Heart Failure Prediction

HarvardX Data Science Professional Certificate

Narendra Kumar Jangid

2022-09-26

Contents

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. As per World Health Organisation (WHO), 17.9 Million people died in the year 2019 from CVDs, an estimated 32% of all deaths worldwide[1].

Heart Failure is an event caused by CVDs where the heart is not able to pump enough blood to meet the body's needs for blood and oxygen. The most common risk factors include High Blood Pressure, Obesity, Age and Genetics.

This project is a requirement for the HarvardX Data Science Professional Certificate Program which aims to develop the Heart Failure Prediction Model using the Heart Failure Clinical Records Data Set.

The report is created using R Markdown in RStudio that covers Data Preparation for Model Building, Data Exploration with common visualization techniques, Model Development using train and test sets, Model Evaluation and Results using validation set and Concluding Remarks.

1.1 Objective

The objective of this project is to develop the Heart Failure Prediction Model using machine learning techniques to assess the likelihood of death by a heart failure event. The model can help identify the at-risk individuals and prevent fatal outcomes with appropriate treatment at an early stage.

The data is split into training, testing and validation sets to develop and validate the model. Models are evaluated using Model Performance metrics.

1.2 Heart Failure Clinical Records Data Set

Heart Failure Clinical Records Data Set is taken from UCI Machine Learning Repository. The data set contains the medical records of 299 patients who have had heart failure, collected during their follow-up period where each patient profile has 13 clinical features[2].

2 Data Preparation

In the data preparation step, Heart Failure Clinical Records Data Set is prepared for exploration, modeling and model evaluation using required packages and libraries.

2.1 Install Required Packages

To prepare and transform the data, required packages are installed and necessary libraries are loaded.

2.2 Create Heart Failure Data Set

The Heart Failure Clinical Records Data Set is read from the source and processed further to create Heart Failure data in the required format. Further processing of data includes having more descriptive data set values and converting required data set fields to factors and integers.

2.3 Create Train and Validation Sets

The Heart Failure data is split into two parts, training set and testing set with 90% and 10% of the original data set respectively. The training set is called edx and the testing set is called the validation set.

The model development is done using the edx set with further split into train and test data sets whereas the validation set is the final test set that is used at the end to check the model performance.

The 90:10 split is taken to have a reasonable number of instances in train, test and validation sets along with the best possible model performance for both test and validation sets.

3 Data Exploration and Analysis

Data exploration and analysis with different visualization techniques help us understand the data and its distribution resulting in a better model building. Density plots and Bar plots are majorly used visuals for Continuous and Categorical variables respectively.

Also, the analysis helps identify the Variable Importance and Correlation that can be taken into account while model building.

3.1 Data Overview

The edx data set is a data.frame comprised of 268 rows and 13 columns. A record represents whether a heart failure patient having certain Gender, Sex and other Clinical Features has survived or not.

An overview of the edx data is tabulated in Table 1.

Table 1: edx Data Overview

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
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

Below is the detailed definition of each of the 13 clinical features[2]:

1. Age: Age of the patient (Years)
2. Anaemia: Decrease of red blood cells or hemoglobin (Yes/No)
3. Creatinine Phosphokinase (CPK): Level of the CPK enzyme in the blood (mcg/L)
4. Diabetes: If the patient has diabetes (Yes/No)
5. Ejection Fraction: Percentage of blood leaving the heart at each contraction (Percentage)
6. High Blood Pressure: If the patient has hypertension (Yes/No)
7. Platelets: Platelets in the blood (kiloplatelets/mL)
8. Serum Creatinine: Level of serum creatinine in the blood (mg/dL)
9. Serum Sodium: Level of serum sodium in the blood (mEq/L)
10. Sex: Male or Female
11. Smoking: If the patient smokes or not (Yes/No)
12. Time: Follow-up period (Days)
13. Death Event: If the patient deceased during the follow-up period (Yes/No)

3.2 Death Event

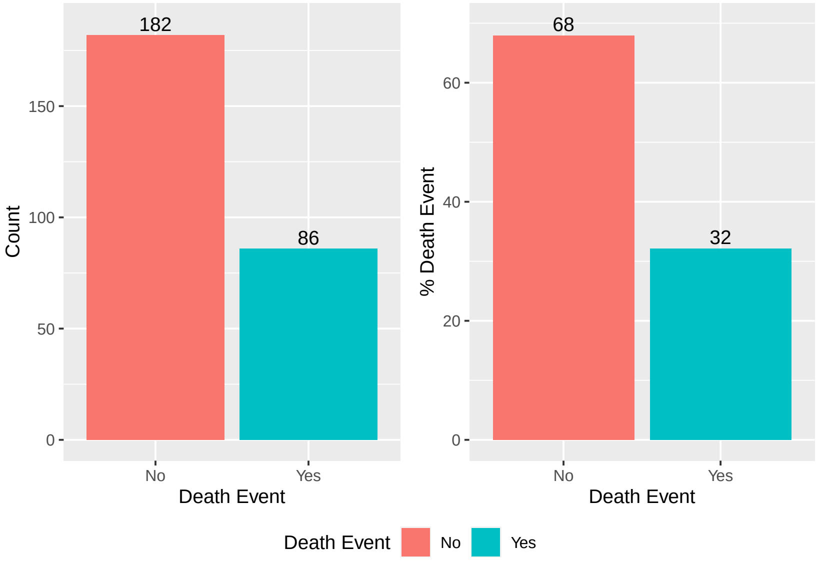

Figure 1: Death Event Distribution

Figure 1 shows that 68% of patients have survived and 32% have died from a clinical record of 268 patients having Heart Failure.

3.3 Age

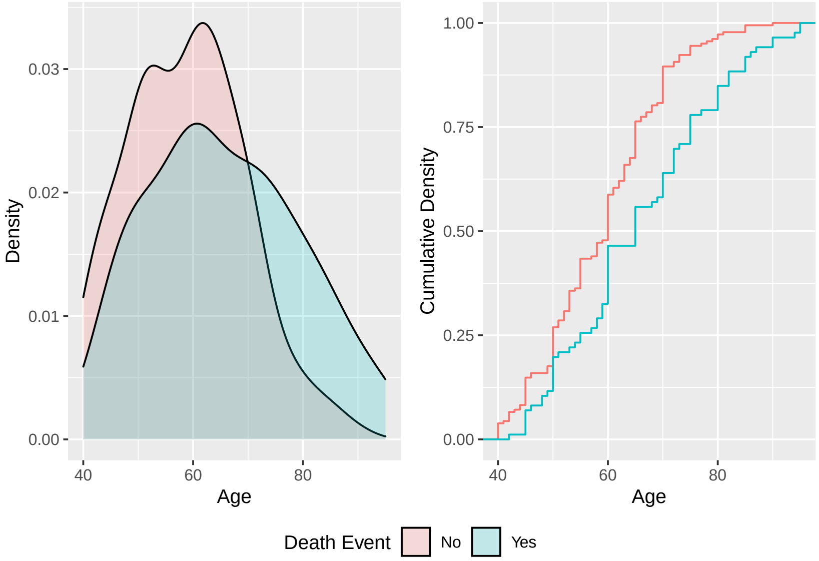

Figure 2: Age Distribution

There is a higher likelihood of Death for patients having an age of 70 Years and above. Cumulative Death likelihood is lower than Survival with bigger differences at the age of 55, 65 and 70. Therefore, the death risk for a patient increases more rapidly as the patient reaches the age range of 50-80 Years.

3.4 Anaemia

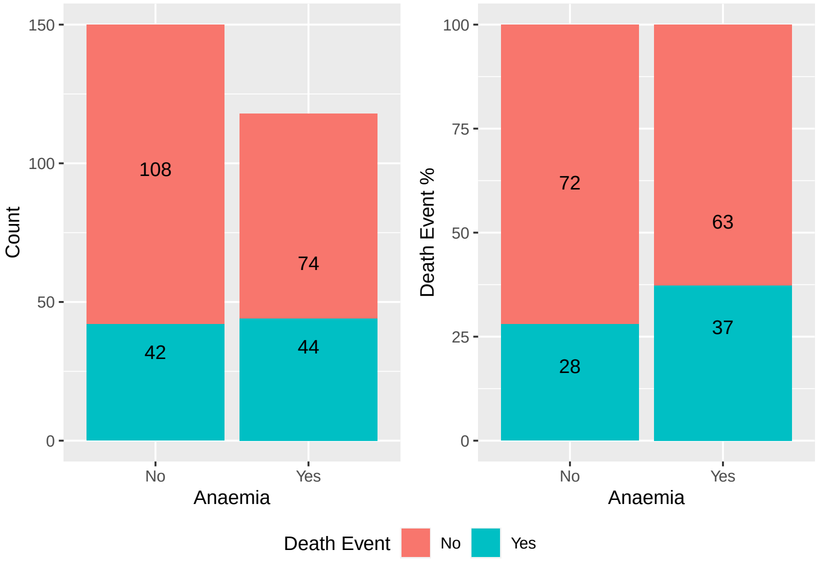

Figure 3: Anaemia Distribution

A Patient having Anaemia has a 37% likelihood of Death, 9% higher than a patient having no Anaemia.

3.5 Creatinine Phosphokinase

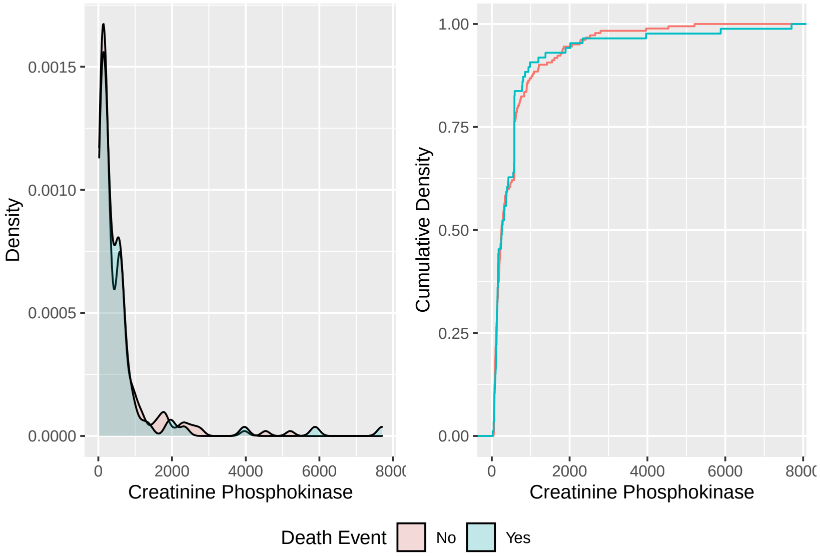

Figure 4: Creatinine Phosphokinase Distribution

No significant difference is observed between Death and Survival likelihood for different Creatinine Phosphokinase levels.

3.6 Diabetes

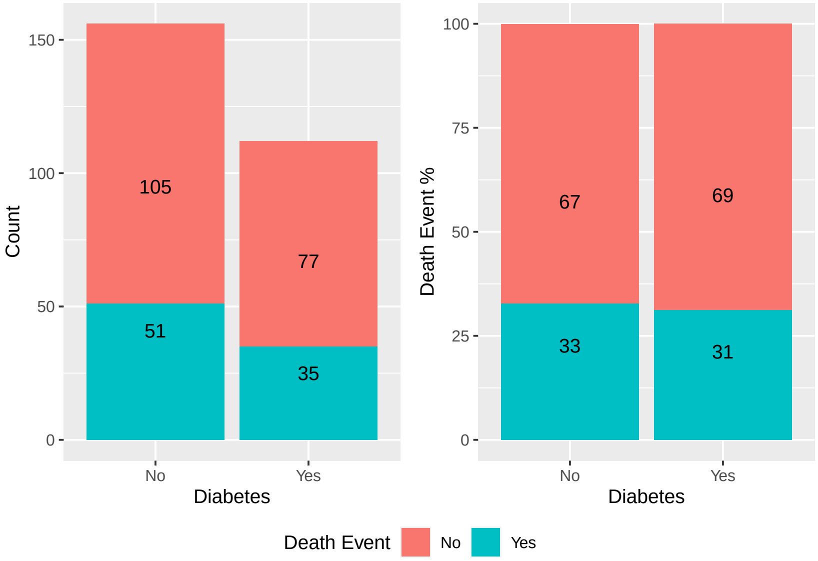

Figure 5: Diabetes Distribution

Diabetes does not make a significant difference in the Survival and Death likelihood of a patient having heart failure.

3.7 Ejection Fraction

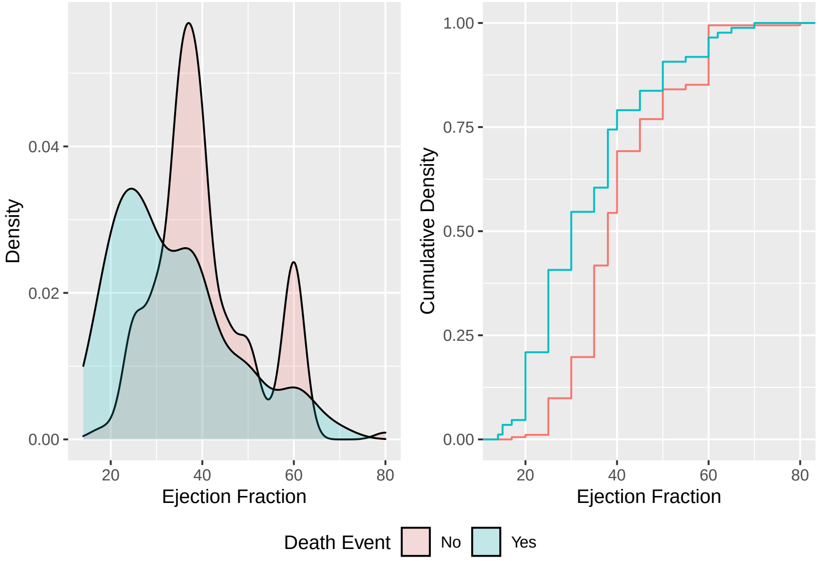

Figure 6: Ejection Fraction Distribution

There is a higher likelihood of a patient‘s Survival for an Ejection Fraction value of 30 and above. Also, the Survival likelihood peaks between the range of 35-40 and at 60 percent. Cumulative Death likelihood is higher than Survival being around three times higher between the Ejection Fraction range 25-30% and 30%-35%, upper and lower values excluded.

3.8 High Blood Pressure

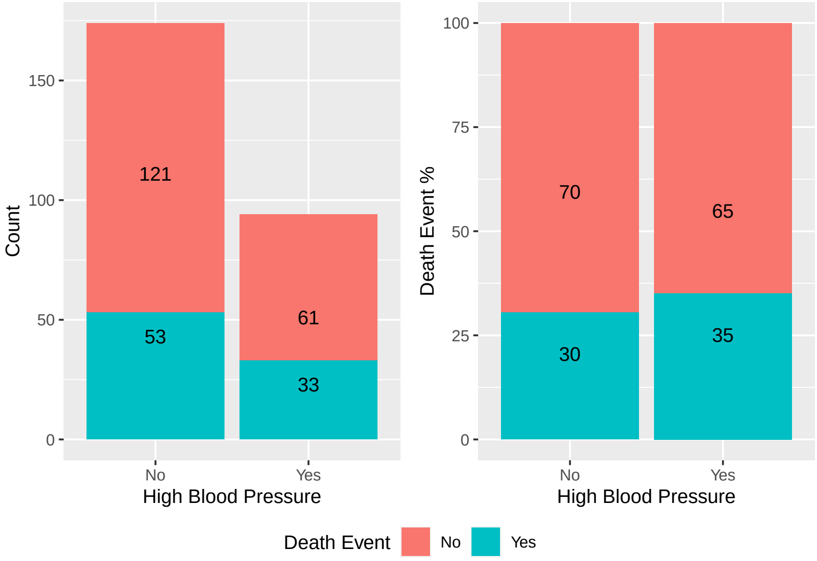

Figure 7: High Blood Pressure Distribution

A Patient having High Blood Pressure has a 35% likelihood of Death, 5% higher than a patient having no High Blood Pressure.

3.9 Platelets

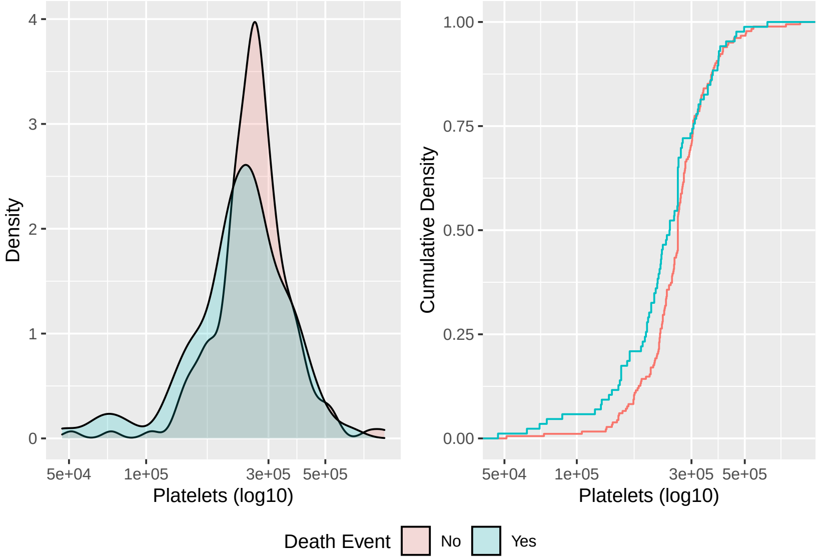

Figure 8: Platelets Distribution

There is a higher likelihood of Death for lower values of platelets whereas no difference is observed for higher values. Also, a higher likelihood of Survival is observed at the peak. Cumulative Death likelihood is higher than Survival for lower values of platelets whereas no significant difference is observed in Cumulative Death and Survival likelihood for higher values of blood platelets.

3.10 Serum Creatinine

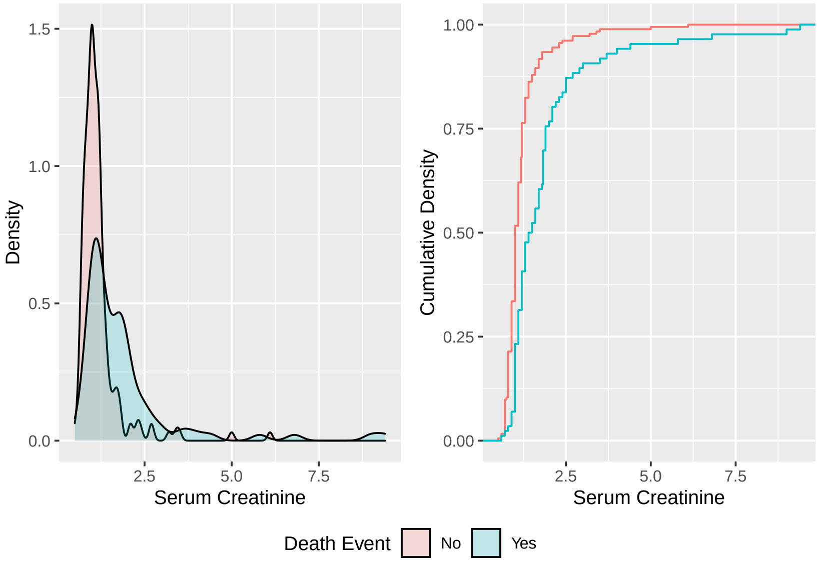

Figure 9: Serum Creatinine Distribution

There is a higher likelihood of Survival for Serum Creatinine levels of 1.25 mg/dL or below. Cumulative Survival likelihood is higher than Death which improves significantly between Serum Creatinine value 1.25 -2.5 mg/dL.

3.11 Serum Sodium

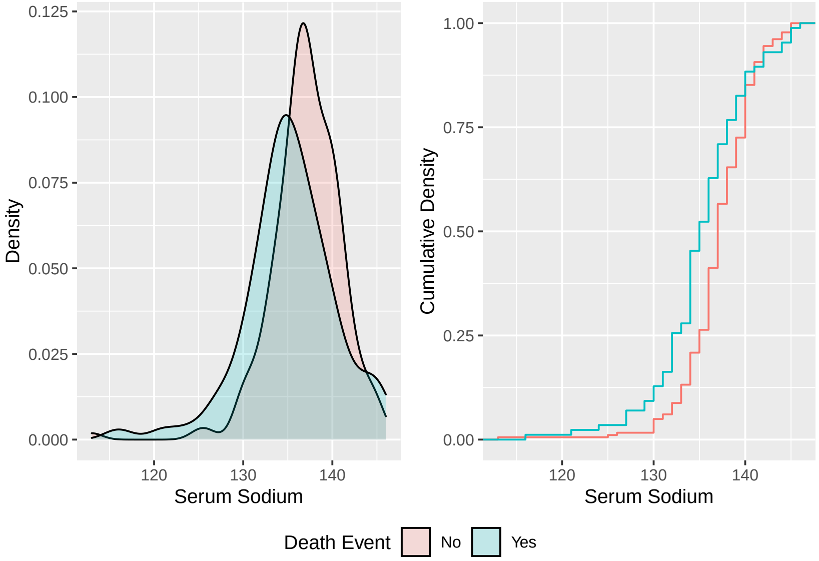

Figure 10: Serum Sodium Distribution

Serum Sodium level of 135 mEq/L or above results in a higher likelihood of Survival. Cumulative Death likelihood is higher than Survival for Serum Sodium levels below 140 mEq/L. Also, the

Cumulative Death likelihood almost becomes double the Survival at Serum Sodium level of 135 mEq/L.

3.12 Sex

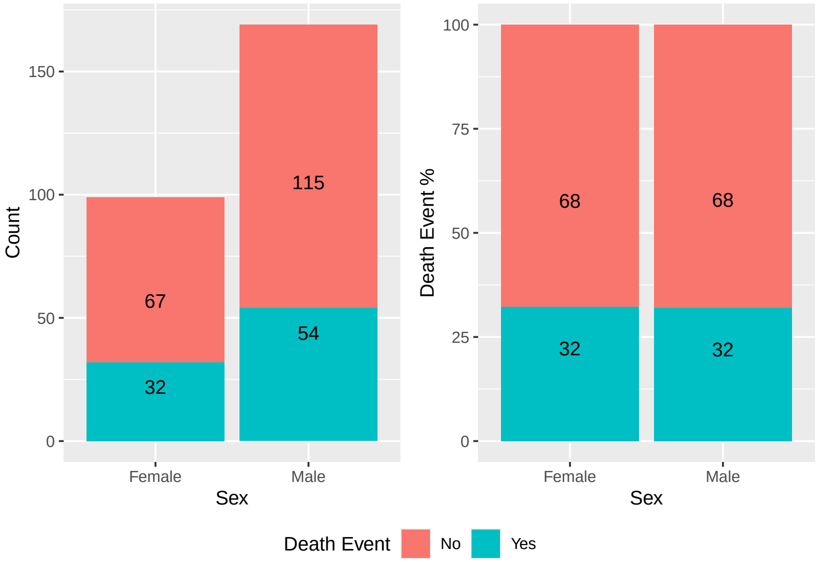

Figure 11: Sex Distribution

The sex of a patient has no significant impact on Survival and Death likelihood.

3.13 Smoking

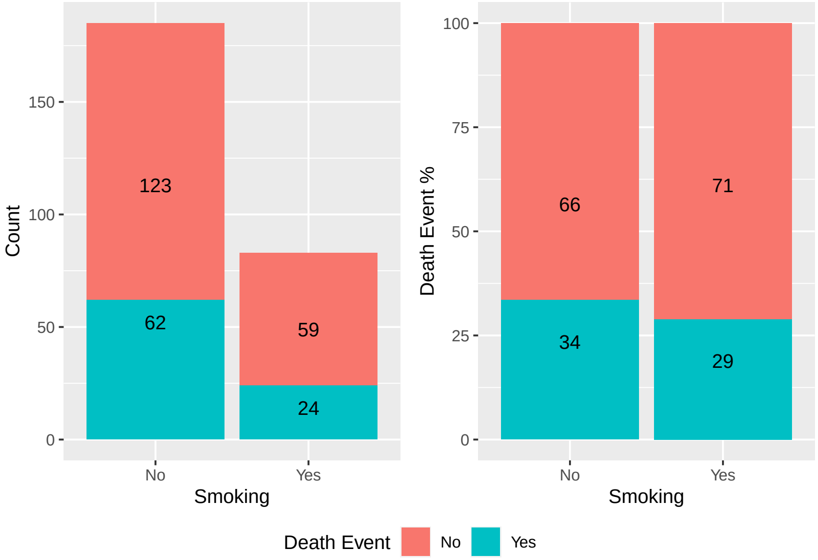

Figure 12: Smoking Distribution

A patient having a Smoking habit has a higher Survival likelihood compared to a non-smoker which is conflicting.

3.14 Time

A follow-up period of around 80 days or above results in a higher likelihood of a patient‘s Survival having Heart Failure.

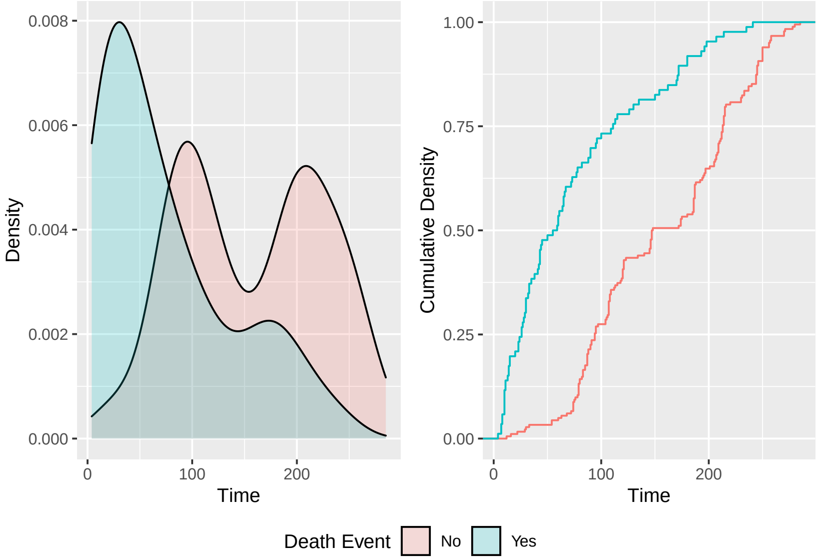

Figure 13: Time Distribution

The Cumulative Death Rate is significantly higher than the Survival Rate being almost 2 times higher at the Time value of 78. Also, the Survival Rate improves over the follow-up period as delta reduces.

3.15 Variable Importance and Correlation

Variable Importance:

Variable (Feature) Importance is estimated using information theory that ranks the best features basis several metrics between features and target variable[3].

Figure 14 shows that time is the most important feature followed by creatinine phosphokinase and platelets whereas sex, diabetes and smoking are the least important ones. The least important features must be excluded while developing the model to reduce complexity and improve accuracy.

Variable Correlation:

Variable correlation helps find variables that are related. Independent Variables or Predictor having strong correlation should be excluded from the model to reduce complexity as a correlated variable does not add any additional information to the model[4].

Figure 15 shows that there is moderate to weak correlation among features and feature exclusion is not required basis the Correlation insights.

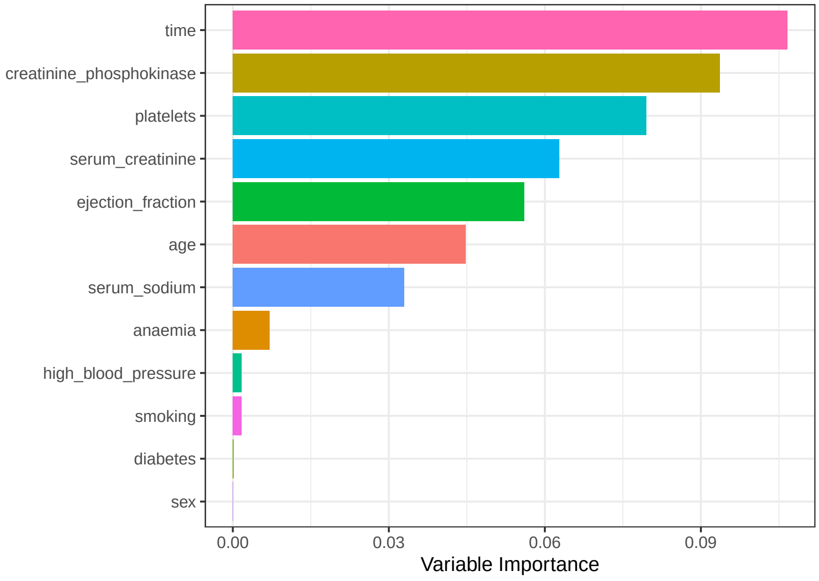

Figure 14: Variable Importance basis Information Gain

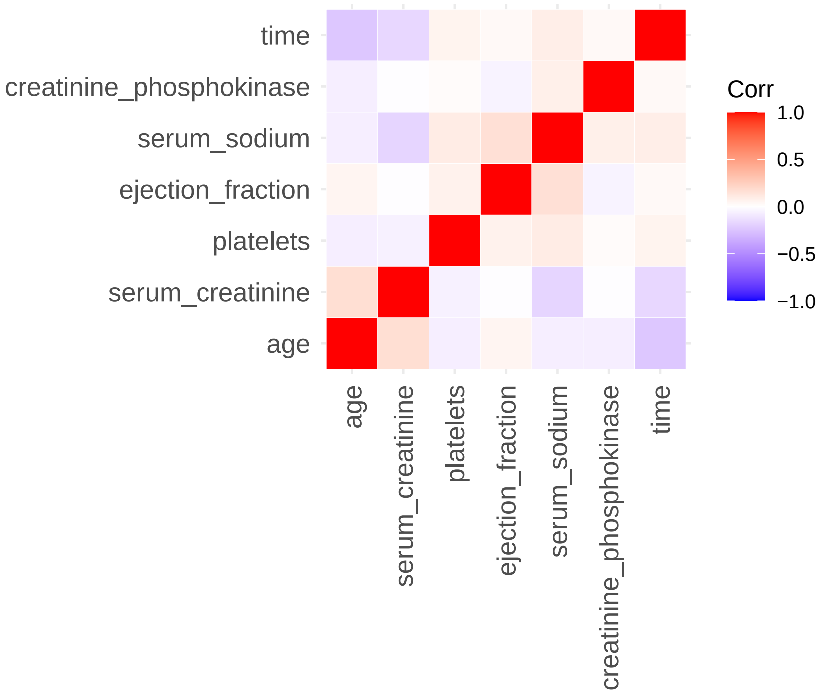

Figure 15: Variable Correlation

4 Model Development and Results

The Heart Failure Prediction Model is developed using the edx set and the final test is performed on the reserved validation set. Three features (sex, diabetes, smoking) are excluded from Model building and Validation to reduce complexity as no significant information is contributed by these features.

4.1 Splitting edx Data into Train and Test Sets

The edx data is split into the training set and testing set with 90% and 10% of the original edx set respectively. The training set is called edx_train_set and the testing set is called the edx_test_set. The model is developed using the edx_train_set and testing is performed on edx_test_set before the final test on the validation set.

4.2 Model Performance Metrics

Confusion Matrix can help estimate the performance or accuracy of a Classification Model. Confusion Matrix tabulates the Actual and Model Predicted Outcomes as shown below.

Table 2: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positives (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- True Positive: Actual = Positive, Predicted = Positive
- True Negative: Actual = Negative, Predicted = Negative
- False Positive: Actual = Negative, Predicted = Positive
- False Negative: Actual = Positive, Predicted = Negative

Accuracy: Accuracy is defined as the percentage of correct predictions. It can be calculated by dividing the number of correct predictions by the number of total predictions[5].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Sensitivity: Sensitivity is defined as True Positive Rate. It can be calculated by dividing the number of True Positive predictions by total number of Positive predictions[5].

$$Sensitivity = \frac{TP}{TP + FP}$$

Specificity: Specificity is defined as True Negative Rate. It can be calculated by dividing the number of True Negative predictions by total number of Negative predictions[5].

$$Specificity = \frac{TN}{FN + TN}$$

4.3 Model Development, Validation and Results

Classification Models are explored for Model Development as the Heart Failure Data has categorical dependent attribute.

4.3.1 Generalized Linear Model (glm)

Generalized Linear Model is an Umbrella term that consists of models like Linear Regression and Logistic Regression. Logistic regression is also called the logistic model or logit model. Logistic Regression analyzes the relationship between multiple independent variables and a categorical dependent variable. It estimates the probability of occurrence of an event by fitting data to a logistic curve[6].

The calculated model performance metrics on the edx test set are listed below in Table 3.

Table 3: Model Accuracy on Test Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.89286	0.77778	0.94737

The calculated model performance metrics on the validation set are listed below in Table 4.

Table 4: Model Accuracy on Validation Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.93548	0.8	1

4.3.2 Naive Bayes (naive_bayes)

The Naive Bayes algorithm, a probabilistic classifier is based on Bayes Theorem and used for solving high dimensional classification problems. It assumes features are independent of each other, such that each feature independently and equally contributes to the probability of a sample belonging to a specific class[7].

The calculated model performance metrics on the edx test set are listed below in Table 5.

Table 5: Model Accuracy on Test Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.89286	0.77778	0.94737
Naive Bayes (naive_bayes)	0.78571	0.44444	0.94737

The calculated model performance metrics on the validation set are listed below in Table 6.

Table 6: Model Accuracy on Validation Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.93548	0.8	1.00000
Naive Bayes (naive_bayes)	0.80645	0.5	0.95238

4.3.3 Decision Tree (rpart)

A Decision Tree is a tree diagram that is used to make a decision based on the choices called branches.

The Decision Tree algorithm is a classification method where decision trees are created recursively by splitting the nodes until a majority of records have been classified under specific class labels or a particular stopping criterion is reached[6].

The calculated model performance metrics on the edx test set are listed below in Table 7.

Table 7: Model Accuracy on Test Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.89286	0.77778	0.94737
Naive Bayes (naive_bayes)	0.78571	0.44444	0.94737
Decision Tree (rpart)	0.85714	0.66667	0.94737

The calculated model performance metrics on the validation set are listed below in Table 8.

Table 8: Model Accuracy on Validation Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.93548	0.8	1.00000
Naive Bayes (naive_bayes)	0.80645	0.5	0.95238
Decision Tree (rpart)	0.93548	0.9	0.95238

4.3.4 Random Forest (rf)

The Random Forest classifier is an ensemble method that trains several decision trees in parallel on various subsets of the training data set using different subsets of available features[6].

The final decision is a result of aggregation of individual tree decisions. The ensemble design of the Random Forest helps avoid over-fitting[6].

The calculated model performance metrics on edx the test set are listed below in Table 9. The tuning parameter mtry = 4 results in the highest accuracy.

Table 9: Model Accuracy on Test Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.89286	0.77778	0.94737
Naive Bayes (naive_bayes)	0.78571	0.44444	0.94737
Decision Tree (rpart)	0.85714	0.66667	0.94737
Random Forest (rf)	0.89286	0.66667	1.00000

The calculated model performance metrics on the validation set are listed below in Table 10.

Table 10: Model Accuracy on Validation Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.93548	0.8	1.00000
Naive Bayes (naive_bayes)	0.80645	0.5	0.95238
Decision Tree (rpart)	0.93548	0.9	0.95238
Random Forest (rf)	0.96774	0.9	1.00000

4.3.5 Support Vector Machines with Linear Kernel (svmLinear2)

Support Vector Machine is a Supervised Learning technique that can be used for Classification.

Support Vector Machines transforms the original feature space into a higher dimensional space based on a user-defined kernel function and then finds support vectors to maximize the separation (margin) between two classes[6].

Different Kernels are implemented in the Support Vector Machine to transform the feature space. The Linear Kernel is the most commonly used Kernel where data is assumed to be linearly separable[6].

The calculated model performance metrics on the edx test set are listed below in Table 11.

Table 11: Model Accuracy on Test Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.89286	0.77778	0.94737
Naive Bayes (naive_bayes)	0.78571	0.44444	0.94737
Decision Tree (rpart)	0.85714	0.66667	0.94737
Random Forest (rf)	0.89286	0.66667	1.00000
Support Vector Machines with Linear Kernel (svmLinear2)	0.85714	0.66667	0.94737

The calculated model performance metrics on the validation set are listed below in Table 12.

Table 12: Model Accuracy on Validation Set

Model	Accuracy	Sensitivity	Specificity
Generalized Linear Model (glm)	0.93548	0.8	1.00000
Naive Bayes (naive_bayes)	0.80645	0.5	0.95238
Decision Tree (rpart)	0.93548	0.9	0.95238
Random Forest (rf)	0.96774	0.9	1.00000
Support Vector Machines with Linear Kernel (svmLinear2)	0.90323	0.7	1.00000

All 5 Models have resulted in good Accuracy, Sensitivity and Specificity. Random Forest is the best performing model among all resulting in a prediction accuracy of 0.96774.

5 Conclusion

Hearth Failure Prediction Clinical Data Set is used to develop the Heart Failure Prediction Model. The data is pre-processed and split into train (edx) and validation sets to develop and validate the model respectively. Data exploration and analysis is performed to understand the data thoroughly and to estimate the importance of different features.

Classification Models - Logistic Regression, Naive Bayes, Decision Tree, Random Forest and Support Vector Machines are considered to develop Heart Prediction Model using edx train set and evaluated using edx test set. The Model performance is evaluated based on Accuracy, Sensitivity and Specificity metrics using a confusion matrix. The Model Accuracy ranges between 0.79-0.89 for the test set and 0.81-0.97 for the validation set. Random Forest results in the best prediction model having an accuracy of 0.96774. The Model metrics Accuracy, Sensitivity and Specificity are in a reasonably good range.

The models can help identify at-risk individuals at an early stage and prevent causalities. Also, it can help hospitals prioritize critical patients.

5.1 Limitation

The used data set is smaller in size but the execution process could become very time-consuming while dealing with large data sets due to limited machine memory. Also, the algorithm has to run again every time if a new patient is included which is complex to perform on a large data set.
The developed models using smaller data sets could not be as reliable as the ones developed using larger data sets. Also, additional related details about the patient like weight and occupation could be useful.

5.2 Future Scope

A larger data set could be explored to build more reliable models.

The Heart Failure Prediction Model is created using the most popular classification models. A more advanced or alternate modeling approach like Boosting Models, Multi-layer Models, Neural Networks etc. can be explored for better prediction.

6 References

1. World Health Organization, Fact Sheet, Cardiovascular Diseases
2. Heart Failure Clinical Records Data Set, UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems
3. Pablo Casas, January 2019, Data Science Live Book
4. Irizarry, Rafael A. 2020. Introduction to Data Science: Data Analysis and Prediction Algorithms with r. CRC Press
5. David Dalpiaz, October 2020, R for Statistical Learning
6. Park and Hyeoun-Ae, April 2013, An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain
7. Siddharth Misra and Hao Li, October 2019, Machine Learning for Subsurface Characterization