TUGAS BESAR BIG DATA

EKSPOLRASI, VISUALISASI DAN KLASIFIKASI DATA

KORBAN TENGGELAMNYA KAPAL TITANIC MENGGUNAKAN BAHASA R

Kelompok:

1. Isep lutpi nur (2113191079)
2. Farhan azis (2113191097)
3. M. Taufiq hidayatuloh (2113191036)
4. Dara atria ferliandini (2113191098)

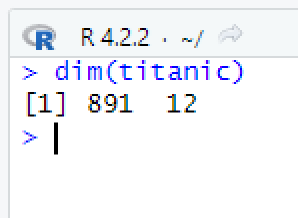
EKSPLORASI DATA

1. Sumber Data

   Data yang kami miliki bersumber dari website Kaggle. Link data:

   https://www.kaggle.com/datasets/brendan45774/test-file

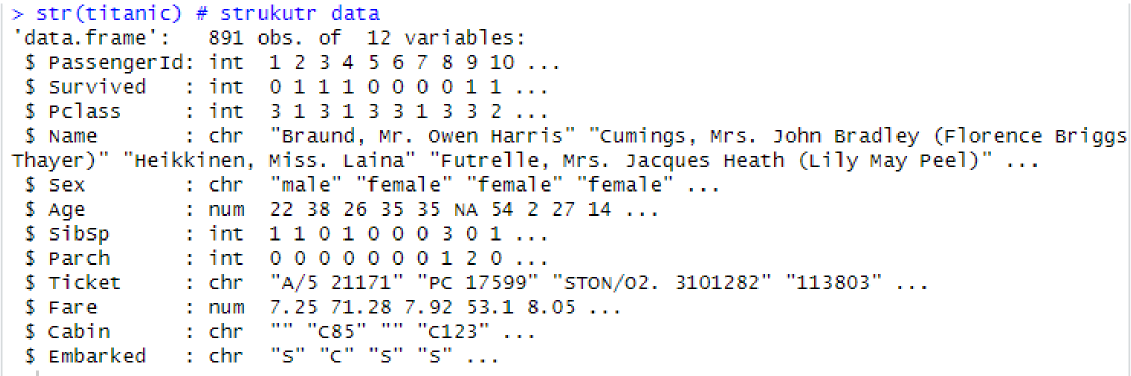
2. Dimensi data



```
> dim(titanic)
[1] 891  12
> |
```

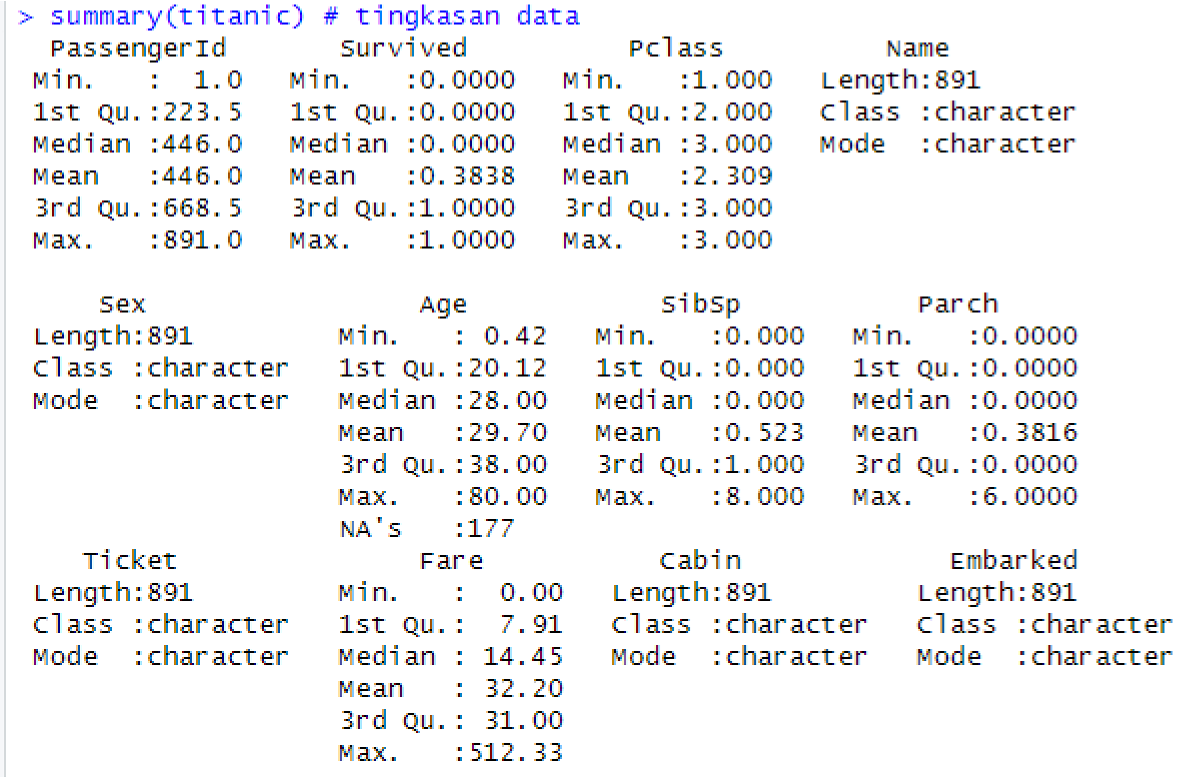
   Mempunyai 12 variable dan 891 jumlah data

3. Struktur Data



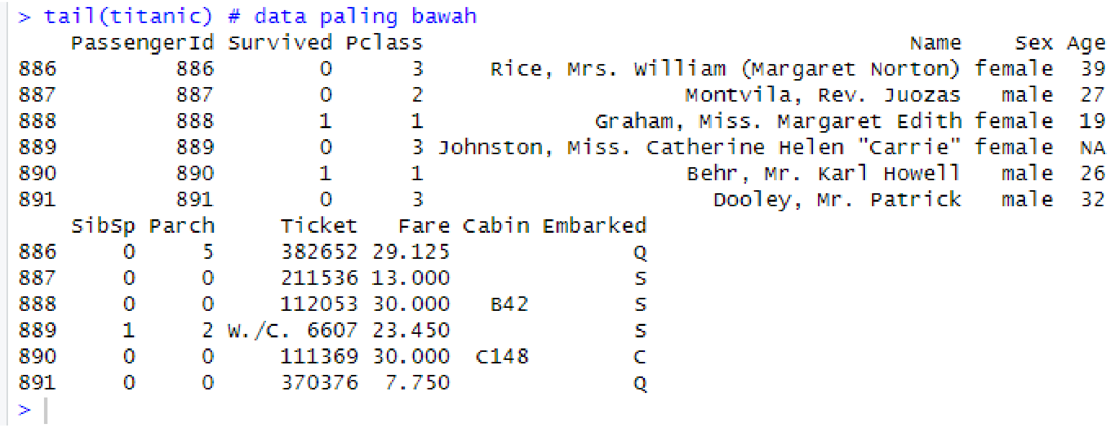
```
> str(titanic) # strukutr data
'data.frame':	891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs
Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

4. Ringkasan Data



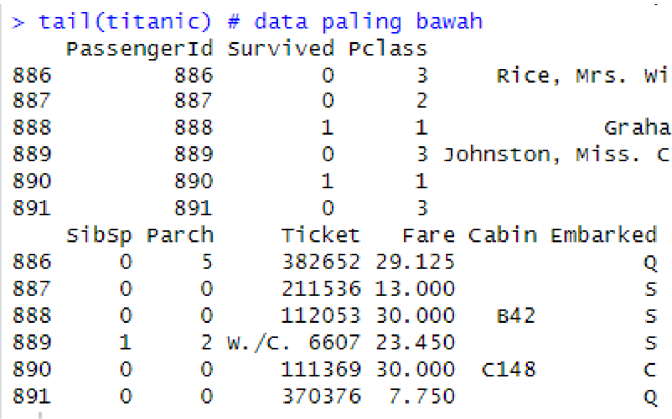
```
> summary(titanic) # tingkasan data
  PassengerId       Survived          Pclass          Name          
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891        
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                     
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                     
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                     
                                                                    
     Sex                 Age            SibSp           Parch       
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000  
 Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000  
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000  
                    Mean   :29.70   Mean   :0.523   Mean   :0.3816  
                    3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000  
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000  
                    NA's   :177                                     
    Ticket               Fare           Cabin             Embarked        
 Length:891         Min.   :  0.00   Length:891         Length:891        
 Class :character   1st Qu.:  7.91   Class :character   Class :character  
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
                    Mean   : 32.20                                        
                    3rd Qu.: 31.00                                        
                    Max.   :512.33                                        
```

5. Data Paling Atas



```
> tail(titanic) # data paling bawah
    PassengerId Survived Pclass                                     Name    Sex Age
886         886        0      3     Rice, Mrs. William (Margaret Norton) female  39
887         887        0      2                    Montvila, Rev. Juozas   male  27
888         888        1      1             Graham, Miss. Margaret Edith female  19
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  NA
890         890        1      1                    Behr, Mr. Karl Howell   male  26
891         891        0      3                      Dooley, Mr. Patrick   male  32
    SibSp Parch     Ticket   Fare Cabin Embarked
886     0     5     382652 29.125              Q
887     0     0     211536 13.000              S
888     0     0     112053 30.000   B42        S
889     1     2 W./C. 6607 23.450              S
890     0     0     111369 30.000  C148        C
891     0     0     370376  7.750              Q
> 
```

6. Data Paling Bawah

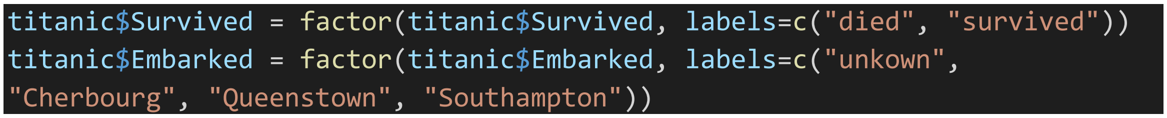


```
> tail(titanic) # data paling bawah
    PassengerId Survived Pclass                                          Name    Sex Age
886         886        0      3          Rice, Mrs. William (Margaret Norton) female  39
887         887        0      2                         Montvila, Rev. Juozas   male  27
888         888        1      1                  Graham, Miss. Margaret Edith female  19
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  NA
890         890        1      1                         Behr, Mr. Karl Howell   male  26
891         891        0      3                           Dooley, Mr. Patrick   male  32
    SibSp Parch     Ticket   Fare Cabin Embarked
886     0     5     382652 29.125              Q
887     0     0     211536 13.000              S
888     0     0     112053 30.000   B42        S
889     1     2 W./C. 6607 23.450              S
890     0     0     111369 30.000  C148        C
891     0     0     370376  7.750              Q
```

VISUALISAI DATA

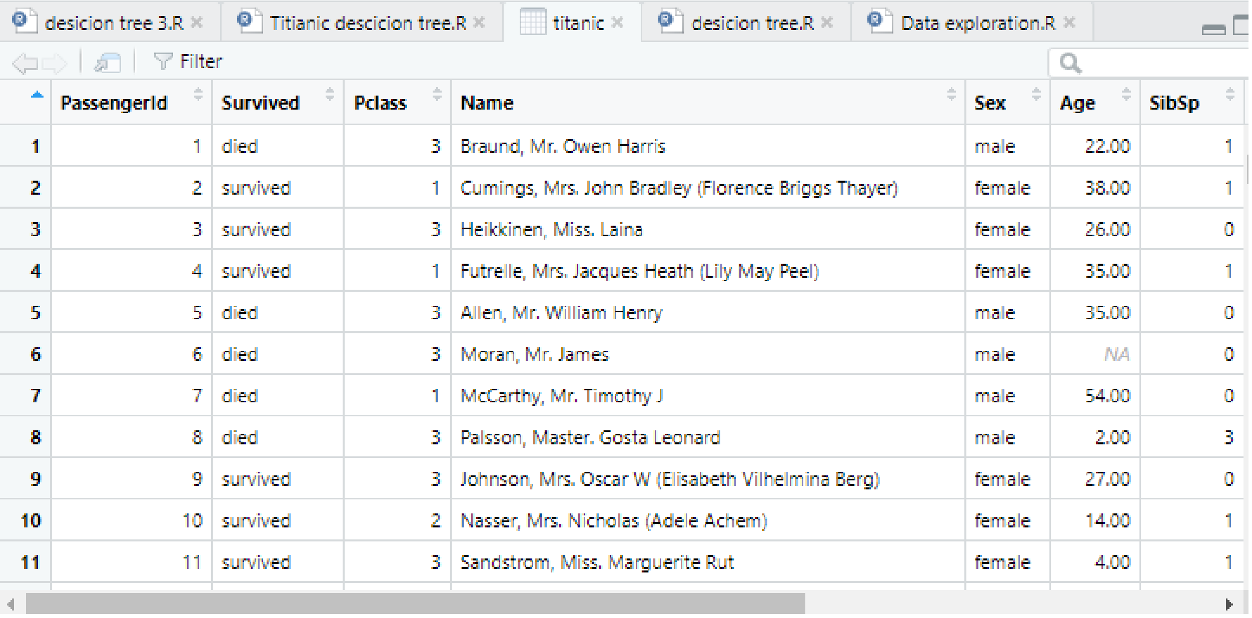
1. Mengubah Kolom yang class menjadi sebuah factor

   untuk keterbacaan contoh nya di data Survied (Orang yang bertahan hidup) hanya nilai 1 untuk selamat dan 0 yang meninggal. Maka di ganti 1 menjadi survived dan 0 menjadi died. Sebelumnya variable dataset utama yaitu di beri nama **titanic.** Mengubah data class menjadi factor dengan perintah di bawah:



```
titanic$Survived = factor(titanic$Survived, labels=c("died", "survived"))
titanic$Embarked = factor(titanic$Embarked, labels=c("unkown",
"Cherbourg", "Queenstown", "Southampton"))
```

Hasil:

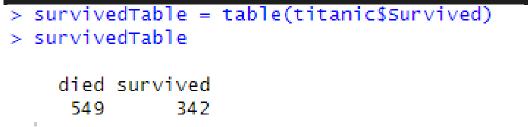


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1	1	died	3	Braund, Mr. Owen Harris	male	22.00	1
2	2	survived	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1
3	3	survived	3	Heikkinen, Miss. Laina	female	26.00	0
4	4	survived	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1
5	5	died	3	Allen, Mr. William Henry	male	35.00	0
6	6	died	3	Moran, Mr. James	male	NA	0
7	7	died	1	McCarthy, Mr. Timothy J	male	54.00	0
8	8	died	3	Palsson, Master. Gosta Leonard	male	2.00	3
9	9	survived	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0
10	10	survived	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1
11	11	survived	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1

2. Distribusi class mengunakan pie chart

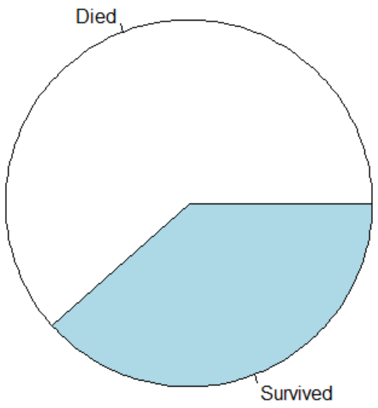
   Untuk melihat perbandingan class(Data bertahan hidup/survived) menggunakan pie chart.

```
survivedTable = table(titanic$Survived)
survivedTable
```



```
> survivedTable = table(titanic$Survived)
> survivedTable

   died survived 
    549      342 
```

```
par(mar = c(0, 0, 0, 0), oma = c(0, 0, 0, 0))
pie(survivedTable,labels=c("Died","Survived"))
```



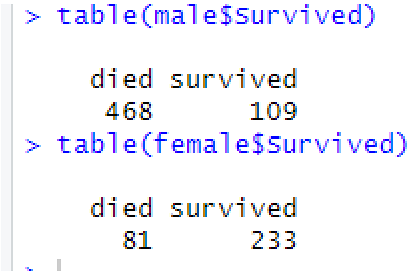
3. Perbandingan korban berdasarkan jenis kelamin menggunakan pie chart

   Data di bagi menjadi berdasarkan jenis kelamin kemudian dilihat perbadingan nya melalui pie chart.

```
male = titanic[titanic$Sex=="male",]
female = titanic[titanic$Sex=="female",]

table(male$Survived)
table(female$Survived)


par(mfrow = c(1, 2), mar = c(0, 0, 2, 0), oma = c(0, 1, 0, 1))
pie(table(male$Survived),labels=c("Dead","Survived"), main="Perbandingan
Korban Penumpang Pira")
```

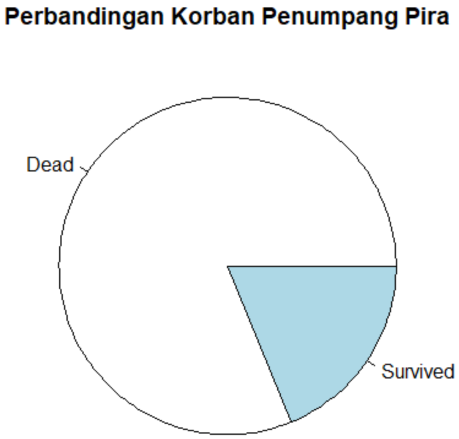
```
pie(table(female$Survived),labels=c("Dead","Survived"), main="Perbandingan
Korban Penumpang Pira")
```

- Rincian

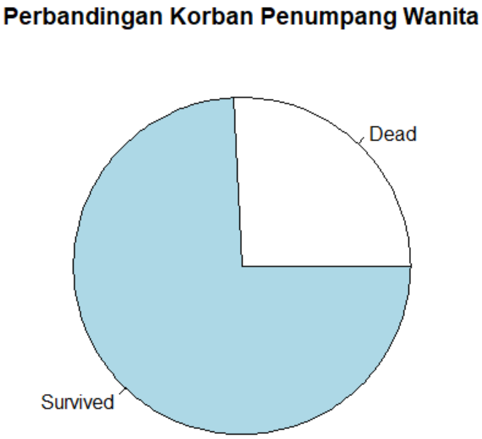


```
> table(male$Survived)

  died survived 
   468      109 
> table(female$Survived)

  died survived 
    81      233 
```

- Pie Chart



Perbandingan Korban Penumpang Pira



Perbandingan Korban Penumpang Wanita

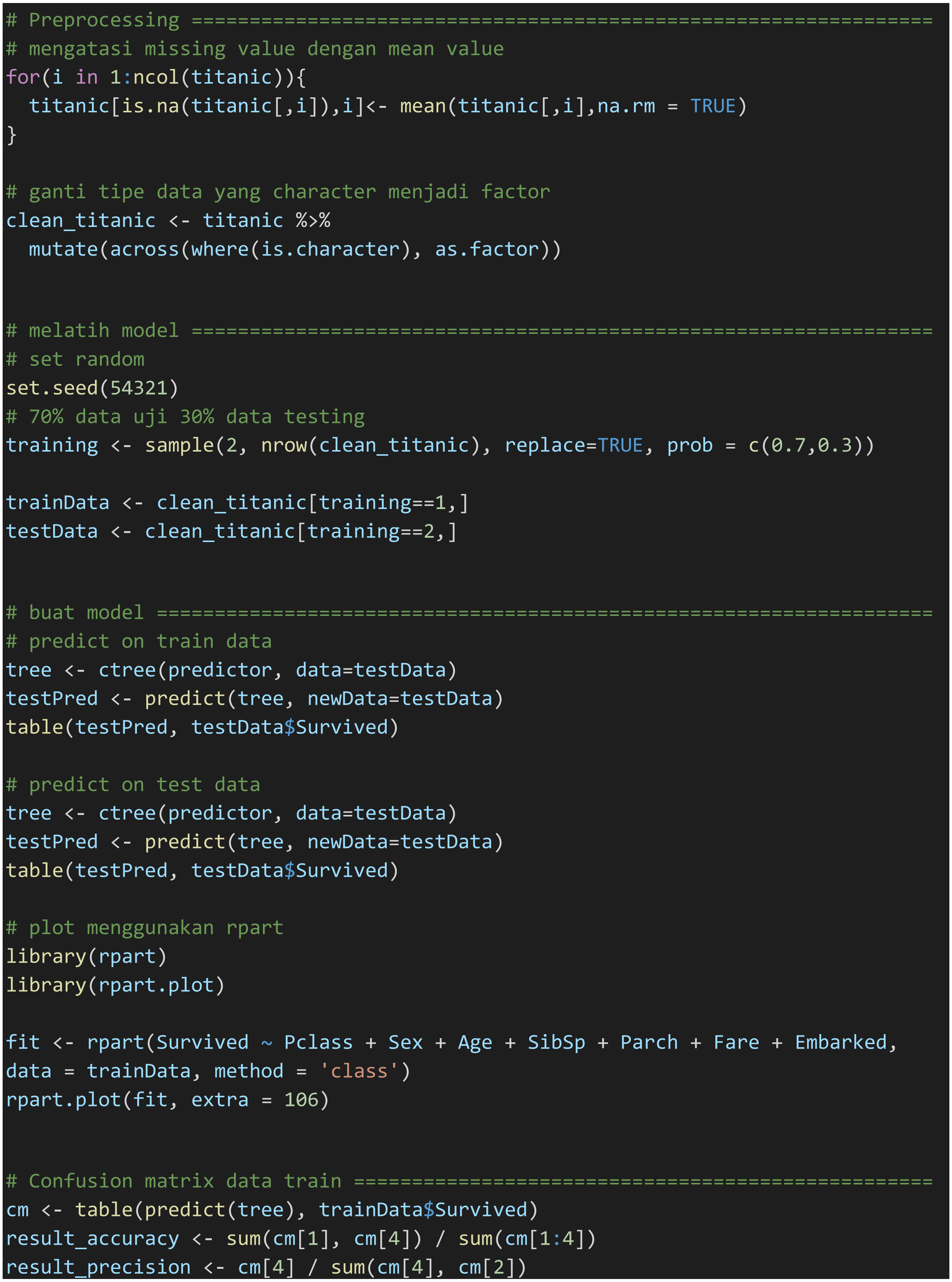
Dari data diatas dapat diambil kesimpulan bahwa yang leibh banyak bertahan hidup yaitu Wanita.

KLASIFIKASI DATA

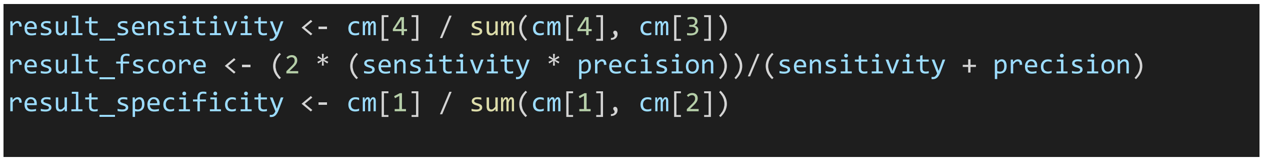
1. Klasifikasi Menggunakan Metode Decision Tree

Syntax:

```
library(dplyr)
library(party)
# clear console
cat("\014")

titanic = read.csv('D:\\Kampus\\big data titanic\\titanic.csv')
titanic$Survived = factor(titanic$Survived, labels=c("died", "survived"))
titanic$Embarked = factor(titanic$Embarked, labels=c("unkown", "Cherbourg",
"Queenstown", "Southampton"))
```



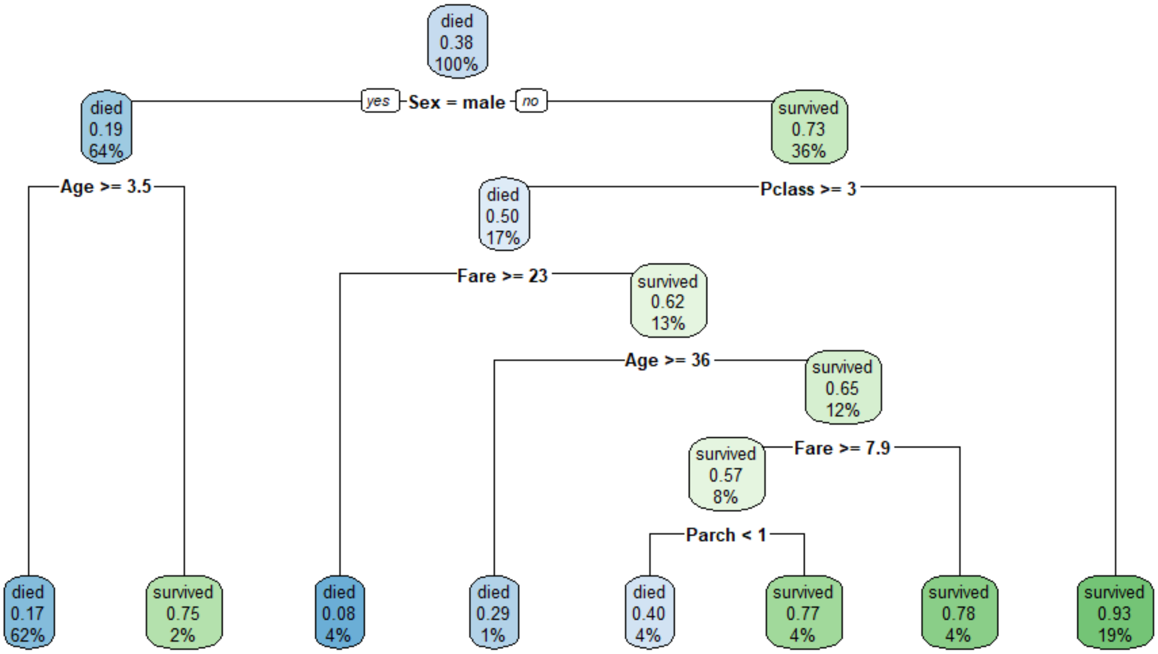
```
# Preprocessing ================================================================
# mengatasi missing value dengan mean value
for(i in 1:ncol(titanic)){
  titanic[is.na(titanic[,i]),i]<- mean(titanic[,i],na.rm = TRUE)
}

# ganti tipe data yang character menjadi factor
clean_titanic <- titanic %>%
  mutate(across(where(is.character), as.factor))


# melatih model ================================================================
# set random
set.seed(54321)
# 70% data uji 30% data testing
training <- sample(2, nrow(clean_titanic), replace=TRUE, prob = c(0.7,0.3))

trainData <- clean_titanic[training==1,]
testData <- clean_titanic[training==2,]


# buat model ===================================================================
# predict on train data
tree <- ctree(predictor, data=testData)
testPred <- predict(tree, newData=testData)
table(testPred, testData$Survived)

# predict on test data
tree <- ctree(predictor, data=testData)
testPred <- predict(tree, newData=testData)
table(testPred, testData$Survived)

# plot menggunakan rpart
library(rpart)
library(rpart.plot)

fit <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
data = trainData, method = 'class')
rpart.plot(fit, extra = 106)


# Confusion matrix data train ==================================================
cm <- table(predict(tree), trainData$Survived)
result_accuracy <- sum(cm[1], cm[4]) / sum(cm[1:4])
result_precision <- cm[4] / sum(cm[4], cm[2])
```



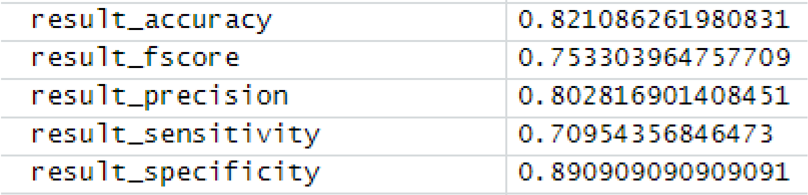
```
result_sensitivity <- cm[4] / sum(cm[4], cm[3])
result_fscore <- (2 * (sensitivity * precision))/(sensitivity + precision)
result_specificity <- cm[1] / sum(cm[1], cm[2])
```

Hasil:

- Pohon keputusan



- Confusion Matrix



result_accuracy	0.821086261980831
result_fscore	0.753303964757709
result_precision	0.802816901408451
result_sensitivity	0.70954356846473
result_specificity	0.890909090909091