

## TUGAS BESAR DATA MINING

### DIAGNOSE BREAST CANCER MENGGUNAKAN METODE CLASSIFICATION AND REGRESSION TREES (CART)

Nama Anggota :

- Agus Riady (3311901002)
- Risma Ananda Harby (3311901006)
- Anissa Nabila (3311901007)

Alat :

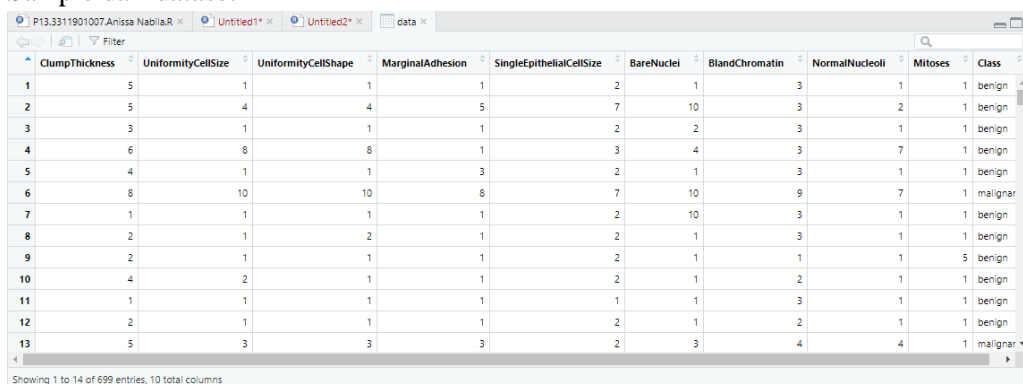
- R dan R Studio

Deskripsi :

- Metode ini dikembangkan oleh Dr. William H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin, USA. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706. Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706.
- CART membangun pohon biner menggunakan fitur dan nilai batasan yang menghasilkan perolehan informasi terbesar (information gain) pada setiap node,
- CART dapat digunakan untuk Classification dan Regression

Dataset :

- **Breast Cancer Wisconsin (Original) Data Set** dari repositori pembelajaran mesin UCI adalah kumpulan data klasifikasi, yang mencatat pengukuran untuk kasus kanker payudara. Ada dua kelas, jinak dan ganas. Set data ini memiliki dimensi 9. Kelas ganas dari set data ini dianggap sebagai outlier, sementara poin di kelas jinak dianggap inlier.
- Jumlah data pada dataset Breast Cancer Wisconsin (Original) yaitu 699 data dan 11 variable
- Number of Instances: 699 (as of 15 July 1992)
- Number of Attributes: 10 plus the class attribute
- Class distribution: Benign: 458 (65.5%), Malignant: 241 (34.5%)
- Sample dari dataset



	ClumpThickness	UniformityCellSize	UniformityCellShape	MarginalAdhesion	SingleEpithelialCellSize	BareNuclei	BlandChromatin	NormalNucleoli	Mitoses	Class
1	5	1	1	1	2	1	3	1	1	benign
2	5	4	4	5	7	10	3	2	1	benign
3	3	1	1	1	2	2	3	1	1	benign
4	6	8	8	1	3	4	3	7	1	benign
5	4	1	1	3	2	1	3	1	1	benign
6	8	10	10	8	7	10	9	7	1	malignant
7	1	1	1	1	2	10	3	1	1	benign
8	2	1	2	1	2	1	3	1	1	benign
9	2	1	1	1	2	1	1	1	5	benign
10	4	2	1	1	2	1	2	1	1	benign
11	1	1	1	1	1	1	3	1	1	benign
12	2	1	1	1	2	1	2	1	1	benign
13	5	3	3	3	2	3	4	4	1	malignant

## Proses Data Mining :

### 1. Mengatur lokasi kerja

```
Console Terminal x Jobs x
D:/wd/TugasBesar/
> # Decision Tree Classification on Breast cancer dataset
> lokasi_kerja <- "D:/wd/TugasBesar"
> setwd(lokalasi_kerja)
> getwd()
[1] "D:/wd/TugasBesar"
```

### 2. Download data dari UCL Machine Learning Repository

```
Console Terminal x Jobs x
D:/wd/TugasBesar/
[1] "D:/wd/TugasBesar"
> # Downloading the file
> fileURL <- "http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data"
> download.file(fileURL, destfile="breast-cancer-wisconsin.data", method="curl")
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 19889 100 19889 0 0 19889 0 0:00:01 --:--:-- 0:00:01 28291
```

### 3. Melihat data dan struktur pada sebuah object

```
Console Terminal x Jobs x
D:/wd/TugasBesar/
> # read the data
> data <- read.table("breast-cancer-wisconsin.data", na.strings = "?", sep=",")
> str(data)
'data.frame': 699 obs. of 11 variables:
 $ v1 : int 1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ...
 $ v2 : int 5 5 3 6 4 8 1 2 2 4 ...
 $ v3 : int 1 4 1 8 1 10 1 1 1 2 ...
 $ v4 : int 1 4 1 8 1 10 1 2 1 1 ...
 $ v5 : int 1 5 1 1 3 8 1 1 1 1 ...
 $ v6 : int 2 7 2 3 2 7 2 2 2 2 ...
 $ v7 : int 1 10 2 4 1 10 10 1 1 1 ...
 $ v8 : int 3 3 3 3 3 9 3 3 1 2 ...
 $ v9 : int 1 2 1 7 1 7 1 1 1 1 ...
 $ v10: int 1 1 1 1 1 1 1 1 5 1 ...
 $ v11: int 2 2 2 2 2 4 2 2 2 2 ...
```

### 4. Menghapus kolom ID

```
Console Terminal x Jobs x
D:/wd/TugasBesar/
> # Remove ID column, col = 1
> data <- data[,-1]
```

### 5. Mengganti nama kolom

```
Console Terminal x Jobs x
D:/wd/TugasBesar/
> data <- data[,-1]
> names(data) <- c("clumpThickness",
+ "uniformityCellsize",
+ "uniformityCellshape",
+ "marginalAdhesion",
+ "singleEpithelialCellsize",
+ "bareNuclei",
+ "blandChromatin",
+ "normalNucleoli",
+ "mitoses",
+ "class")
```

## 6. Mengkonversi nilai numerik dalam variabel respon menjadi label

```

Console Terminal x Jobs x
D:/wd/TugasBesar/
> # Numerical values in the response variable are converted to labels
> data$class <- factor(data$class, levels=c(2,4), labels=c("benign", "malignant"))
> print(summary(data))
ClumpThickness    UniformityCellSize    UniformityCellShape    MarginalAdhesion
Min.   : 1.000    Min.   : 1.000    Min.   : 1.000    Min.   : 1.000
1st Qu.: 2.000    1st Qu.: 1.000    1st Qu.: 1.000    1st Qu.: 1.000
Median : 4.000    Median : 1.000    Median : 1.000    Median : 1.000
Mean   : 4.418    Mean   : 3.134    Mean   : 3.207    Mean   : 2.807
3rd Qu.: 6.000    3rd Qu.: 5.000    3rd Qu.: 5.000    3rd Qu.: 4.000
Max.   :10.000    Max.   :10.000    Max.   :10.000    Max.   :10.000

SingleEpithelialCellSize    BareNuclei    BlandChromatin    NormalNucleoli
Min.   : 1.000    Min.   : 1.000    Min.   : 1.000    Min.   : 1.000
1st Qu.: 2.000    1st Qu.: 1.000    1st Qu.: 2.000    1st Qu.: 1.000
Median : 2.000    Median : 1.000    Median : 3.000    Median : 1.000
Mean   : 3.216    Mean   : 3.545    Mean   : 3.438    Mean   : 2.867
3rd Qu.: 4.000    3rd Qu.: 6.000    3rd Qu.: 5.000    3rd Qu.: 4.000
Max.   :10.000    Max.   :10.000    Max.   :10.000    Max.   :10.000

```

## 7. Membagi dataset kedalam training dan validation

```

Console Terminal x Jobs x
D:/wd/TugasBesar/
> # Dividing the dataset into training and validation sets. There are many ways to do this.
> # Alternate method is also listed here.
> set.seed(123)
> ind <- sample(2, nrow(data), replace=TRUE, prob=c(0.7, 0.3))
> trainData <- data[ind==1,]
> validationData <- data[ind==2,]
> table(trainData$class)

    benign malignant 
      322       166 
> prop.table(table(trainData$class))

    benign malignant 
0.6598361 0.3401639 

```

## 8. Memprediksi Benign Breast Cancer, Malignant Breast Cancer. Root Node Error dan Tingkat Akurasi

```

Console Terminal x Jobs x
D:/wd/TugasBesar/
> dataFormula <- Class ~ ClumpThickness + UniformityCellSize + UniformityCellShape + MarginalAdhesion + SingleEpithelialCellSize + BareNuclei + BlandChromatin + NormalNucleoli + Mitoses
> Breast_Cancer_rpart <- rpart(dataFormula, data = trainData, control = rpart.control(minsplit = 10))
> rpartMod <- rpart(Class ~ ClumpThickness + UniformityCellSize + UniformityCellShape + MarginalAdhesion + SingleEpithelialCellSize + BareNuclei + BlandChromatin + NormalNucleoli + Mitoses, data = trainData, method = "class")
> printcp(rpartMod)

Classification tree:
rpart(formula = Class ~ ClumpThickness + UniformityCellSize + UniformityCellShape + MarginalAdhesion + SingleEpithelialCellSize + BareNuclei + BlandChromatin + NormalNucleoli + Mitoses, data = trainData, method = "class")

variables actually used in tree construction:
[1] UniformityCellShape UniformityCellSize

Root node error: 166/488 = 0.34016

n= 488

      CP nsplit rel error  xerror   xstd
1 0.837349      0  1.00000 1.00000 0.063047
2 0.042169      1  0.16265 0.16265 0.030424
3 0.010000      2  0.12048 0.12651 0.027005
> plotcp(rpartMod)
> print(Breast_Cancer_rpart)
n= 488

node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 488 166 benign (0.659836066 0.340163934)
2) UniformityCellSize< 2.5 303 4 benign (0.986798680 0.013201320)
4) NormalNucleoli< 3.5 299 1 benign (0.996655518 0.003344482) *
5) NormalNucleoli>=3.5 4 1 malignant (0.250000000 0.750000000) *
3) UniformityCellSize>=2.5 185 23 malignant (0.124324324 0.875675676)
6) UniformityCellShape< 2.5 13 3 benign (0.769230769 0.230769231)

```

9. Menggunakan `sample.split()` untuk membuat vektor dengan dua nilai yaitu TRUE dan FALSE. Dengan mengatur `SplitRatio` menjadi 70% data training dan 30% data testing.

```
Console Terminal Jobs
D:/wd/TugasBesar/
> library(caTools)
> # Alternate method
> set.seed(123)
> split = sample.split(data$class, splitRatio = 0.7)
> split
[1] TRUE FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE FALSE TRUE
[14] TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[27] TRUE TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE FALSE
[40] TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[53] TRUE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE FALSE
[66] FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE
[79] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE
[92] FALSE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[105] TRUE TRUE TRUE FALSE FALSE TRUE FALSE TRUE TRUE FALSE FALSE TRUE TRUE TRUE
[118] TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE
[131] TRUE FALSE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[144] TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE
[157] FALSE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE
```

10. Membuat data training dan data testing

```
Console Terminal Jobs
D:/wd/TugasBesar/
> # Create training and testing sets
> dataTrain = subset(data, split == TRUE)
> dataTest = subset(data, split == FALSE)
```

11. Install library `rpart`, `rpart.plot` dan `party`

```
install.packages("rpart")
install.packages("rpart.plot")
install.packages("party")

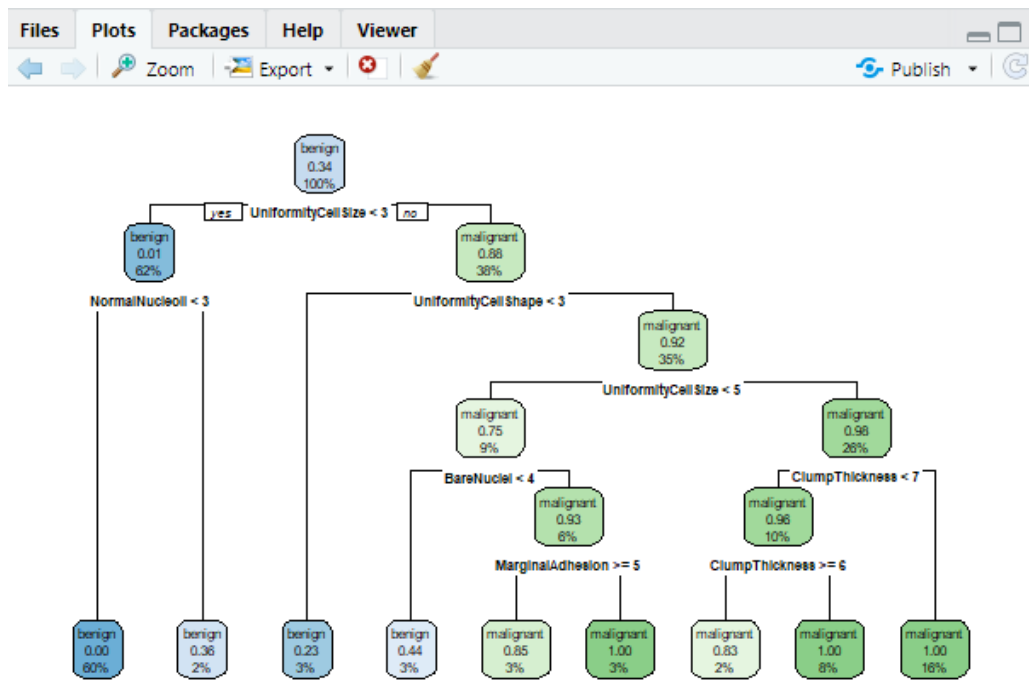
library(rpart)
library(rpart.plot)
library(party)
```

12. Membangun model

```
> set.seed(123)
> model = rpart(class ~ ., data=trainData, method="class", minsplit = 10, minbucket = 10, cp = -1)
```

13. Menampilkan pohon model

```
> par(xpd = NA)
> rpart.plot(model)
```



14. Memprediksi dan mengevaluasi kinerja pada trained tree model

```
> predicted.classes = predict(model, validationData, type = "class")
```

Environment	History	Connections	Tutorial
Global Environment	Import Dataset		
data	450 obs. of 10 variables		
model	List of 14		
trainData	488 obs. of 10 variables		
validationData	211 obs. of 10 variables		
values			
fileURL	"http://archive.ics.uci.edu/ml/machine-learning-d...		
ind	int [1:699] 1 2 1 2 2 1 1 2 1 1 ...		
lokasi_kerja	"D:/wd/TugasBesar"		
predicted.classes	Factor w/ 2 levels "benign","malignant": 2 2 1 1 ...		
split	logi [1:699] TRUE FALSE TRUE FALSE FALSE TRUE ...		

15. Menampilkan beberapa nilai dari object yang telah diprediksi

```
> head(predicted.classes)
      2      4      5      8     11     16
malignant malignant benign benign benign benign
Levels: benign malignant
> head(validationData$class)
[1] benign benign benign benign benign malignant
Levels: benign malignant
```

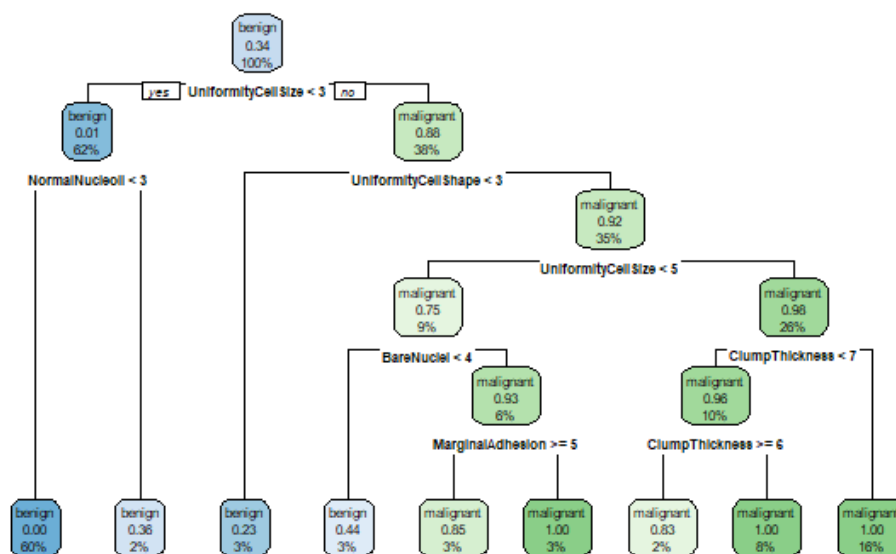
16. Membandingkan nilai yang diprediksi dengan nilai yang sebenarnya, lalu dihitung rata-ratanya

```
> mean(predicted.classes == validationData$class)
[1] 0.9099526
> |
```

## Hasil data mining :

- Ditemukan 699 data dengan 11 variabel.
- Data yang dapat diproses sebanyak 488 data
- Benign Breast Cancer yang benar diprediksi sebanyak 166 data
- Malignant Breast Cancer yang benar diprediksi sebanyak 322 data
- Tingkat root node error Breast Cancer rpart adalah sebesar 0.34016
- Tingkat kepercayaan (akurasi) Breast Cancer rpart adalah sebesar **65.983%**
- Data Training memiliki 488 baris data dan 10 kolom
- Data Validation memiliki 211 baris data dan 10 kolom
- Variable untuk menentukan apakah seseorang mengalami kanker payudara adalah Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, dan Mitoses

## Hasil analisa data mining :



Probabilitas benign sebesar 0.34

- Apabila UniformityCellSize lebih kecil dari 3 maka probabilitas benign sebesar 0.01
- Apabila UniformityCellSize lebih besar dari 3 maka probabilitas malignant sebesar 0.88
- Apabila NormalNucleoli lebih kecil dari 3 maka probabilitas benign sebesar 0.00
- Apabila NormalNucleoli lebih besar dari 3 maka probabilitas benign sebesar 0.36
- Apabila UniformityCellShape lebih kecil dari 3 maka probabilitas benign sebesar 0.23
- Apabila UniformityCellShape lebih besar dari 3 maka probabilitas malignant sebesar 0.92
- Apabila UniformityCellSize lebih kecil dari 5 maka probabilitas malignant sebesar 0.75
- Apabila UniformityCellSize lebih besar dari 5 maka probabilitas malignant sebesar 0.96
- Apabila BareNuclei lebih kecil dari 4 maka probabilitas benign sebesar 0.44

- Apabila BareNuclei lebih besar dari 4 maka probabilitas malignant sebesar 0.93
- Apabila ClumpThickness lebih kecil dari 7 maka probabilitas malignant sebesar 0.96
- Apabila ClumpThickness lebih besar dari 7 maka probabilitas malignant sebesar 1.00
- Apabila MarginalAdhesion lebih besar sama dengan 5 maka probabilitas malignant sebesar 0.85
- Apabila MarginalAdhesion lebih kecil sama dengan 5 maka probabilitas malignant sebesar 1.00
- Apabila ClumpThickness lebih besar sama dengan 6 maka probabilitas malignant sebesar 0.83
- Apabila ClumpThickness lebih kecil sama dengan 6 maka probabilitas malignant sebesar 1.00

Penjelasan diatas dapat disimpulkan dalam tabel sebagai berikut ;

UniformityCellSize	NormalNucleoli	UniformityCellShape	UniformityCellSize	BareNuclei	ClumpThickness	MarginalAdhesion	ClumpThickness	Benign	Malignant
< 3	-	-	-	-	-	-	-	0.01	
< 3	< 3	-	-	-	-	-	-	0	
< 3	> 3	-	-	-	-	-	-	0.36	
> 3	-	-	-	-	-	-	-		0.88
> 3	-	< 3	-	-	-	-	-	0.23	
> 3	-	> 3	-	-	-	-	-		0.92
> 3	-	> 3	< 5	-	-	-	-		0.75
> 3	-	> 3	> 5	-	-	-	-		0.98
> 3	-	> 3	< 5	< 4	-	-	-	0.44	
> 3	-	> 3	< 5	> 4	-	-	-		0.93
> 3	-	> 3	> 5	-	< 7	-	-		0.96
> 3	-	> 3	> 5	-	> 7	-	-		1
> 3	-	> 3	< 5	> 4	-	> = 5	-		0.85
> 3	-	> 3	< 5	> 4	-	< = 5	-		1
> 3	-	> 3	> 5	-	> 7	-	> = 6		0.83
> 3	-	> 3	> 5	-	> 7	-	< = 6		1

#### Referensi :

- [UCI Machine Learning Repository: Breast Cancer Wisconsin \(Original\) Data Set](#)[R Pubs - Decision Tree modeling with Breast cancer Dataset](#)
- [Breast w - Dataset - DataHub - Frictionless Data](#)
- [rpart.plot function | R Documentation](#)
- [BreastCancer function | R Documentation](#)
- [Splitting a data frame into training and testing sets in R – Stories Data Speak \(duttashi.github.io\)](#)
- [Breast Cancer Wisconsin \(Original\) dataset – ODDS \(stonybrook.edu\)](#)
- [Breast Cancer Analysis \(rstudio-pubs-static.s3.amazonaws.com\)](#)