BIG DATA ANALYTICS FOR BUSINESS

PRESENTED BY GROUP 3
ABATE GAIA ,
CASADIDIO LETIZIA,
JOSEPH KHAGEMBA MOIRANGTHEM ,
SAMBHAJI KADAMRAJVAIBHAV

OUR PROJECT WORKS

1. Regression
2. Classification

1

REGRESSION WORK PROJECT

REGRESSION

Problem Statement

01. CONTEXT

Analysis of student achievement in secondary education at two Portuguese schools.
The dataset includes a variety of attributes related to student performance, such as grades, demographic information, social factors, and school-related features.

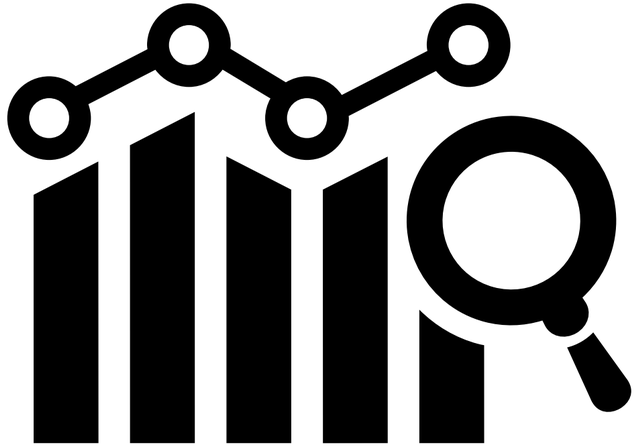
We focus on student performance in Mathematics.

02. PROBLEM

Which are the **variables that contributed to the dynamics of student performance and the various factors that contribute to academic success in secondary education?**
We want to investigate the **impact of alcohol consumption on academic performance among students.**

REGRESSION

Problem Statement



03. DATA

395 Observations
31 Starting Predictors

04. GOAL

Find the best model to know the alcohol effects on studies.
This study aims to determine if there is a significant correlation between alcohol use and academic outcomes such as grades and study habits.

REGRESSION

Main Steps

1

EXPLORATORY DATA ANALYSIS ON THE DEPENDENT VARIABLE (G3)

2

STATISTICAL SIGNIFICANCE AND MULTICOLLINEARITY

3

MODEL SELECTION

4

REGRESSION TREES

5

FINAL CONCLUSIONS

1 - Exploratory Data Analysis on the Dependent Variable G3

Data Pre-Proceeding:

- We have removed the missing data (N/A) from the original data, which was extracted from Kaggle.

- We have removed some columns (variables) not needed for our analysis, in specific ID column.

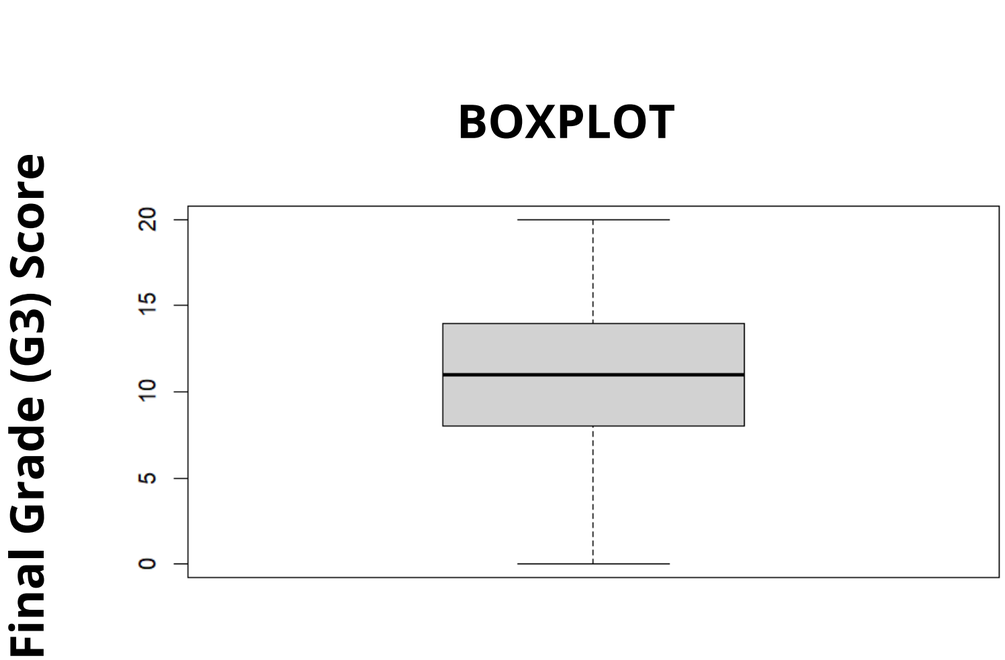
- We have done some data cleaning.
- G3 is the response variable and we have checked for outliers in the this variable

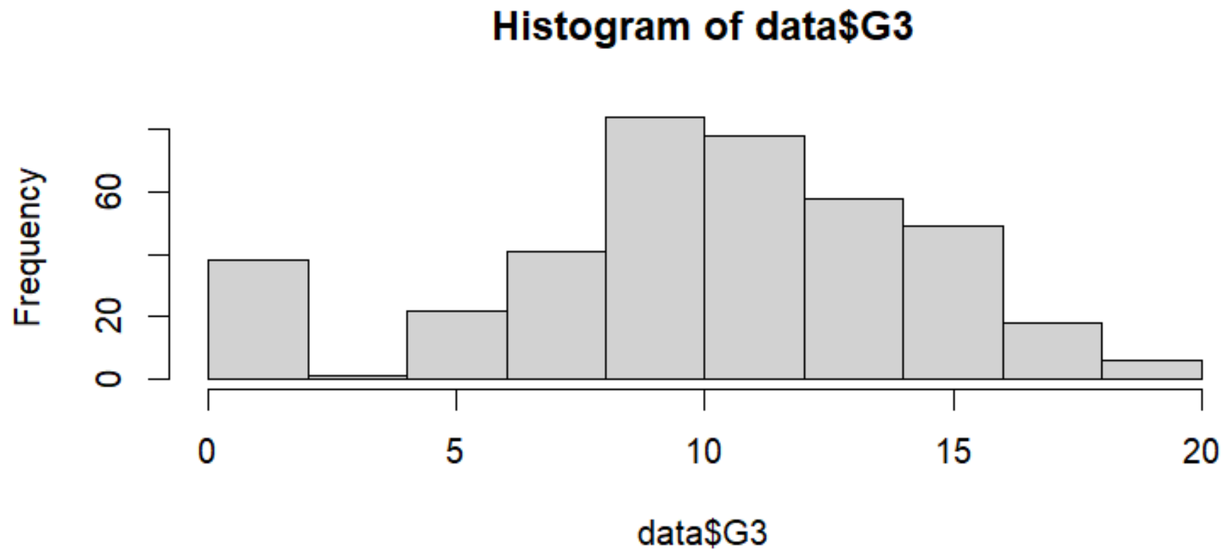
```
> summary(data$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
```

1 - Exploratory Data Analysis on the the Dependent Variable G3

Data Pre-Proceeding:

- We do not have **any outliers**





1 - Exploratory Data Analysis

Descriptive Statistics

```
> summary(data)
      sex              age          address            famsize            Pstatus
 Min.   :0.0000   Min.   :15.0   Length:395         Length:395         Length:395
 1st Qu.:0.0000   1st Qu.:16.0   Class :character   Class :character   Class :character
 Median :0.0000   Median :17.0   Mode  :character   Mode  :character   Mode  :character
 Mean   :0.4734   Mean   :16.7
 3rd Qu.:1.0000   3rd Qu.:18.0
 Max.   :1.0000   Max.   :22.0
      Medu            Fedu            Mjob              Fjob              reason
 Min.   :0.000   Min.   :0.000   Length:395         Length:395         Length:395
 1st Qu.:2.000   1st Qu.:2.000   Class :character   Class :character   Class :character
 Median :3.000   Median :2.000   Mode  :character   Mode  :character   Mode  :character
 Mean   :2.749   Mean   :2.522
 3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :4.000   Max.   :4.000
   guardian          traveltime      studytime        failures       schoolsup          famsup
 Length:395         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Median :0.0000   Median :1.0000
                    Mean   :1.448   Mean   :2.035   Mean   :0.3342   Mean   :0.1291   Mean   :0.6127
                    3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
                    Max.   :4.000   Max.   :4.000   Max.   :3.0000   Max.   :1.0000   Max.   :1.0000
     paid          activities          nursery            higher            internet
 Min.   :0.0000   Length:395         Length:395         Length:395         Min.   :0.0000
 1st Qu.:0.0000   Class :character   Class :character   Class :character   1st Qu.:1.0000
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Median :1.0000
 Mean   :0.4582                                                           Mean   :0.8329
 3rd Qu.:1.0000                                                           3rd Qu.:1.0000
 Max.   :1.0000                                                           Max.   :1.0000
   romantic            famrel          freetime         goout            Dalc            Walc
 Length:395         Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 Class :character   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Mode  :character   Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000
                    Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
                    3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
                    Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
     health         absences            G3
 Min.   :1.000   Min.   : 0.000   Min.   : 0.00
 1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00
 Median :4.000   Median : 4.000   Median :11.00
 Mean   :3.554   Mean   : 5.709   Mean   :10.42
 3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:14.00
 Max.   :5.000   Max.   :75.000   Max.   :20.00
```

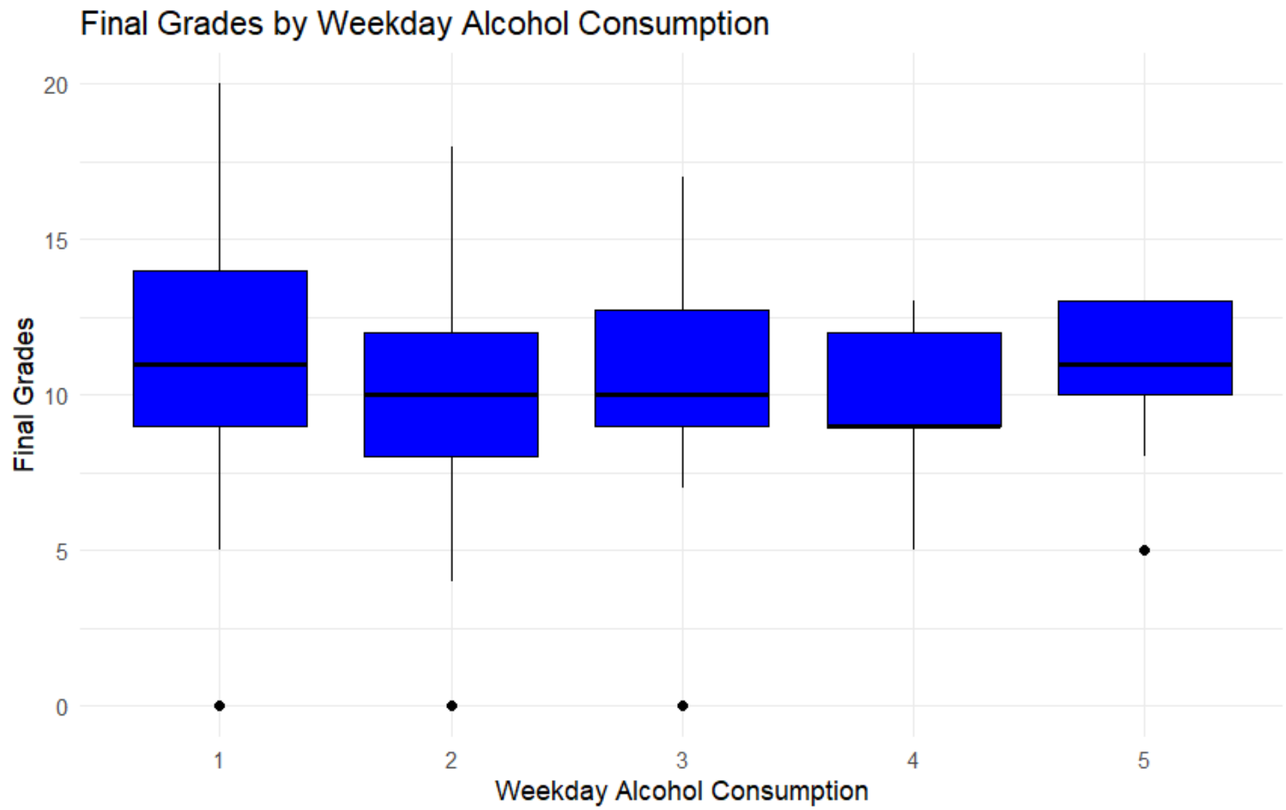
1 - Exploratory Data Analysis

Descriptive Statistics

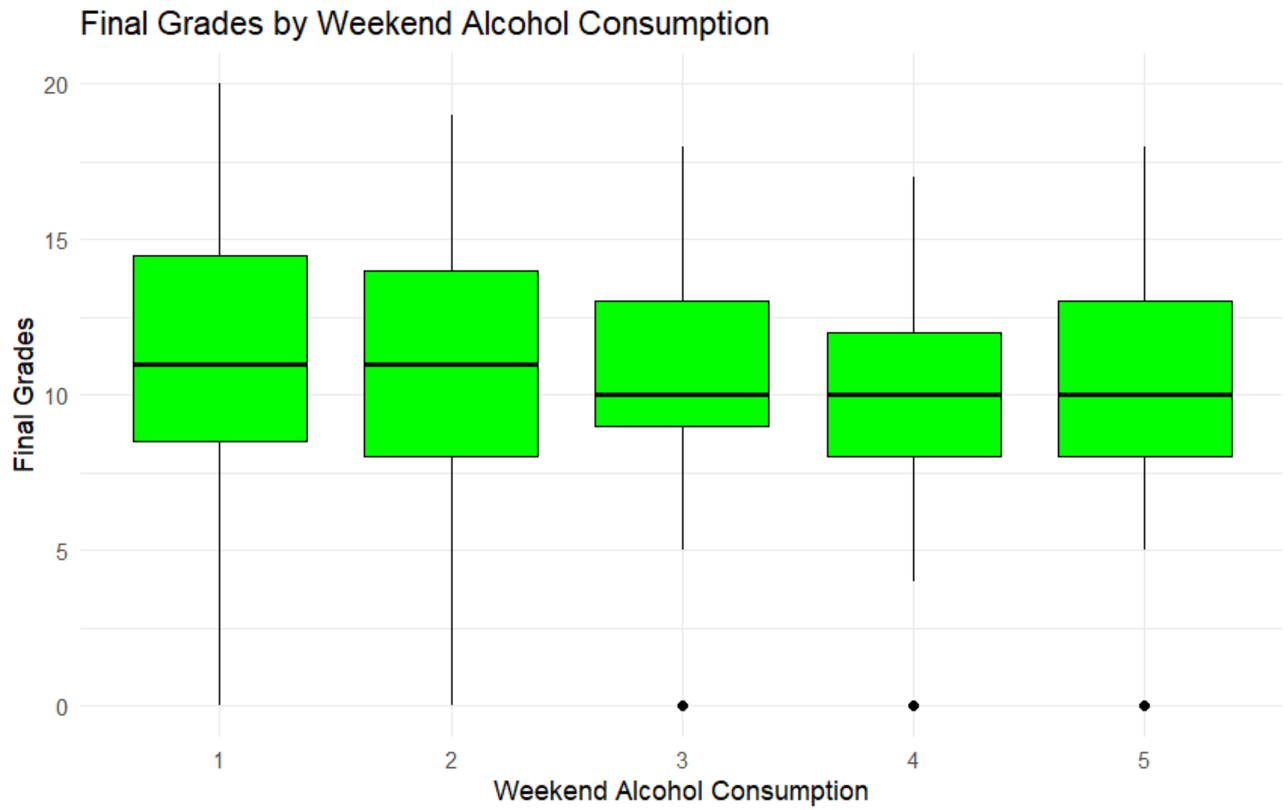
- We have converted categorial variables into dummy variables.

- We consider **sex, age, student’s home address, family size, parent's cohabitation status, mother’s education, father’s education, mother’s job, father job’s, reason to chose this school, student’s guardian, home to school travel time, study time, failure, school support, family support, extra paid classes within the course subject , extra-curricular activities, attended nursery school, wants to take higher education, Internet access at home, with a romantic relationship, quality of family relationships, free time after school, going out with friends, workday alcohol consumption, weekend alcohol consumption, number of school absences, final grade**

Boxplot for Weekday Alcohol Consumption vs Final Grade



Boxplot for Weekend Alcohol Consumption vs Final Grade



2 - Statistical Significance and Multicollinearity

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.2544  -1.8733   0.4061   2.6915   8.6700

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.13449    4.36015   3.012  0.00278 **
sexM              1.23790    0.49920   2.480  0.01361 *
age              -0.30554    0.20345  -1.502  0.13405
addressU          0.44206    0.57169   0.773  0.43989
famsizeLE3        0.73139    0.48713   1.501  0.13413
PstatusT         -0.31174    0.72368  -0.431  0.66689
Medu              0.45495    0.32309   1.408  0.15997
Fedu             -0.09257    0.27725  -0.334  0.73865
Mjobhealth        1.01979    1.11768   0.912  0.36217
Mjobother        -0.32421    0.71199  -0.455  0.64913
Mjobservices      0.66830    0.79759   0.838  0.40265
Mjobteacher      -1.21198    1.03748  -1.168  0.24351
Fjobhealth        0.30640    1.43693   0.213  0.83127
Fjobother        -0.67336    1.02113  -0.659  0.51005
Fjobservices     -0.45690    1.05669  -0.432  0.66572
Fjobteacher       1.24283    1.29306   0.961  0.33713
reasonhome        0.08012    0.55368   0.145  0.88503
reasonother       0.84663    0.81386   1.040  0.29892
reasonreputation  0.57117    0.57462   0.994  0.32090
guardianmother    0.04999    0.54505   0.092  0.92698
guardianother     0.66279    0.99469   0.666  0.50563
traveltime       -0.18190    0.33285  -0.546  0.58508
studytime         0.52860    0.28668   1.844  0.06604 .
failures         -1.73161    0.33273  -5.204  3.3e-07 ***
schoolsupyes     -1.36781    0.66651  -2.052  0.04088 *
famsupyes        -0.90818    0.47590  -1.908  0.05715 .
paidyes           0.35859    0.47720   0.751  0.45288
activitiesyes    -0.37279    0.44233  -0.843  0.39991
nurseryyes       -0.21365    0.54776  -0.390  0.69674
higheryes         1.47826    1.07112   1.380  0.16842
internetyes       0.47741    0.61900   0.771  0.44107
romanticyes      -1.08274    0.46897  -2.309  0.02153 *
famrel            0.21248    0.24499   0.867  0.38636
freetime          0.31991    0.23653   1.353  0.17707
goout            -0.59904    0.22438  -2.670  0.00794 **
Dalc             -0.26663    0.33074  -0.806  0.42069
Walc              0.25105    0.24759   1.014  0.31128
health           -0.18193    0.16088  -1.131  0.25887
absences          0.05244    0.02865   1.830  0.06806 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.107 on 356 degrees of freedom
Multiple R-squared:  0.2739,    Adjusted R-squared:  0.1964
F-statistic: 3.533 on 38 and 356 DF,  p-value: 2.357e-10
```

Performing Multiple Linear Regression (Preliminary) Model

ANALYZING THE MOST STATISTICALLY SIGNIFICANT PREDICTORS

- Following the order, **the most statistically significant predictors are number of past class failures and going out with friends.** We can also consider **being male student, school support and being in a romantic relationship.**

WHILE

- All the rest of the predictors will be removed.

2 - Statistical Significance and Multicollinearity

- From the model age, address, family size, parental status, mother education, father education, mother's job, father's job,reason, guardian, travel time, study time, family support, paid, activities, nursery, higher, internet, family relationship , freetime, health, absences **do NOT really affect the final grades**, therefore, we **remove** them

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.5253  -1.8725   0.4802   2.8328   9.4091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.41883    0.70292  17.667  < 2e-16 ***
sexM          0.98355    0.45281   2.172   0.0305 *
failures     -2.11596    0.29134  -7.263  2.1e-12 ***
schoolsupyes -1.07083    0.64449  -1.662   0.0974 .
romanticyes  -0.89358    0.45715  -1.955   0.0513 .
goout        -0.41746    0.21119  -1.977   0.0488 *
Dalc         -0.10087    0.31696  -0.318   0.7505
Walc          0.05323    0.23333   0.228   0.8197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.219 on 387 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1521
F-statistic: 11.09 on 7 and 387 DF,  p-value: 8.367e-13
```

2 - Statistical Significance and Multicollinearity

Checking for Multicollinearity

- We used **VIF (Variance Inflation Factor)** which is a measure used to detect the presence of multicollinearity among the predictors in a regression model. **Multicollinearity** occurs when **predictors are highly correlated with each other**, which can make it difficult to determine the individual effect of each predictor on the response variable.

```
> vif(lm(G3 ~ ., data))
      sex  failures schoolsup  romantic     goout      Dalc      Walc
1.134411  1.039111  1.036556  1.031996  1.223696  1.764619  1.999004
```

- Based on the VIF values provided, we do not have a multicollinearity issue with any of the predictors, so we can keep them all.
- This low multicollinearity implies that the standard errors of the estimated coefficients are not inflated, and the estimates are reliable, which means these **predictors are not strongly correlated with the others.**

3 - Model Selection

Training and Test

- Divide the data into **a Train and a Test set.**
- We will need the test set to compare mean-square errors between regression models and trees
- Sample split in **Train (70%) and Test (30%) observations** → Train with 279 obs and 8 variables, Test with 116 obs and 8 variables

```
> index = caret::createDataPartition(data$G3, times=1, p=0.7, list=FALSE)
> head(index)
     Resample1
[1,]         3
[2,]         5
[3,]         6
[4,]         7
[5,]         9
[6,]        10
> train = data[index,]
> test  = data[-index,]
```

Best Subset Selection

```
> summ.subset$which
  (Intercept)  sexM failures schoolsupyes romanticyes goout  Dalc  Walc
1        TRUE FALSE     TRUE         FALSE        FALSE FALSE FALSE FALSE
2        TRUE  TRUE     TRUE         FALSE        FALSE FALSE FALSE FALSE
3        TRUE  TRUE     TRUE         FALSE        FALSE  TRUE FALSE FALSE
4        TRUE  TRUE     TRUE         FALSE         TRUE  TRUE FALSE FALSE
5        TRUE  TRUE     TRUE          TRUE         TRUE  TRUE FALSE FALSE
6        TRUE  TRUE     TRUE          TRUE         TRUE  TRUE  TRUE FALSE
7        TRUE  TRUE     TRUE          TRUE         TRUE  TRUE  TRUE  TRUE
```

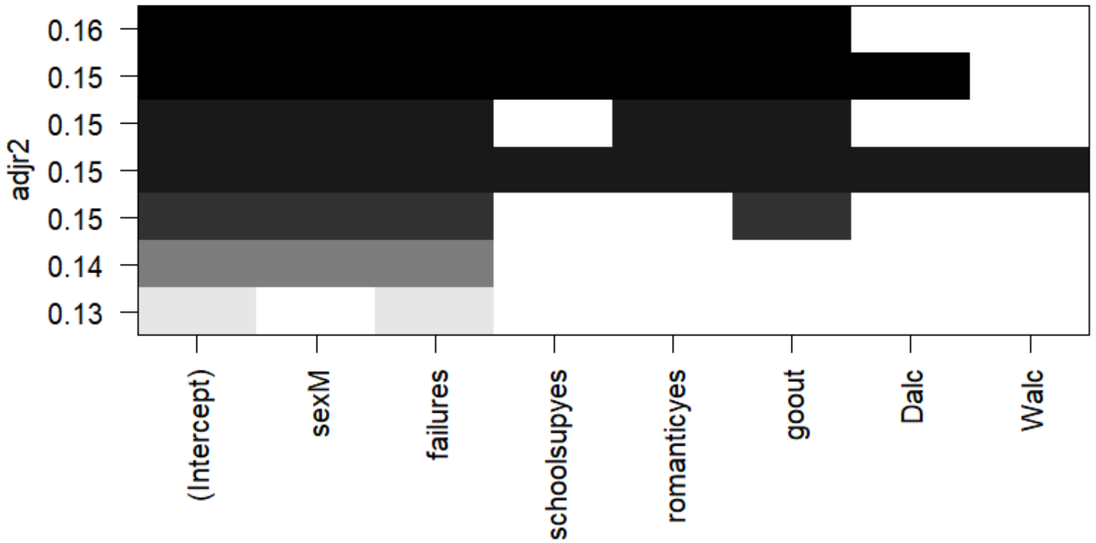
- Since that we have a large number of potential predictors, we want to explore which ones are most relevant for predicting the response variable.

3- Model Selection

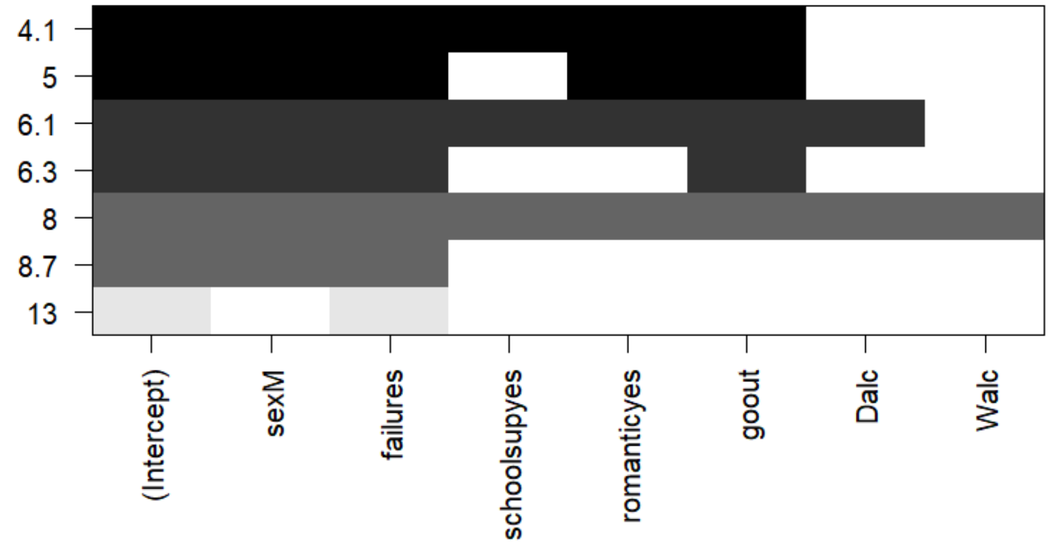
Stepwise Model Selection

In this phase we look at the **adjusted R-squares, CP and BIC** to locate the "best" model.

Adj R^2



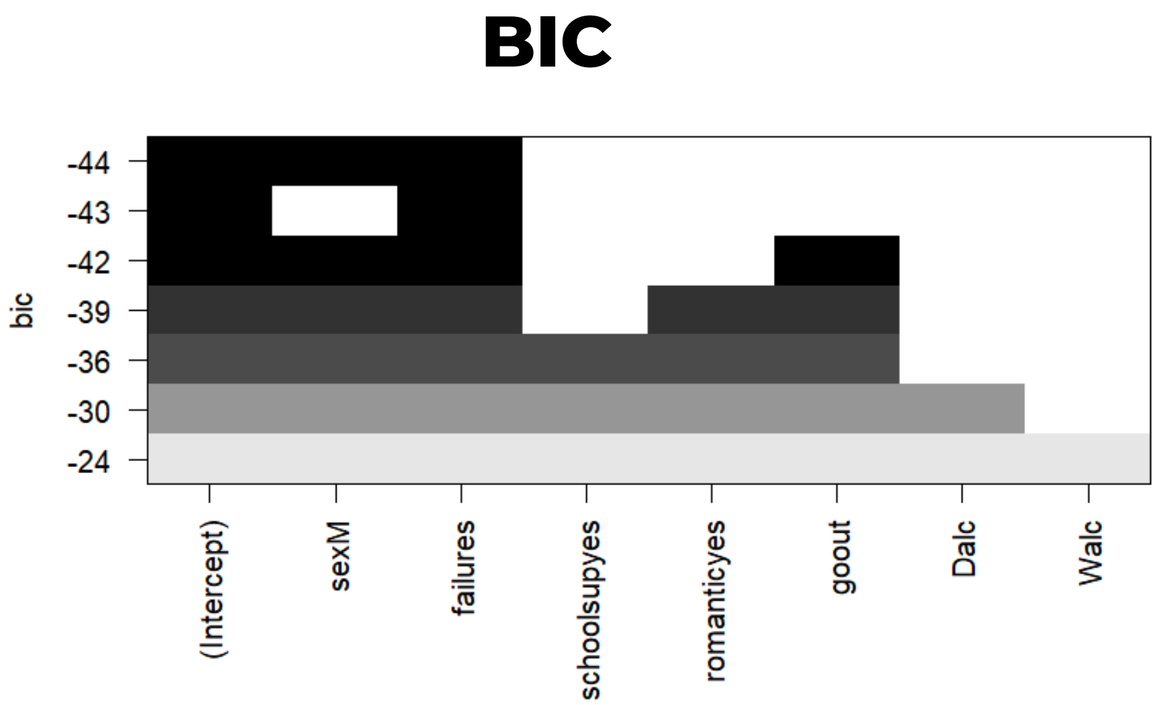
CP



According to Adj R^2 and CP statistics, the best model is with 5 predictors

3- Model Selection

Stepwise Model Selection



According to BIC statistics,
the best model is with 2 predictors

3- Model Selection

Stepwise Model Selection

```
> mod2 = glm(G3 ~ . - Dalc - Walc -schoolsup - romantic - goout, data = train)
> mse2.K5 = cv.glm(train, mod2, K = 5)$delta[2]
>
>
> c(mse5.K5, mse2.K5)
[1] 19.16210 18.81131
> # If we run the above code many times, results may vary because the
> # sampling seed for K-fold CV is changed each time
>
> # To avoid this problem, we could use LOOCV
> mse5.loocv = cv.glm(train, mod5)$delta[2]
> mse2.loocv = cv.glm(train, mod2)$delta[2]
> c(mse5.loocv, mse2.loocv)
[1] 18.92356 18.97901
```

- In this phase, in order to simplify matters compare only the models selected by the adjusted R^2 and BIC.
- We have used **5 -FOLD CROSS-VALIDATION** and **LEAVE-ONE-OUT CROSS VALIDATION (LOOCV).**
- They share almost **the MSE approximately around 19%** → The **model 5** and the **model 2 have almost the same MSE after cross-validation**

3- Model Selection

Stepwise Model Selection

```
> pred2 = predict(mod2, test)
> test.mse.lm =  mean((test$G3 - pred2)^2)
>
> summary(mod2)

Call:
glm(formula = G3 ~ . - Dalc - Walc - schoolsup - romantic - goout,
    data = train)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6794     0.3751  28.467  < 2e-16 ***
sexM          1.1097     0.5204   2.132   0.0339 *
failures     -2.3199     0.3664  -6.331 9.84e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 18.76914)

    Null deviance: 5983.6  on 278  degrees of freedom
Residual deviance: 5180.3  on 276  degrees of freedom
AIC: 1614.8

Number of Fisher Scoring iterations: 2
```

- In this phase, we predict the **test set (pred2)** and c**alculating MSE (test.mse.lm)** in order to evaluate **how well the model (mod2) generalizes to new, unseen data.**

- It helps us in selecting the best-performing model and gaining insights into the relationships between predictors and the response variable.

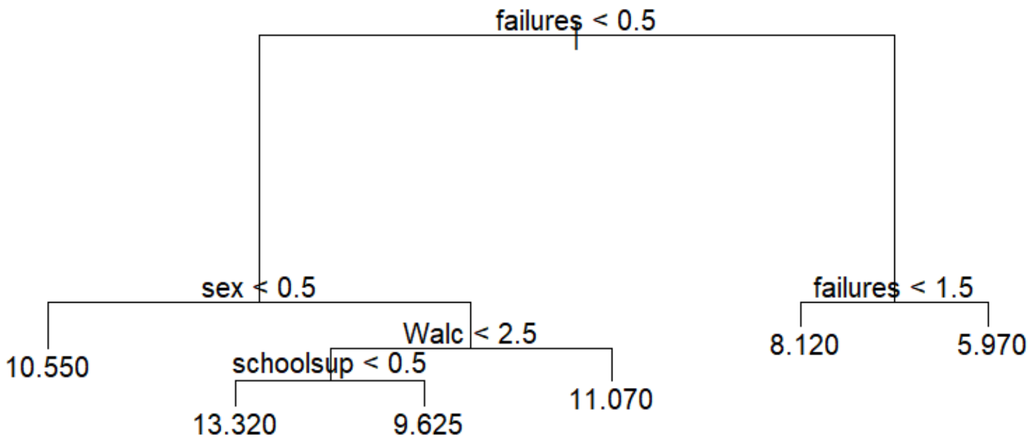
4 - Regression Trees

```
> test.mse.tree = mean((yhat.5 - test$G3)^2)
> c(test.mse.lm, test.mse.tree)
[1] 18.23548 18.09520
```

- We have run the **MSE for the Linear Regression Model and for the Regression Trees**
- The **Regression Trees performs slightly better in terms of prediction accuracy** compared to the linear regression model .

4 - Regression Trees

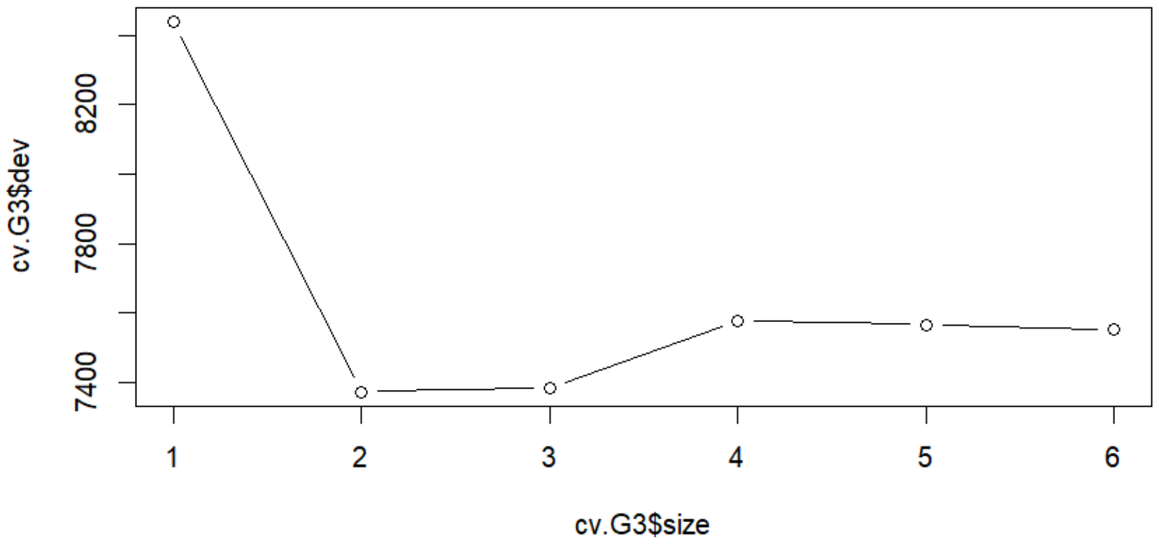
```
> summary(tree.G3)

Regression tree:
tree(formula = G3 ~ ., data = data)
Variables actually used in tree construction:
[1] "failures"  "sex"       "Walc"      "schoolsup"
Number of terminal nodes:  6
Residual mean deviance:  17.31 = 6735 / 389
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.3200  -2.0750   0.4464   0.0000   2.7810   9.8800
```



4 - Regression Trees

```
> cv.G3
$size
[1] 6 5 4 3 2 1

$dev
[1] 7551.280 7565.611 7576.206 7384.969 7373.079 8436.668

$k
[1]       -Inf   91.91898   97.81482  124.00337  178.17338 1042.74339

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Cross-Validation for cost-complexity pruning



Based on the Cross-Validation, 2 **is lowest:**

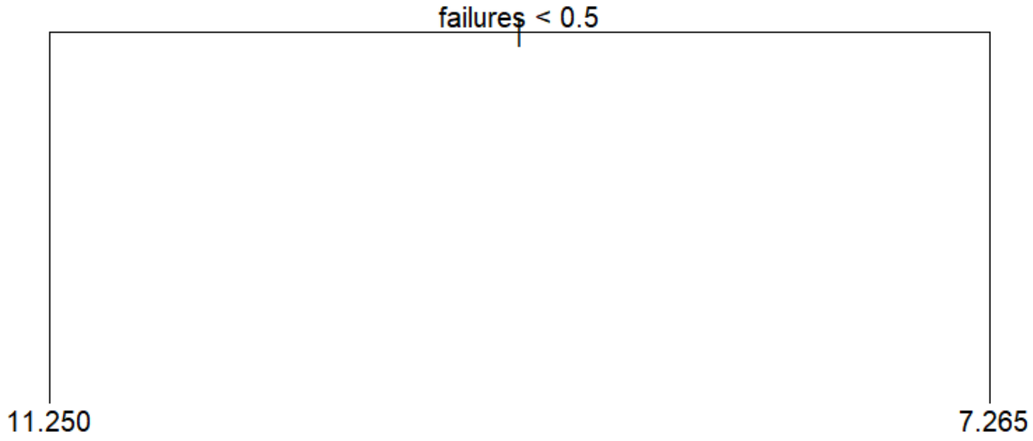
The tree with 2 terminal nodes has the lowest deviance (7373.079), indicating it has the best fit among the considered sizes.

4 - Regression Trees

```
> summary(prune.G3)

Regression tree:
snip.tree(tree = tree.G3, nodes = 3:2)
Variables actually used in tree construction:
[1] "failures"
Number of terminal nodes:  2
Residual mean deviance:  18.39 = 7227 / 393
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11.2500  -2.2530   0.7349   0.0000   2.7470  10.7300
```

TEST SET PERFORMANCE:

TREE WITH 2 TERMINAL NODES



- The tree has been pruned from 6 to 2 terminal nodes.
- In this way, having a small tree we get a lower variance and a better interpretation.
- Based on this regression tree, students with failures higher than 0.5 have a lower final grade

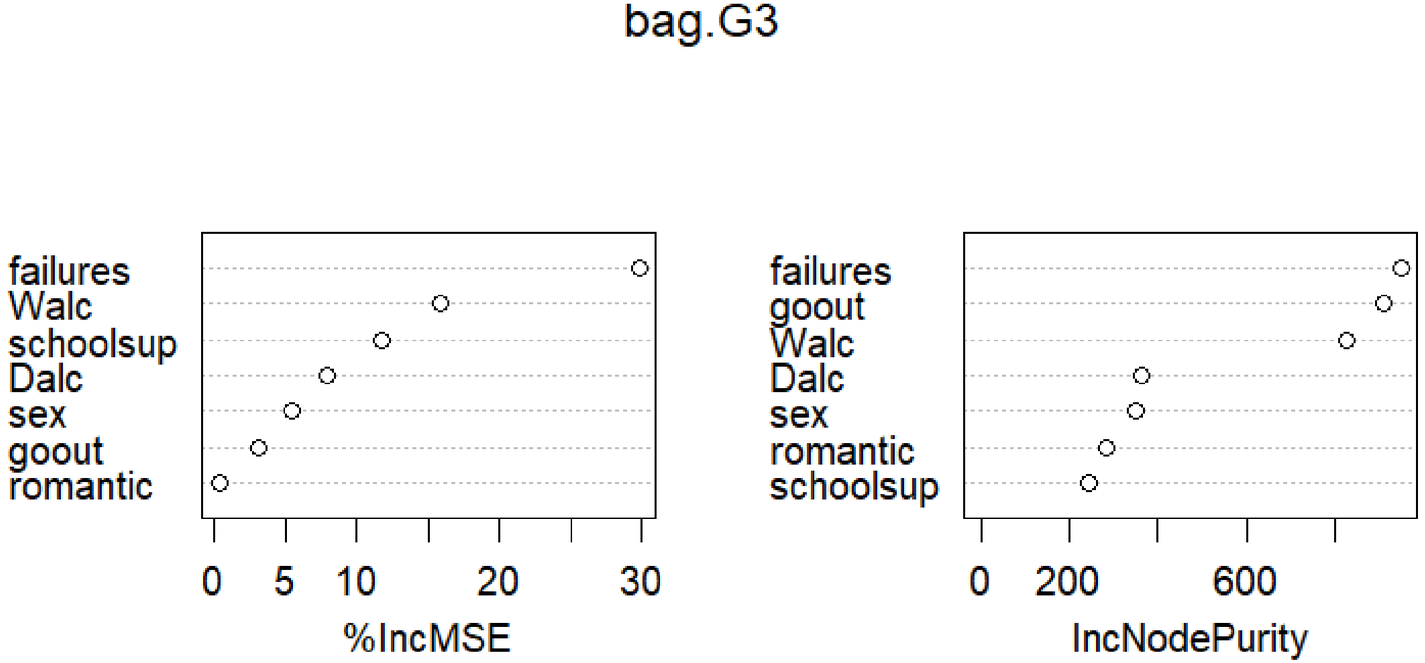
4 - Regression Trees

Bagging

```
> bag.G3

Call:
 randomForest(formula = G3 ~ ., data = train, mtry = 8, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 21.41562
                    % Var explained: 0.15
> importance(bag.G3)
             %IncMSE IncNodePurity
sex        5.4692730      349.4085
failures  29.7996549      949.7279
schoolsup 11.7327533      244.1866
romantic   0.3371959      285.3116
goout      3.1419440      909.8510
Dalc       7.9280320      365.6434
Walc      15.7919376      828.0576
```

VARIABLE IMPORTANCE PLOT

- Left Plot: **Failures** is the one that generate larger MSE if you do not use it
- Right Plot: **Failures** is important in determining the purity of nodes in the trees

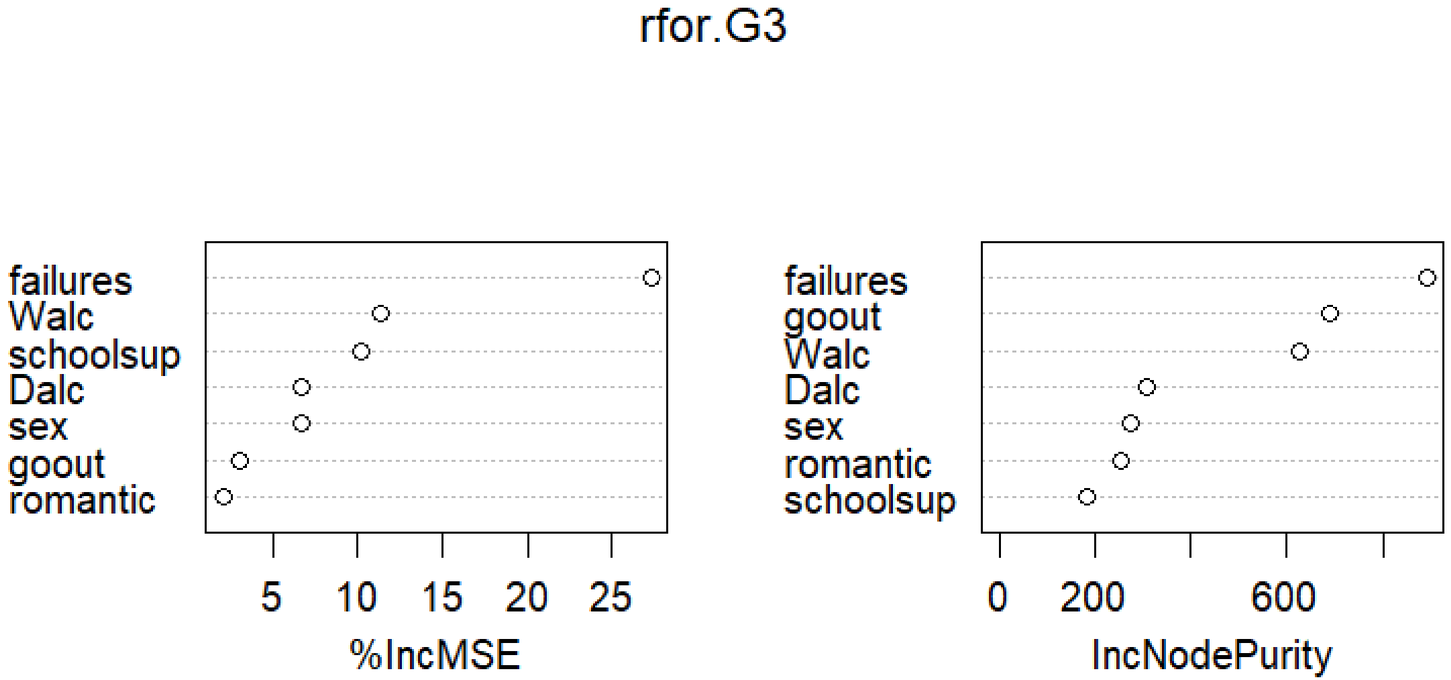



```
> c(test.mse.lm, test.mse.tree, test.mse.bag)
[1] 16.36341 16.24523 20.33390
```

4 - Regression Trees

Random Forest

```
Call:
 randomForest(formula = G3 ~ ., data = train, mtry = 3, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 19.88118
                    % Var explained: 7.3
> importance(rfor.G3)
             %IncMSE IncNodePurity
sex         6.628906      277.2131
failures   27.310048      892.2104
schoolsup  10.208891      183.2505
romantic    2.032943      253.0451
goout       3.025795      691.6584
Dalc        6.711849      309.7741
Walc       11.377672      627.6377
```



Final Conclusions

FAILURES seem to be a strong predictor.
We can declare that there is a clear directly proportional relation between **number of past class failures and Final Grade.**

```
> c(test.mse.lm, test.mse.tree, test.mse.bag, test.mse.rf)
[1] 16.36341 16.24523 20.33390 18.38977
```

To conclude, the **REGRESSION TREES** seems to be the best model in order to **predict Final Grades in students.**

2

CLASSIFICATION WORK PROJECT

CLASSIFICATION

Problem Statement

01. CONTEXT

This dataset centers on heart failure, providing detailed medical records for 5,000 patients. The information was gathered during a follow-up period, allowing researchers to track the patients' health over time. Each patient's profile encompasses 13 distinct clinical features that offer a comprehensive view of their heart health.

.

02. PROBLEM

This study investigates a dataset of 5,000 heart failure patients with the following objectives:

1. Identify Key Clinical Features: We aim to understand which clinical characteristics, such as age, bloodwork results, and lifestyle habits, significantly influence the progression of heart failure.
2. Predict Patient Mortality: By analyzing the rich clinical data, we want to develop models to predict patient mortality (indicated by the DEATH_EVENT variable). This will allow for better risk stratification of patients.

CLASSIFICATION
Problem Statement

03. DATA

5000 Observations
13 Starting Predictors

04. GOAL

Analyze data from 5,000 heart failure patients. Identify key factors affecting disease progression and predict patient risk using their health information. This aims to improve patient care and outcomes

CLASSIFICATION

Main Steps TO DO

1. Data Loading and Preprocessing
2. EXPLORATORY DATA ANALYSIS (EDA)
3. DATA PARTITIONING
4. MODEL SELECTION&EVALUATION
5. FINAL CONCLUSIONS

1 - Exploratory Data Analysis on the Dependent Variable Death Event

Death Event is our response variable

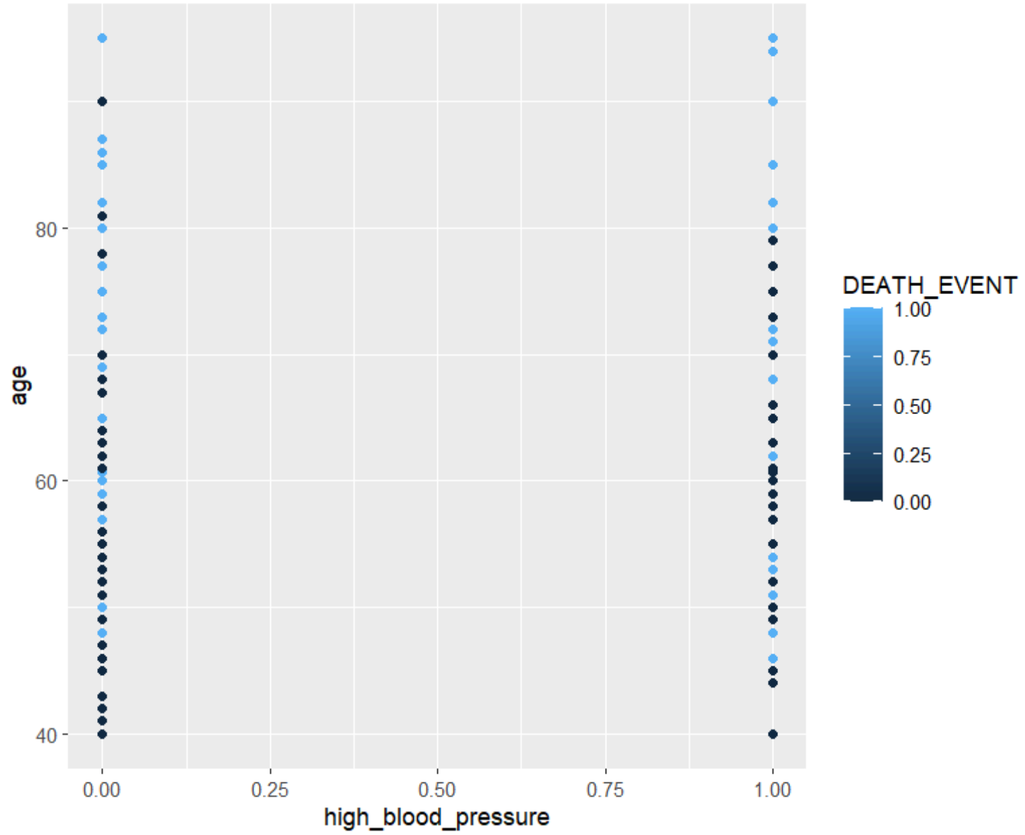
Death Event did NOT occur is coded as 0

Death Event occur is coded as 1

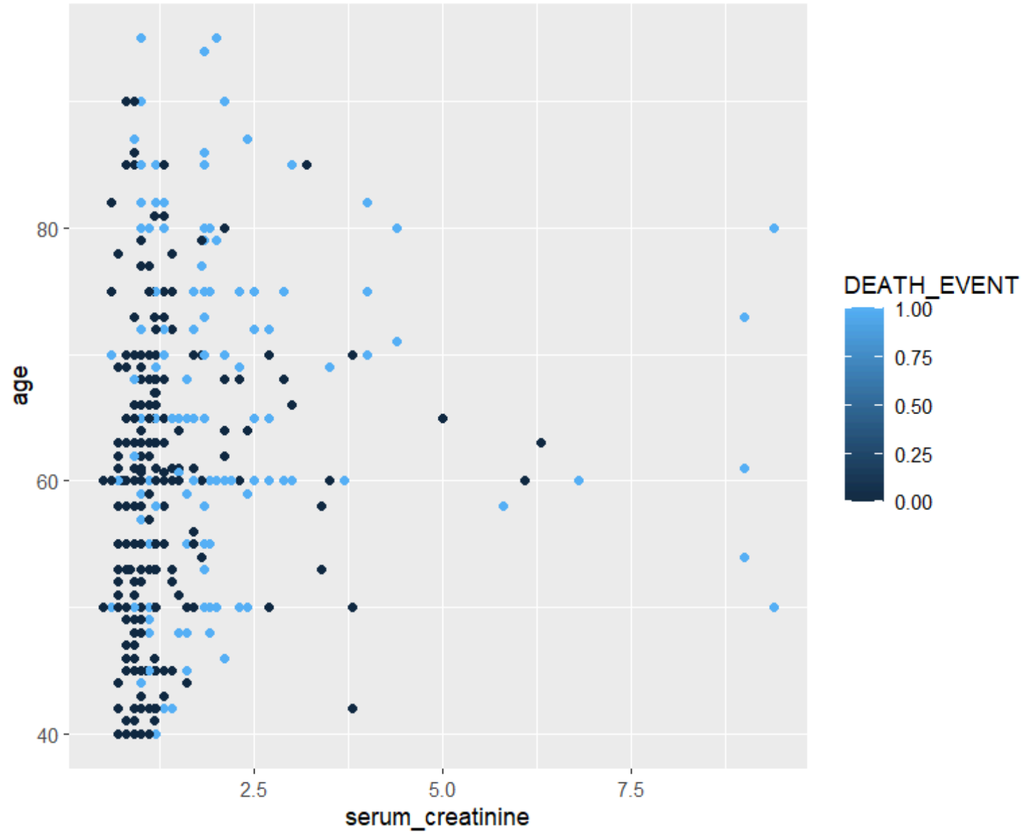
```
> table(data$DEATH_EVENT)

   0    1
3432 1568
> prop.DEATH_EVENT = prop.table(table(data$DEATH_EVENT)); prop.DEATH_EVENT

     0      1
0.6864 0.3136
> summary(data$DEATH_EVENT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3136  1.0000  1.0000
```

1 - Exploratory Data Analysis

Data Visualization



Relationship between high blood pressure and age

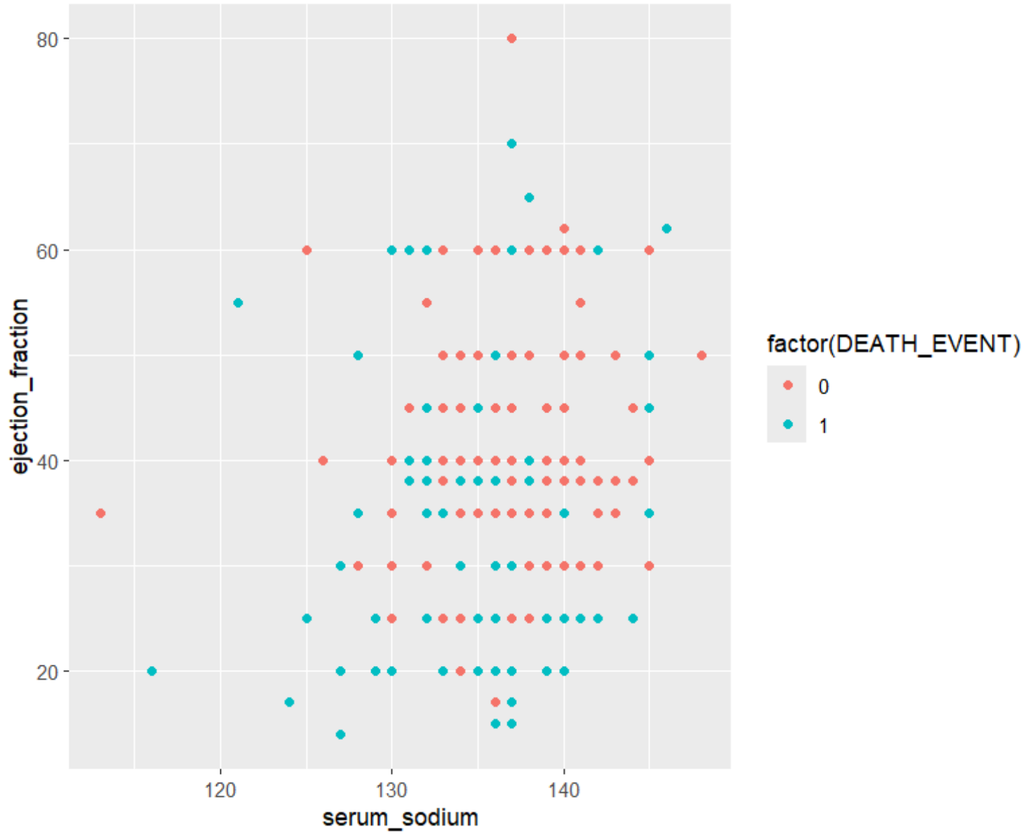


Relationship between serum creatinine and age

1 - Exploratory Data Analysis

Data Visualization

Relationship between age and platelets



Relationship between serum sodium and ejection function

2 - Preliminary Linear Probability Model

Dependent Variable: DEATH_EVENT

Significant variables

- Age
- Creatinine Phosphokinase
- Ejection Fraction
- Serum Creatinine
- Serum Sodium
- Time

Non-significant Variables: Anaemia, Diabetes, High Blood Pressure, Platelets, Sex, Smoking

Model Fit:

- Residual Standard Error: 0.3544
- Multiple R-squared: 41.81%
- Adjusted R-squared: 41.67%

Overall Model Significance:

- F-statistic p < 2.2e-16

```
Call:
lm(formula = DEATH_EVENT ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90461 -0.26912 -0.03258  0.22708  1.19704

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.837e+00  1.653e-01  11.112  < 2e-16 ***
age                       5.599e-03  4.526e-04  12.370  < 2e-16 ***
anaemia                   4.452e-03  1.041e-02   0.428    0.669
creatinine_phosphokinase  4.043e-05  5.269e-06   7.672 2.02e-14 ***
diabetes                 -1.288e-03  1.045e-02  -0.123    0.902
ejection_fraction        -9.896e-03  4.516e-04 -21.915  < 2e-16 ***
high_blood_pressure       1.754e-02  1.072e-02   1.635    0.102
platelets                -5.544e-08  5.212e-08  -1.064    0.287
serum_creatinine          8.009e-02  5.239e-03  15.287  < 2e-16 ***
serum_sodium             -9.312e-03  1.197e-03  -7.782 8.61e-15 ***
sex                      -2.079e-03  1.200e-02  -0.173    0.862
smoking                  -3.193e-03  1.226e-02  -0.260    0.794
time                     -2.588e-03  6.912e-05 -37.445  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3544 on 4987 degrees of freedom
Multiple R-squared:  0.4181,    Adjusted R-squared:  0.4167
F-statistic: 298.6 on 12 and 4987 DF,  p-value: < 2.2e-16
```

3 - Multicollinearity_

- **No Multicollinearity Issues**: All VIF values are well below the threshold of 5, indicating that there is no significant multicollinearity among the predictors in your model.
- **Model Stability**: Since multicollinearity is not an issue, the model's coefficients are stable, and the interpretations of the predictors are reliable.

age	anaemia	creatinine_phosphokinase
1.115624	1.075941	1.054400
diabetes	ejection_fraction	high_blood_pressure
1.071676	1.076308	1.061125
platelets	serum_creatinine	serum_sodium
1.038389	1.113964	1.135818
sex	smoking	time
1.310853	1.284042	1.137036

- VIF = 1: No correlation between the predictor and other predictors.
- 1 < VIF < 5: Moderate correlation that is typically not problematic.
- VIF > 5: High correlation, potentially problematic, and may indicate multicollinearity.
- VIF > 10: Very high correlation, serious multicollinearity, and suggests that the predictor may need to be removed or the model needs to be adjusted

4 - Data Partition: Train and Test

```
> indexes = createDataPartition(data$DEATH_EVENT, times=1, p=0.7, list=FALSE)
> trainData = data[indexes,]
> testData = data[-indexes,]
> head(indexes)
     Resample1
[1,]         1
[2,]         2
[3,]         3
[4,]         4
[5,]         5
[6,]         7
```

- Train data- Contains 70% of the original data, specifically rows specified by indexes
- Test data - Contains 30% of the original data not specified in the indexes

The data in the train and test are well retained from the original data as seen here →

```
> prop.table(table(data$DEATH_EVENT))

     0      1
0.6864 0.3136
> prop.table(table(trainData$DEATH_EVENT))

        0         1
0.6845714 0.3154286
> prop.table(table(testData$DEATH_EVENT))

        0         1
0.6906667 0.3093333
```

5 - K-Nearest Neighbors (KNN)

- We perform cross-validation to choice the n° of neighbors for KNN
- Using 10-fold cross-validation (cv_10) to find the best k value.
- **The higher accuracy is achieved with a 3 n° of neighbors.**

	k	Accuracy	Kappa	AccuracySD	KappaSD
1	3	0.9468569	0.8774246	0.006057589	0.01396104

```
> table(testData$DEATH_EVENT, pred.knn)
   pred.knn
      0   1
  0 995  41
  1  34 430
> tab = prop.table(table(pred.knn, testData$DEATH_EVENT)); tab

pred.knn          0          1
       0 0.66333333 0.02266667
       1 0.02733333 0.28666667
>
> err.knn = tab[1,2] + tab[2,1]; err.knn
[1] 0.05
```

We then fit in the k-nn with k=k.best on the whole training set and compute the error rate using the test data. We concluded with a **5% misclassification rate**

6 - Generalized Linear Model

Binary Choice Model: Logit Model

```
> mod.logit = glm(DEATH_EVENT ~ ., data = trainData, family=binomial)
> summary(mod.logit)

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

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.187e+01  1.649e+00   7.195 6.25e-13 ***
age                      4.393e-02  4.689e-03   9.369  < 2e-16 ***
anaemia                  1.277e-01  1.067e-01   1.197   0.2315
creatinine_phosphokinase 2.937e-04  5.403e-05   5.436 5.45e-08 ***
diabetes                -7.548e-02  1.042e-01  -0.724   0.4689
ejection_fraction       -7.995e-02  4.966e-03 -16.100  < 2e-16 ***
high_blood_pressure      2.581e-01  1.059e-01   2.436   0.0149 *
platelets               -6.533e-07  5.327e-07  -1.226   0.2201
serum_creatinine         7.140e-01  5.531e-02  12.909  < 2e-16 ***
serum_sodium            -8.269e-02  1.169e-02  -7.074 1.51e-12 ***
sex                     -3.196e-02  1.213e-01  -0.264   0.7921
smoking                  6.300e-02  1.235e-01   0.510   0.6100
time                    -2.052e-02  8.854e-04 -23.181  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4363.6  on 3499  degrees of freedom
Residual deviance: 2479.5  on 3487  degrees of freedom
AIC: 2505.5

Number of Fisher Scoring iterations: 6
```

- In a binary choice model such as Logistic Regression, the coefficients represent the log odds of the outcome occurring for a one-unit change in the response variable, keeping other things being equal.
- **Most statistically variable** are: age, anemia, creatinine phosphokinase, ejection fraction, serum creatinine, serum sodium and time.
- We can also consider high blood pressure.

7 - Predictions for the test dataset

- We assign labels to the test set using this model.
- We use the predictive function

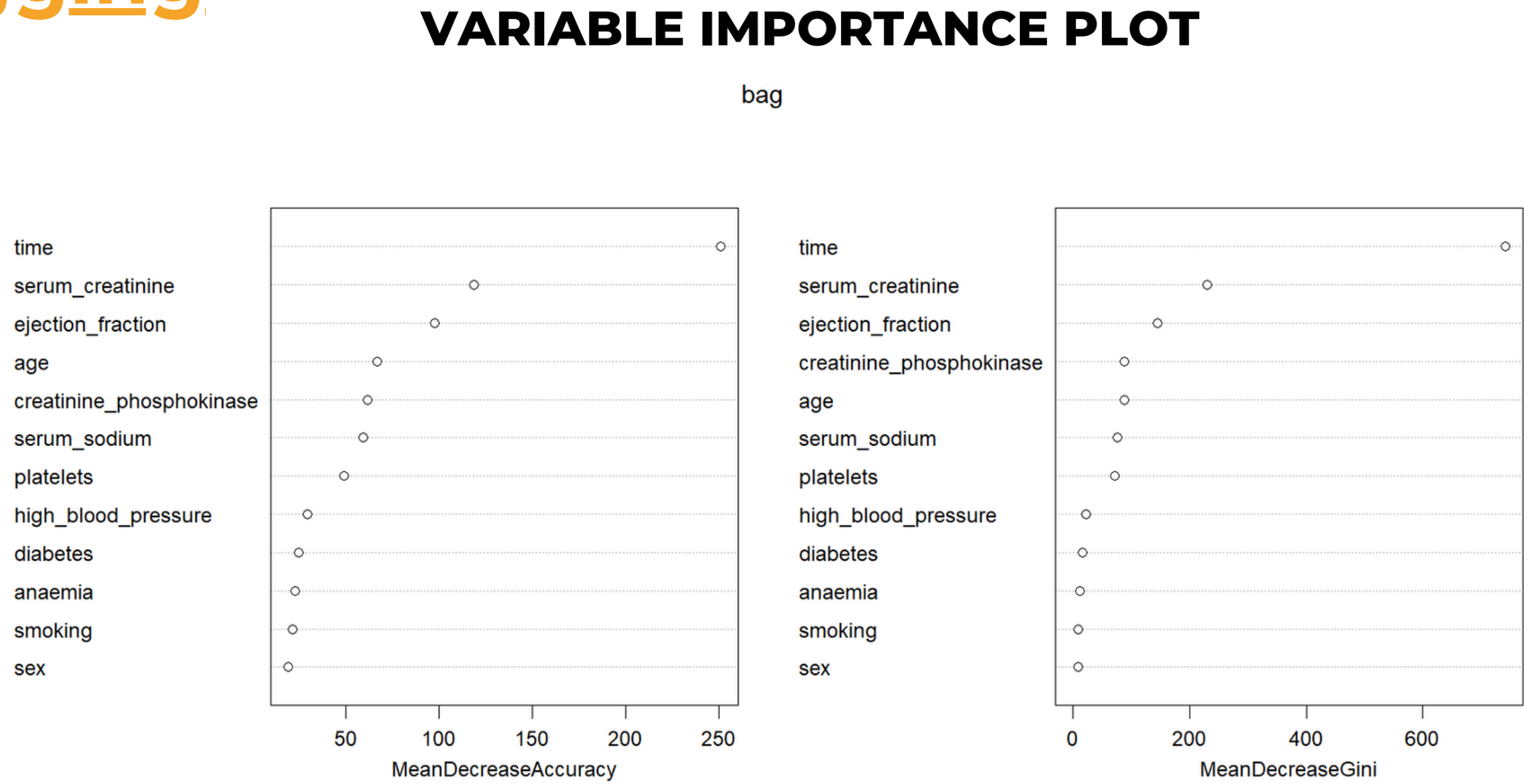
```
> prob.logit = predict(mod.logit, newdata=testData, type="response")
> summary(prob.logit)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0004899 0.0366782 0.1794450 0.3074694 0.5649625 0.9997564
```

- We compute **Confusion Matrix** for test set.
- We assign value 1 to all observations for which the fitted probabilities in the test set is greater than 0.5

```
> pred.logit = 1*(prob.logit>0.5)
>
> table(testData$DEATH_EVENT, pred.logit)
   pred.logit
      0   1
  0 941  95
  1 154 310
> tab = prop.table(table(pred.logit, testData$DEATH_EVENT)); tab

pred.logit          0          1
         0 0.62733333 0.10266667
         1 0.06333333 0.20666667
```

We have **95 False Positive and 154 False Negative**

8 - Bagging



- Important predictors:

Time (250.77844) - *Mean decrease accuracy*
Ejection fraction (97.47412)
Serum creatinine (118.30541)

8 - Bagging

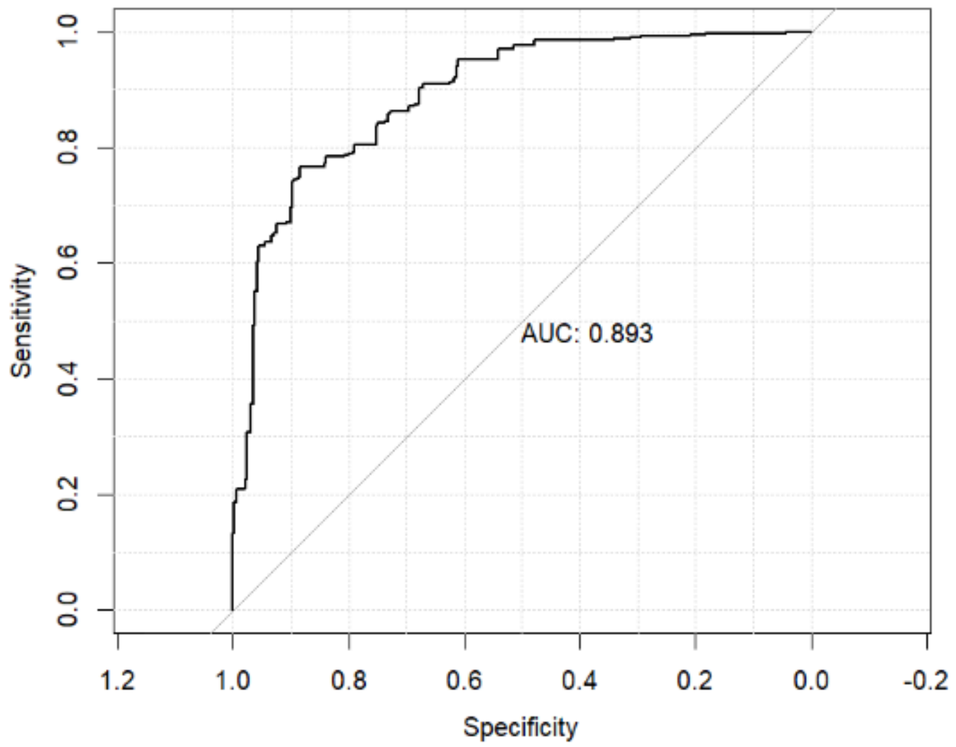
```
> table(testData$DEATH_EVENT, pred.bag)
   pred.bag
      0    1
  0 1032    4
  1    5  459
> tab = prop.table(table(pred.bag, testData$DEATH_EVENT)); tab

pred.bag           0           1
       0 0.688000000 0.003333333
       1 0.002666667 0.306000000
> err.bag = tab[1,2] + tab[2,1]
> cbind(err.knn, err.logit,err.tree, err.bag)
     err.knn err.logit   err.tree err.bag
[1,]    0.05     0.166 0.08666667   0.006
```

CONFUSION MATRIX:

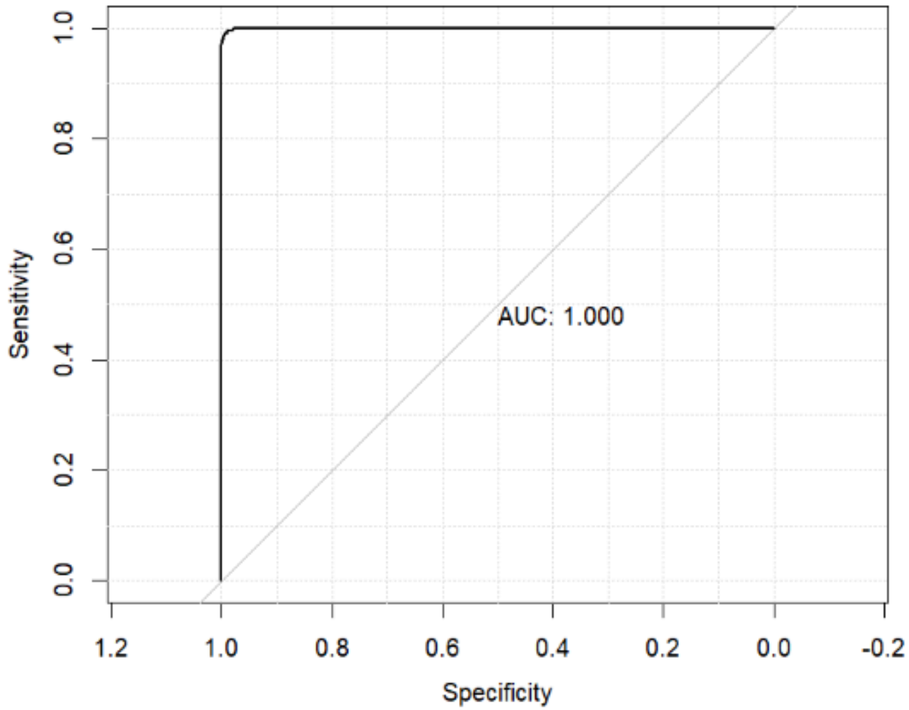
- True Negative (TN): Predicted 0 when actually 0 = 0.688
- False Positive (FP): Predicted 1 when actually 0 = 0.003
- False Negative (FN): Predicted 0 when actually 1 = 0.002
- True Positive (TP): Predicted 1 when actually 1 = 0.306

After checking with other various models bagging has the least error rate with 0.6%

9 - ROC Curve and AUC



At the beginning using the Logit Model the AUC is 0.893



At the end using the Bagging with random forest , AUC is 1.000

Final Conclusions

- Begging is the best model identified with a error rate of 0.6%.
- With time, Ejection fraction, Serum creatinine as the most important variables in predicting DEATH EVENT


THANK YOU