**Praktikum 4**

**Prediksi Hasil Menggunakan Decision Tree**

Tugas #1

Gunakan Titanic dataset yang terbagi atas training set (train.csv) dan test set (test.csv). Lakukan analisis frekuensi dan bangun model decision tree berdasarkan data training untuk kemudian memprediksi keselamatan penumpang (survived) berdasarkan data test. Tuliskan output pada kotak yang tersedia berikut interpretasinya.

> setwd("F:/Kuliah Data Mining gasal 1819/Kaggle/Titanic")

> train <- read.csv('train.csv')

> test <- read.csv('test.csv')

> table(train$Survived)

|  |
| --- |
| 0 1  549 342 |

> prop.table(table(train$Survived))

|  |
| --- |
| 0 1  0.6161616 0.3838384 |

> prop.table(table(train$Sex, train$Survived), 1)

|  |
| --- |
| 0 1  female 0.2579618 0.7420382  male 0.8110919 0.1889081 |

> fit <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data = train,method = "class")

> plot(fit)

> text(fit)

|  |
| --- |
|  |

> Prediction <- predict(fit, test, type = "class")

> submit <- data.frame(PassengerId = test$PassengerId, Survived = Prediction)

> write.csv(submit, file = 'decisiontree.csv', row.names = FALSE)

|  |
| --- |
|  |

Tugas #2

Gunakan data training dari Titanic dataset di atas dan pilah kembali menjadi 75% datalatih dan 25% datauji secara random. Gunakan library(rpart), library(rpart.plot) dan library(Caret) untuk membangun model decision tree dan membuat confusion matrix terkait prediksi keselamatan penumpang (survived). Tuliskan script R yang digunakan, interpretasikan modelnya dan nilai akurasi yang didapatkan.

|  |
| --- |
| > train=read.delim("clipboard")  > library(rpart)  > library(rpart.plot)  > library(caret)  > sampel\_train<-sample(1:nrow(train),0.75\*nrow(dataprak4))  > datalatih<-data.frame(train)[sampel\_train,]  > datauji<-data.frame(train)[-sampel\_train,]  > pohon\_train = rpart(Survived~Pclass+Sex+Age, data=datalatih,method="class",control = rpart.control(minsplit=10,cp=0))  > prp(pohon\_train,extra=4)    > prediksi\_titanic=predict(pohon\_train,datauji)  > pred.respon<-colnames(prediksi\_titanic)[max.col(prediksi\_titanic,ties.method = c("random"))]  > library(e1071)  Warning message:  package ‘e1071’ was built under R version 3.4.4  > class=table(pred.respon,datauji$Survived)  > confusionMatrix(class)  Confusion Matrix and Statistics    pred.respon 0 1  0 104 33  1 18 68    Accuracy : 0.7713  95% CI : (0.7105, 0.8247)  No Information Rate : 0.5471  P-Value [Acc > NIR] : 2.938e-12    Kappa : 0.5325  Mcnemar's Test P-Value : 0.04995    Sensitivity : 0.8525  Specificity : 0.6733  Pos Pred Value : 0.7591  Neg Pred Value : 0.7907  Prevalence : 0.5471  Detection Rate : 0.4664  Detection Prevalence : 0.6143  Balanced Accuracy : 0.7629    'Positive' Class : 0 |

Tugas #3

Gunakan iris dataset dan pilah menjadi 75% datalatih dan 25% datauji secara random. Gunakan library(rpart), library(rpart.plot) dan library(Caret) untuk membangun model decision tree dan membuat confusion matrix terkait prediksi jenis species. Tuliskan script R yang digunakan, interpretasikan modelnya dan nilai akurasi yang didapatkan.

|  |
| --- |
| > library(rpart)  > library(rpart.plot)  > library(caret)  > sampel\_iris<-sample(1:nrow(iris),0.75\*nrow(iris))  > datalatih<-data.frame(iris)[sampel\_iris,]  > datauji<-data.frame(iris)[-sampel\_iris,]  > pohon\_train = rpart(Species~Sepal.Width+Sepal.Length+Petal.Length+Petal.Width, data=iris,method="class",control = rpart.control(minsplit=10,cp=0))  > prp(pohon\_train,extra=4)    > prediksi\_iris=predict(pohon\_train,datauji)  > pred.respon<-colnames(prediksi\_iris)[max.col(prediksi\_iris,ties.method = c("random"))]  > library(e1071)  > class=table(pred.respon,datauji$Species)  > confusionMatrix(class)  Confusion Matrix and Statistics    pred.respon setosa versicolor virginica  setosa 16 0 0  versicolor 0 11 0  virginica 0 0 11  Overall Statistics    Accuracy : 1  95% CI : (0.9075, 1)  No Information Rate : 0.4211  P-Value [Acc > NIR] : 5.306e-15    Kappa : 1  Mcnemar's Test P-Value : NA  Statistics by Class:  Class: setosa Class: versicolor Class: virginica  Sensitivity 1.0000 1.0000 1.0000  Specificity 1.0000 1.0000 1.0000  Pos Pred Value 1.0000 1.0000 1.0000  Neg Pred Value 1.0000 1.0000 1.0000  Prevalence 0.4211 0.2895 0.2895  Detection Rate 0.4211 0.2895 0.2895  Detection Prevalence 0.4211 0.2895 0.2895  Balanced Accuracy 1.0000 1.0000 1.0000 |

**Lampiran : Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| 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 |

**Variable Notes**

**pclass**: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower  
  
**age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
**sibsp**: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
**parch**: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

Sumber :

<https://rpubs.com/zscore/titanicml>

<https://www.kaggle.com/c/titanic/data>

<http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/>

<https://rpubs.com/abhaypadda/k-nn-decision-tree-on-IRIS-dataset>