Biostatistics: Application of Biomedical Data Science

Application of Machine Learning and Survival Analysis on Predicting Heart Failure

Instructor: Prof. Shuangge Steven Ma

Yifan LI

Contents

1. Introduction and background 3

2. Methods 3

2.1.Data Preparation 3

2.2 Methods Summary 4

3. Results 5

3.1 Expornary Data Analysis (EDA) 5

3.1.1 Baseline Characteristic Table 5

3.1.2 Correlation Matrix 6

3.2 Machine Learning Models 6

3.2.1 Classification of training and test datas 7

3.2.2 Logistic Regression Model 7

3.2.3 Support Vector Machine 8

3.2.4 Decision Tree 9

3.2.5 Random Forest 10

3.2.6 Model Comparison 10

3.3 Survival Analysis 11

3.3.1 KM Estimator 11

3.3.2 Propotional Ratio Hazard Model 13

4. Discussion 15

5. Conclusion 16

References 17

1 Introduction and Background

Heart failure is a condition where the heart muscle is unable to adequately pump blood to meet the needs of the body. The primary fluid that moves throughout the body and carries oxygen to every cell is blood. The leading cause of hospital admissions and fatalities, acute heart failure places a growing burden on healthcare systems[1].

A very prevalent cardiovascular disease is heart failure. In the world, cardiovascular diseases claim the lives of 17.9 million people annually, or 31% of all fatalities. The ongoing threat of cardiovascular issues is growing as a result of unhealthy lifestyle choices and a careless attitude toward health. For the majority of people who experience mental health issues, the general public has embraced behaviors like smoking, unhealthy eating and obesity, sedentary lifestyles, and harmful alcohol use. A machine learning model and survival analysis can be very useful in the early diagnosis and treatment of people with high cardiovascular risk[2].

Fortunately, behavioral risk factors can be addressed through population-level strategies to prevent the majority of cardiovascular diseases. By preventing the overtreatment of low-risk patients and the premature discharge of high-risk patients, appropriate patient risk stratification can enhance clinical outcomes and resource allocation.

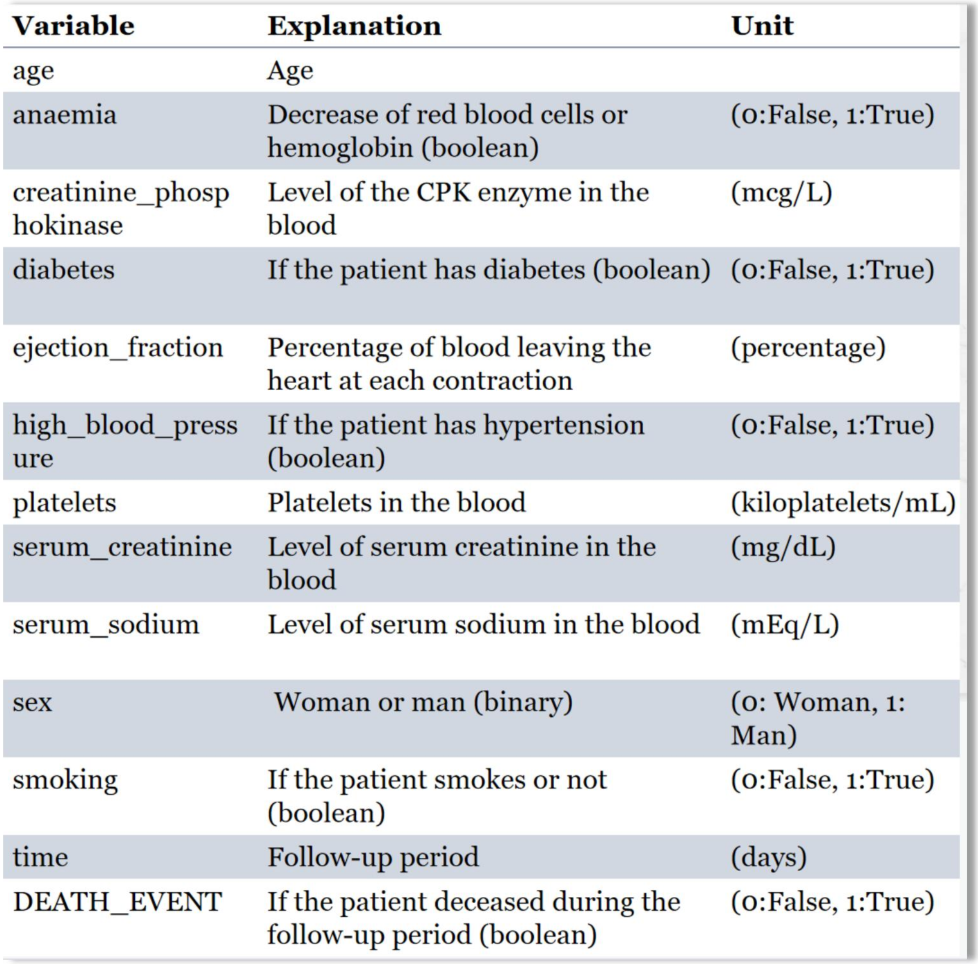
In order to predict heart failure events and mortality, this project will perform exploratory data analysis, use various machine learning models to identify key features, and then apply the Cox model, survival analysis, and risk ratio to validate the outcome.

2 Methods

2.1 Data Preparation

The data set is available as a CSV file. It is stored directly as a data frame that requires no cleanup and does not contain missing data, so no imputation needs to be done. The data set is from Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Health Informatics and Decision Making 20, 16 (2020)[3].

The dataset presented in this project contains the medical records of 299 heart failure patients (lines) collected between April and December 2015 at the Faisalabad Heart Institute and Faisalabad Allied Hospital (Punjab, Pakistan).The patients consisted of 105 women and 194 men and their ages ranged from 40 to 95 years. Contains 13 columns (traits): age, anemia,hypertension, creatine phosphokinase (CPK), diabetes, ejection fraction, gender, platelets, serum creatinine, serum sodium, smoking, time point, and death, since the latter is the characteristic target of the predictive model. See Table 1 for a brief description of each variable:



Variable	Explanation	Unit
age	Age	
anaemia	Decrease of red blood cells or hemoglobin (boolean)	(0:False, 1:True)
creatinine_phosphokinase	Level of the CPK enzyme in the blood	(mcg/L)
diabetes	If the patient has diabetes (boolean)	(0:False, 1:True)
ejection_fraction	Percentage of blood leaving the heart at each contraction	(percentage)
high_blood_pressure	If the patient has hypertension (boolean)	(0:False, 1:True)
platelets	Platelets in the blood	(kiloplatelets/mL)
serum_creatinine	Level of serum creatinine in the blood	(mg/dL)
serum_sodium	Level of serum sodium in the blood	(mEq/L)
sex	Woman or man (binary)	(0: Woman, 1: Man)
smoking	If the patient smokes or not (boolean)	(0:False, 1:True)
time	Follow-up period	(days)
DEATH_EVENT	If the patient deceased during the follow-up period (boolean)	(0:False, 1:True)

Table 1:Data Descriptions

As is shown in table 1, some features are numeric: years, Creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium and days; an others are boolean (binary), so they ranges are 0 or 1: anaemia, high blood pressure, diabetes, sex, smoking and death event.

However, binary features should be changed to factors, and where 0 represents "no" and 1 "yes" in anaemia, High blood pressure, diabetes, smocking and death event; and in sex feature "female" or "male" respectively.

2.2 Methods Summary

The methods will used is summarized below:

1. Exploratory Data Analysis

- Baseline Characteristic Table
- Correlation Matrix

2. Machine Learning Models

- Logistic Model
- Decision Tree
- Random Forest
- Support Vector Machine

3. Survival Analysis

- KM Estimator
- Proportional Ratio Hazard Model

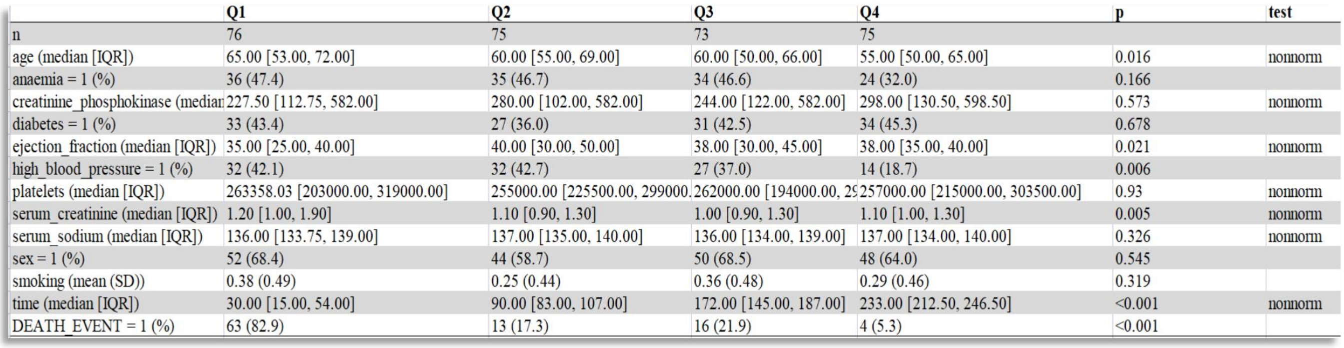
3 Results

3.1 EDA (Exploratory Data Analysis)

Exploratory data analysis (EDA) is used by data scientists to analyze and examine data sets and summarize their key characteristics, often using data visualization methods. It helps find the best way to manipulate data sources to get the answers they need and makes it easier for data scientists to spot patterns, find anomalies, test hypotheses, or test hypotheses[4].

The main purpose of EDA is to help you examine the data before forming hypotheses. It can help you identify obvious errors, better understand patterns in your data, spot outliers or unusual events, and find interesting relationships between variables.

This is the Baseline Characteristic Table for this dataset, which disaggregates the quartiles of the variable of time and the results are shown below:

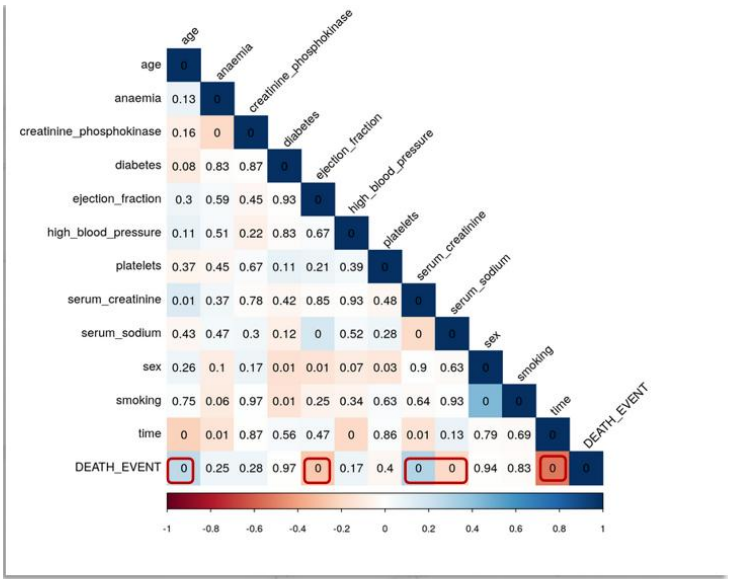


	Q1	Q2	Q3	Q4	p	test
n	76	75	73	75		
age (median [IQR])	65.00 [53.00, 72.00]	60.00 [55.00, 69.00]	60.00 [50.00, 66.00]	55.00 [50.00, 65.00]	0.016	nonnorm
anaemia = 1 (%)	36 (47.4)	35 (46.7)	34 (46.6)	24 (32.0)	0.166	
creatinine_phosphokinase (median	227.50 [112.75, 582.00]	280.00 [102.00, 582.00]	244.00 [122.00, 582.00]	298.00 [130.50, 598.50]	0.573	nonnorm
diabetes = 1 (%)	33 (43.4)	27 (36.0)	31 (42.5)	34 (45.3)	0.678	
ejection_fraction (median [IQR])	35.00 [25.00, 40.00]	40.00 [30.00, 50.00]	38.00 [30.00, 45.00]	38.00 [35.00, 40.00]	0.021	nonnorm
high_blood_pressure = 1 (%)	32 (42.1)	32 (42.7)	27 (37.0)	14 (18.7)	0.006	
platelets (median [IQR])	263358.03 [203000.00, 319000.00]	255000.00 [225500.00, 299000.	262000.00 [194000.00, 2!	257000.00 [215000.00, 303500.00]	0.93	nonnorm
serum_creatinine (median [IQR])	1.20 [1.00, 1.90]	1.10 [0.90, 1.30]	1.00 [0.90, 1.30]	1.10 [1.00, 1.30]	0.005	nonnorm
serum_sodium (median [IQR])	136.00 [133.75, 139.00]	137.00 [135.00, 140.00]	136.00 [134.00, 139.00]	137.00 [134.00, 140.00]	0.326	nonnorm
sex = 1 (%)	52 (68.4)	44 (58.7)	50 (68.5)	48 (64.0)	0.545	
smoking (mean (SD))	0.38 (0.49)	0.25 (0.44)	0.36 (0.48)	0.29 (0.46)	0.319	
time (median [IQR])	30.00 [15.00, 54.00]	90.00 [83.00, 107.00]	172.00 [145.00, 187.00]	233.00 [212.50, 246.50]	<0.001	nonnorm
DEATH_EVENT = 1 (%)	63 (82.9)	13 (17.3)	16 (21.9)	4 (5.3)	<0.001	

Table 2: Baseline Characteristic Table

3.1.1 Correlation Matrix

The Correlation Matrix for this dataset is shown below, and the last row shows the correlation indices for death event and several other variables.



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	0												
anaemia	0.13	0											
creatinine_phosphokinase	0.16	0	0										
diabetes	0.08	0.83	0.87	0									
ejection_fraction	0.3	0.59	0.45	0.93	0								
high_blood_pressure	0.11	0.51	0.22	0.83	0.67	0							
platelets	0.37	0.45	0.67	0.11	0.21	0.39	0						
serum_creatinine	0.01	0.37	0.78	0.42	0.85	0.93	0.48	0					
serum_sodium	0.43	0.47	0.3	0.12	0	0.52	0.28	0	0				
sex	0.26	0.1	0.17	0.01	0.01	0.07	0.03	0.9	0.63	0			
smoking	0.75	0.06	0.97	0.01	0.25	0.34	0.63	0.64	0.93	0	0		
time	0	0.01	0.87	0.56	0.47	0	0.86	0.01	0.13	0.79	0.69	0	
DEATH_EVENT	0	0.25	0.28	0.97	0	0.17	0.4	0	0	0.94	0.83	0	0

Figure 1:Correlation Matrix

We can see that Death Event is highly correlated with serum creatinine, age, serum sodium, ejection fraction and time.

More importantly, outcome (death) has a significant correlation with age, ejection fraction, serum creatinine, and serum sodium predictors, which must be accounted for in the final model. The only strong correlation between predictors alone was found for gender and smoking scores. No significant values or outliers could be identified.

3.2 Machine Learning Models

Prediction of patients' mortality (binary outcome) based on several predictors presents a typical classification problem. Machine learning algorithms that will be used in this project are as following:

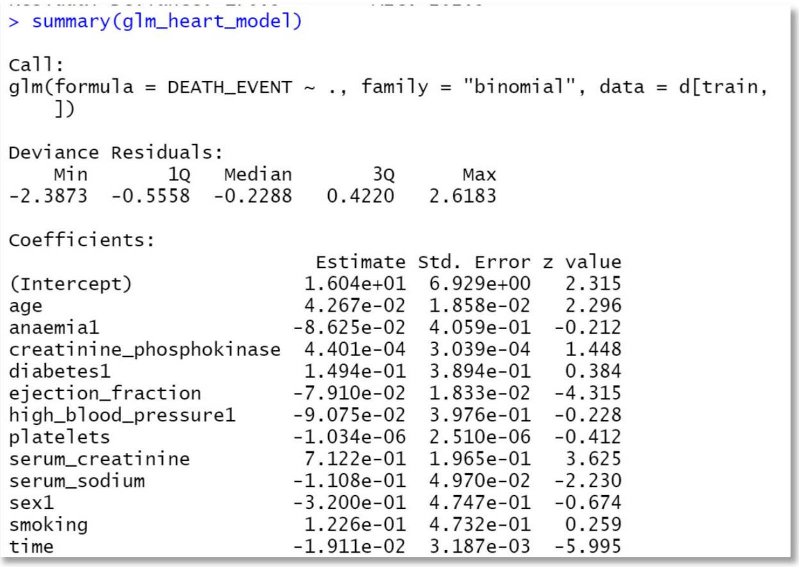
- Random Forest
- Decision Tree
- Logistic regression Model
- Support vector machine (SVM)

3.2.1 Classification of training and test set

Before building the prediction and classification models, we need to partition the dataset into a training set and a test set. Based on the resulting data set, train and test sets are created. Due to the small size of the data set, 80% of data will be used for training. We randomly partitioned the dataset with a total of 299 observations in a ratio of 80:20, resulting in 239 training sets and 60 test sets. It is shown that the both sets have a similar proportion of positive and negative outcomes.

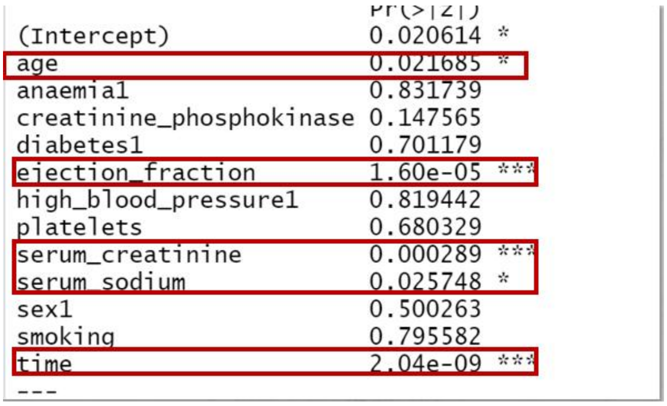
3.2.2 Logistic Regression Model

We now perform logistic regression on the training set and take the 12 variables as predictors. We then look at the model summary and then make predictions on the test data set. We store the success rate of the test data in a vector for later comparison and also look at the confusion matrix to understand the numbers where the predictions worked or didn't work.



```
> summary(glm_heart_model)

Call:
glm(formula = DEATH_EVENT ~ ., family = "binomial", data = d[train,
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3873  -0.5558  -0.2288   0.4220   2.6183

Coefficients:
                           Estimate Std. Error z value
(Intercept)               1.604e+01  6.929e+00   2.315
age                       4.267e-02  1.858e-02   2.296
anaemia1                 -8.625e-02  4.059e-01  -0.212
creatinine_phosphokinase  4.401e-04  3.039e-04   1.448
diabetes1                 1.494e-01  3.894e-01   0.384
ejection_fraction        -7.910e-02  1.833e-02  -4.315
high_blood_pressure1     -9.075e-02  3.976e-01  -0.228
platelets                -1.034e-06  2.510e-06  -0.412
serum_creatinine          7.122e-01  1.965e-01   3.625
serum_sodium             -1.108e-01  4.970e-02  -2.230
sex1                     -3.200e-01  4.747e-01  -0.674
smoking                   1.226e-01  4.732e-01   0.259
time                     -1.911e-02  3.187e-03  -5.995
```

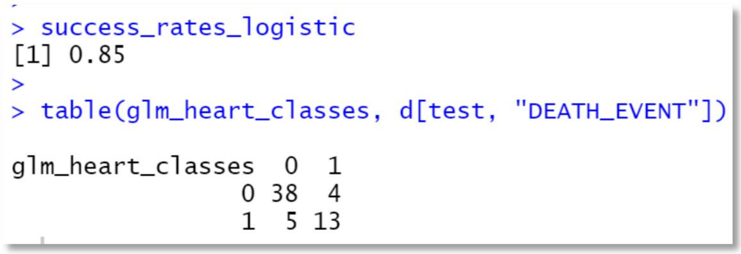
Figure 2:Summary Outcome of Logistic Model



```
                         Pr(>|z|)
(Intercept)              0.020614 *
age                      0.021685 *
anaemia1                 0.831739
creatinine_phosphokinase 0.147565
diabetes1                0.701179
ejection_fraction        1.60e-05 ***
high_blood_pressure1     0.819442
platelets                0.680329
serum_creatinine         0.000289 ***
serum_sodium             0.025748 *
sex1                     0.500263
smoking                  0.795582
time                     2.04e-09 ***
---
```

Figure 3:P-value of Logistic Model

As can be seen from the above summary statistics that age, ejection fraction, serum creatinine, serum sodium and time (follow up time) are some of the significant predictors of heart failure. We now use this model to predict the event of death due to heart failure on the test data and then calculate and store the success rate of the predicted values.



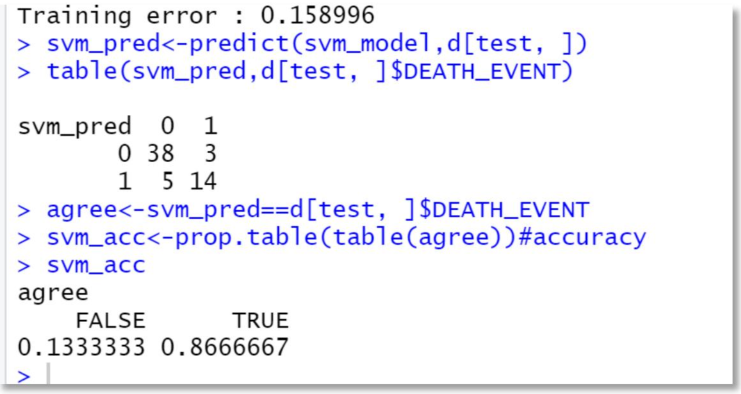
```
> success_rates_logistic
[1] 0.85
>
> table(glm_heart_classes, d[test, "DEATH_EVENT"])

glm_heart_classes  0  1
                0 38  4
                1  5 13
```

Figure 4:Confusion Matrix and Success Rate of Logistic Model

The confusion matrix above shows the hits and misses of the model. Also we observe the success rate of the logistic model as about 85%. The model with 12 predictors and 239 training observations does a fairly good job in predicting the response and surely performs better than tossing a coin.

3.2.3 Support Vector Machine(SVM)

For classification and other uses, the SVM (Support Vector Machine) model draws a line (or separation hyperplane) that clearly divides two or more classes[5]. A support vector machine will actually be trained using training data, and the ROC curve, confusion matrix, and predict accuracy will all be displayed.



```
Training error : 0.158996
> svm_pred<-predict(svm_model,d[test, ])
> table(svm_pred,d[test, ]$DEATH_EVENT)

svm_pred  0  1
       0 38  3
       1  5 14
> agree<-svm_pred==d[test, ]$DEATH_EVENT
> svm_acc<-prop.table(table(agree))#accuracy
> svm_acc
agree
    FALSE      TRUE
0.1333333 0.8666667
>
```

Figure 5:Confusion Matrix and Success Rate of SVM Model

We then used the Support Vector Machine (SVM) model to make predictions and the results showed that the success rate of SVM was 86.67%.

AUC value is the area under the ROC curve, and the the ROC curve is shown below:

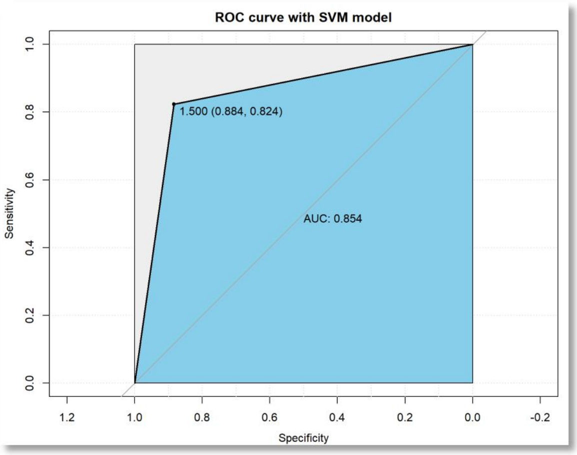


Figure 6:ROC Curve of SVM model

We can see that both accuracy and recall have increased compared to the logistic regression analysis.

3.2.4 Decision Tree

The non-parametric supervised learning algorithm used for classification and regression tasks is the decision tree. It has a tree-like hierarchy with a root node, branches, internal nodes, and leaf nodes[6].

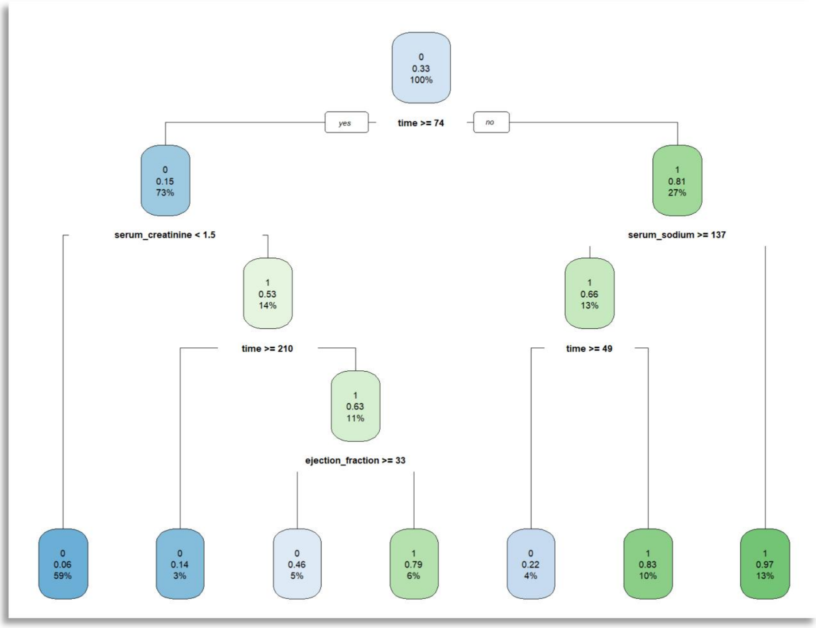
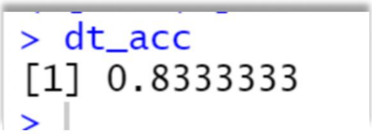


Figure 7: Decision Tree Plot

When we look at this decision tree, we notice that features with high p-values, such as time, ejection_fraction, and serum_creatinine, are used in the decision to branch.



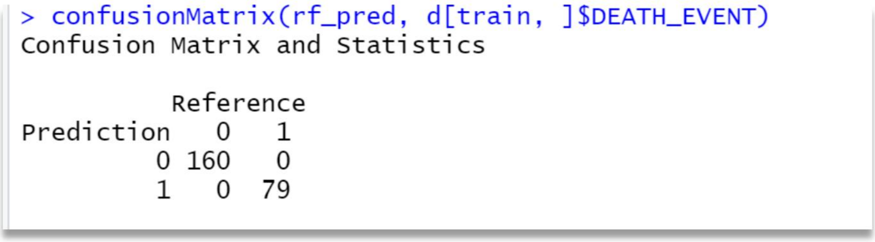
```
> dt_acc
[1] 0.8333333
>
```

Figure 8: Success Rate of Decision Tree

Similarly, we used the decision tree classifier to draw a decision tree and found a success rate of 83.33% for the model.

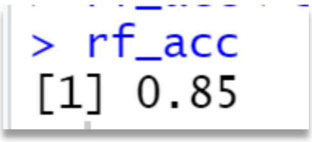
3.2.5 Random Forest

The Random Forest model extracts test data from samples, builds decision trees for each sample in parallel, and then predicts based on the average or majority vote of these decision trees. The model will be trained, and predictions will be made as before.



```
> confusionMatrix(rf_pred, d[train, ]$DEATH_EVENT)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  79
```

Figure 9:Confusion Matrix of Random Forest



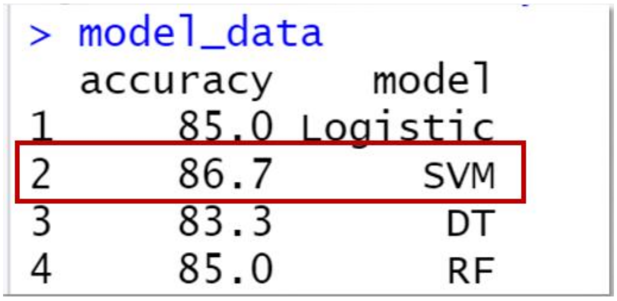
```
> rf_acc
[1] 0.85
```

Figure 10: Success Rate of Random Forest

Finally, we applied the random forest prediction to find its Confusion matrix and a success rate of 85%, which is better than the performance of decision tree.

3.2.6 Model Comparison

We compared the accuracy values of the previous four predictive classification models and we ultimately found that the SVM model performed best in predicting heart failure events with the highest predictive accuracy. The results are shown below:



```
> model_data
  accuracy    model
1     85.0 Logistic
2     86.7      SVM
3     83.3       DT
4     85.0       RF
```

Figure 11:Comparison of Success Rate for four models

3.3 Survival Analysis

The most effective statistical modeling technique for working with censored, time-to-event data, which is the most popular statistical method in the medical literature, is survival analysis. It's crucial to understand the time of event for each patient. The moment of heart failure occurs in our case[7].

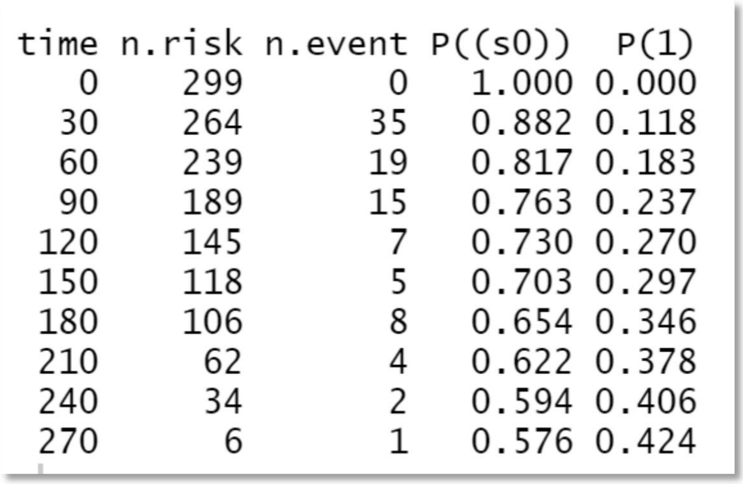
Given a cohort of patients and an observation window, we would monitor the event for some of the patients while censoring the data for the others because the observation window had passed or they had prematurely left the study.

Applying survival analysis techniques to the current Heart Failure data set is highly recommended. And we will use 2 models below for model building and fitting:

- Kaplan-Meier Estimator
- Cox Proportional Hazard Model

3.3.1 Kaplan-Meier Estimator

The Kaplan-Meier estimator is a test statistic that provides an approximate representation of the population's true survival function, with the approximation improving with sample size. Maximum Likelihood Estimation can be used to derive this estimator from the Hazard Function, which can robustly handle censoring[8].



```
time n.risk n.event P((s0))  P(1)
   0    299       0   1.000 0.000
  30    264      35   0.882 0.118
  60    239      19   0.817 0.183
  90    189      15   0.763 0.237
 120    145       7   0.730 0.270
 150    118       5   0.703 0.297
 180    106       8   0.654 0.346
 210     62       4   0.622 0.378
 240     34       2   0.594 0.406
 270      6       1   0.576 0.424
```

Figure 12:Cumulative Survival Probability Table

The cumulative survival probability broken down by time is provided in a table by the Kaplan-Meier model. Life insurance companies frequently use such "life tables" when developing their products.

Graphical representations of Kaplan-Meier plots are also possible:

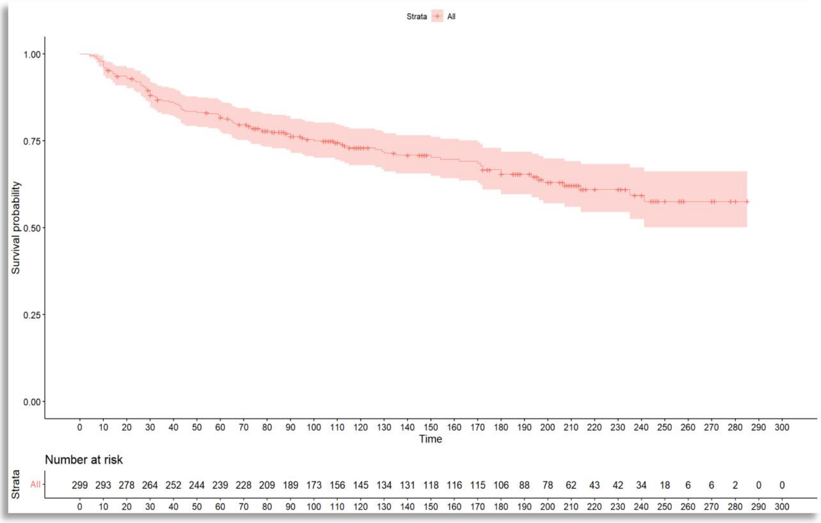
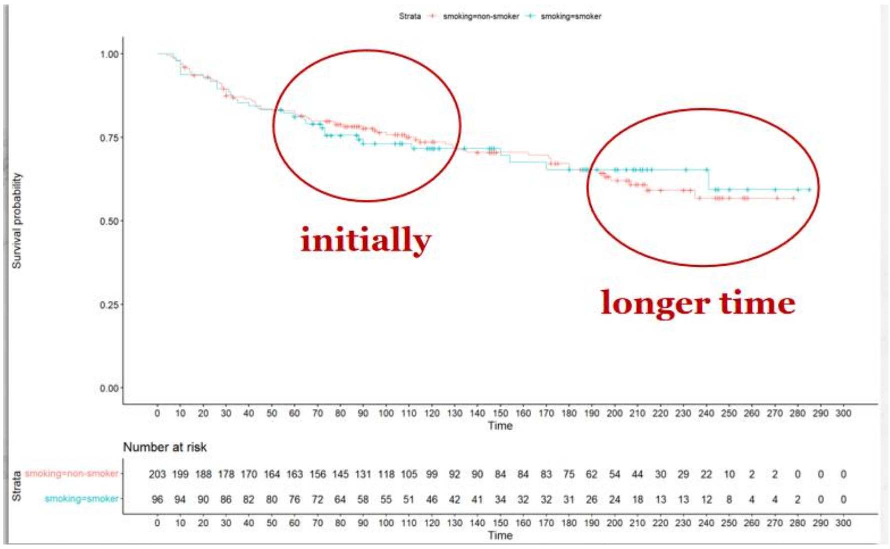


Figure 13:Kaplan-Meier plot

We can infer from this plot that as sample size rises, the Kaplan-Meier plot gets closer to the population's actual survival curve. The "+" tick marks on the plot indicate a censoring event.

It is also possible to analyze the effect of categorical features on survival using a Kaplan-Meier plot. This estimator can be used to quickly compare group survival rates. compare the survival rates of smokers and non-smokers, for instance.



Strata	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
smoking=non-smoker	203	199	188	178	170	164	163	156	145	131	118	105	99	92	90	84	84	83	75	62	54	44	30	29	22	10	2	2	0	0	0
smoking=smoker	96	94	90	86	82	80	76	72	64	58	55	51	46	42	41	34	32	32	31	26	24	18	13	13	12	8	4	4	2	0	0

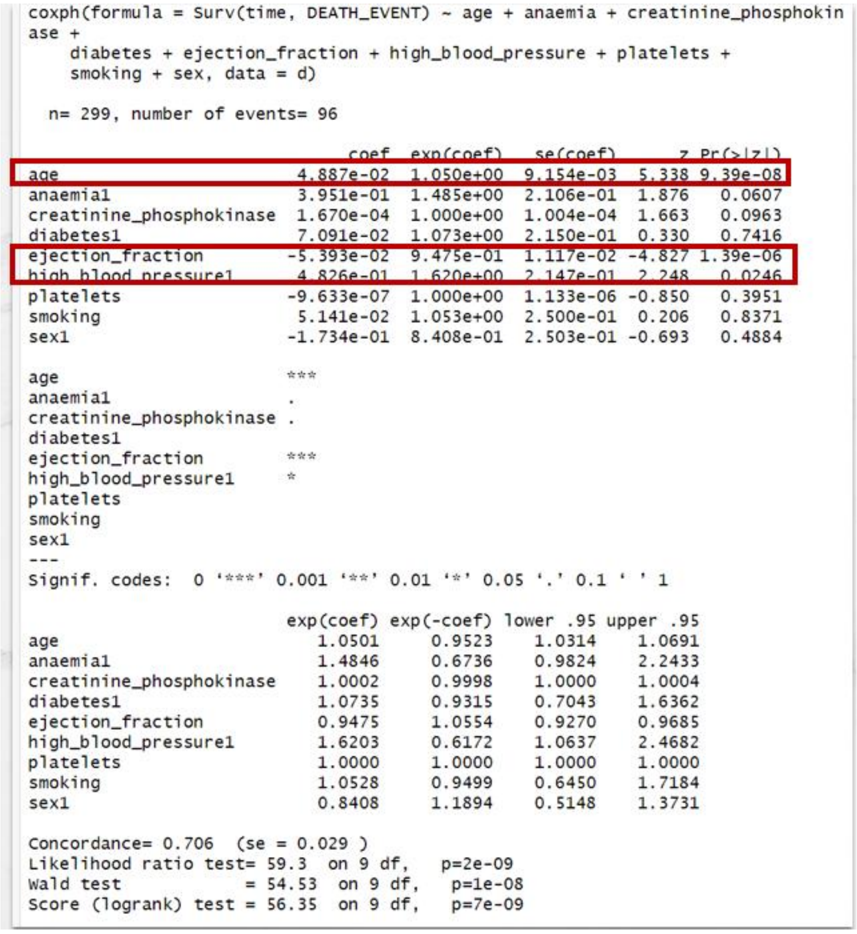
Figure 14:Comparison KM Plot of Smoker and Non-smoker

We can see that non-smokers have a higher survival rate than smokers in the early stages, and that non-smokers have a lower survival rate than smokers in the later stages.

As a result, insurance companies can design business rules for life insurance products using such straightforward models. A smoker's insurance premiums will be raised to reflect the decreased chance of survival.

3.3.2 Cox Proportional Hazard Model

The Cox Proportional Hazard Model is a survival analysis model that makes the assumption that the covariates multiply affect the population's baseline hazard function. Hazard rates can be modelled using the Cox Proportional Hazard model based on a variety of features, either categorical or numerical[9].



```
coxph(formula = Surv(time, DEATH_EVENT) ~ age + anaemia + creatinine_phosphokin
ase +
    diabetes + ejection_fraction + high_blood_pressure + platelets +
    smoking + sex, data = d)

  n= 299, number of events= 96

                              coef  exp(coef)   se(coef)      z Pr(>|z|)
age                      4.887e-02  1.050e+00  9.154e-03  5.338 9.39e-08
anaemia1                 3.951e-01  1.485e+00  2.106e-01  1.876   0.0607
creatinine_phosphokinase 1.670e-04  1.000e+00  1.004e-04  1.663   0.0963
diabetes1                7.091e-02  1.073e+00  2.150e-01  0.330   0.7416
ejection_fraction       -5.393e-02  9.475e-01  1.117e-02 -4.827 1.39e-06
high_blood_pressure1     4.826e-01  1.620e+00  2.147e-01  2.248   0.0246
platelets               -9.633e-07  1.000e+00  1.133e-06 -0.850   0.3951
smoking                  5.141e-02  1.053e+00  2.500e-01  0.206   0.8371
sex1                    -1.734e-01  8.408e-01  2.503e-01 -0.693   0.4884

age                      ***
anaemia1                 .
creatinine_phosphokinase .
diabetes1
ejection_fraction        ***
high_blood_pressure1     *
platelets
smoking
sex1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
age                         1.0501     0.9523    1.0314    1.0691
anaemia1                    1.4846     0.6736    0.9824    2.2433
creatinine_phosphokinase    1.0002     0.9998    1.0000    1.0004
diabetes1                   1.0735     0.9315    0.7043    1.6362
ejection_fraction           0.9475     1.0554    0.9270    0.9685
high_blood_pressure1        1.6203     0.6172    1.0637    2.4682
platelets                   1.0000     1.0000    1.0000    1.0000
smoking                     1.0528     0.9499    0.6450    1.7184
sex1                        0.8408     1.1894    0.5148    1.3731

Concordance= 0.706  (se = 0.029 )
Likelihood ratio test= 59.3  on 9 df,   p=2e-09
Wald test            = 54.53  on 9 df,   p=1e-08
Score (logrank) test = 56.35  on 9 df,   p=7e-09
```

Figure 15: Summary Outcome of Cox Proportional Hazard Model

The p-values allow us to know that the variables age, ejection fraction and high blood pressure have a greater effect on survival time.

We can also analyse the strength of the effect of each variable. As an example, if the variable anaemial increases by one unit, the risk of failure for that patient will increase by 48%, and if the variable anaemial decreases by one unit, the risk of failure for that patient will decrease by 33%.

Now that the model has been fit, we can also use it to plot the cumulative survival probability of a population. The graph below shows the results of the plotting:

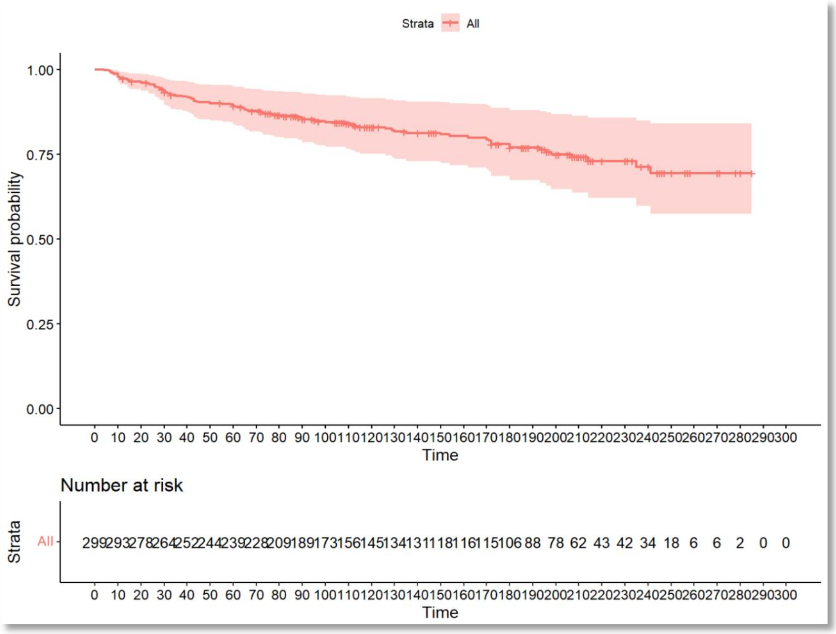


Figure 16:Cumulative Survival Probability Plot

4 Discussion

✓ Significance

The significance of this project is that our finding can be used as a reference for which indicators are more significant in the study of the probability of survival and death in heart failure, to maximize survival among cardiovascular disease patients and can then be analyzed and tested in more depth in a clinical setting.

✓ Finding and Comparison

While previous studies have shown that smoking increases the probability of cardiovascular disease, this project did not show a significant association and at a later stage non-smokers had a higher probability of survival compared to smokers.

✓ Limitation

The dataset for this project is small, and larger datasets with more data are needed for training and prediction models to produce better simulations and predictions. More information would have provided a strong base for cross-validation and training, which would have improved and deepened learning. It is likely that several of the tested algorithms would perform better with more data.

- ✓ Outlook

The selection of algorithms used to train the models included both straightforward and more sophisticated methods of various kinds. Therefore, to achieve higher accuracy rates and ultimately more accurate model predictions, training and testing algorithms for future research can be chosen that are better suited for smaller datasets.

5 Conclusion

- ➢ For EDA:

In the first part of the EDA, we found that Death Event is highly correlated with serum creatinine, age, serum sodium, ejection fraction, according to the correlation matrix.

- ➢ For Machine Learning Classification and Prediction models:

In the second part of the machine learning, we found that the SVM model performed best among the four models, with the best prediction and an accuracy of 86.7%.

- ➢ For the Survival Analysis:

In the last part of the survival analysis, we found that the survival time of the patient was highly correlated with covariate age, ejection fraction, high blood plessures.

Reference

[1] Cleveland Clinic. (2022, January 21). *Understanding Heart Failure | Cleveland Clinic*. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/17069-heart-failure-understanding-heart-failure

[2] World. (2019, June 11). *Cardiovascular diseases*. Who.int; World Health Organization: WHO. https://www.who.int/health-topics/hypertension/cardiovascular-diseases#tab=tab_1

[3] 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). https://doi.org/10.1186/s12911-020-1023-5

[4] Patil, P. (2018, March 23). *What is Exploratory Data Analysis?* Towards Data Science; Towards Data Science. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[5] Saini, A. (2021, October 12). *Support Vector Machine(SVM): A Complete guide for beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[6] IBM. (2022). *What is a Decision Tree | IBM.* Www.ibm.com. https://www.ibm.com/topics/decision-trees

[7] Singh, R., & Mukhopadhyay, K. (2011). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, *2*(4), 145. https://doi.org/10.4103/2229-3485.86872

[8] Wikipedia Contributors. (2019, October 23). *Kaplan–Meier estimator*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator

[9] Deo, S. V., Deo, V., & Sundaram, V. (2021). Survival analysis—part 2: Cox proportional hazards model. *Indian Journal of Thoracic and Cardiovascular Surgery*, *37*(2), 229–233. https://doi.org/10.1007/s12055-020-01108-7