Application of Machine Learning and Survival Analysis on Predicting Heart Failure

Supervisor: Prof.Ma

Speaker：Yifan LI

Content

01

Introduction & Background

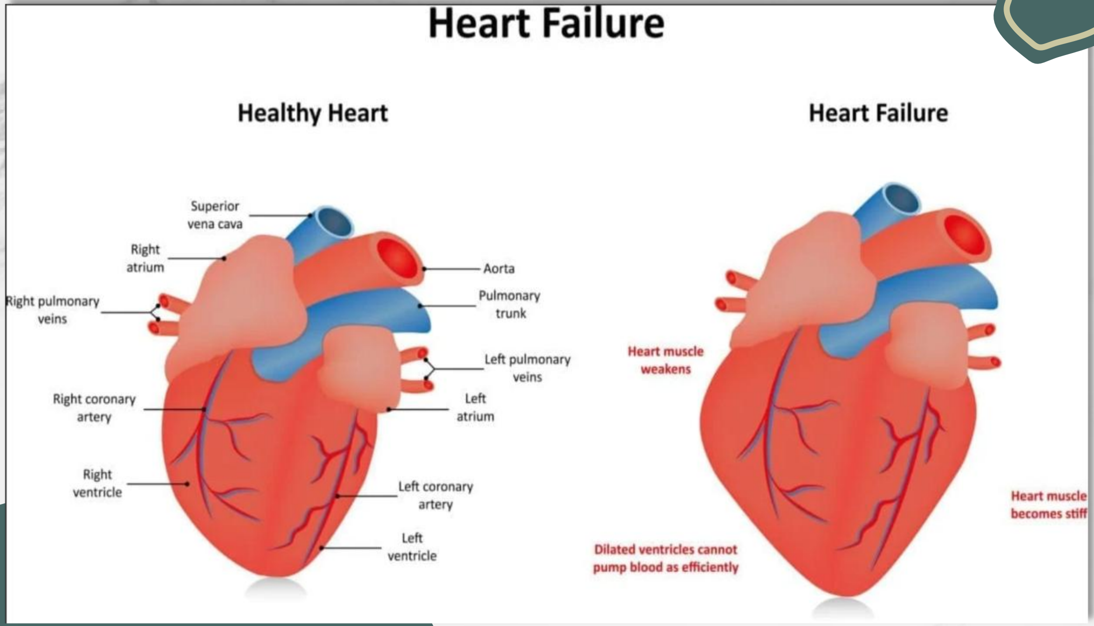
Background

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help[1].

What is Heart Failure?



Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure[2].

This project is aiming to do an exploratory data analysis, utilize various machine learning models to detect the most crucial features to predict the heart failure event and apply Cox model, Survival Analysis, and Hazard Ratio to validate the result.

02 Methodology

Dataset

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)

Dataset from Davide Chicco, Giuseppe Jurman[3]

What will do next?

Exploratory Data Analysis (EDA)

Baseline Characteristic Table

Correlation Matrix

Modelling

Logistic Model

Decision Tree

Random Forest

SVM

Survival Analysis

KM Estimator

Propotional Ratio Hazard Model

03

Results

Results --EDA

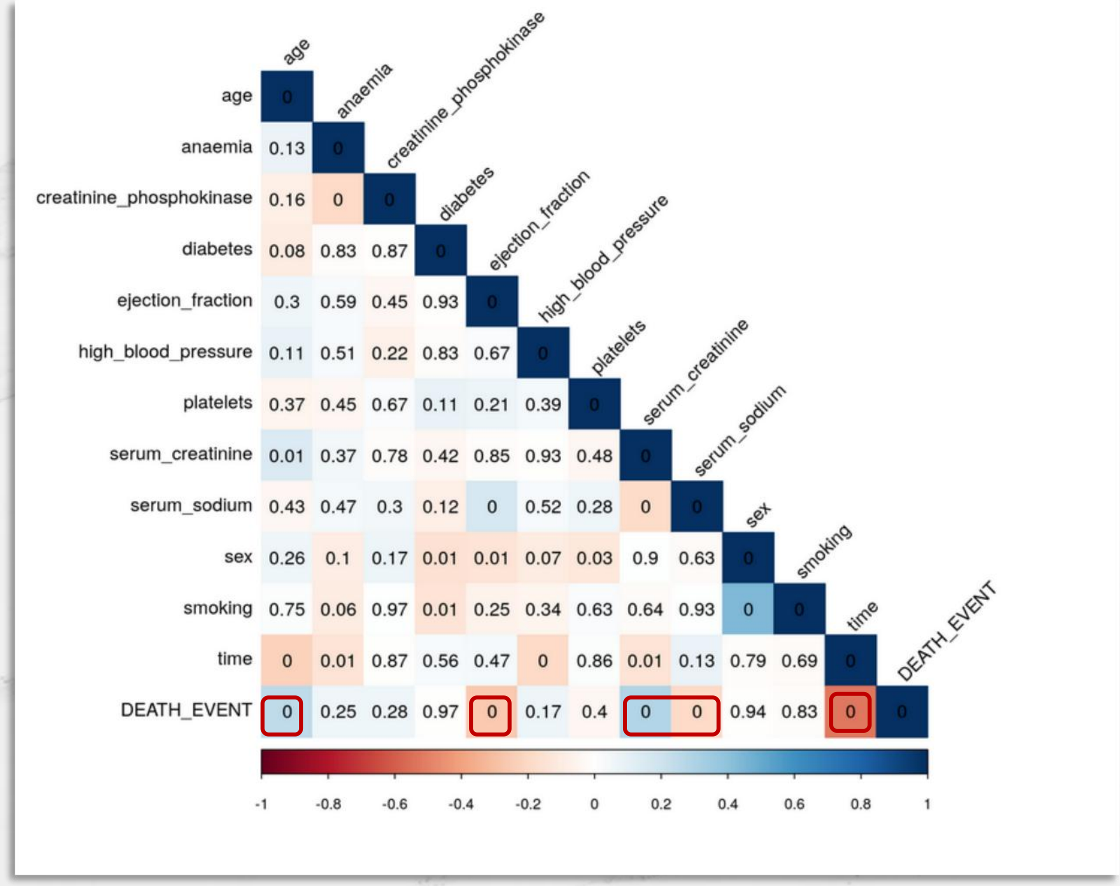
Baseline Characteristic Table

Clinical characteristics of the study population disaggregated by quartiles of **time**:

	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, 29	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	

Correlation Matrix

Results
--EDA



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

Creation of Training and Test Data

Training Data Set: 239 observations

Test Data Set: 60 observations

On the set of 299 observations and a 80:20 random split

Logistic Regression Model

```
> 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
```

```
                         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 ***
---
```

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 siginifcant predictors of heart failure.

Logistic Regression Model

```
> 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
```

Success rate of the logistic model : 85%

The model in predicting the response: performs better than tossing a coin

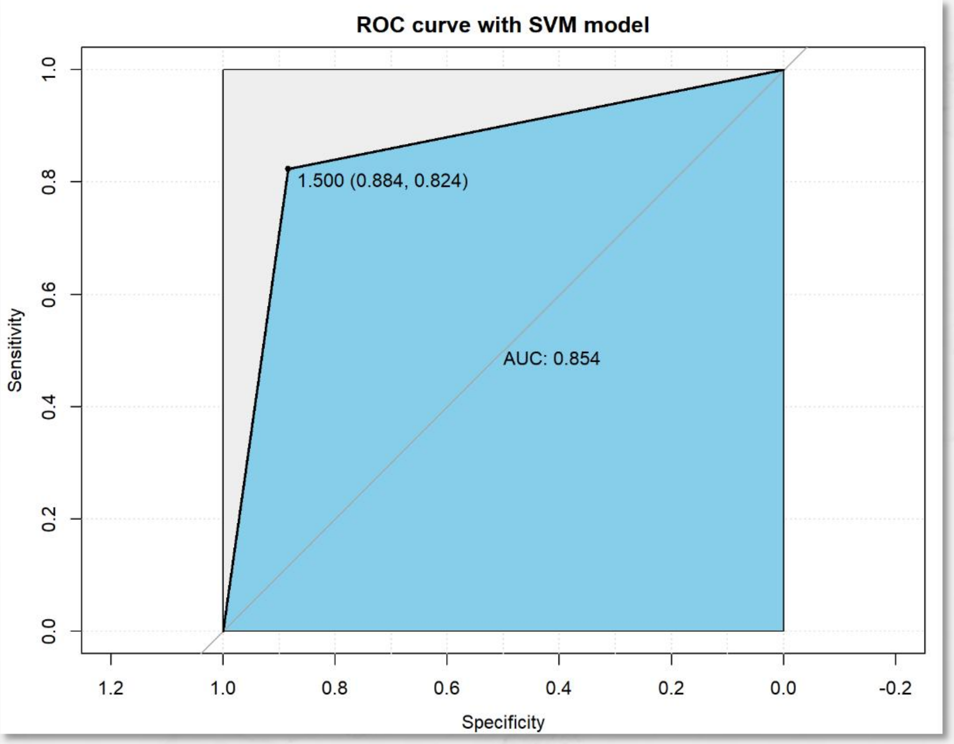
Support Vector Machine(SVM)

Results--Modelling

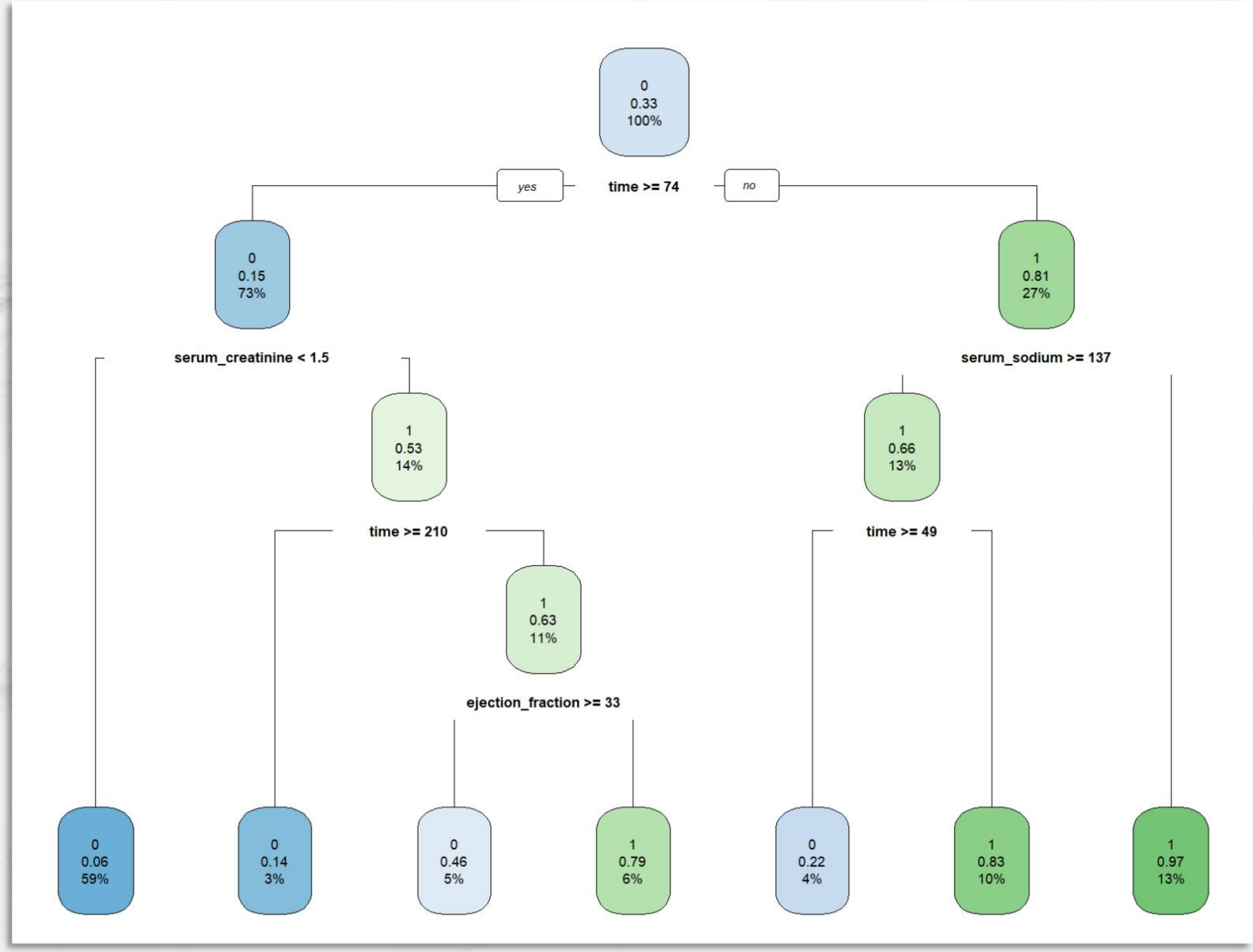
```
Training error : 0.158996
> svm_pred<-predict(svm_model,d[test, ])
> table(svm_pred,d[test, ]$DEATH_EVENT)

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
>
```

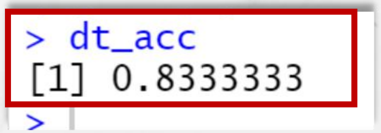
Success rate of the SVM model : 86.67%

ROC curve



Decision Tree





Success rate of the DT model : 83.33%

Random Forest

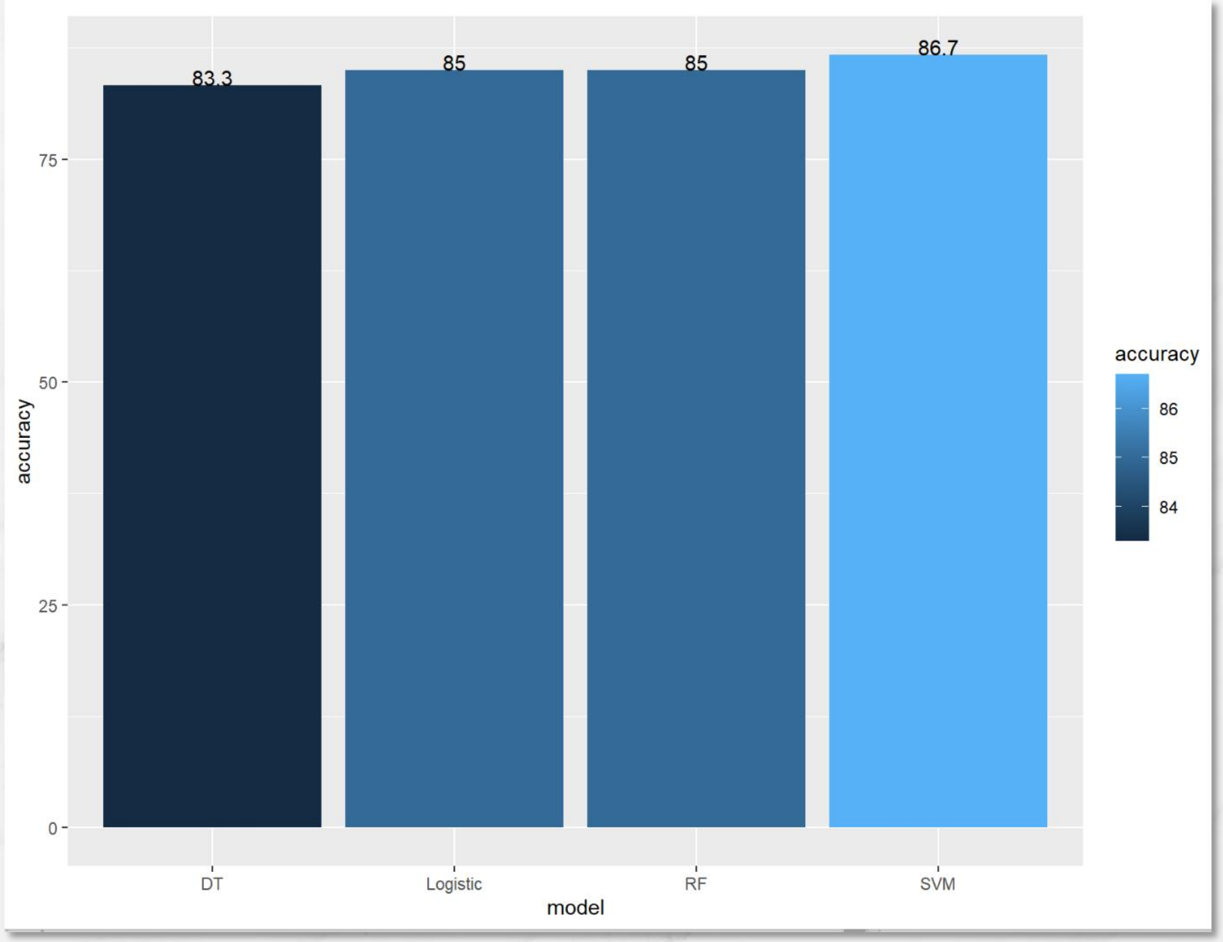
```
> confusionMatrix(rf_pred, d[train, ]$DEATH_EVENT)
Confusion Matrix and Statistics

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

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

Success rate of the RF model : 85%

Model Comparison

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



Survival Analysis

01

Definition

Survival Analysis is a branch of statistical modelling that is optimal for working with censored, time-to-event data[4].

Modelling Method

02

Kaplan-Meier estimator
Cox Proportional Hazard Model

Kaplan-Meier Estimator

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

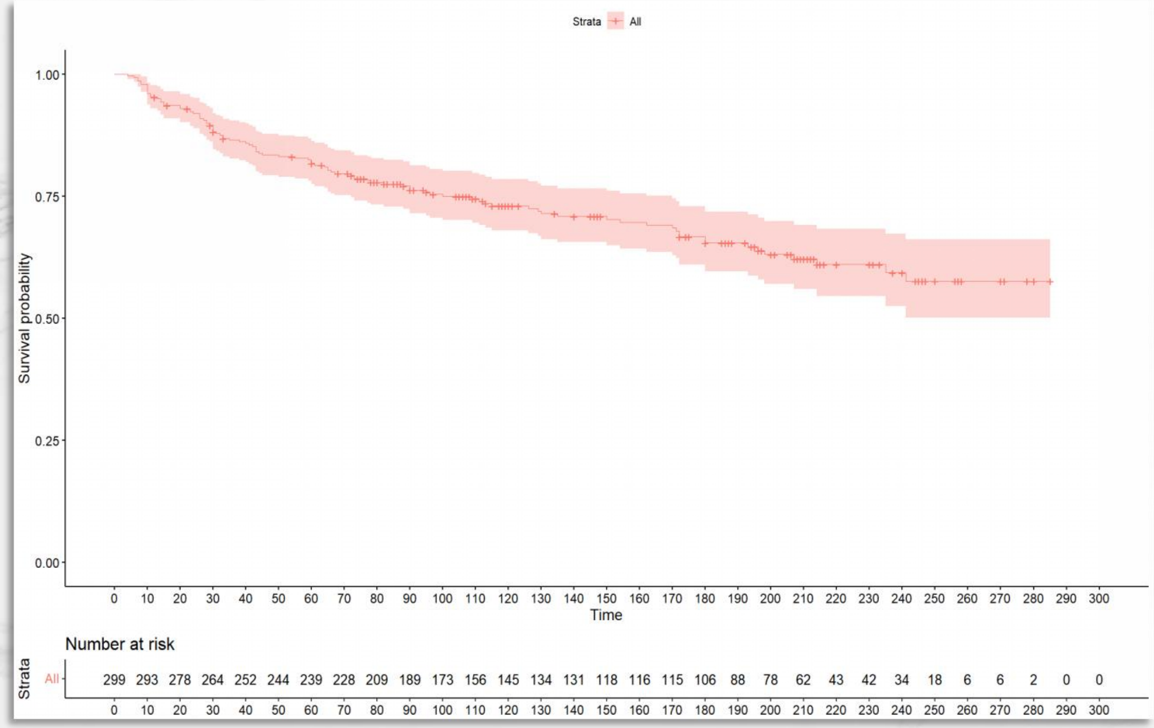
table: the cumulative survival probability

- a test statistic that gives us an approximation of the true survival function of a population [5]

- can be used for simple comparison of survival rates between groups

Kaplan-Meier Estimator

Results--Survival Analysis



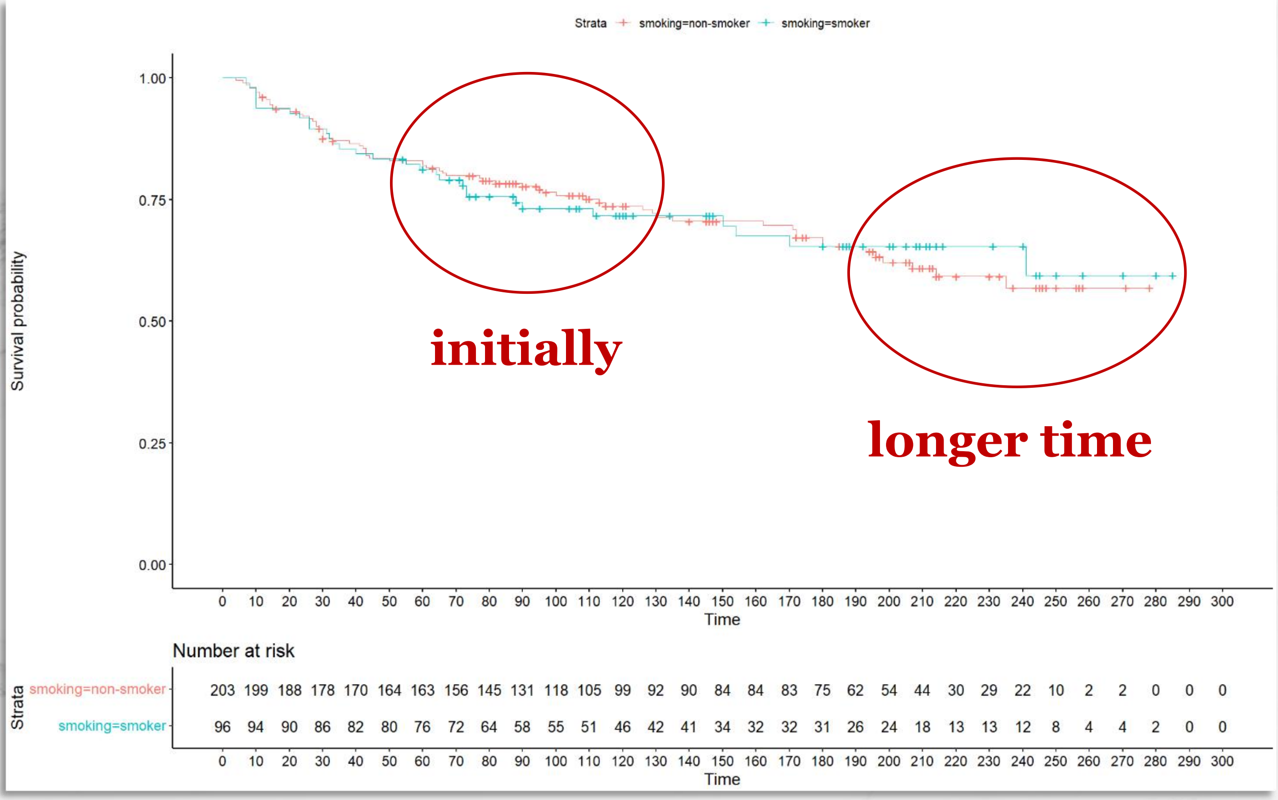
- "+" tick marks: a censoring event

- A Kaplan-Meier plot :

  1. approaches the true survival curve of the population
  2. analyze impact of categorical features on survival

Kaplan-Meier Estimator

Results-- Survival Analysis



non-smokers: a higher probability of survival initially but a lower survival for longer time horizons

smokers: a lower probability of survival initially but a higher survival for longer time horizons

Cox Proportional Hazard Model

Results--Survival Analysis

```
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
```

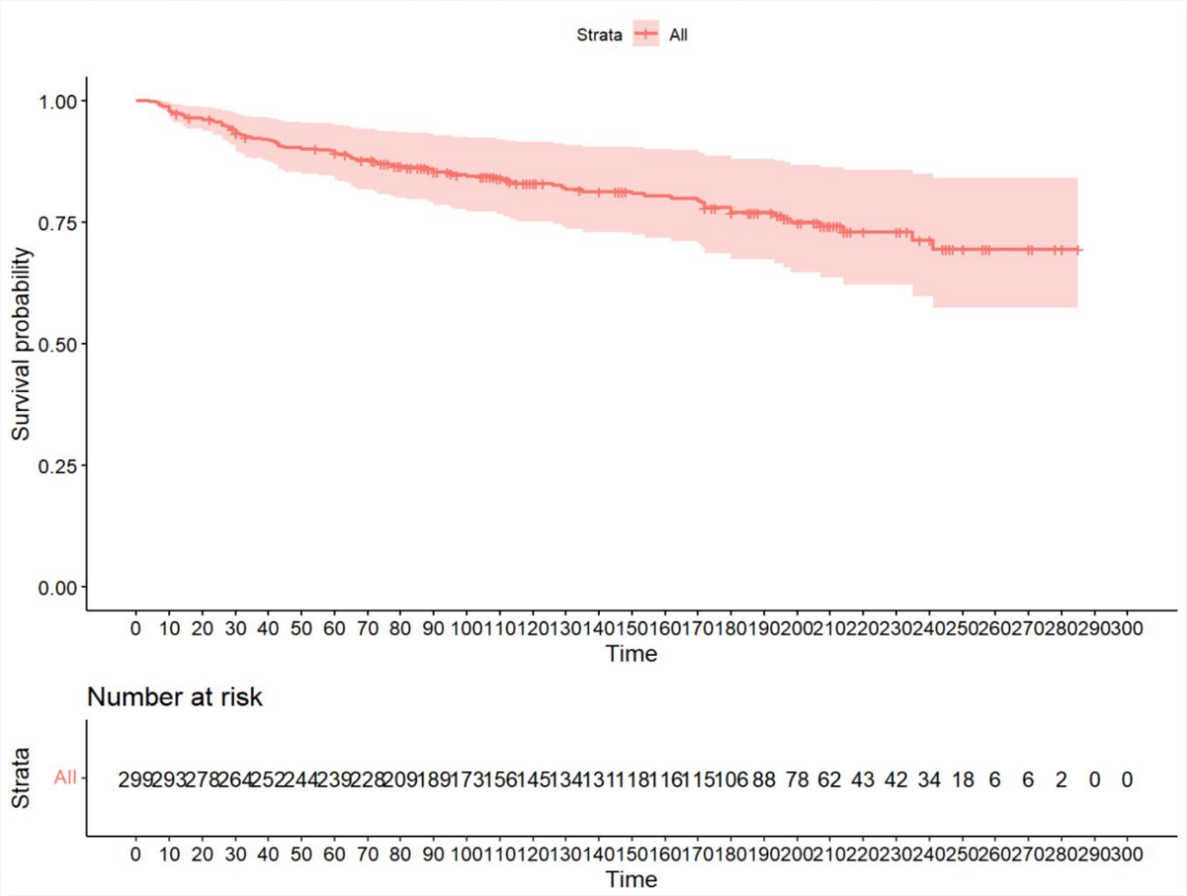
Definition:
a survival analysis model that assumes that the baseline hazard function of a population is multiplicatively influenced by the covariates[6].

Example: anaemial
Increase by 1 unit: hazard 48% higher(1.48-1)
Decrease by 1 unit: hazard 33% lower (1 - 0.67)

Cumulative Survival Probability

Results--Survival Analysis

Use Cox Proportional Hazard Model to plot the cummulative survival probability:



04

Discussion

Discussion

01 Significance

Our findings can be used by medical workers to identify which factors are most significant in addressing to maximize survival among cardiovascular disease patients.

02 Finding and Comparison

A quick scan of the available research reveals that smoking does raise the risk of heart problems, including the likelihood of a heart attack. However, since everyone in this data set has already experienced a heart attack, smoking status does not seem to have an impact on the result.

Discussion

03 Limitation

The project's main constraint, as already mentioned, was the project's small data set. More data would have made for a solid foundation for training and cross-validation, resulting in better and more in-depth learning. With enough data, it is likely that several of the tested algorithms would perform better.

04 Outlook

Simple as well as more complex techniques of various types were included in the selection of algorithms used to train the models. A more precise selection based on in-depth literature research can be made for future work. The prediction may be enhanced by using algorithms that are better suited for small data sets.

05

Conclusion

Conclusion

- **For EDA:**

From the correlation matrix, we can see Death Event is highly correlated with serum creatinine, age, serum sodium, ejection fraction.

- **For Machine Learning Classification and Prediction models：**

The Support Vector Machine(SVM) has the highest prediction power among the models, with the accuracy of 86.7%.

- **For the survival analysis:**

The covariate age, ejection fraction, high blood plessures had a more significant effect on survival time.

Reference

[1]fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[2]Kim H. C. (2021). Epidemiology of cardiovascular disease and its risk factors in Korea. Global health & medicine, 3(3), 134–141. https://doi.org/10.35772/ghm.2021.01008

[3]Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[4]Boyang C.(May 2022).Time Series Survival Analysis: Implementation in Python, Retrieved from https://medium.com/@boyangchen02/time-series-survival-analysis-implementation-in-python-f31c43b3099d

[5][6]UCLA: Statistical Consulting Group(August 22, 2021). Introduction to SAS. Retrieved from https://stats.oarc.ucla.edu/sas/modules/introduction-to-the-features-of-sas/.

Thanks for your listening!

Supervisor: Prof.Ma

Speaker：Yifan LI