

# Tugas Besar Data Mining

## Klasifikasi C5.0



Menerapkan algoritma C5.0 dengan R dan R Studio





# Nama Kelompok;

Evriska Dayanti - 3311901003

Ayusda Renjani - 3311901019

Gabriella Joice Sitompul - 3311901029

## —Klasifikasi C5.0

Pada tugas besar ini kelompok kami menggunakan metode klasifikasi C5.0 dan memakai dataset Cryotherapy Dataset yang memiliki Kumpulan data memberikan informasi yang berkaitan dengan pasien, yang karakteristiknya seperti Jumlah Kutil, Area kutil, jenis kelamin dan usia, dll. Digunakan untuk menentukan tingkat ekstremitas kanker, yaitu 0 jinak atau 1 ganas

# Dataset: Cryotherapy.csv Terdiri dari 90 data 7 variabel

	A	B	C	D	E	F
1	sex,"age","Time","Number_of_Warts","Type","Area","Result_of_Treatment"					
2	1,"35","12","5","1","100","0"					
3	1,"29","7","5","1","96","1"					
4	1,"50","8","1","3","132","0"					
5	1,"32","11.75","7","3","750","0"					
6	1,"67","9.25","1","1","42","0"					
7	1,"41","8","2","2","20","1"					
8	1,"36","11","2","1","8","0"					
9	1,"59","3.5","3","3","20","0"					
10	1,"20","4.5","12","1","6","1"					
11	2,"34","11.25","3","3","150","0"					
12	2,"21","10.75","5","1","35","0"					
13	2,"15","6","2","1","30","1"					
14	2,"15","2","3","1","4","1"					
15	2,"15","3.75","2","3","70","1"					
16	2,"17","11","2","1","10","0"					
17	2,"17","5.25","3","1","63","1"					
18	2,"23","11.75","12","3","72","0"					
19	2,"27","8.75","2","1","6","0"					
20	2,"15","4.25","1","1","6","1"					
21	2,"18","5.75","1","1","80","1"					
22	1,"22","5.5","2","1","70","1"					
23	2,"16","8.5","1","2","60","1"					
24	1,"28","4.75","3","1","100","1"					
25	2,"40","9.75","1","2","80","0"					

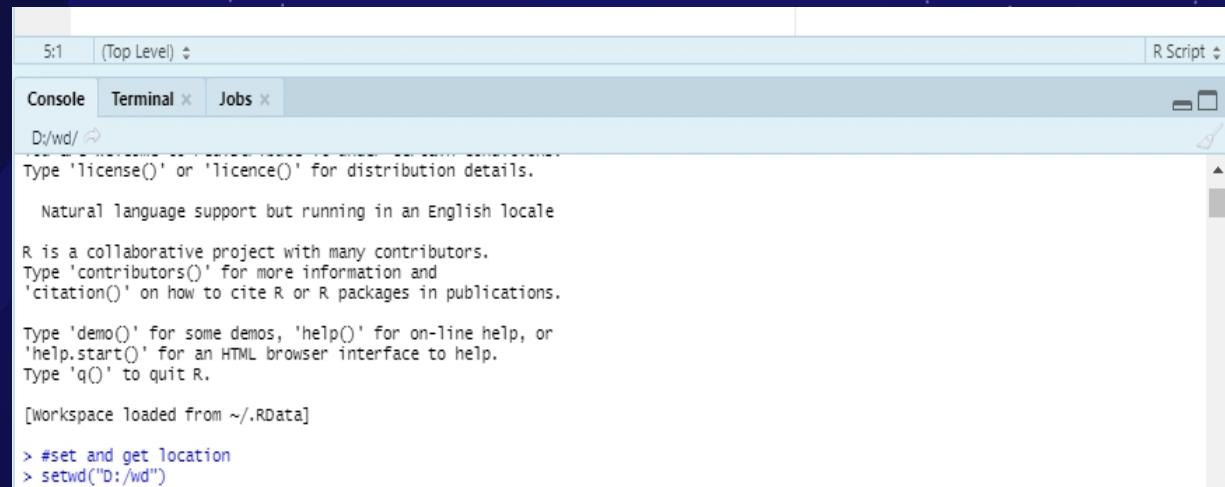
25	2,"40","9.75","1","2","80","0"
26	1,"30","2.5","2","1","115","1"
27	2,"34","12","3","3","95","0"
28	1,"20","0.5","2","1","75","1"
29	2,"35","12","5","3","100","0"
30	2,"24","9.5","3","3","20","0"
31	2,"19","8.75","6","1","160","1"
32	1,"35","9.25","9","1","100","1"
33	1,"29","7.25","6","1","95","1"
34	1,"50","8.75","11","3","132","0"
35	1,"31","4.25","1","1","60","1"

# -Langkah-langkah-

## • Pengaturan lokasi direktori



```
1 #set and get location
2 setwd("D:/wd")
3 rm(list =ls())
4
5 |
```



5:1 (Top Level) R Script

Console Terminal Jobs

D:/wd/ ↵

Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

[workspace loaded from ~/.RData]

```
> #set and get location
> setwd("D:/wd")
```

# Membaca Dataset

```
dataset <- read.csv("Cryotherapy.csv", sep = ";")
```

```
1 #set and get location  
2 setwd("D:/wd")  
3 rm(list=ls())  
4  
5 #pembacaan dataset  
6 dataset <- read.csv("Cryotherapy.csv", sep = ",")  
7  
8 |
```

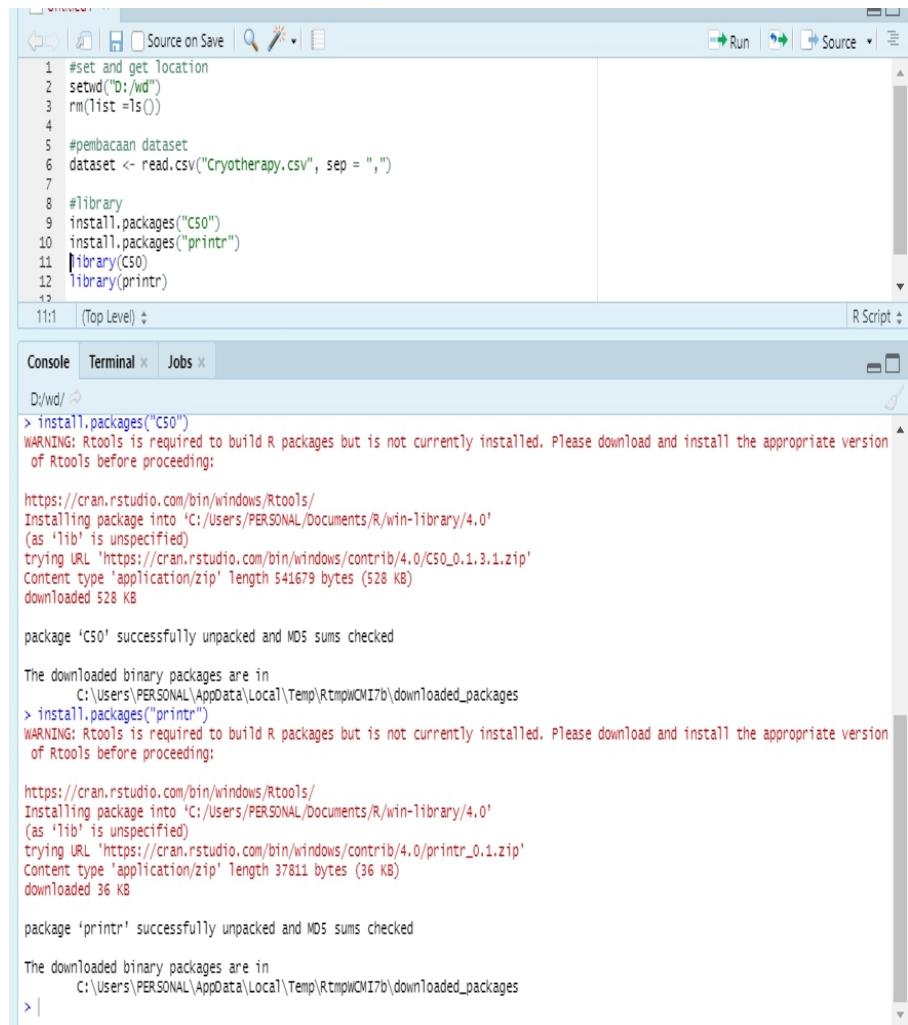
The screenshot shows the RStudio interface. The top bar indicates it's at line 8:1 of a script named '(Top Level)'. The tabs at the bottom are 'Console', 'Terminal X', and 'Jobs X'. The main area displays the R code for reading a CSV file. The 'Console' tab shows the command history, including the code entered and its execution results.

```
8:1 (Top Level) ▾ R Script ▾  
Console Terminal X Jobs X  
D:/wd/  
> #set and get location  
> setwd("D:/wd")  
> rm(list=ls())  
> #pembacaan dataset  
> dataset <- read.csv("Cryotherapy.csv", sep = ",")  
> 14 observations
```

# Instalasi package install.packages("C50") install.packages("printr")

Menggunakan package  
**library(C50)** **library(printr)**

```
C:\Users\PERSONAL> library(C50)
> library(printr)
>
```



The screenshot shows the RStudio interface. The code editor window at the top contains the following R script:

```
1 #set and get location
2 setwd("D:/wd")
3 rm(list = ls())
4
5 #pembacaan dataset
6 dataset <- read.csv("Cryotherapy.csv", sep = ",")
7
8 #library
9 install.packages("C50")
10 install.packages("printr")
11 library(C50)
12 library(printr)
13
```

The terminal window below shows the execution of the script and the resulting output. It includes messages about Rtools being required for package building, URLs for package downloads, and confirmation of successful unpacking and MD5 sum checks.

```
D:/wd/ >
> install.packages("C50")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding.

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/PERSONAL/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/C50_0.1.3.1.zip'
Content type 'application/zip' length 541679 bytes (528 KB)
downloaded 528 KB

package 'C50' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/PERSONAL/AppData/Local/Temp/RtmpwCM17b/downloaded_packages
> install.packages("printr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding.

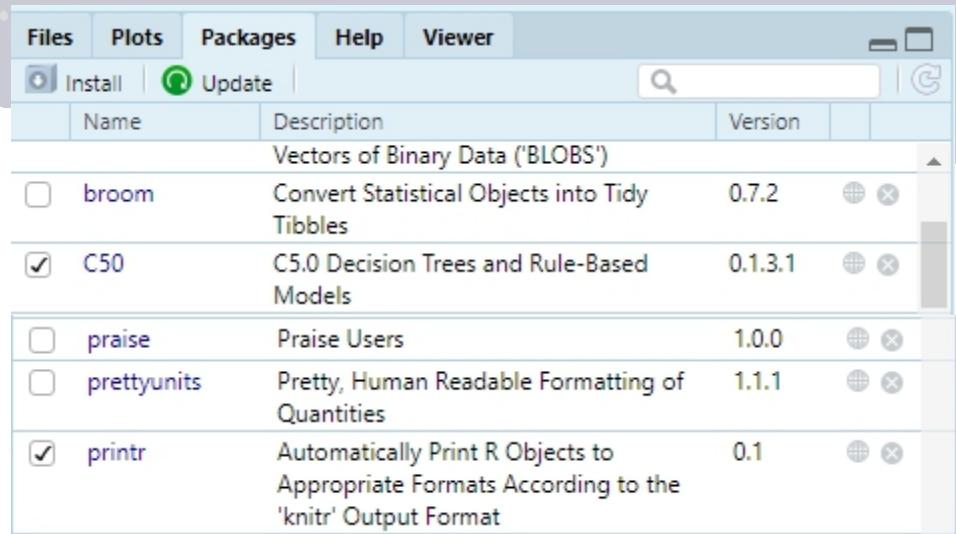
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/PERSONAL/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/printr_0.1.zip'
Content type 'application/zip' length 37811 bytes (36 KB)
downloaded 36 KB

package 'printr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/PERSONAL/AppData/Local/Temp/RtmpwCM17b/downloaded_packages
> |
```

01

## Melihat package yang sudah terinstall



## 02 Pembuatan model decision tree menggunakan algoritman C5.0

```
> model <- C5.0(Result_of_Treatment ~., data=dataset)
Error: C5.0 models require a factor outcome
>
```

03

Terjadi error karena outputnya bukan factor, factor adalah tipe data. Sehingga untuk melihat tipe kita ketikkan class(dataset\$Result\_of\_Treatment)

```
> #jika model error cek class/tipe dari kolom Result_of_Treatment
> class(dataset$Result_of_Treatment)
[1] "integer"
>
```

Setelah di cek tipe datanya, ternyata tipe datanya adalah character. Sehingga kita mengonversinya ke factor dengan mengetikkan **dataset\$Result\_of\_treatment <- as.factor(dataset\$Result\_of\_Treatment)**. Selanjutnya jalankan perintah untuk membuat model. Pembuatan model sudah berhasil.

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Shows R code for creating a decision tree model. It includes comments like "#Membuat model decision tree menggunakan C5.0" and "#mengubah tipe class ke factor".
- Console Output:**
  - A warning message: "WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding: https://cran.rstudio.com/bin/windows/Rtools/".
  - The command `install.packages("printr")` is run, which downloads and installs the 'printr' package from CRAN.
  - The package 'printr' is successfully unpacked and MD5 sums checked.
  - The downloaded binary packages are listed in the temporary directory: "C:\Users\PERSONAL\AppData\Local\Temp\RtmpWCM17b\downloaded\_packages".
  - The R code continues to run, including `library(C50)` and `model <- C5.0(Result\_of\_Treatment ~., data=dataset)`. An error occurs: "Error: C5.0 models require a factor outcome".
  - The user then runs `dataset\$Result\_of\_Treatment <- as.factor(dataset\$Result\_of\_Treatment)` to convert the column to a factor.
  - The final command `model <- C5.0(Result\_of\_Treatment ~., data=dataset)` is run again without errors.

# Model

```
> model1  
  
Call:  
C5.0(formula = Result_of_Treatment ~ ., data = dataset)  
  
Classification Tree  
Number of samples: 90  
Number of predictors: 6  
  
Tree size: 8  
  
Non-standard options: attempt to group attributes
```

## Classification Tree

- Number of samples: 90
- Number of predictors: 6

# Summary Model

```
> summary(model1)  
  
Call:  
C5.0(formula = Result_of_Treatment ~ ., data = dataset)  
  
C5.0 [Release 2.07 GPL Edition]      Tue Jan 12 13:29:37 2021  
-----  
Class specified by attribute 'outcome'  
Read 90 cases (7 attributes) from undefined.data  
Decision tree:  
  
Time <= 8:  
:...age <= 41: 1 (39)  
: age > 41: 0 (4)  
Time > 8:  
:...age <= 16: 1 (4)  
: age > 16:  
  :...Type > 2: 0 (19)  
  : Type <= 2:  
    :...Area <= 10: 0 (9)  
    : Area > 10:  
      :...Area <= 20: 1 (3)  
      : Area > 20:  
        :...Area <= 96: 0 (9)  
        : Area > 96: 1 (3/1)  
  
Evaluation on training data (90 cases):  
  Decision Tree  
  -----  
  Size   Errors  
  8     1( 1.1%)  <<
```

```
(a) (b) <-classified as  
... ...  
41 1 (a): class 0  
48 0 (b): class 1  
  
Attribute usage:  
  
100.0% age  
100.0% Time  
47.78% Type  
26.67% Area  
  
Time: 0.0 secs
```

# Plot Model

```
18 #cek class/tipe dari kolom Result_of_Treatment
19 class(dataset$Result_of_Treatment)
20
21 #mengubah tipe class ke faktor
22 dataset$Result_of_Treatment <- as.factor(dataset$Result_of_Treatment)
23 model <- C5.0(Result_of_Treatment ~., data=dataset)
24
25 #melihat hasil model
26 model
27 summary(model)
28
29 #menampilkan gambar/pohon model
30 plot(model)
31
32
```

31:1 (Top Level) ▾

Console Terminal ▾ Jobs ▾

D:/wd/ ↵ 8 1( 1.1% ) <<

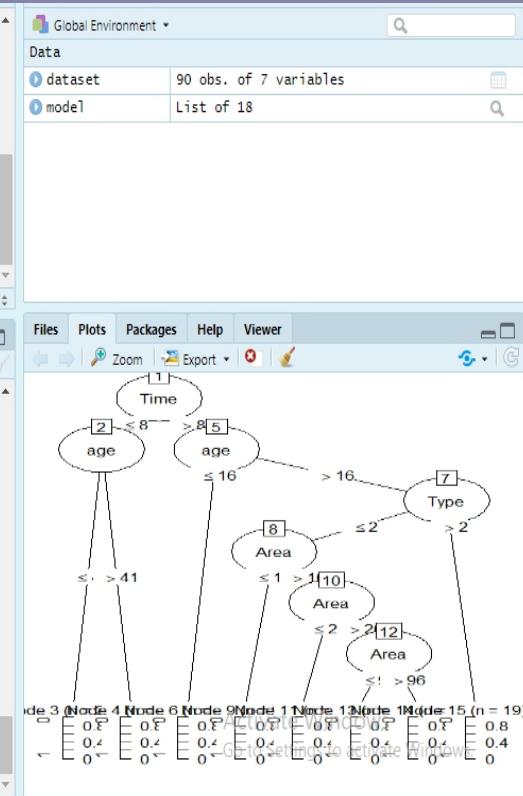
(a) (b) <-classified as  
----  
41 1 (a): class 0  
48 (b): class 1

Attribute usage:

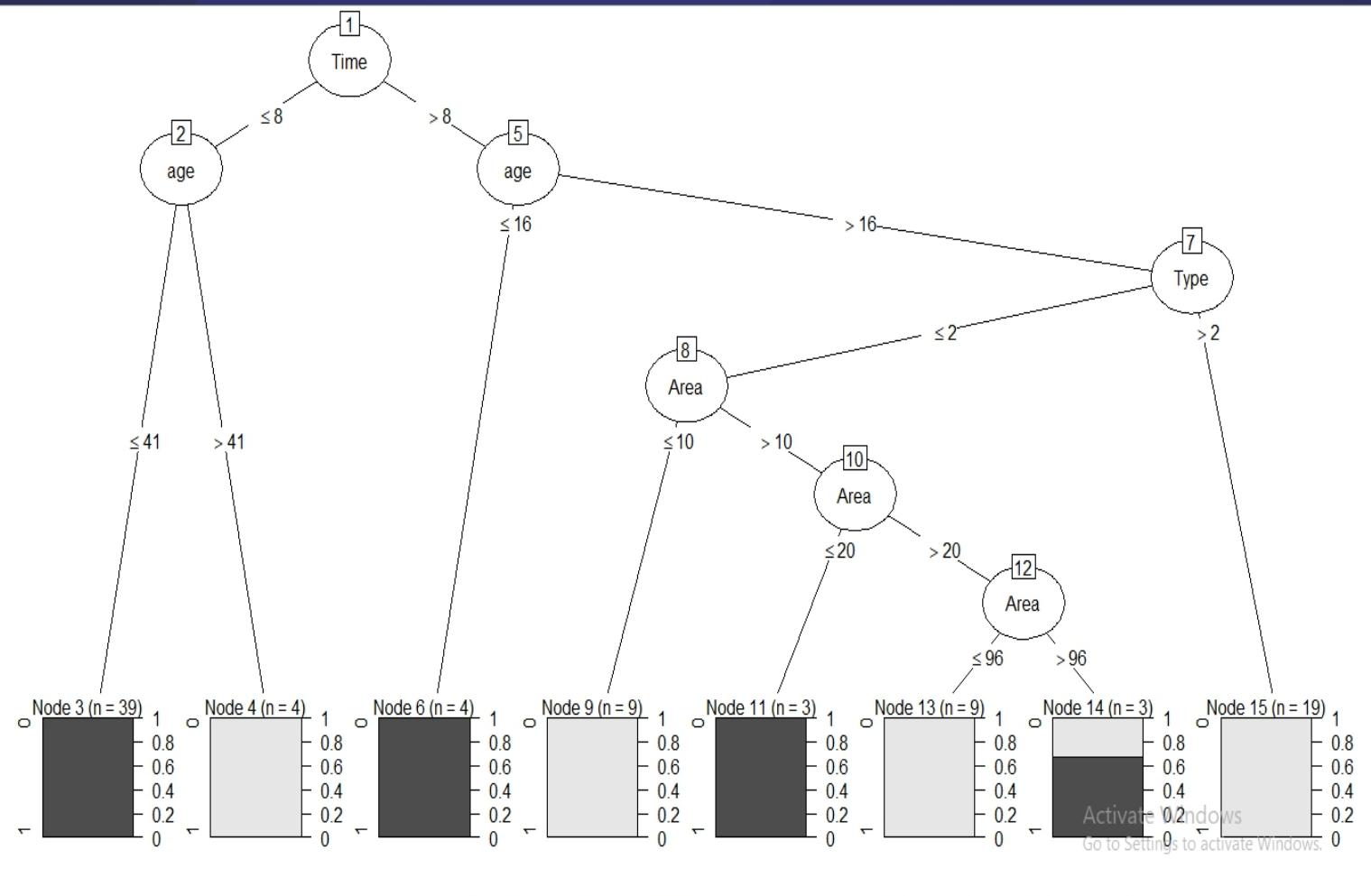
100.00% age  
100.00% Time  
47.78% Type  
26.67% Area

Time: 0.0 secs

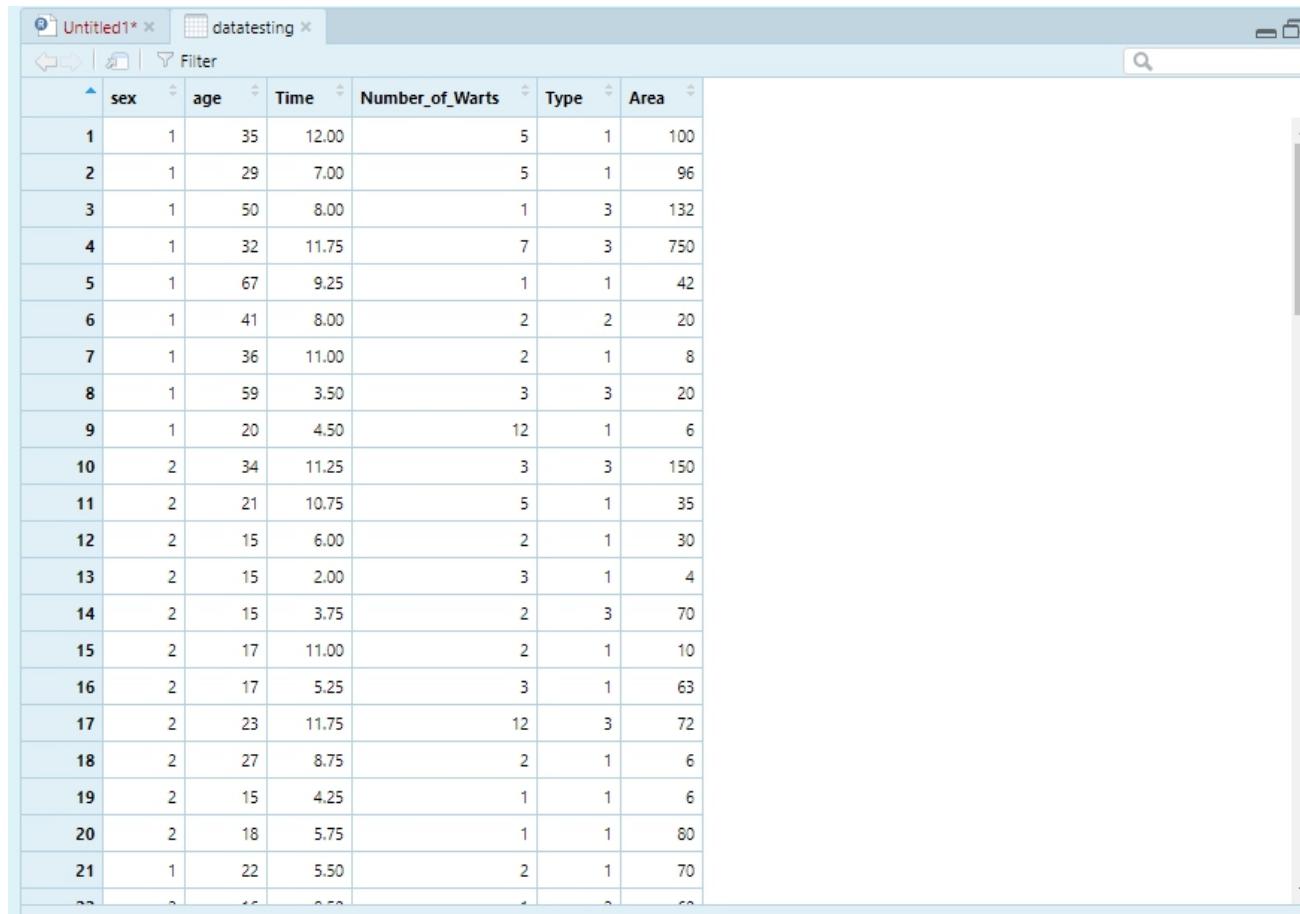
```
> #menampilkan gambar/pohon model
> plot(model)
> |
```



# Menampilkan pohon yang sudah dibangun



Menjadikan dataset sebagai data testing.  
Namun hanya kolom 1, 2, 3, 4 saja dan  
tanpa label. **datatesting <- dataset[,1:6]**



	sex	age	Time	Number_of_Warts	Type	Area
1	1	35	12.00	5	1	100
2	1	29	7.00	5	1	96
3	1	50	8.00	1	3	132
4	1	32	11.75	7	3	750
5	1	67	9.25	1	1	42
6	1	41	8.00	2	2	20
7	1	36	11.00	2	1	8
8	1	59	3.50	3	3	20
9	1	20	4.50	12	1	6
10	2	34	11.25	3	3	150
11	2	21	10.75	5	1	35
12	2	15	6.00	2	1	30
13	2	15	2.00	3	1	4
14	2	15	3.75	2	3	70
15	2	17	11.00	2	1	10
16	2	17	5.25	3	1	63
17	2	23	11.75	12	3	72
18	2	27	8.75	2	1	6
19	2	15	4.25	1	1	6
20	2	18	5.75	1	1	80
21	1	22	5.50	2	1	70
22	2	15	8.50	1	2	60

```
predictions <- predict(model, datatesting)
```

The screenshot shows the RStudio interface. The top panel displays an R script with code for fitting a decision tree model, visualizing it, and making predictions on a testing dataset. The bottom panel shows the R Console output, which includes the predicted classes for two observations, attribute usage percentages, and execution time.

```
Untitled1* datatesting
Source on Save Run Source

24
25 #melihat hasil model
26 model
27 summary(model)
28
29 #menampilkan gambar/pohon model
30 plot(model)
31
32 #membuat dataset
33 datatesting <- dataset[,1:6]
34
35 #prediksi
36 predictions <- predict(model, datatesting)
37 |
38

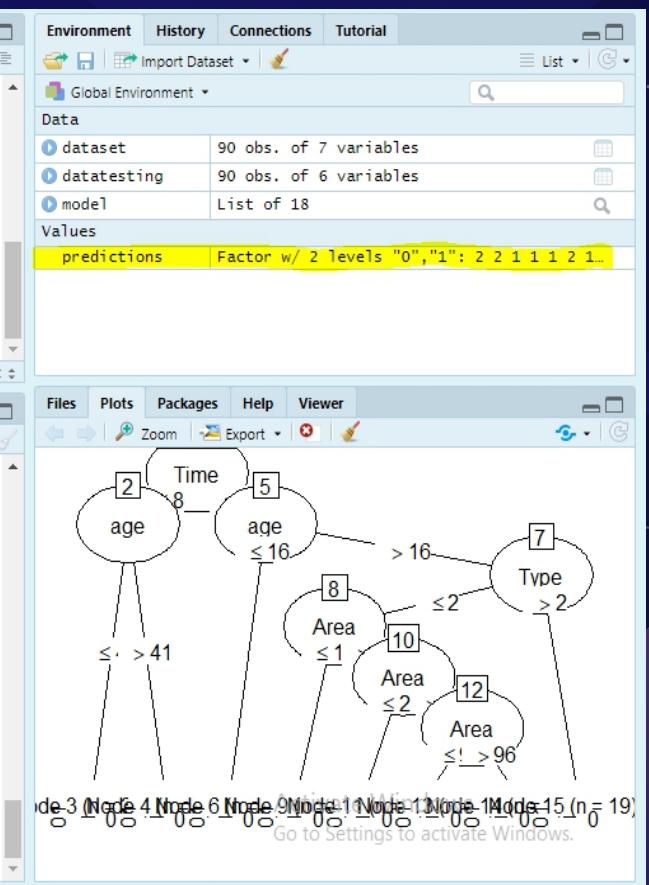
37:1 (Top Level) R Script

Console Terminal Jobs
D:/wd/ ↵
41     1   (a): class 0
48     48   (b): class 1

Attribute usage:
100.00% age
100.00% Time
47.78% Type
26.67% Area

Time: 0.0 secs

> #menampilkan gambar/pohon model
> plot(model)
> #membuat dataset
> datatesting <- dataset[,1:6]
> View(datatesting)
> #prediksi
> predictions <- predict(model, datatesting)
> |
```



Membandingkan hasil prediksi dengan dataset  
**table(predictions, dataset\$Result\_of\_Treatment)**

```
> table(predictions, dataset$Result_of_Treatment)

predictions  0   1
          0 41  0
          1  1 48
> |
```

Melihat tingkat akurasi  
**mean(predictions == dataset\$Result\_of\_Treatment)**

```
> #Melihat tingkat akurasi
> mean(predictions == dataset$Result_of_Treatment)
[1] 0.9888889
```

# RULES YANG DIHASILKAN

```
> rulemodel <- C5.0(Result_of_Treatment ~., data = dataset, rules = TRUE)
> rulemodel

Call:
C5.0.formula(formula = Result_of_Treatment ~ ., data = dataset, rules = TRUE)

Rule-Based Model
Number of samples: 90
Number of predictors: 6

Number of Rules: 8

Non-standard options: attempt to group attributes
> |
```

```
> summary(rulemodel)

Call:
C5.0.formula(formula = Result_of_Treatment ~ ., data = dataset, rules = TRUE)

C5.0 [Release 2.07 GPL Edition]      Tue Jan 12 13:40:59 2021
-----
Class specified by attribute 'outcome'

Read 90 cases (7 attributes) from undefined.data

Rules:

Rule 1: (19, lift 2.0)
    Time > 8
    Type > 2
    -> class 0 [0.952]

Rule 2: (16, lift 2.0)
    age > 16
    Time > 8
    Area > 20
    Area <= 96
    -> class 0 [0.944]

Rule 3: (9, lift 1.9)
    Time > 8
    Area <= 10
    -> class 0 [0.909]

Rule 4: (9, lift 1.9)
    age > 41
    -> class 0 [0.909]

Rule 5: (39, lift 1.8)
    age <= 41
    Time <= 8
    -> class 1 [0.976]
```

# RULES YANG DIHASILKAN

Rule 6: (15, lift 1.8)  
age <= 16  
-> class 1 [0.941]

Rule 7: (5, lift 1.6)  
Type <= 2  
Area > 10  
Area <= 20  
-> class 1 [0.857]

Rule 8: (11/1, lift 1.6)  
Type <= 2  
Area > 96  
-> class 1 [0.846]

Default class: 1

Evaluation on training data (90 cases):

Rules		
No	Errors	
8	1( 1.1%)	<<
(a)	(b)	<-classified as
41	1	(a): class 0
48		(b): class 1

Attribute usage:

84.44% Time  
72.22% age  
45.56% Area  
38.89% Type

## RULES YANG DIHASILKAN

➤ rule 1 :

time > 8

type > 2

class -> 0 (jinak)

➤ rule 2 :

age > 16

Time > 8

Area > 20

Area <= 96

class -> 0 (jinak)

➤ rule 3 :

Time > 8

Area <= 10

class -> 0 (jinak)

➤ rule 4 :

age > 41

Area <= 10

Class -> 0 (Jinak)

➤ rule 5 :

age <= 41

Time <= 8

class -> 1 (Ganas)

➤ rule 6 :

age <= 16

class -> 1 (Ganas)

➤ rule 7 :

Type <= 2

Area > 10

Area <= 20

class -> 1 (Ganas)

➤ rule 8 :

Type <= 2

Area > 96

class -> 1 (Ganas)

Sekian Terimakasih :)

