심부전에 예측

12가지 임상적 특징들에 의한 사망 예측

과목명	기계학습 및 실습
제출일	2020.6.26
학과	통계학과
학번	2017580033
이름	길경주

1.서론

a. 연구 주제 소개

심혈관계 질환(Cardiovascular diseases)은 심장과 순환계에 영향을 주는 여러 질 환으로, 여기에는 관상동맥질환(Coronary artery disease), 고혈압(Hypertension), 심부전증(heart failure), 선천성심혈관결함(Congenital cardiovascular defects), 뇌혈관질환(Cerebrovascular disease)등이 포함되게 되며 심혈관 질환은 만성질환으로 나이를 먹으면서 상태가 더 악화되게 된다. 심혈관 질환은 미국에서 주요 사망 원인의 하나로 매 33초인 하루 2,600여 명의 사망을 보이게 된다. 암 사망률에 비해 거의 두 배의 수준이며, 약 60,8 백면망이 심혈관 질환으로 고통받고 있다. 이러한 심혈관계 질환 중 하나인 심부전(heart disease)은 심장의 기능 저하로 신체에 혈액을 제대로 공급하지 못해서 생기는 질환을 말한다. 각종 심장질환으로 인해 심장의 펌프기능이 떨어지면 심장에 들어오는 혈액을 퍼낼 수 없으므로 심장이 커지고 혈액순환이 원활하지 못해 체액이 연약한 폐조직으로 스며들게 되어 폐부종이 발생하게 되는 심부전에 이르게 되고, 심부전 상태가 되면 움직일 때 숨찬 증상이 가장 먼저 나타난다.

매우 다양한 원인에 의해 심부전이 초래할 수 있는데, 심장 혈관(관상동맥) 질환(예, 심근경색 등)이 2/3 정도로 가장 흔한 원인이고, 심장 근육(심근) 질환(예, 원인 미상이거나 유전적 원인인 심근병증, 바이러스 감염 등), 고혈압, 판막 질환 등이 주요 원인이다. 그 밖에도 장기간의 빠른 맥박(빈맥), 지속적인 과도한 음주, 극심한 스트레스 등도 원인을 제거하면 좋아지는 가역적인 심부전의 원인이 될 수 있고, 드물지만 출산 전후에 원인 미상의 심부전이 발생하는 산후(또는 임신성) 심근(병)증도 있다. 항암제 중 일부도 누적되는 사용 용량에 비례하여 심부전을 발생시키는 경우가 있다.

b.데이터 획득 방법

데이터는 Davide Chicco, Giuseppe Jurman에 의해 만들어졌으며, BMC Medical Informatics and Decision Making이라는 의학/건간 관련 오픈 저널에 실렸다.

이 데이터는 캐글에서 볼 수 있으며 출처는 다음과 같다.

https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

c. 데이터 설명

심부전으로 인한 사망에 영향을 주는 특질에 대한 데이터로 변수는 다음과 같고, 관측치는 총 299개이다.

Age : 나이

Anaemia : 적혈구 또는 헤모글로빈의 감소

Creatinine_phosphokinase : 혈액에서 CPK 효모의 수준 (mcg/L)

Diabetes : 환자들의 당뇨병 여부 (1: 당뇨병 있음, 0 : 당뇨병 없음)

Ejection_fraction : 수축 시 심장에서 빠져나가는 혈액의 퍼센티지 (percentage)

High_blood_pressure : 환자의 고혈압 여부

Platelets : 혈엑에서 혈소판 (kiloplatelets/ml)

Serum_creatinine : 혈액에서 혈청크레아티닌의 수준 (mg/dl)

Serum_sodium : 혈액에서 혈청 나트륨의 수준 (meq/l)

Sex : 성별

d. 분석 목표

심혈관계 질환들은(Cardiovascular diseases)는 전세계적으로 사망의 가장 첫번째 원인이며, 매년 전세계 모든 사망의 약 30%를 차지하는 약 17.9 백만의 목숨을 앗아가고 있다. 심부전(heart disease)는 심혈관계 질환들에 의해 유발되는 공통전인 질환이며, 이 데이터 셋은 심부전에 의한 사망을 예측하기 위해 사용될 수 있는 12가지 특질들을 포함하고 있다.

대부분의 심혈관계 질환은 흡연, 건강하지 않은 식이요법과 비만, 그리고 신체적 비활동과 과도한 음주와 같은 행동적인 위험 요인들을 다룸으로써 예방할 수 있다.

심혈관계 질환을 가진 사람들 또는 높은 심혈관계 질환을 가진 사람 누구든지 이 데이터 분석 결과를 통해 어떤 요인이 심혈관계 질환에 많은 위험성을 가지는지 알고 예방하는데 큰 도움이 될 수 있다.

2. 본론

a. 분석 방법 소개

분석에서 사용되는 데이터의 반응변수(DEATH_EVENT)가 binary로서 1과 0의 값을 갖는다. 따라서 0과 1의 값을 모두 가질 수 있는 변수에 대한 분석이므로 로지스틱 회귀분석이 적절하다. 로지스틱 회귀모델의 정확도를 평가하기 위해 데이터의 일부를 이용하여 모델을 적합하고, 그 다음에 모델이 나머지 데이터를 얼마나 잘 예측하는지 알 수 있다. 즉, 데이터를 training data와 test data로 나누어 분석한 후 분석 결과에 대해 혼동행렬을 이용하여 검정오차율을 알면 그 모델의 정확도를 알 수 있다. 마찬가지로 훈련 데이터 셋과 검정 데이터 셋으로 나누어 서포트 벡터 머신을 이용해 모형을 적합한다. 마찬가지로 혼동행렬을 통해 검정오차율에 대해서 알아보고, 검정셋 예측에 대한 ROC 곡선을 나타낸다. 마지막으로 랜덤포레스트를 통한 모델적합을 실시한다. 세 모형의 정확도를 비교해 더 높은 것을 최종 모형으로 선택한다.

b. 데이터 분석

i . eda

```
'data.frame':   299 obs. of  13 variables:
 $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
 $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
 $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
 $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
 $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
 $ platelets               : num  265000 263358 162000 210000 327000 ...
 $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
 $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
 $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
 $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
 $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
 $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
```

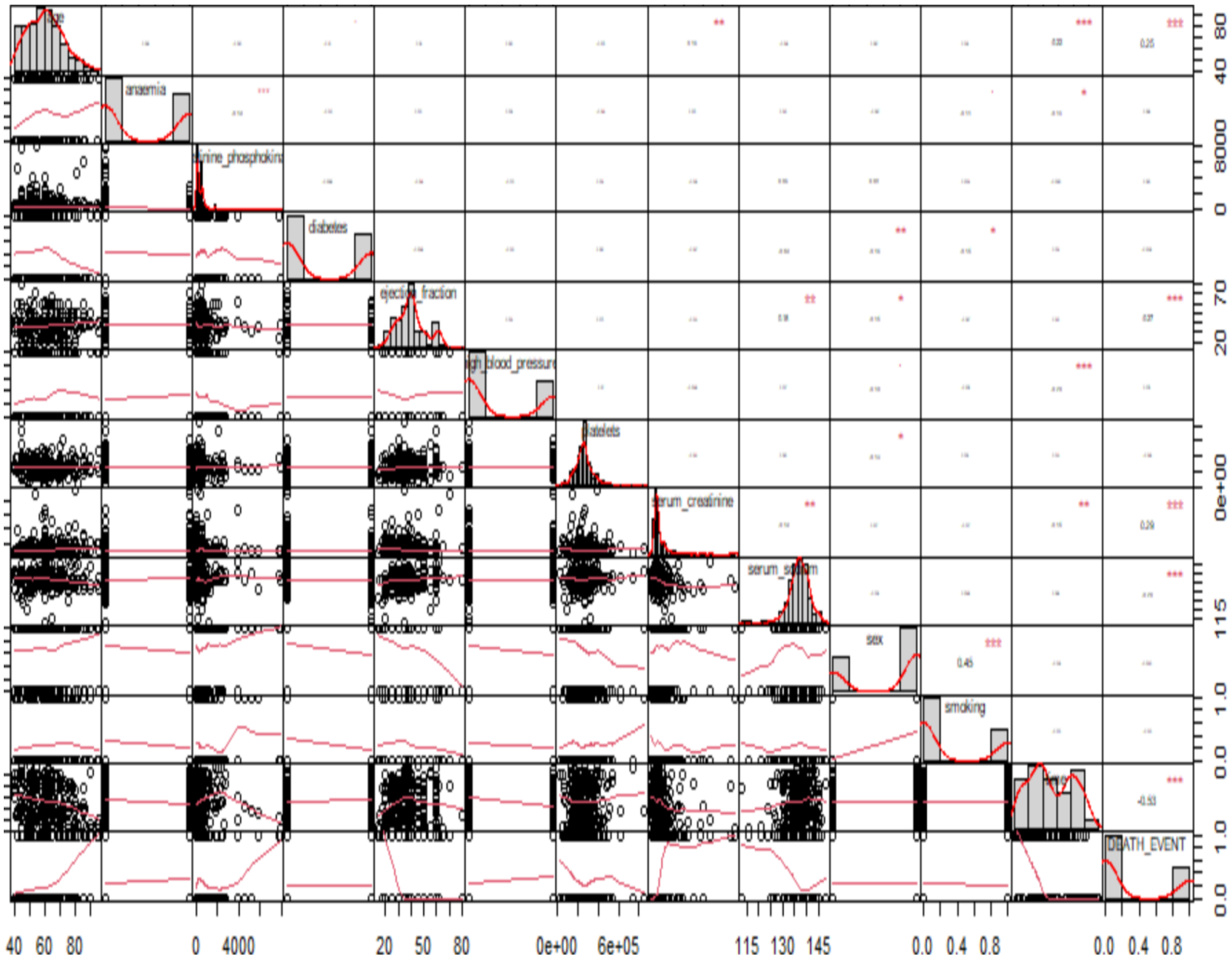
데이터 탐색 결과 총 12개의 설명변수, 1개의 binary 반응변수가 존재하며 총 관측값은 299개임을 알 수 있다.

```
> table(is.na(df))

FALSE
 3887
```

결측치가 있는 관측지가 존재하지 않아 모든 관측치를 사용할 수 있다.

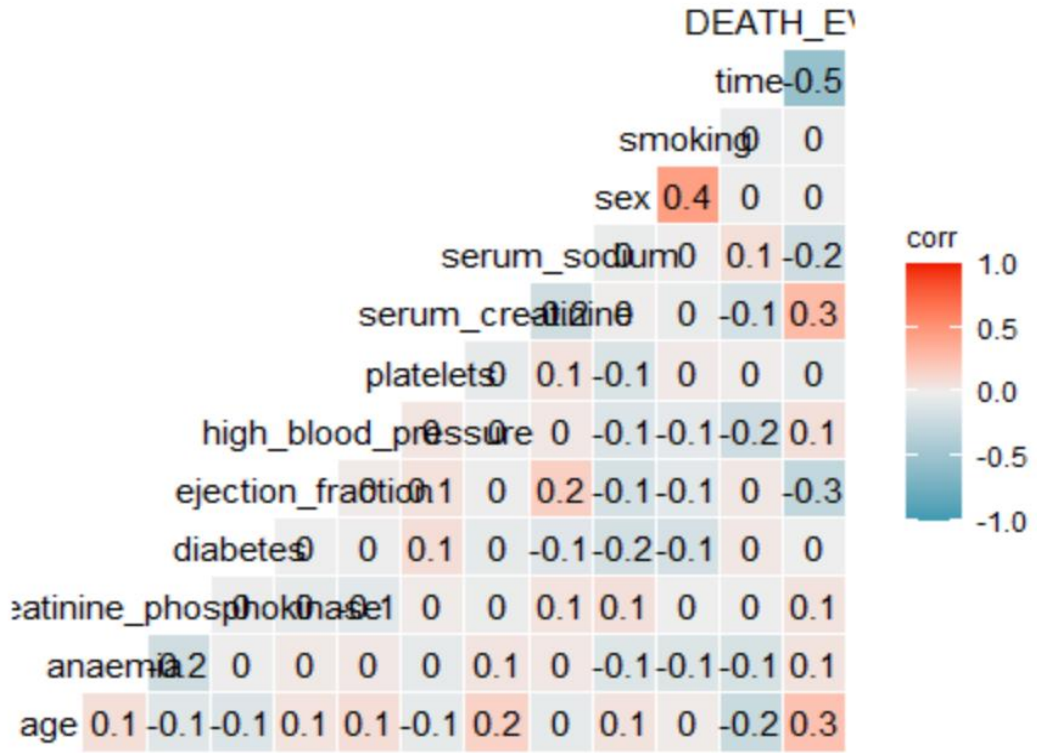
```
'data.frame':   299 obs. of  13 variables:
 $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anaemia                 : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...
 $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
 $ diabetes                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
 $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
 $ high_blood_pressure     : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
 $ platelets               : num  265000 263358 162000 210000 327000 ...
 $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
 $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
 $ sex                     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...
 $ smoking                 : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
 $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
 $ DEATH_EVENT             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Binary인 변수가 모두 integer로 표기되어 있으므로, factor형으로 바꿔준다.



변수들 간 코릴레이션 차트를 그려 변수간의 상관관계를 알아볼 수 있다. 더불어 각 변수의 히스토그램도 함께 나타냄으로써 변수들의 분포와 연속형 변수들의 정규성 여부를

알아볼 수 있다.



변수들 간 다중공산성을 조사한 결과 뚜렷한 상관관계를 가지는 변수들은 없는 것으로 드러났다.

```
      age            creatinine_phosphokinase ejection_fraction
 Min.   :-1.75151   Min.   :-0.575952        Min.   :-2.034976
 1st Qu.:-0.82674   1st Qu.:-0.479589        1st Qu.:-0.683035
 Median :-0.07011   Median :-0.342001        Median :-0.007065
 Mean   : 0.00000   Mean   : 0.000000        Mean   : 0.000000
 3rd Qu.: 0.77060   3rd Qu.: 0.000165        3rd Qu.: 0.584409
 Max.   : 2.87235   Max.   : 7.502063        Max.   : 3.541779
   platelets         serum_creatinine     serum_sodium
 Min.   :-2.43607   Min.   :-0.864061   Min.   :-5.35423
 1st Qu.:-0.52000   1st Qu.:-0.477404   1st Qu.:-0.59500
 Median :-0.01388   Median :-0.284076   Median : 0.08489
 Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.00000
 3rd Qu.: 0.41043   3rd Qu.: 0.005916   3rd Qu.: 0.76478
 Max.   : 5.99812   Max.   : 7.739045   Max.   : 2.57782
      time          anaemia diabetes high_blood_pressure sex
 Min.   :-1.6268   0:170   0:174    0:194                0:105
 1st Qu.:-0.7378   1:129   1:125    1:105                1:194
 Median :-0.1966
 Mean   : 0.0000
 3rd Qu.: 0.9372
 Max.   : 1.9937
```

변수들 간에 단위가 모두 다름으로 이를 스케일링 시켜준다.

ii. 로지스틱 회귀분석

```
> train<-x[1:200, ]
> test<-x[201:299, ]
```

먼저, 변수를 훈련 데이터 셋과 검정 데이터 셋으로 나누어 준다. 훈련 데이터 셋과 검정 데이터 셋의 비율은 약 70: 30 정도로 설정했다.

```
> glm.fit <- glm( DEATH_EVENT~ ., data =train, family = "binomial")
> summary(glm.fit)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1941  -0.7489  -0.2701   0.7142   2.7503

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.02069    0.50282  -2.030  0.04236 *
age                       0.67256    0.20905   3.217  0.00129 **
creatinine_phosphokinase  0.16081    0.17179   0.936  0.34921
ejection_fraction        -0.90973    0.21129  -4.306 1.67e-05 ***
platelets                -0.05410    0.19491  -0.278  0.78133
serum_creatinine          0.58173    0.26206   2.220  0.02643 *
serum_sodium             -0.20521    0.18739  -1.095  0.27349
time                     -1.75999    0.34180  -5.149 2.62e-07 ***
anaemia1                 -0.19947    0.39243  -0.508  0.61125
diabetes1                 0.20366    0.37803   0.539  0.59007
high_blood_pressure1     -0.13453    0.39572  -0.340  0.73389
sex1                     -0.58726    0.44732  -1.313  0.18923
smoking1                  0.04089    0.45056   0.091  0.92769
```

훈련데이터 셋에 로지스틱 회귀분석을 실시한 결과 유의수준 0.05하에서 유의하지 않은 변수가 너무 많이 존재함을 알 수 있다. 따라서, p-value가 가장 높은 값을 순차적으로 제거해줌으로써 모든 변수가 유의할 수 있도록 변수선택을 실시한다.

```
> glm.fit8<- glm(DEATH_EVENT~age+ejection_fraction+
+                  serum_creatinine+time
+                ,data =train, family = "binomial")
>
> summary(glm.fit8) #최종

Call:
glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatini

    time, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1715  -0.7283  -0.3044   0.7669   2.9222

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.4004     0.2795  -5.010 5.44e-07 ***
age                 0.6062     0.1966   3.084  0.00204 **
ejection_fraction  -0.8668     0.2002  -4.329 1.50e-05 ***
serum_creatinine    0.6513     0.2506   2.599  0.00935 **
time               -1.7303     0.3333  -5.192 2.08e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 274.83  on 199  degrees of freedom
Residual deviance: 186.85  on 195  degrees of freedom
AIC: 196.85

Number of Fisher Scoring iterations: 5
```

8번째 변수선택 결과 모든 설명변수들이 유의수준 0.05하에서 유의한 것을 알 수 있다.

```
> LR_statistic=274.83-186.85
> pchisq(LR_statistic,df=(199-195),lower.tail = F) #유의미
[1] 3.535914e-18
```

최종적으로 선택한 모형에 대한 LR테스트를 실시한 결과 유의미한 결과가 나온 것을 알 수 있다. 따라서 로지스틱 회귀분석에서는 반응변수에 유의미한 영향을 미치는 네 개의 설명변수 age, ejection_fraction, serum_creatinine, time를 가지는 로지스틱 회귀모델을 채택할 수 있다.

```
> table(glm.pred,DEATH_EVENT)
         DEATH_EVENT
glm.pred   0   1
       0 199  94
       1   4   2
> mean(glm.pred==DEATH_EVENT)
[1] 0.6722408
```

이 모델의 정확도를 알기 위해 검정 오차율을 계산해보면 검정 오차율은 약 0.33 정도인 것을 알 수 있다.

iii. Svm

```
> tune.out=tune(svm,DEATH_EVENT~., data=train, kernel="linear",
+               ranges=list(cost=c(0.001,0.01,0.1,1,5,10,100)))
> summary(tune.out)  #cost=0.1

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 cost
  0.1

- best performance: 0.2

- Detailed performance results:
   cost error dispersion
1 1e-03 0.445 0.14424131
2 1e-02 0.250 0.10000000
3 1e-01 0.200 0.07453560
4 1e+00 0.205 0.06433420
5 5e+00 0.205 0.06433420
6 1e+01 0.210 0.06582806
7 1e+02 0.210 0.06582806
```

tune함수를 이용해 모델 셋에 대해 10-fold 교차검증을 수행한다. 그 결과 cost가 0.1인 경우에 교차검증 오차율이 가장 낮은 것을 알 수 있다.

```
> svm.fit=svm(DEATH_EVENT~.,data=train, kernel="linear",cost= 0.1,scale=
TRUE)
> summary(svm.fit)

Call:
svm(formula = DEATH_EVENT ~ ., data = train, kernel = "linear",
    cost = 0.1, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  116

 ( 58 58 )


Number of Classes:  2

Levels:
 0 1
```

Linear 커널을 이용해 분석한 결과 총 116개의 서포트 벡터들이 사용되었다.

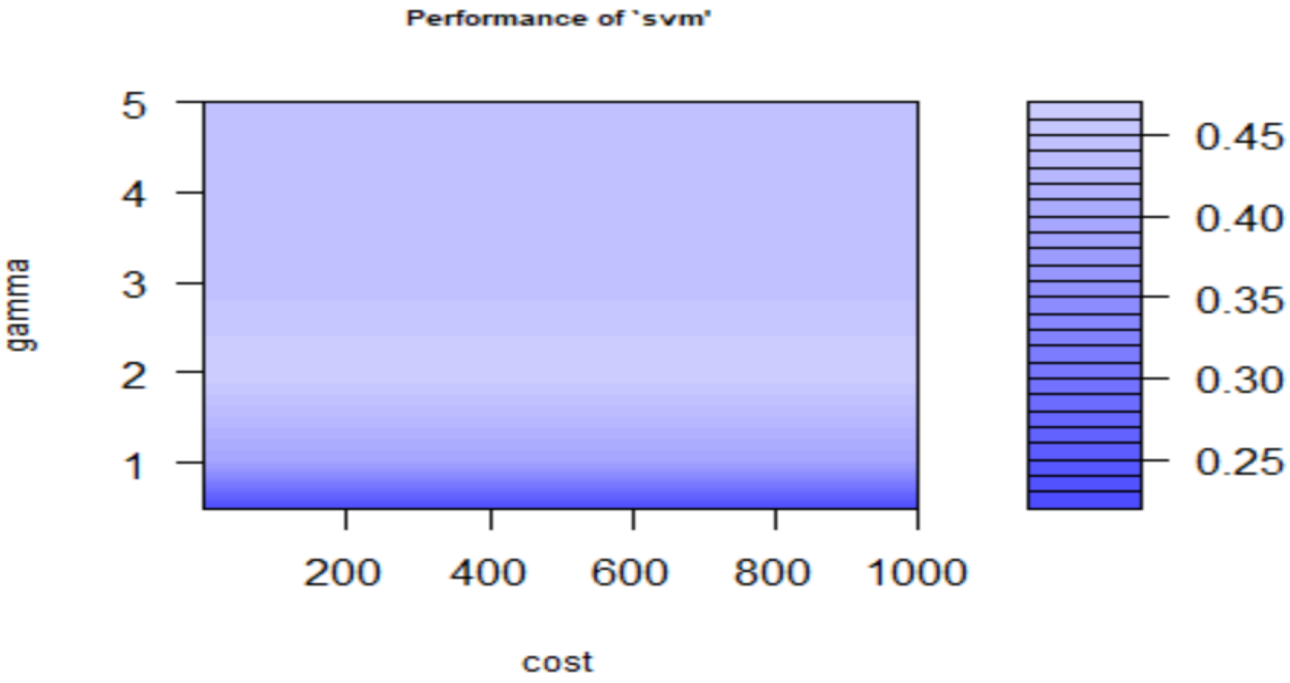
```
> confusionMatrix(data=ypred, reference = test$DEATH_EVENT, positiv
e="1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 92  7
         1  0  0
```

```
               Accuracy : 0.9293
                 95% CI : (0.8597, 0.9711)
    No Information Rate : 0.9293
    P-Value [Acc > NIR] : 0.59878

                  Kappa : 0

 Mcnemar's Test P-Value : 0.02334

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.92929
             Prevalence : 0.07071
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

이 혼동행렬을 보면 정확도는 92.93%로 매우 높지만, 민감도는 0%, 특이도는 100%인 다소 극단적인 모델임을 알 수 있다.

```
> tune.out=tune(svm,DEATH_EVENT~.,data=train, kernel="radial",
+               range=list(cost=c(0.001,0.01,0.1,1,5,10,100,1000),
+                          gamma=c(0.5,1,2,3,4,5)))
> summary(tune.out)

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
    1   0.5

- best performance: 0.225
```

tune함수를 이용한 10-fold 교차검증을 수행하여 방사커널의 svm에 대한 최상의 감마와 cost를 선택한 결과 각각 cost=1, gamma=0.5임을 알 수 있다. 이를 바탕으로 svm을 적합해본다.



파란색이 가장 짙은 영역은 가장 최적의 성능을 나타내는 파라미터 값을 보여준다.

```
> svm.fit2=svm(DEATH_EVENT~., data=train ,kernal="radical",gamma=0.5, co
st=1,scale=TRUE)
> summary(svm.fit2)

Call:
svm(formula = DEATH_EVENT ~ ., data = train, kernal = "radical",
    gamma = 0.5, cost = 1, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  184

 ( 88 96 )


Number of Classes:  2

Levels:
 0 1
```

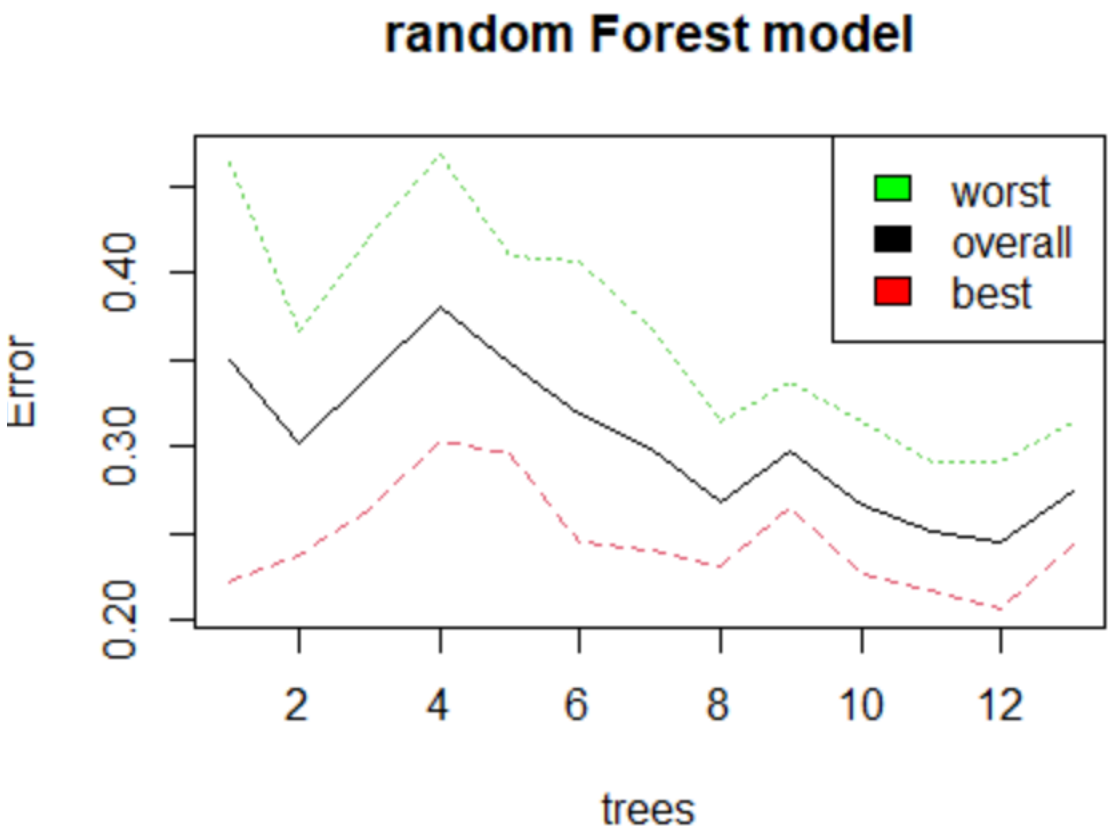
방사형 커널을 이용해 분석한 결과 총 184개의 서포트 벡터들이 사용되었다.

```
> confusionMatrix(data=ypred2, reference = test$DEATH_EVENT, positive
="1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 27  1
         1 65  6

               Accuracy : 0.3333
                 95% CI : (0.2418, 0.4352)
    No Information Rate : 0.9293
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0288

 Mcnemar's Test P-Value : 8.851e-15

            Sensitivity : 0.85714
            Specificity : 0.29348
         Pos Pred Value : 0.08451
         Neg Pred Value : 0.96429
             Prevalence : 0.07071
         Detection Rate : 0.06061
   Detection Prevalence : 0.71717
      Balanced Accuracy : 0.57531

       'Positive' Class : 1
```

이 혼동행렬을 보면 정확도는 33.33%로 linear커널을 이용한 svm보다 정확도는 현저히 낮지만, 민감도와 예측도는 각각 85%와 29%로 나아졌음을 알 수 있다.

iii. 랜덤 포레스트 모델

```
> rf_model <- randomForest(DEATH_EVENT ~ ., data=train, ntree=13, mtr
y=2)
> rf_model

Call:
 randomForest(formula = DEATH_EVENT ~ ., data = train, ntree = 13,
   mtry = 2)
               Type of random forest: classification
                     Number of trees: 13
No. of variables tried at each split: 2

        OOB estimate of  error rate: 27.5%
Confusion matrix:
   0  1 class.error
0 84 27   0.2432432
1 28 61   0.3146067
```




```
> importance(rf_model)
                          MeanDecreaseGini
age                           11.975076
creatinine_phosphokinase       7.254793
ejection_fraction             11.515320
platelets                      8.343327
serum_creatinine              11.330084
serum_sodium                  10.816307
time                          25.649366
anaemia                        1.741747
diabetes                       1.079897
high_blood_pressure            1.666120
sex                            2.254929
smoking                        1.737150
```

변수의 중요도를 봤을 때 상위 중요도를 갖는 두 변수는 age와 creatinine_phosphokinase임을 알 수 있다.

```
> confusionMatrix(test$rf_pred, test$DEATH_EVENT)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 82  4
         1 10  3

               Accuracy : 0.8586
                 95% CI : (0.7741, 0.9205)
    No Information Rate : 0.9293
    P-Value [Acc > NIR] : 0.9959

                  Kappa : 0.2291

 Mcnemar's Test P-Value : 0.1814

            Sensitivity : 0.8913
            Specificity : 0.4286
         Pos Pred Value : 0.9535
         Neg Pred Value : 0.2308
             Prevalence : 0.9293
         Detection Rate : 0.8283
   Detection Prevalence : 0.8687
      Balanced Accuracy : 0.6599

       'Positive' Class : 0
```

혼동행렬 결과 정확도는 85%, 민감도는 89%, 특이도는 42% 임을 알 수 있다.

3. 결론

최적 모델 선택

각각 적합한 로지스틱 회귀모델, svm, 랜덤포레스트 결과 모델이 적합하며 정확도가 가장 높은 모델은 랜덤포레스트였다. 그러나 그와 비슷한 정확도와 변수선택을 가지는 로지스틱 회귀분석이 해석이 더 쉬움으로 최종적으로는 우리는 설명변수 age, ejection_fraction, serum_creatinine, time를 가지는 로지스틱 회귀모델을 채택할 수 있다.

따라서 심부전에 의한 사망에 가장 많이 영향을 미치는 요인들은 age, ejection_fraction, serum_creatinine, time로 다음과 같은 요인들을 관리함으로써 심부전에 의한 사망률을 낮출 수 있을 것으로 보인다.

4. 출처

https://terms.naver.com/entry.nhn?docId=926912&cid=51007&categoryId=51007

캐글