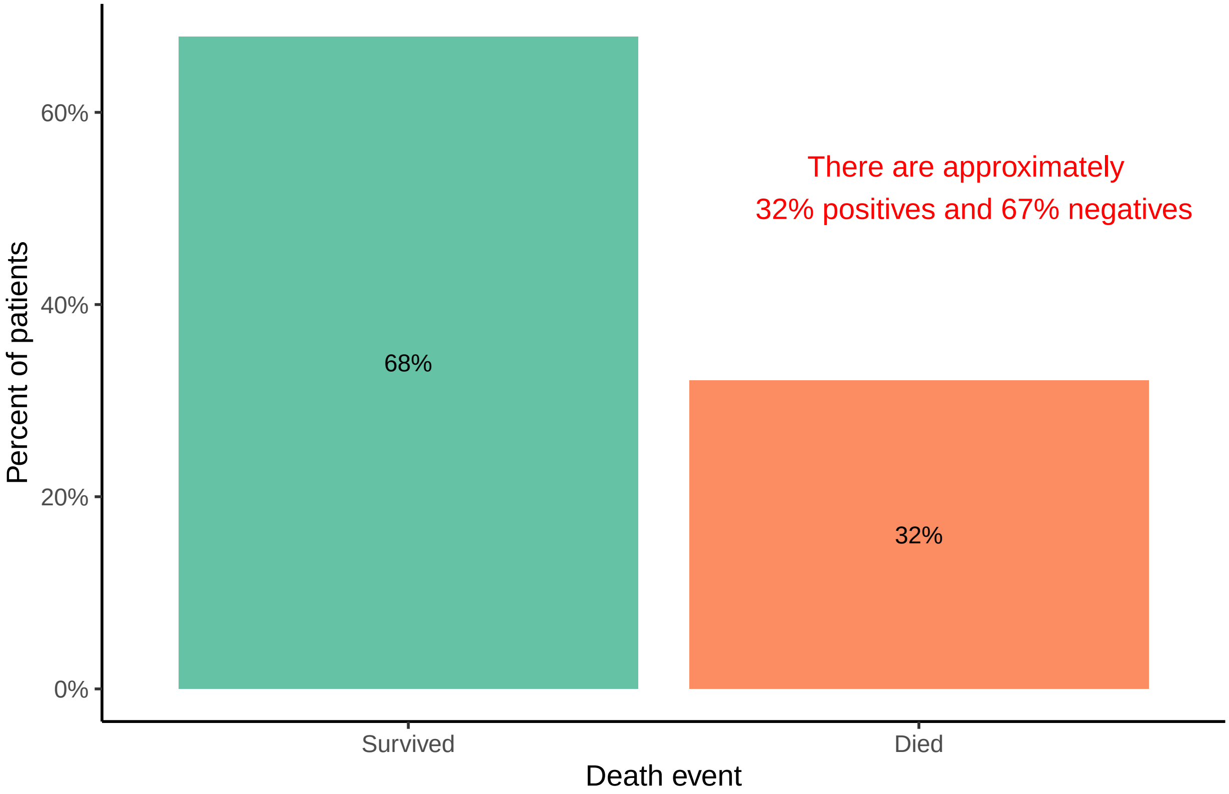
2 Eploratory Data Analysis

Before proceeding to fit the models, it is important to gain some initial insight about the data. To this end, we look at the distributions of our target variable (death event) and all the predictor variables.

2.1 Distribution of Target Variable

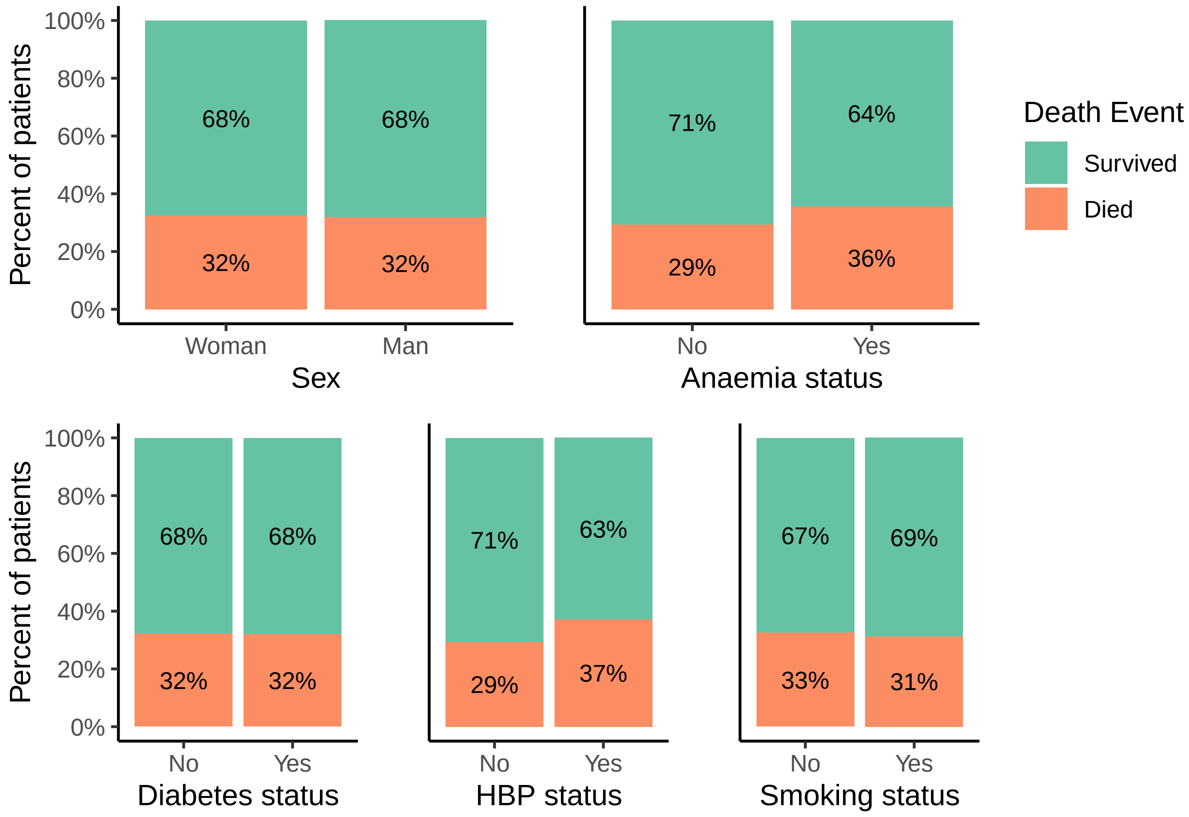
Figure 1 shows how the death event (survival) is distributed among the patients involved in the study. As can be seen the data used is not evenly distributed between patients who died and patients who survived heart failures.

Figure 1: Distribution of Death Event



2.2 Distribution of Categorical Predictors by Target Variable

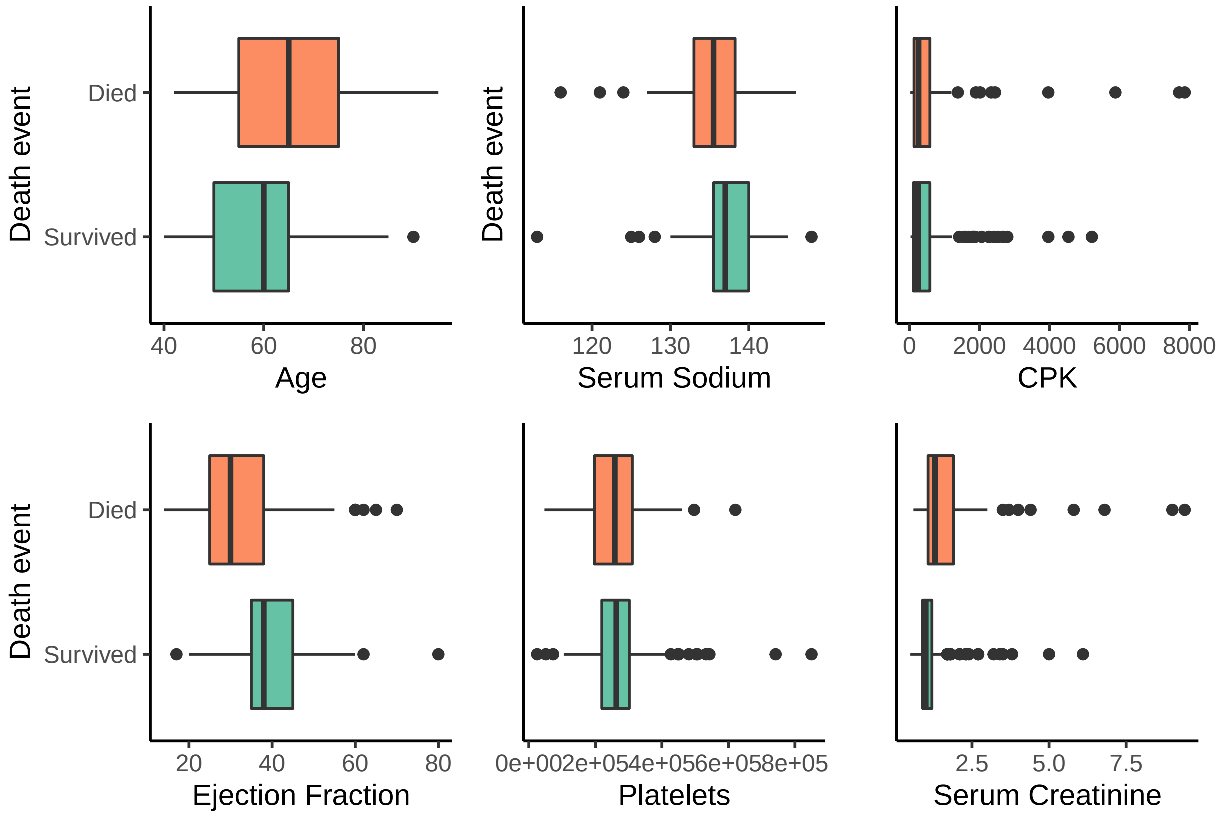
Figure 2: The effect of *Gender*, *Anaemia*, *Diabetes*, *High Blood Pressure*, and *Smoking* on survival.



From **Figure 2**, we can see that, whether a patient died or survived does not appear to depend on their sex orientation, diabetes and smoking status since the death event is distributed equally across the two levels of these variables. However, the anaemia and high blood pressure status of the patients seem to play a role in their survival, though not very significant. Hence these two predictor variables are likely to have some impact on our predictive model. Overall, there appears to be little effect of all these categorical predictor variables on the target variable, `death event`.

2.3 Distribution of Continuous Predictors by Target Variable

Figure 3: Distribution of continuous predictors grouped by the Death Event

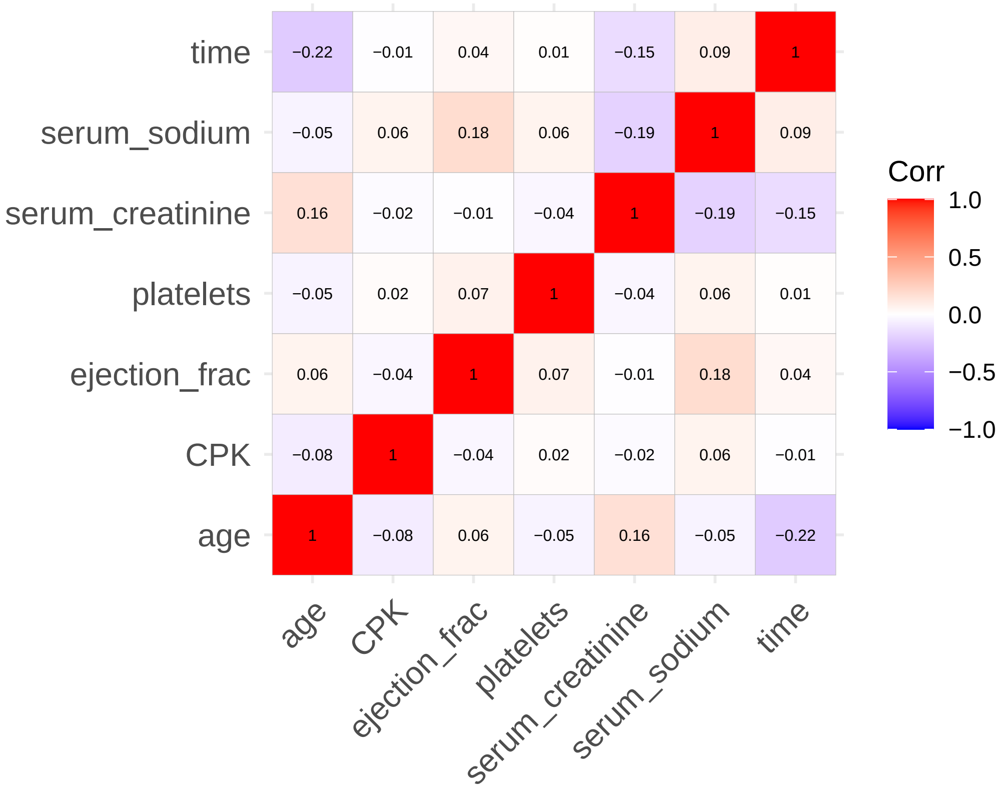


While there is only one high outlier on th Age predictor, the rest of the continuous predictors have many outliers. Additionally, apart from the Age variable, all the other predictors are highly skewed with CPK, Ejection Fraction, Platelets, and Serum Creatinine skewed to the right, and Serum Creatinine skewed to the left.

The median Age, Serum Sodium, Ejection Fraction, and Serum Creatinine differ between the patients who died and those that survived, suggesting that these variables might play an important role in predicting the `death event`. There is high variability in Age, followed by Ejection Fraction, Serum Sodium, and the Platelets, with CPK and Serum Creatinine having low variability. As should be expected, the risk of dying is high among older patients. Also, low levels of Ejection Fraction and Serum Sodium appear to be associated with high death risk, whereas high levels of Serum Creatinine seem to be associated with high death risk.

From these observations, Serum Creatinine is likely to be the most important predictor, followed by Ejection Fraction. So, in all, the EDA has revealed four variables, including Serum Creatinine, Ejection Fraction, Anaemia status, and High Blood Pressure status, that appear to have much contribution on the survival of patients with hearth failure. These variables will be explored further in the Predictive Modeling section.

Figure 4: Correlation between the continuous predictors



According to Figure 4, there is a weak correlation among the continuous predictors. This suggests no issue of multicolinearity.

3 Predictive Modeling

In this part of the analysis, we fit several classification models to the death event as a function of all the other 12 variables in the dataset. For model validation purposes, the heart failure data were split into 80% training set and 20% testing set. This meant that the models were trained on 240 observations and the remaining 59 observations were used for predictions from which the trained models were evaluated for predictive performance. In training the models, a 5-fold cross validation (CV) was adopted, partly due to the relatively small number of observations in the data and computation time. However, a leave-one-out cross-validation (LOOCV) would have been preferred to eliminate the randomness in results but it turned out to be too computationally expensive, especially, for the Tree-based models and the Support Vector Classifiers.

Due to the sensitivity of the KNN model to different scales, the variables were normalized and kept across the other models for consistency.

3.1 Classification Models

Six unique models, including Logistic regression, Linear Discriminant Analysis (LDA), K-nearest neighbors (KNN), 2 Regularized models, 2 Tree-based models and 3 Support Vector Classifiers, were considered as candidate models for predicting whether a patient will die from a heart failure. A full model including all the 12 predictor variables and a reduced were fitted. The reduced models were constructed based on `Ejection Fraction`, `Serum Creatinine`, `Anaemia`, and `High Blood Pressure`, which appeared to be the four most important predictors from the EDA section. **Table 3** provides information about these models regarding the values of parameters used in the fitting process. 1000 trees were used in training the two Tree-based models.

Table 3: Table 3: Candidate models and their parameter values

Model	Parameters used
KNN	K: 1 - 10 by 1
Logistic Regression	
LDA	
LASSO Regression	Lambda: 0 - 0.1 by 0.1
Ridge Regression	Lambda: 0 - 0.1 by 0.1
Bagging	mtry: 11; ntree: 1000
Random Forest	mtry: 1-10 by 1; ntree: 1000
SVC	Cost: 0.1, 0.2, 1, 1.4, 1.6, 10
SVM (Polynomial Kernel)	Cost: 0.1, 0.2, 1, 1.4, 1.6, 10; Degree: 2 - 5 by 1; Scale: 1
SVM (Radial Kernel)	Cost: 0.1, 0.2, 1, 1.4, 1.6, 10; Sigma: 0.01, 0.05, 0.1, 0.5, 1

3.2 Results and Model Comparison

The results obtained from training the various classification models with all 11 predictors and a subset of 4 predictors are presented in Table 4 and Table 5, respectively. Specifically, we report the testing misclassification rates of all the models as well as the best values chosen for some of the model parameters that were tuned. The misclassification rates were computed from predictions made on the testing set. Misclassification rate is the proportion of wrongly predicted classes (survived or died), and hence, lower values are indicative of better predictive performance.

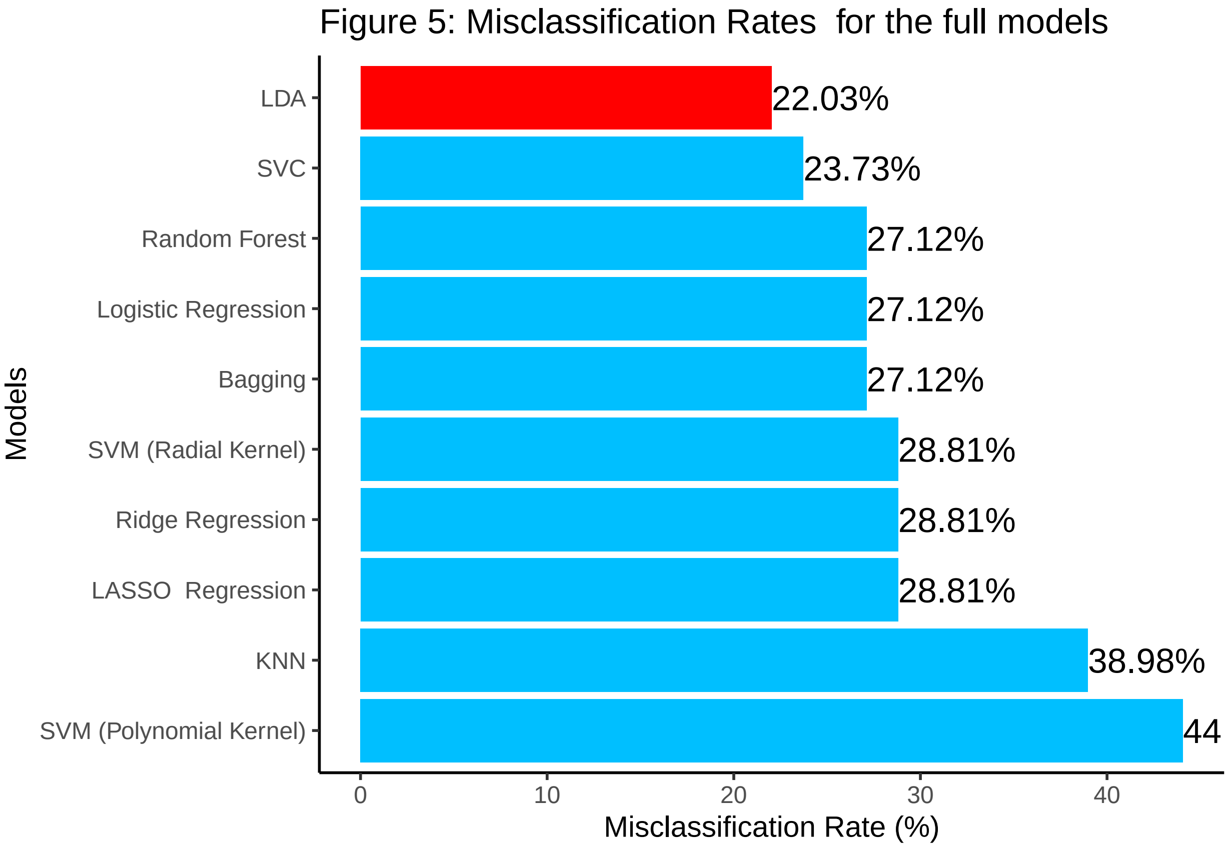
Table 4: Results from the full models

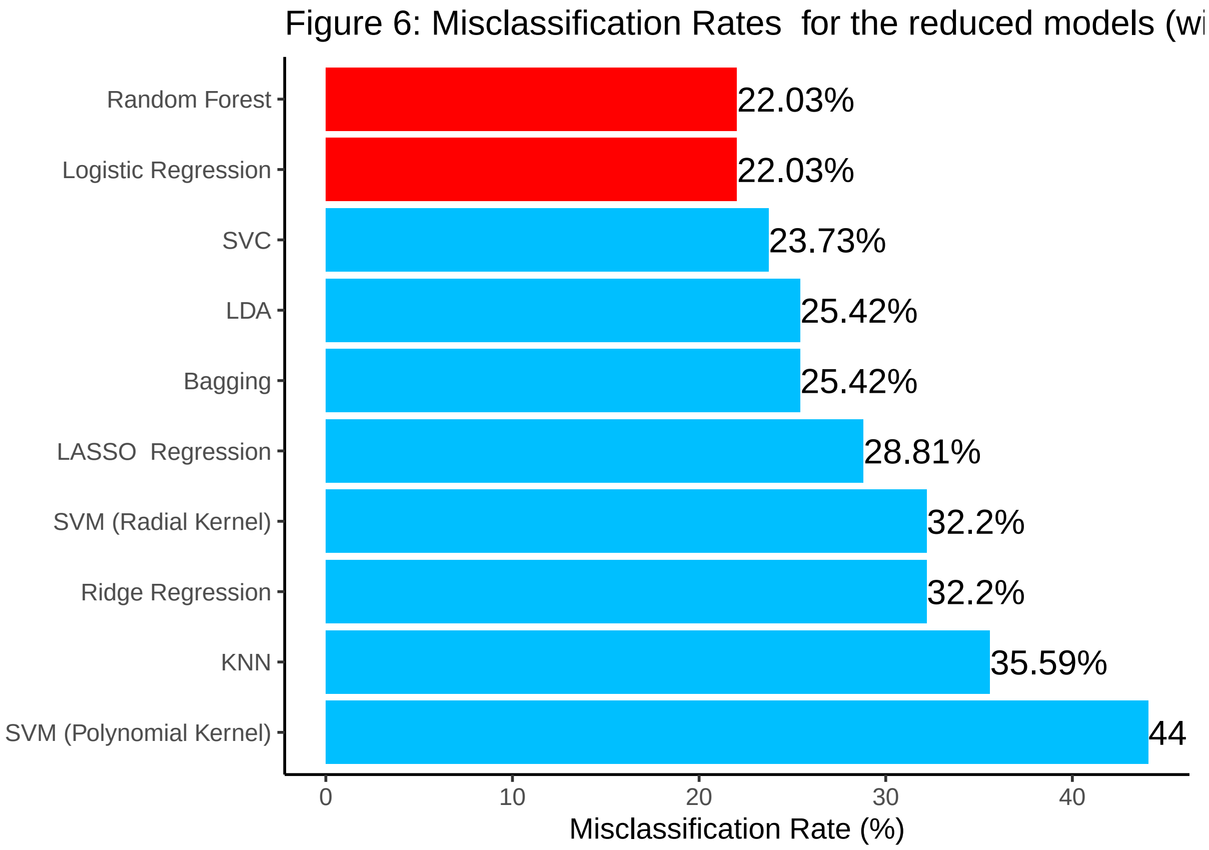
Model	K	Lambda	Cost	Degree	Sigma	Mtry	Misclassification Rate (%)
KNN	9	-	-	-	-	-	38.98
Logistic Regression	-	-	-	-	-	-	27.12
LDA	-	-	-	-	-	-	22.03
LASSO Regression	-	0.02	-	-	-	-	28.81
Ridge Regression	-	0.04	-	-	-	-	28.81
Bagging	-	-	-	-	-	-	27.12
Random Forest	-	-	-	-	-	8	27.12
SVC	-	-	1.4	-	-	-	23.73
SVM (Polynomial Kernel)	-	-	0.1	5	-	-	44.07
SVM (Radial Kernel)	-	-	0.1	-	0.1	-	28.81

Table 5: Results from the reduced models

Model	K	Lambda	Cost	Degree	Sigma	Mtry	Misclassification Rate (%)
KNN	10	-	-	-	-	-	35.59
Logistic Regression	-	-	-	-	-	-	22.03
LDA	-	-	-	-	-	-	25.42
LASSO Regression	-	0.02	-	-	-	-	28.81
Ridge Regression	-	0.09	-	-	-	-	32.2
Bagging	-	-	-	-	-	-	25.42
Random Forest	-	-	-	-	-	2	22.03
SVC	-	-	1.4	-	-	-	23.73
SVM (Polynomial Kernel)	-	-	0.1	5	-	-	44.07
SVM (Radial Kernel)	-	-	0.1	-	0.01	-	32.2

Figure 5: Misclassification Rates for the full models (Left) and the reduced models (Right)





From Figure 5, the model with the least misclassification rate is the Linear Discriminant (LDA) model. Therefore, LDA turns out to be the best model for predicting death event due to heart failure when all 11 predictors are taken into account. As can be seen, the Support Vector Classifier (SVC) came very close to our proposed model, with KNN and SVM with a polynomial kernel performing very poorly in this situation. Interestingly, the two regularization models (LASSO and Ridge) as well as the two Tree-based models performed equally. On the other hand, according to Figure 6, the Logistic regression and Random Forest happen to be the best models among the with only 4 predictors. SVM, KNN, LASSO, and Ridge continued to perform poorly.

Clearly, there are three competing models all with the lowest misclassification rate of 22.03% - the LDA model (based on 11 predictors), and the Logistic Regression model and Random Forest model both based on 4 predictors. Since simple models are preferable, the Random Forest and Logistic Regression models will be selected at this stage and studied furthered for the selection of an overall best model in the next section.

3.3 Best Models: Logistic Regression and Random Forest

As noted from the previous section, the best models were selected based on the misclassification rates. However, it is not clear what specific types of misclassifications were made, so we present confusion matrices. with the "Died" class as the positive response, using information from the confusion matrices, model performance metrics as included in Table 6 can be computed. In Figure 6, the confusion matrix on the left relates to the Logistic Regression model while the one on the right is for the Random Forest model.

Figure 6: Confusion Matrices

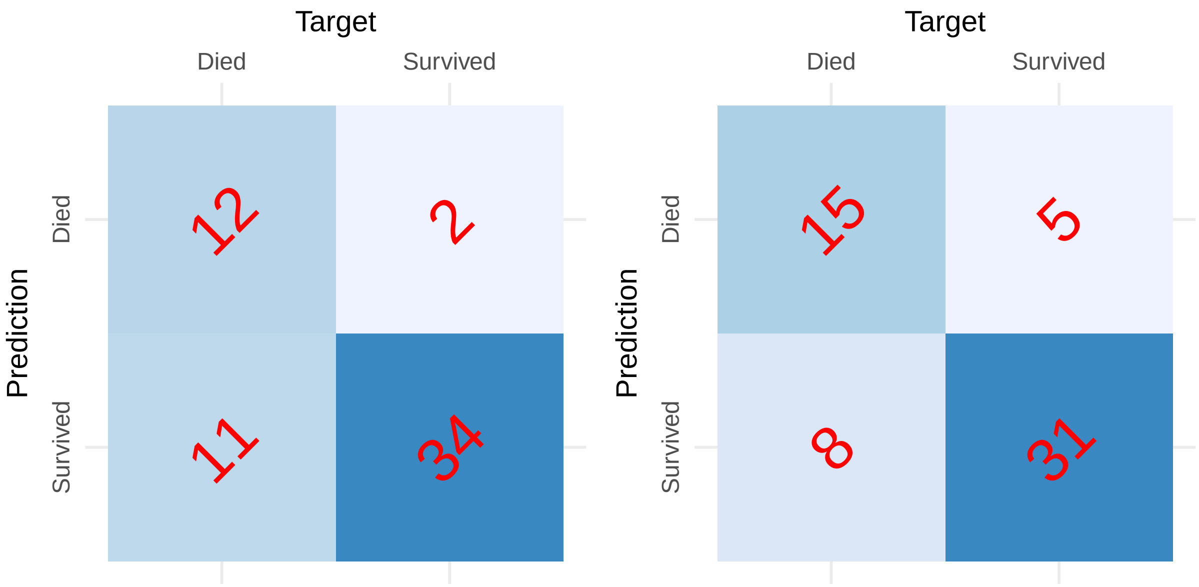


Table 6: Performance Metrics

Metrics	Logistic Regression	Random Forest
Accuracy	77.9660	77.9660
Sensitivity	0.5217	0.6522
Specificity	0.9444	0.8611
Positive Predictive Value	0.8571	0.7500
Negative Predictive Value	0.7556	0.7949

All the measures are relatively high, which suggests the Logistic Regression and Random Forest models did a good job at predicting the survival/death of patients with heart failures given the various predictors considered. However, the Logistic Regression model beats the Random Forest model in terms of specificity and positive predictive value, while the random forest model performed better in terms of sensitivity and negative predictive value. Since our goal is to obtain a model with the highest accuracy, both of these models can be considered, I will choose the Random Forest model over the Logistic Regression model for the reason that Decision Trees are non-linear classifiers which do not require the data to be linearly separable and can handle both cases.

4 Conclusion

In this analysis ten different models were fitted to the heart failure data set. From these models, three classifiers - Linear Discriminant Analysis (LDA) model, Logistic Regression model, and Random Forest model - emerged as best models with the same predictive accuracy. The LDA model was dropped because it was the most complex model among the three since it was based on 11 predictors. The other two models were each based on only 4 predictors but the Random Forest model was chosen as the overall best model because of some advantages it has over the competing Logistic Regression model. That is, unlike the Logistic Regression model, the Random Forest model is non-linear classifier which does not require the data to be linearly separable. Therefore, I will conclude that a Random Forest model with `Ejection Fraction`, `Serum Creatinine`, `Anaemia`, and `High Blood Pressure` is the best model for predicting the death event of a patient with a heart failure.

However, I believe more can be done to improve upon the predictive power. For instance, treating the issue of class imbalance and handling outliers could enhance model performance. Other models such as ones that consider interaction and dimension reduction methods can also be pursued in future analysis for possible improvement in performance.

References

- Source of Data (Kaggle): https://www.kaggle.com/andrewmvd/heart-failure-clinical-data
- 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(1), 16.
- Cardiovascular diseases (World Health Organization, 2017): https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
- https://cran.r-project.org/web/packages/cvms/vignettes/Creating_a_confusion_matrix.html