**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>

1. Dimensi data

Text

Description automatically generated

Mempunyai 12 variable dan 891 jumlah data

1. Struktur Data

Text

Description automatically generated

1. Ringkasan Data

A screenshot of a computer

Description automatically generated with medium confidence

1. Data Paling Atas

Text

Description automatically generated

1. Data Paling Bawah

Text

Description automatically generated

# **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:



1. Distribusi class mengunakan pie chart

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

survivedTable = table(titanic$Survived)

survivedTable

A picture containing graphical user interface

Description automatically generated

par(mar = c(0, 0, 0, 0), oma = c(0, 0, 0, 0))

pie(survivedTable,labels=c("Died","Survived"))

Chart, pie chart

Description automatically generated

1. 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

Text

Description automatically generated

* Pie Chart

Chart

Description automatically generated

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

Diagram

Description automatically generated

* Confusion Matrix

Table

Description automatically generated