Pokok Bahasan IX
Titanic Dataset

Kode Pokok Bahasan: TIK.RPL03.001.006.01

Deskripsi Pokok Bahasan:
Membahas tentang Studi Kasus Data Mining pada R dengan Titanic Dataset

No	Elemen Kompetensi	Indikator Kinerja	Jml Jam	Hal
1	Memahami proses analisis group means	Mampu Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable pada R	1	12
2	Menerapkan decision tree untuk membangun model	Mampu melakukan pemodelan prediksi dengan decision tree	2	15

TUGAS PENDAHULUAN
Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :
1. Menginstal R pada PC masing-masing praktikan.
2. Menginstal R Studio pada PC masing-masing praktikan.

DAFTAR PERTANYAAN
1. Apa tujuan melakukan analisis group means?
   Untuk mengklasifikasi data berdasarkan nilai yang dijadikan sebagai grouping variable.
2. apa keunggulan decision tree?
   Membuat keputusan-keputusan yang semula sangat kompleks menjadi lebih simple dan mengerucut.

TEORI SINGKAT

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include **gender_submission.csv**, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Data Dictionary

VariableDefinitionKey survival Survival 0 = No, 1 = Yes pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd sex Sex Age Age in years sibsp # of siblings / spouses aboard the Titanic parch # of parents / children aboard the Titanic ticket Ticket number fare Passenger fare cabin Cabin number embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

LAB SETUP

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.
2. Menjalankan R Studio.

ELEMEN KOMPETENSI I

Deskripsi:
Memahami proses analisis group means

Kompetensi Dasar:
Mampu Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable pada R

Latihan 1.1.1

Penjelasan Singkat :
Pada latihan ini anda akan diminta untuk Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable. Dan menjelaskan maknanya untuk setiap variable predictor yang bersifat numerik

Langkah-Langkah Praktikum:

Target (class) : *Survival (1=survived; 0= not survived)*
Data :*titanic.csv*,

```
dataku=read.csv("C:/USers/Sectio/Downloads/titanic.csv")
#analisis group mean (predictor numeric)
by(dataku$Age, dataku$Survived, mean)
by(dataku$Fare, dataku$Survived, mean)
```

Output :

```
> #analisis group mean (predictor numeric)
> dataku$Fare = as.numeric(as.character(dataku$Fare))
> str(dataku)
'data.frame':   680 obs. of  8 variables:
 $ PassengerId: int  1 2 3 4 5 7 8 9 10 11 ...
 $ Survived   : int  0 1 1 1 0 0 0 1 1 1 ...
 $ Pclass     : int  3 1 3 1 3 1 3 3 2 3 ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 54 2 27 14 4 ...
 $ SibSp      : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch      : int  0 0 0 0 0 0 1 2 0 1 ...
 $ Fare       : num  7.25 712.83 7.92 53.1 8.05 ...
> by(dataku$Age, dataku$Survived, mean)
dataku$Survived: 0
[1] 30.70074
------------------------------------------------------------------------------------------------
dataku$Survived: 1
[1] 28.20682
> by(dataku$Fare, dataku$Survived, mean)
dataku$Survived: 0
[1] NA
------------------------------------------------------------------------------------------------
dataku$Survived: 1
[1] NA
>
```

Penjelasan :

Pada output praktikum diatas mendapatkan hasil mean umur dari penumpang yang tidak survived dan survived. Rata rata umur yang tidak survived yaitu 30. 13853 sedangkan umur yang survived yaitu 28.40839 dan mean dari fare yang tidak survived yaitu 22,20858 dan yang survived yaitu 48,39541.

Lakukan analisis tabulasi silang (cross tabulation) berdasarkan nilai survival sebagai grouping variable. Jelaskan maknanya untuk setiap variable predictor yang bersifat kategorik

```
analisa cross tabulation (predictor kategorik)
table(dataku$Sex, dataku$Survived)
table(dataku$Pclass, dataku$Survived)
```

Output :

```
> ##analisa cross tabulation (predictor kategorik)
> table(dataku$Sex, dataku$Survived)

          0   1
  female  60 184
  male   346  90
> table(dataku$Pclass, dataku$Survived)

     0   1
  1  62 114
  2  85  79
  3 259  81
>
```

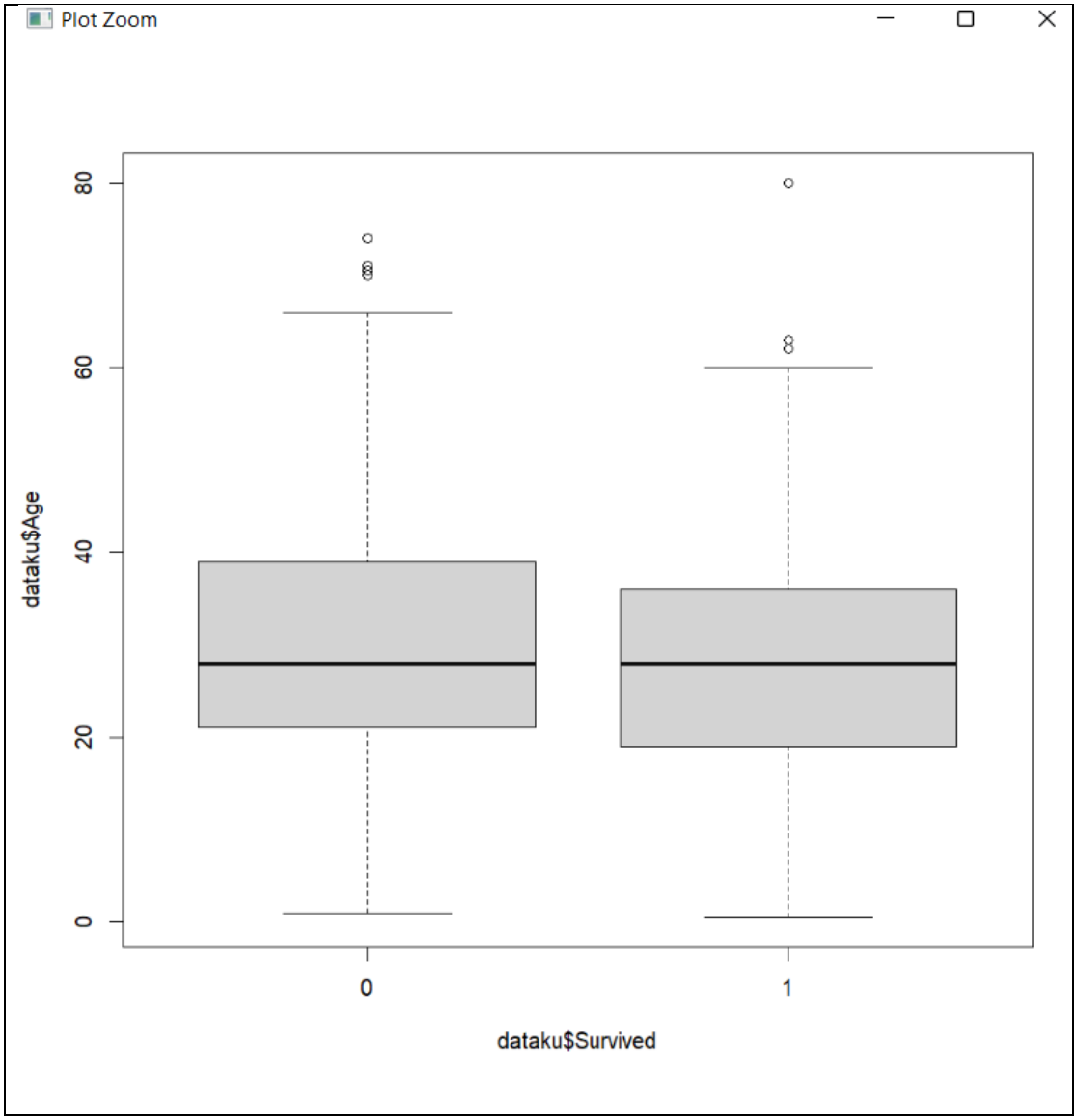
Penjelasan :

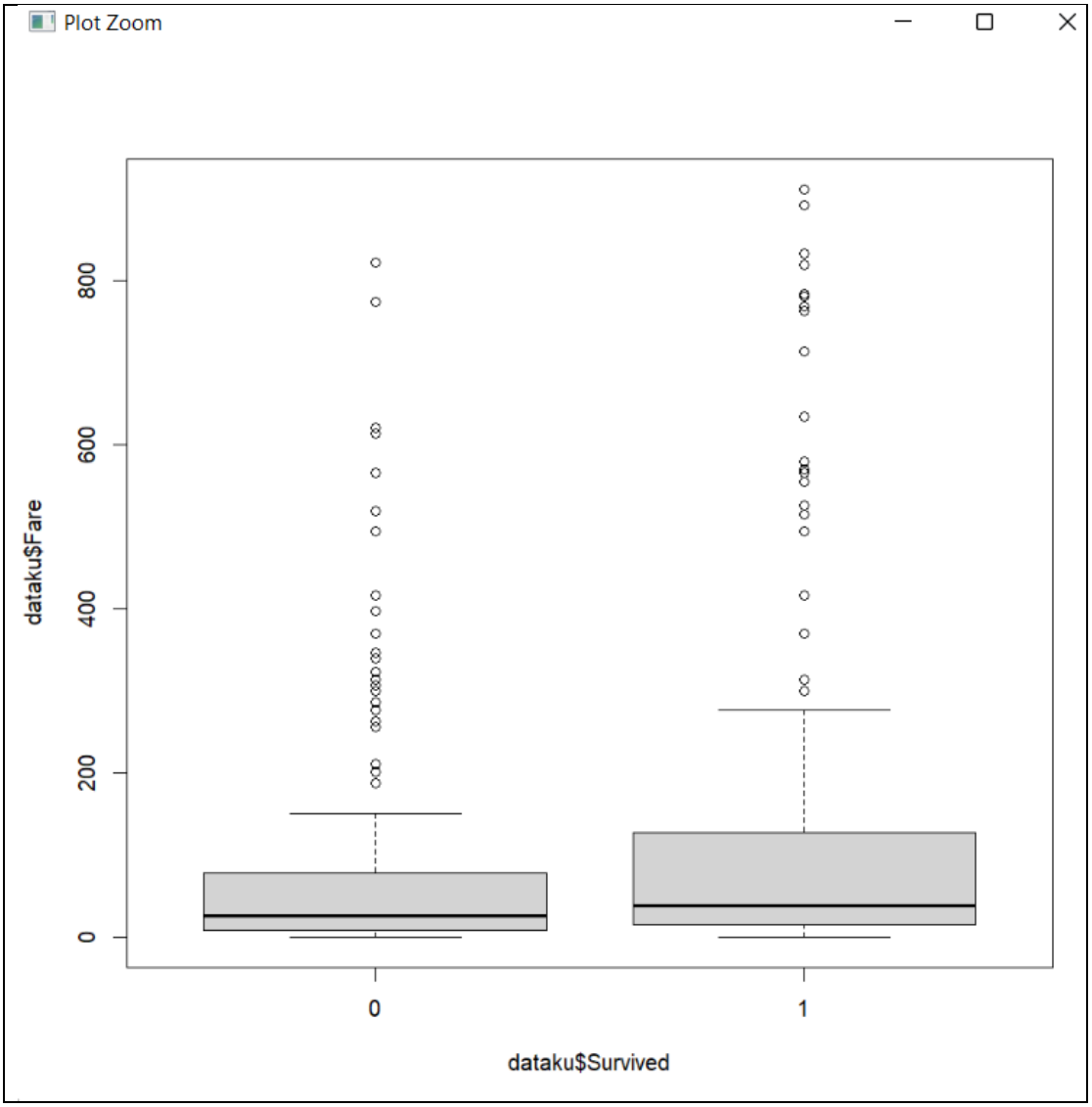
Pada output praktikum diatas mendapatkan hasil jumlah penumpang yang selamat dan tidak selamat berdasarkan gender dan PClass. Pada gender male merupakan jumlah penumpang yang tidak selamat sedangkan pada gender female merupakan jumlah penumpang yang selamat. Pada PClass kelas dengan penumpang yang paling banyak tidak selamat ada di kelas 3 dan penumpang yang selamat paling banyak ada pada kelas 1.

Lakukan analisis boxplot untuk setiap variable predictor yang bersifat numerik. Jelaskan maknanya

```
#boxplot
boxplot(dataku$Age~dataku$Survived)
boxplot(dataku$Fare~dataku$Survived)
```

Output :


Plot Zoom
dataku$Age
0
20
40
60
80
0
1
dataku$Survived



Penjelasan :

Pada output praktikum diatas menampilkan boxplot rata-rata penumpang yang selamat dilihat berdasarkan umur dan berdasarkan fare.

ELEMEN KOMPETENSI II

Deskripsi:
Menerapkan decision tree untuk membangun model

Kompetensi Dasar:
Mampu melakukan pemodelan prediksi dengan decision tree

Latihan 1.2.1

Penjelasan Singkat :
Pada latihan ini anda akan diminta untuk Lakukan analisis group means berdasarkan nilai survival sebagai grouping variable untuk pemodelan prediksi dengan decision tree.

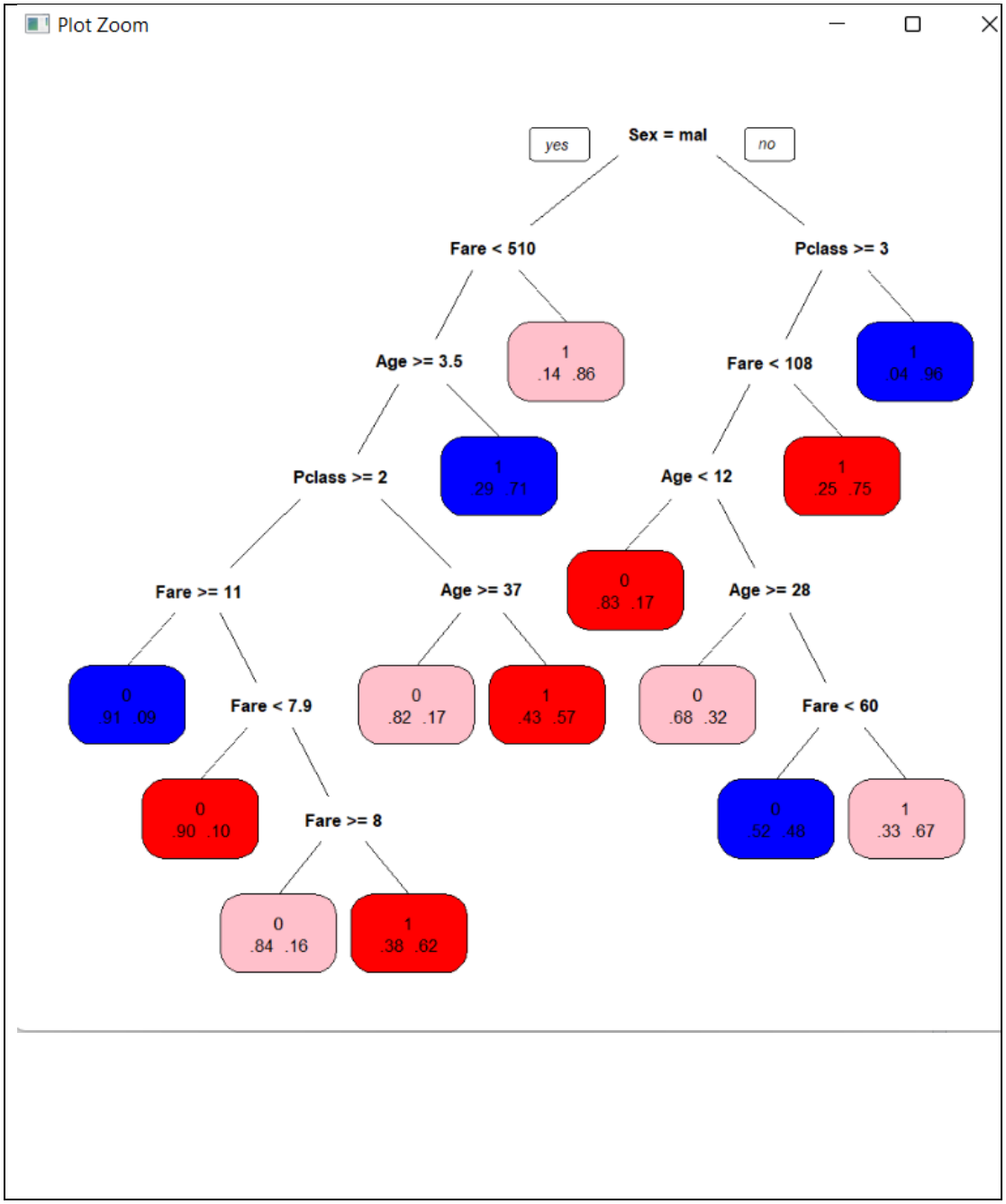
Langkah-Langkah Praktikum:

Buatlah decision tree menggunakan data training untuk membangun model yang dapat digunakan untuk memprediksi kelas survive. Bagi data menjadi 75% untuk training dan 25% untuk uji (tes). Jelaskan maknanya serta nilai confusion matrix-nya

```
library(rpart)
library(rpart.plot)
library(caret)
head(dataku)
sampleTitanic<-sample(1:nrow(dataku),0.75*nrow(dataku))
trainingTitanic<-data.frame(dataku)[sampleTitanic,]
testingTitanic<-data.frame(dataku)[-sampleTitanic,]
pohonTitanic <- rpart(Survived ~ Sex + Age + Fare + Pclass,data=trainingTitanic,
method = "class", control = rpart.control(minsplit = 25, cp = 0))
prp(pohonTitanic,extra=4,box.col=c("pink","red","blue"))
```

Output :

```
> ##ek2
> library(rpart)
> library(rpart.plot)
> library(caret)
> head(dataku)
  PassengerId Survived Pclass    Sex Age SibSp Parch    Fare
1           1        0      3   male  22     1     0   7.250
2           2        1      1 female  38     1     0 712.833
3           3        1      3 female  26     0     0   7.925
4           4        1      1 female  35     1     0  53.100
5           5        0      3   male  35     0     0   8.050
6           7        0      1   male  54     0     0 518.625
> sampleTitanic<-sample(1:nrow(dataku),0.75*nrow(dataku))
> trainingTitanic<-data.frame(dataku)[sampleTitanic,]
> testingTitanic<-data.frame(dataku)[-sampleTitanic,]
> pohonTitanic <- rpart(Survived ~ Sex + Age + Fare + Pclass,data=trainingTitanic, method = "class", control = rpart.contro
(minsplit = 25, cp = 0))
> prp(pohonTitanic,extra=4,box.col=c("pink","red","blue"))
>
```



Penjelasan :

Pada output praktikum diatas menampilkan hasil prediksi respon penumpang yang selamat dan tidak selamat. Prediksi respon 0 merupakan prediksi yang memiliki penumpang yang tidak selamat dengan jumlah 118 sedangkan prediksi respon 1 merupakan prediksi yang memiliki jumlah penumpang yang selamat dengan jumlah 57.

prediksi=predict(pohonTitanic,testingTitanic)

```
pred.respon<- colnames(prediksi)[max.col(prediksi, ties.method = c("random"))]
class=table(pred.respon,testingTitanic$Survived)
class
```

Output :

```
> prediksi=predict(pohonTitanic,testingTitanic)
> pred.respon<- colnames(prediksi)[max.col(prediksi, ties.method = c("random"))]
> class=table(pred.respon,testingTitanic$Survived)
> class

pred.respon  0  1
          0 89 18
          1 21 42
> |
```

Penjelasan :

Membuat decision tree untuk menentukan keputusa-keputusan, kemudian didapat keputusan berdasarkan sex=male yang kemudian dibreakdown lagi berdasarkan fare dan pclass menjadi lebih keputusan yang lebih spesifik untuk menentukan penumpang hidup atau meninggal, kemudian ditemukan bahwa keputusan yaitu penumpang meninggal.

Tugas :
Berdasarkan data yang sama (titanic.csv) dengan 887 observasi (row), lakukan prediksi survival menggunakan teknik naive bayes. Data dibagi menjadi 75% data latih dan 25% data uji.

```
library(e1071)

View(dataku)

sampleTitanic<-sample(1:nrow(dataku),0.75*nrow(dataku))

trainingTitanic<-data.frame(dataku)[sampleTitanic,]

testingTitanic<-data.frame(dataku)[-sampleTitanic,]

BayesTitanicModel<-naiveBayes(as.factor(Survived)~., trainingTitanic)

BayesPrediction<-predict(BayesTitanicModel, testingTitanic)

summary(BayesPrediction)

output<-data.frame(testingTitanic, BayesPrediction)

str(output)

colnames(output)<-cbind("Survived")
```

```
Loading required package: RMySQL
Loading required package: DBI
> dataku=read.csv("D:/File Kuliah Semester 5/Data Analitik/Prak-7/titanic.csv", sep = ";")
Warning message:
package 'RMySQL' was built under R version 4.2.2
> library(e1071)
> View(dataku)
> sampleTitanic<-sample(1:nrow(dataku),0.75*nrow(dataku))
> trainingTitanic<-data.frame(dataku)[sampleTitanic,]
> testingTitanic<-data.frame(dataku)[-sampleTitanic,]
> BayesTitanicModel<-naiveBayes(as.factor(Survived)~., trainingTitanic)
> BayesPrediction<-predict(BayesTitanicModel, testingTitanic)
> summary(BayesPrediction)
 0  1
94 76
> output<-data.frame(testingTitanic, BayesPrediction)
> str(output)
'data.frame':   170 obs. of  9 variables:
 $ PassengerId    : int  2 5 9 11 13 15 21 24 35 38 ...
 $ Survived       : int  1 0 1 1 0 0 0 1 0 0 ...
 $ Pclass         : int  1 3 3 3 3 3 2 1 1 3 ...
 $ Sex            : chr  "female" "male" "female" "female" ...
 $ Age            : num  38 35 27 4 20 14 35 28 28 21 ...
 $ SibSp          : int  1 0 0 1 0 0 0 0 1 0 ...
 $ Parch          : int  0 0 2 1 0 0 0 0 0 0 ...
 $ Fare           : chr  "712.833" "8.05" "111.333" "16.7" ...
 $ BayesPrediction: Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 2 2 1 ...
> colnames(output)<-cbind("Survived")
>
```

CEK LIST

Elemen Kompetensi	No Latihan	Penyelesaian	
		Selesai	Tidak selesai
1	1.1.1	✓	
2	1.2.1	✓	

FORM UMPAN BALIK

Elemen Kompetensi	**Tingkat Kesulitan**	**Tingkat Ketertarikan**	**Waktu Penyelesaian dalam menit**
Memahami proses analisis group means	☐ Sangat Mudah ☐ Mudah ☐ ✓ Biasa ☐ Sulit ☐ Sangat Sulit	☐ Tidak Tertarik ☐ Cukup Tertarik ☐ Tertarik ☐ ✓ Sangat Tertarik	30
Menerapkan decision tree untuk membangun model	☐ Sangat Mudah ☐ Mudah ☐ ✓ Biasa ☐ Sulit ☐ Sangat Sulit	☐ Tidak Tertarik ☐ Cukup Tertarik ☐ Tertarik ☐ ✓ Sangat Tertarik	30