

**TUGAS BESAR DATA MINING**  
**PENDEKATAN DATA MINING UNTUK SURVEI KEBAHAGIAAN**  
**PENDUDUK SOMERVILLE MENGGUNAKAN**  
**METODE KLASIFIKASI**



Disusun oleh:

<b>Ardita Hardi</b>	<b>: 3311801029</b>
<b>Maulita Rizchita Putri</b>	<b>: 3311801025</b>
<b>Rizky Indriati</b>	<b>: 3311801005</b>

Teknik Informatika 3 A

Disusun untuk memenuhi salah satu syarat kelulusan  
mata kuliah IF317 Data Mining

**PROGRAM STUDI TEKNIK INFORMATIKA**  
**JURUSAN TEKNIK INFORMATIKA**  
**POLITEKNIK NEGERI BATAM**  
**BATAM**  
**2019**

**HALAMAN PENGESAHAN**

**PENDEKATAN DATA MINING UNTUK SURVEI KEBAHAGIAAN  
PENDUDUK SOMERVILLE MENGGUNAKAN  
METODE KLASIFIKASI**

**Disusun oleh:**

**Ardita Hardi : 3311801029**

**Maulita Rizchita Putri : 3311801025**

**Rizky Indriati : 3311801005**

Teknik Informatika 3 A

Batam, Desember 2019

Disetujui dan disahkan oleh:

Dosen pengajar,

**MUHAMMAD NASHRULLAH, SST., M.Sc**

**NIK/NIP. 0009049101**

## DAFTAR ISI

Halaman Judul .....	1
Halaman Pengesahan .....	2
Daftar Isi .....	3
BAB I    PENDAHULUAN .....	4
1.1      Instruksi.....	4
1.2      Setoran .....	4
BAB II   DATA MINING.....	5
2.1      Link Repository .....	5
2.2      Penjelasan Dataset .....	5
2.3      Proses Data Mining.....	6
2.4      Hasil Data Mining.....	14
BAB III  KESIMPULAN DAN SARAN.....	15
3.1      Kesimpulan .....	15
3.2      Saran .....	15
Daftar Pustaka .....	16

# **BAB I**

## **PENDAHULUAN**

### **1.1 Instruksi**

1. Mahasiswa diminta untuk membentuk sebuah tim, terdiri 3 orang (jika mahasiswa tidak mendapatkan tim, boleh maksimal 4 orang).
2. Diskusi dengan tim mengenai dataset dan metode yang ingin digunakan.
  - <https://archive.ics.uci.edu/ml/datasets.php>
  - <https://www.kaggle.com/datasets>

### **1.2 Setoran**

1. Upload dokumen pada learning-if.polibatam.ac.id, paling lambat 4 Desember 2019
2. Mahasiswa diwajibkan memiliki akun github/gitlab/bitbucket
  - a. Buatlah repository baru, beri nama nim\_datamining\_polibatam
  - b. Upload dokumen pada repository tersebut
3. Dokumen
  - a. Dataset
  - b. Full code
  - c. File presentasi, terdapat informasi mengenai:
    - Judul dan link github project
    - Dataset
    - Data mahasiswa seperti nim, nama, link github (akun mahasiswa)
    - Proses data mining
    - Hasil data mining
    - Referensi

\*Presentasi akan dilakukan pada saat UAS Praktikum

## BAB II

### DATA MINING

#### 2.1 Link Repository

- Berikut adalah link repository kami:

[https://github.com/arditahrd/029-025-005\\_DataMining\\_Polibatam](https://github.com/arditahrd/029-025-005_DataMining_Polibatam)

#### 2.2 Penjelasan Dataset

- Source Link : <https://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey>

 Somerville Happiness Survey		Classification	Integer	143	7	2018
---	--	----------------	---------	-----	---	------

##### Informasi Kumpulan Data:

Dataset dari Kota Somerville yang melakukan survei kepada penduduk Somerville untuk menilai kebahagiaan pribadi dan kepuasan mereka terhadap layanan Kota. Kasus pembelajaran yang terawasi dengan menggunakan *Receiver Operating Characteristic* (ROC) untuk memilih set minimal atribut yang mempertahankan atau meningkatkan prediktabilitas data.

##### Informasi Atribut:

D = Atribut keputusan (D) dengan nilai Happy dan Unhappy

X1 = ketersediaan informasi tentang layanan kota

X2 = biaya perumahan

X3 = kualitas keseluruhan sekolah umum

X4 = kepercayaan Anda pada polisi setempat

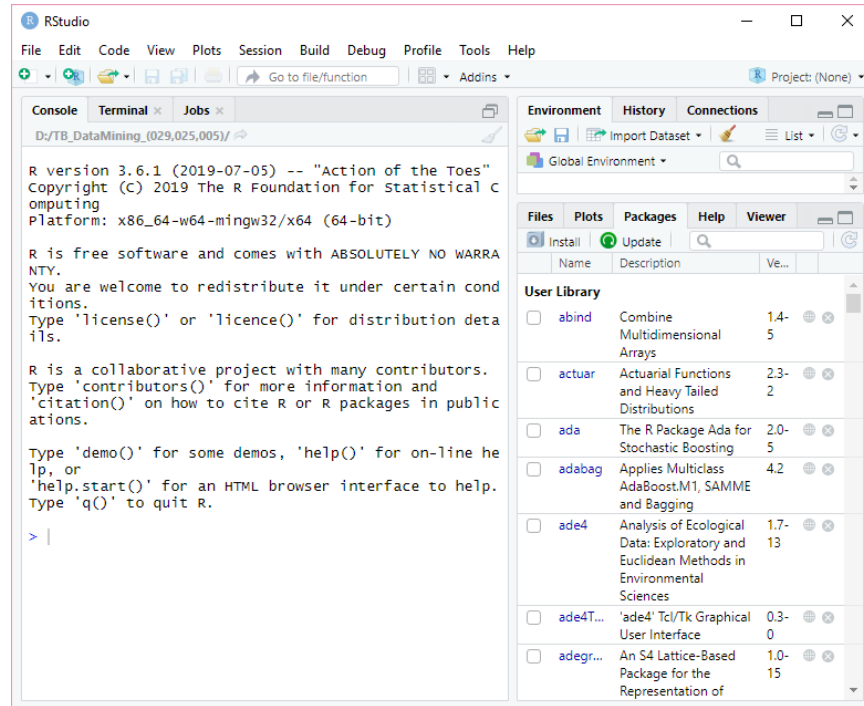
X5 = pemeliharaan jalan dan trotoar

X6 = ketersediaan acara komunitas sosial

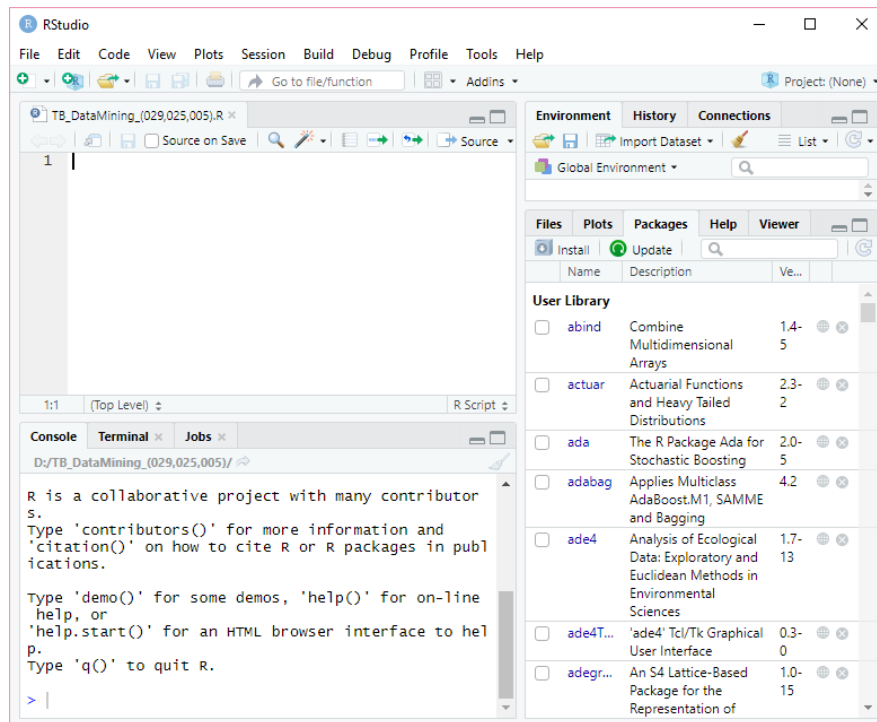
Atribut X1 hingga X6 memiliki nilai 1 hingga 5

## 2.3 Proses Data Mining

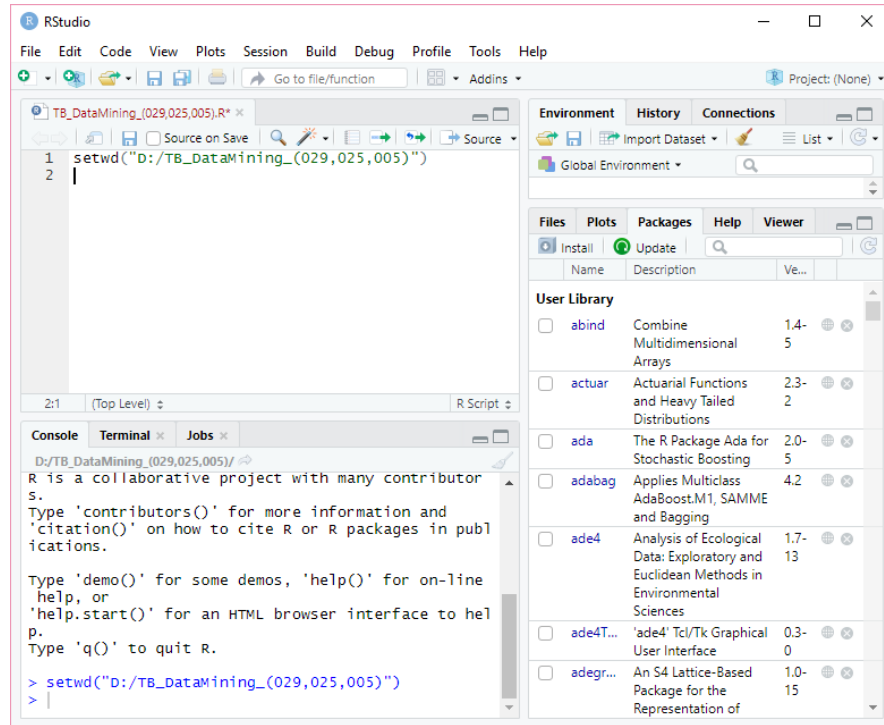
### 1. Buka Rstudio



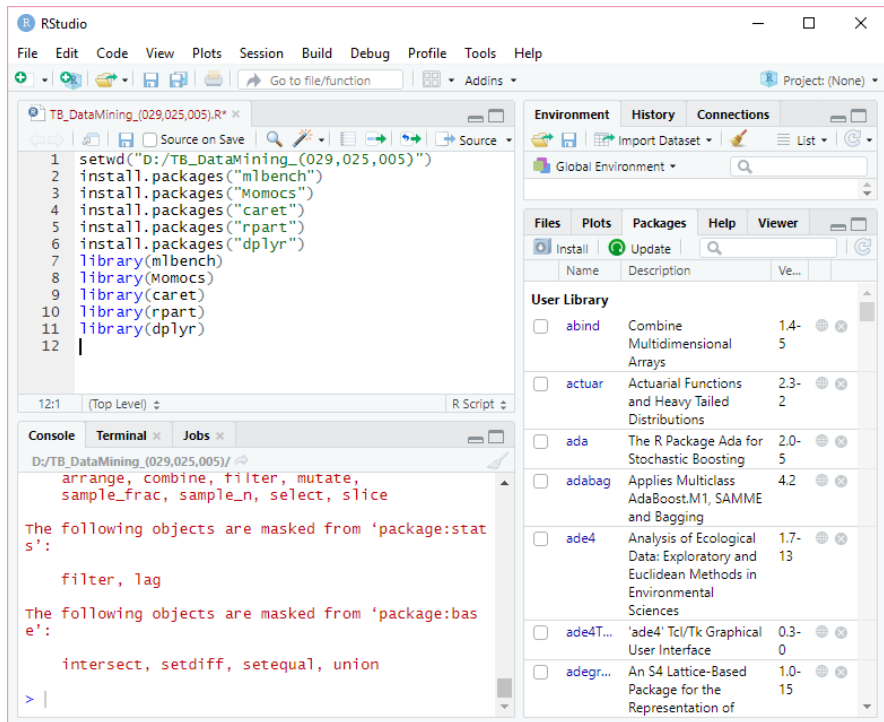
### 2. Membuat Script Baru



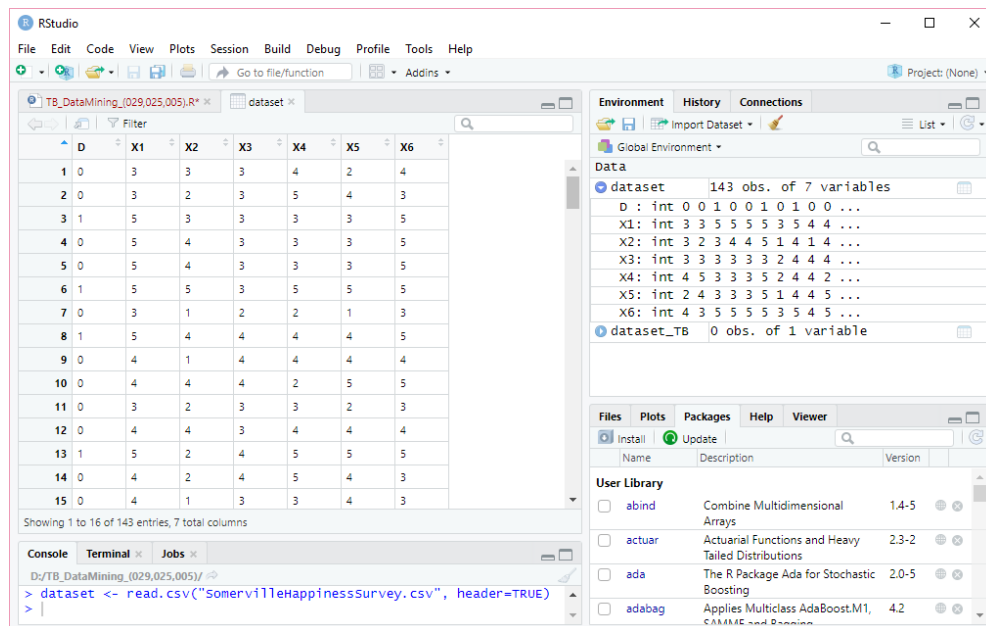
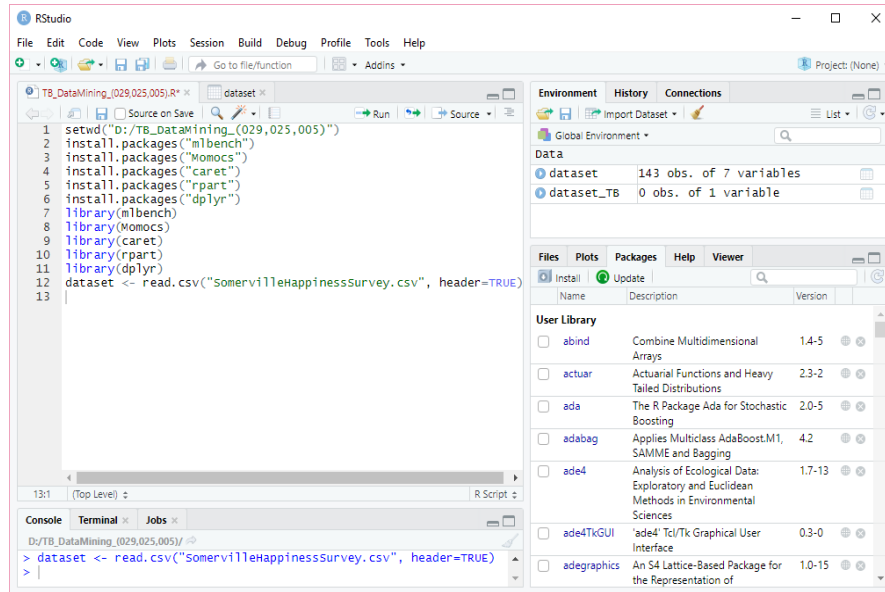
### 3. Mengatur lokasi directory



### 4. Instalasi package dan gunakan package

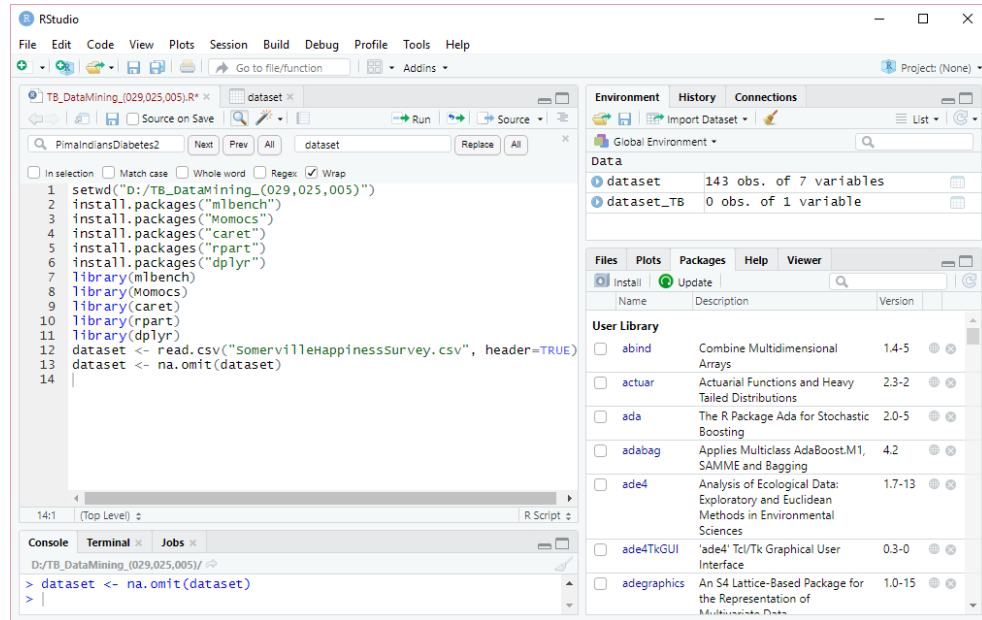


## 5. Membaca data dengan format csv

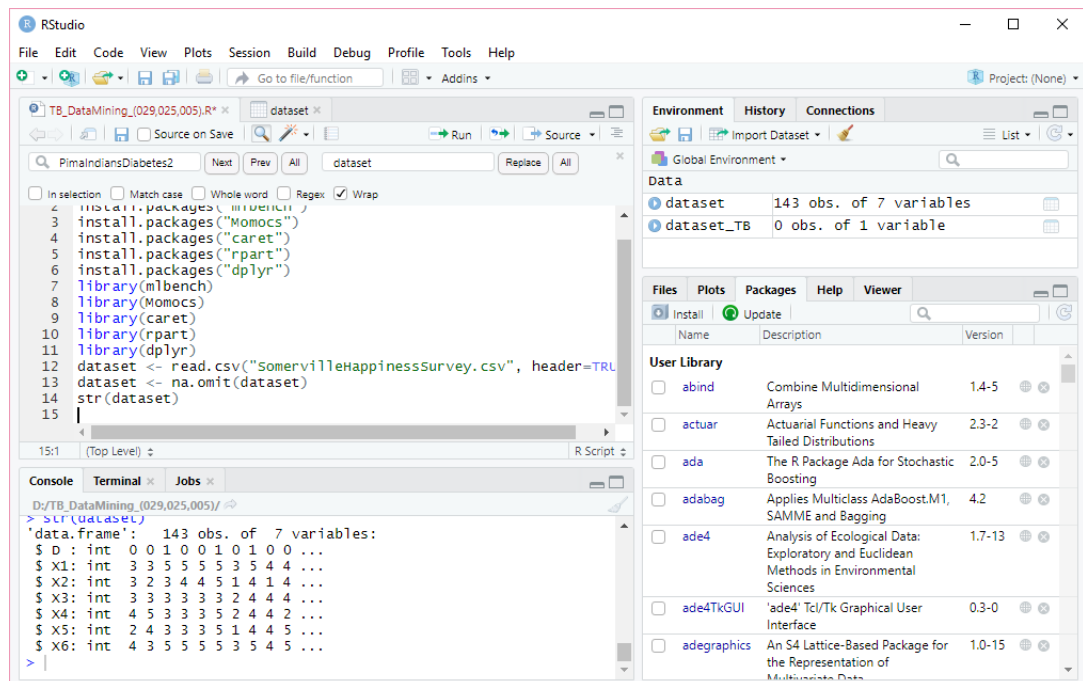




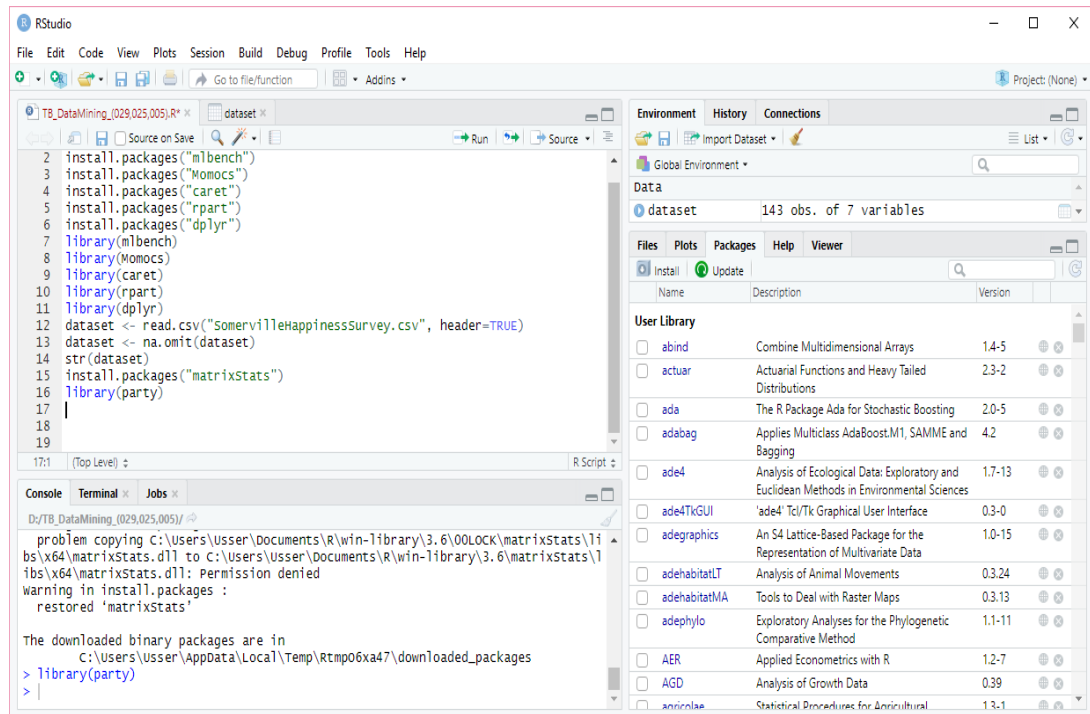
## 6. Menghilangkan baris yang memiliki nilai NA



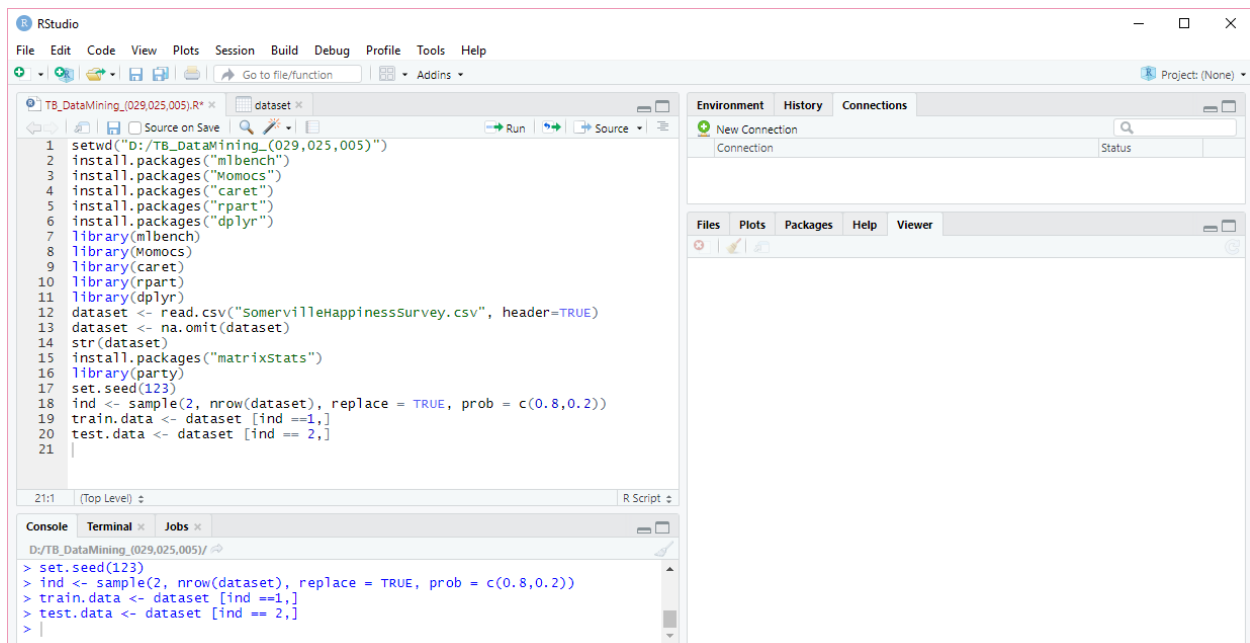
## 7. Melihat struktur pada sebuah object



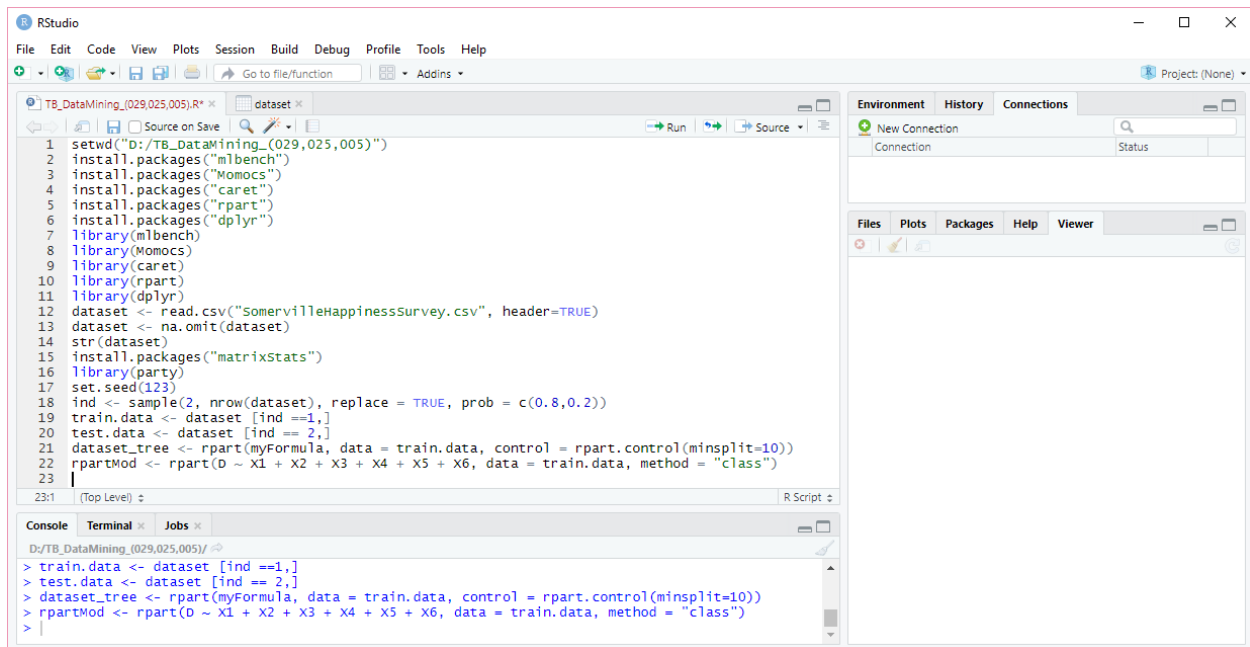
## 8. Instalasi package yang belum diinstall dan gunakan package



## 9. Membagi data menjadi data training dan data testing



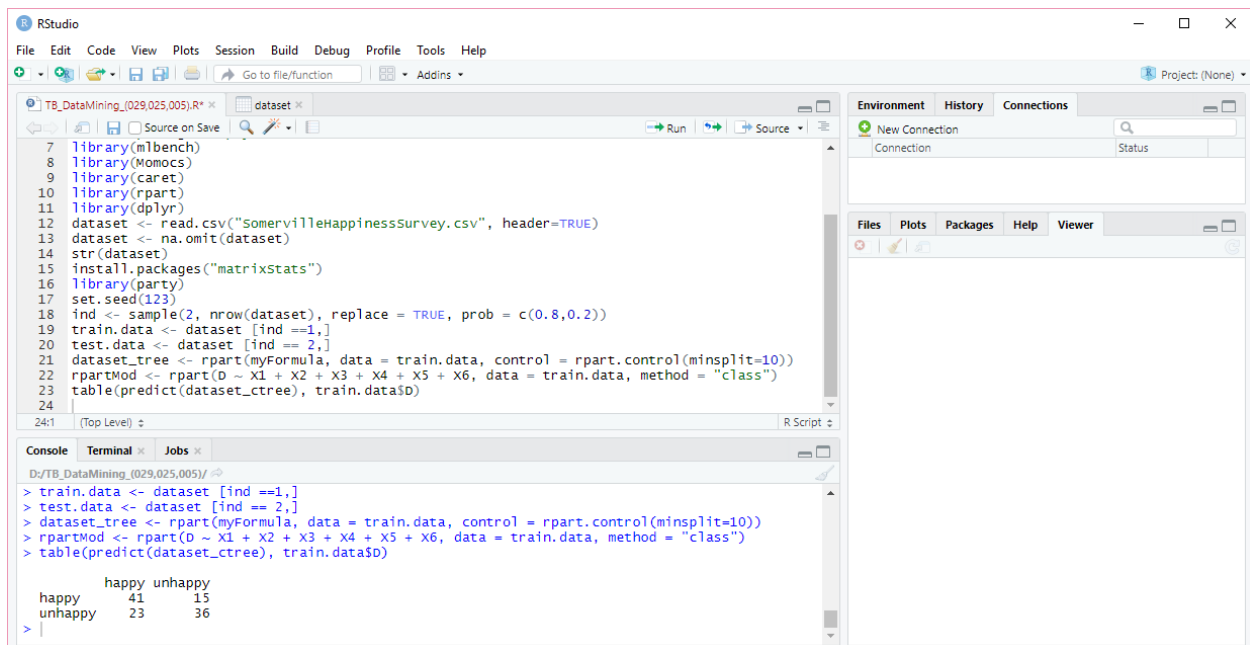
## 10. Membangun model



```
1 setwd("D:/TB_DataMining_029,025,005/R")
2 install.packages("mlbench")
3 install.packages("Momocs")
4 install.packages("caret")
5 install.packages("rpart")
6 install.packages("dplyr")
7 library(mlbench)
8 library(Momocs)
9 library(caret)
10 library(rpart)
11 library(dplyr)
12 dataset <- read.csv("SomervilleHappinessSurvey.csv", header=TRUE)
13 dataset <- na.omit(dataset)
14 str(dataset)
15 install.packages("matrixStats")
16 library(party)
17 set.seed(123)
18 ind <- sample(2, nrow(dataset), replace = TRUE, prob = c(0.8,0.2))
19 train.data <- dataset[ind == 1,]
20 test.data <- dataset[ind == 2,]
21 dataset_tree <- rpart(myFormula, data = train.data, control = rpart.control(minsplit=10))
22 rpartMod <- rpart(D ~ X1 + X2 + X3 + X4 + X5 + X6, data = train.data, method = "class")
23
```

```
> train.data <- dataset[ind == 1,]
> test.data <- dataset[ind == 2,]
> dataset_tree <- rpart(myFormula, data = train.data, control = rpart.control(minsplit=10))
> rpartMod <- rpart(D ~ X1 + X2 + X3 + X4 + X5 + X6, data = train.data, method = "class")
>
```

## 11. Memprediksi Data

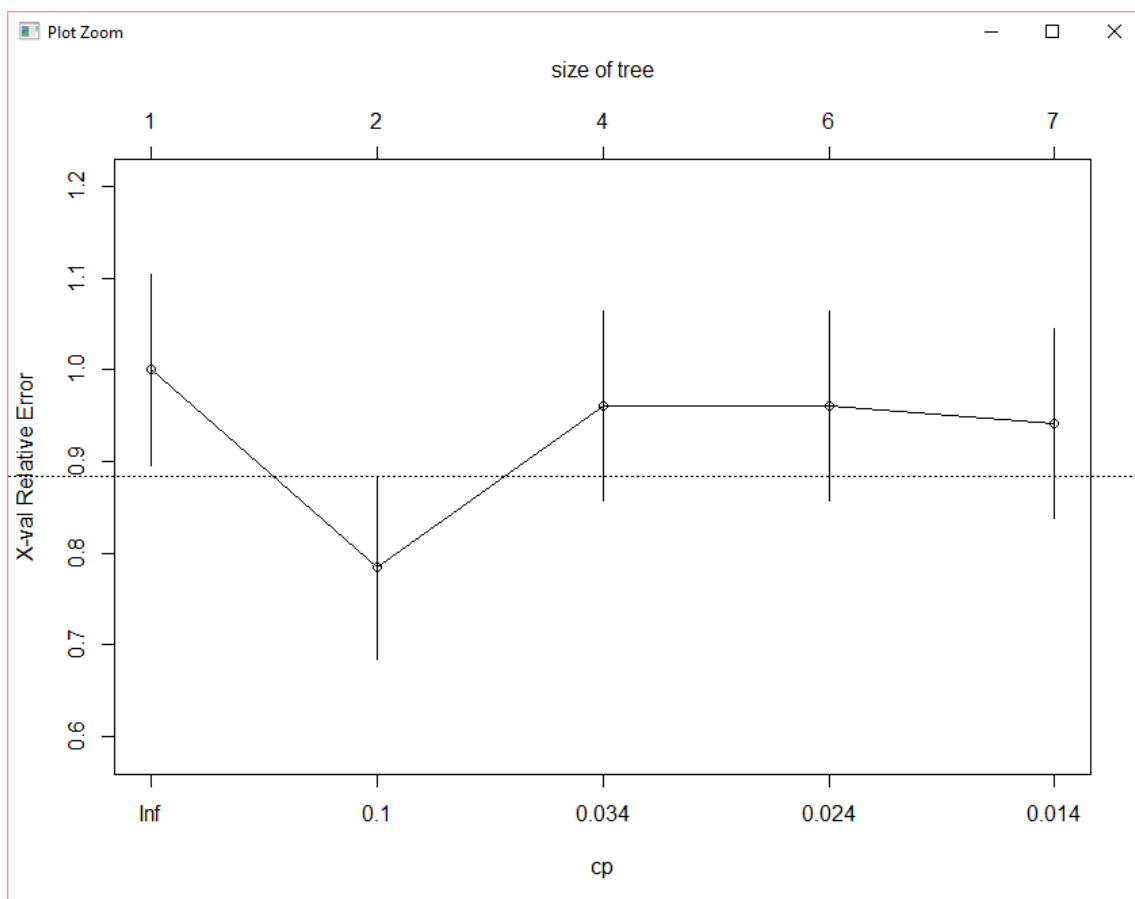
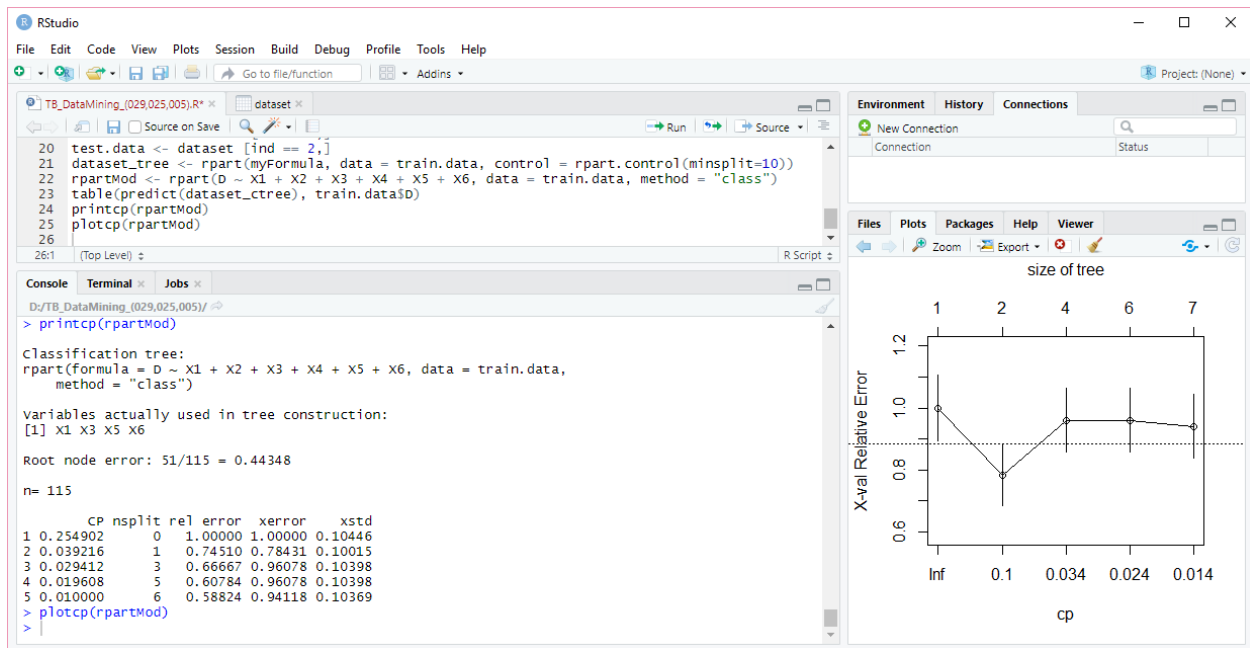


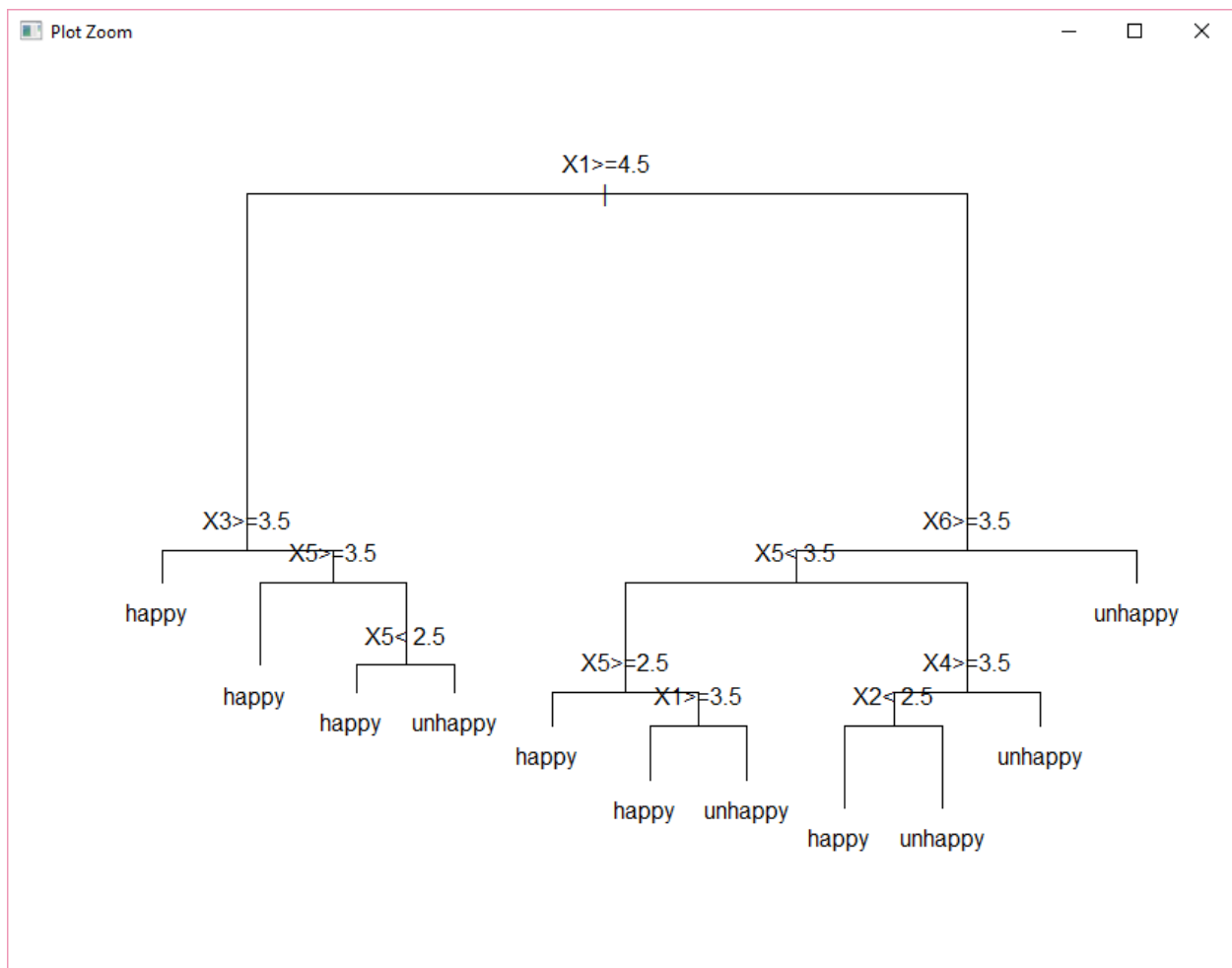
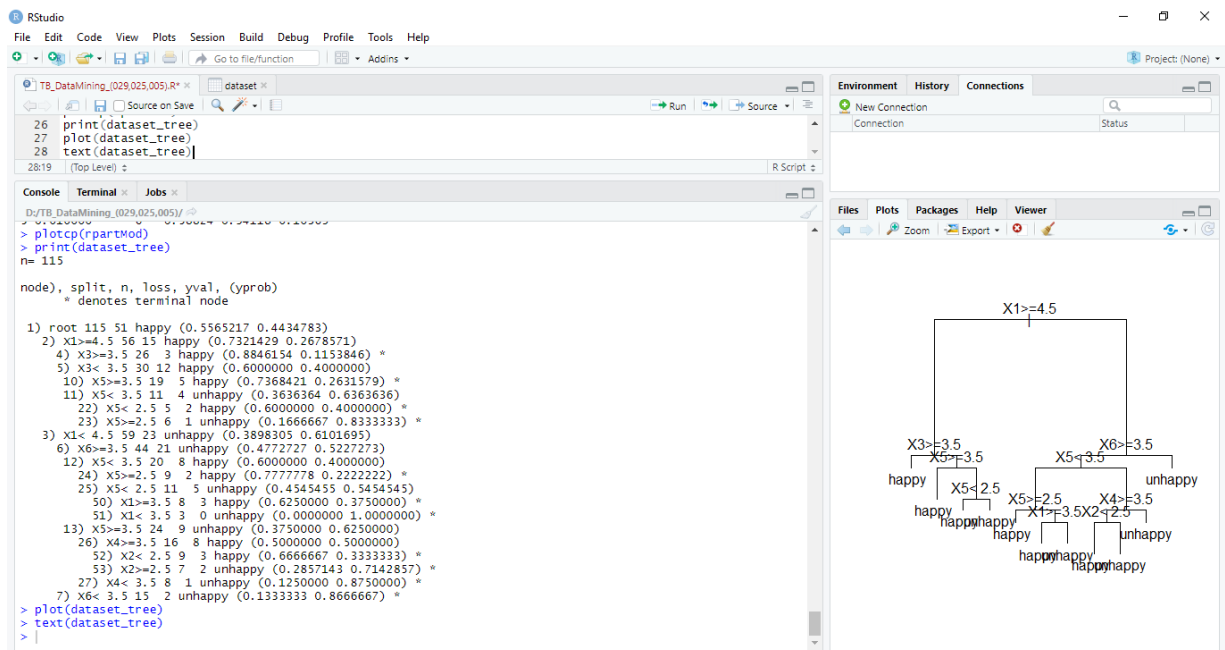
```
7 library(mlbench)
8 library(Momocs)
9 library(caret)
10 library(rpart)
11 library(dplyr)
12 dataset <- read.csv("SomervilleHappinessSurvey.csv", header=TRUE)
13 dataset <- na.omit(dataset)
14 str(dataset)
15 install.packages("matrixStats")
16 library(party)
17 set.seed(123)
18 ind <- sample(2, nrow(dataset), replace = TRUE, prob = c(0.8,0.2))
19 train.data <- dataset[ind == 1,]
20 test.data <- dataset[ind == 2,]
21 dataset_tree <- rpart(myFormula, data = train.data, control = rpart.control(minsplit=10))
22 rpartMod <- rpart(D ~ X1 + X2 + X3 + X4 + X5 + X6, data = train.data, method = "class")
23 table(predict(dataset_tree), train.data$D)
24
```

```
> train.data <- dataset[ind == 1,]
> test.data <- dataset[ind == 2,]
> dataset_tree <- rpart(myFormula, data = train.data, control = rpart.control(minsplit=10))
> rpartMod <- rpart(D ~ X1 + X2 + X3 + X4 + X5 + X6, data = train.data, method = "class")
> table(predict(dataset_tree), train.data$D)

      happy unhappy
happy    41      15
unhappy  23      36
>
```

## 12. Menampilkan data





## 2.3 Hasil Data Mining

1. Ditemukan 143 data dengan 7 variabel
2. Penduduk dengan kategori Happy yang benar diprediksi sebanyak 41 data
3. Penduduk dengan kategori Unhappy yang benar diprediksi sebanyak 36 data
4. Penduduk dengan kategori Happy yang salah diprediksi menjadi Unhappy sebanyak 23 data
5. Penduduk dengan Kategori Unhappy yang salah diprediksi menjadi Happy sebanyak 15 data
6. Maka, tingkat kepercayaan (akurasi) Dataset Ctree adalah sebesar  $(41+36)/(41+36+23+15) = 66.96\%$
7. Hasil pohon keputusan dataset:
  - a.  $X1 \geq 4.5$  jika TRUE maka " $X3 \geq 3.5$ ", jika FALSE maka " $X6 \geq 3.5$ "
  - b.  $X3 \geq 3.5$  jika TRUE maka "happy", jika FALSE maka " $X5 \geq 3.5$ "
  - c.  $X6 \geq 3.5$  jika TRUE maka " $X5 < 3.5$ ", jika FALSE maka "unhappy"
  - d.  $X5 \geq 3.5$  jika TRUE maka "happy", jika FALSE maka " $X5 < 2.5$ "
  - e.  $X5 < 3.5$  jika TRUE maka " $X5 \geq 2.5$ ", jika FALSE maka " $X4 \geq 3.5$ "
  - f.  $X5 < 2.5$  jika TRUE maka "happy", jika FALSE maka "unhappy"
  - g.  $X5 \geq 2.5$  jika TRUE maka "happy", jika FALSE maka " $X1 \geq 3.5$ "
  - h.  $X4 \geq 3.5$  jika TRUE maka " $X2 < 2.5$ ", jika FALSE maka "unhappy"
  - i.  $X1 \geq 3.5$  jika TRUE maka "happy", jika FALSE maka "unhappy"
  - j.  $X2 < 2.5$  jika TRUE maka "happy", jika FALSE maka "unhappy"

## **BAB III**

### **KESIMPULAN DAN SARAN**

#### **3.1 Kesimpulan**

Sesuai dengan laporan yang disajikan, kami membahas tentang dataset dari Kota Somerville yang melakukan survei kepada penduduk Somerville untuk menilai kebahagiaan pribadi dan kepuasan mereka terhadap layanan Kota. Survei tersebut terkait dengan ketersediaan informasi tentang layanan kota, biaya perumahan, kualitas keseluruhan sekolah umum, kepercayaan penduduk pada polisi setempat, pemeliharaan jalan dan trotoar, serta ketersediaan acara komunitas sosial. Data survei menggunakan *Receiver Operating Characteristic* (ROC) untuk memilih set minimal atribut yang mempertahankan atau meningkatkan prediktabilitas data dan juga menggunakan metode klasifikasi yang bertujuan untuk memudahkan dalam proses mendeskripsikan data melalui pohon keputusan yang telah dihasilkan.

#### **3.2 Saran**

Menyadari bahwa laporan ini masih jauh dari kata sempurna, kedepannya kami akan lebih fokus dalam menjelaskan tentang laporan dengan sumber - sumber yang benar dan dapat di pertanggung jawabkan. Oleh karena itu, kritik dan saran yang membangun dari para pembaca sangat kami harapkan sebagai bahan evaluasi untuk kedepannya.

## DAFTAR PUSTAKA

- [1] Ramadhani, Dian Puteri. 2016. *A Data Mining Approach To Diagnose Breast Cancer Using Classification Methods Machine Learning Algorithm*.  
<https://dianrdntelkomuniversity.wordpress.com/2016/11/02/a-data-mining-approach-to-diagnose-breast-cancer-using-classification-methods-machine-learning-algorithm/>.
- [2] Hidayatullah, Ade Ihsan. 2017. *Klasifikasi Menggunakan Decision Tree di R*.  
<http://www.masterstatistik.com/2017/10/klasifikasi-menggunakan-decision-tree.html>.