**TUGAS BESAR DATA MINING**

**DIAGNOSE BREAST CANCER MENGGUNAKAN METODE CLASSIFICATION AND REGRESSION TRESS (CART)**

**Nama Anggota :**

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

**Alat :**

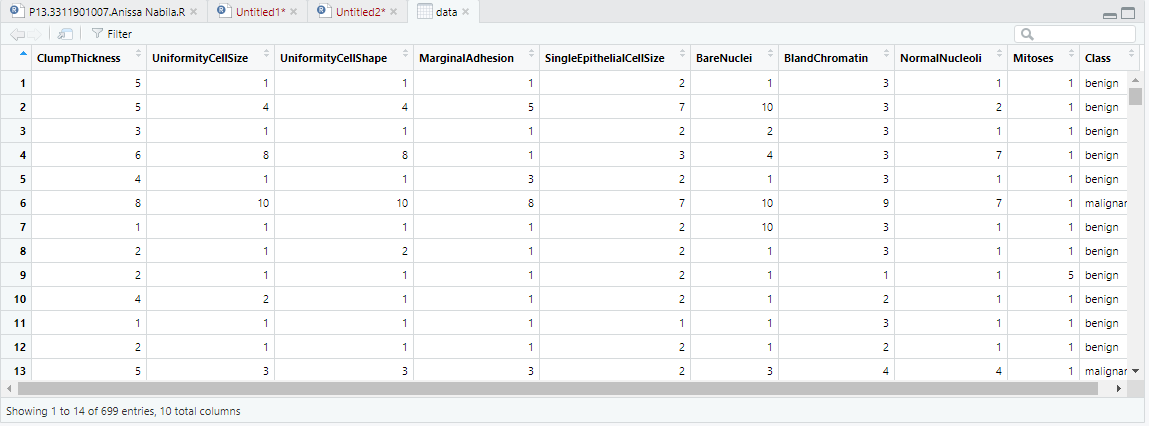
* R dan R Studio

**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

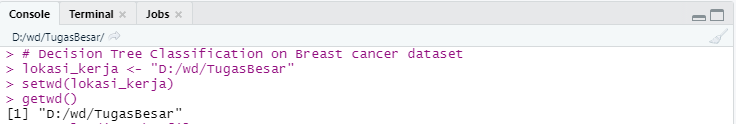
**Dataset :**

* **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

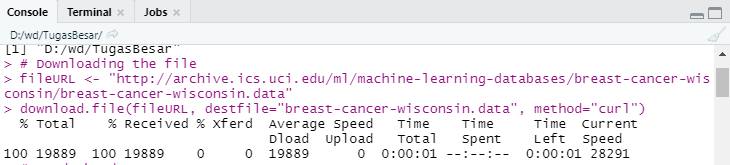


**Proses Data Mining :**

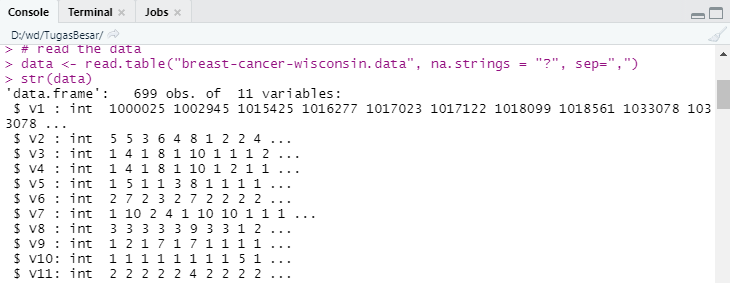
1. Mengatur lokasi kerja



1. Download data dari UCL Machine Learning Repository



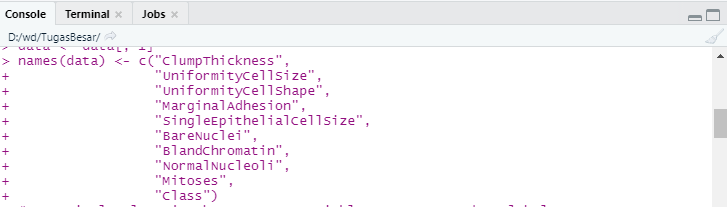
1. Melihat data dan struktur pada sebuah object



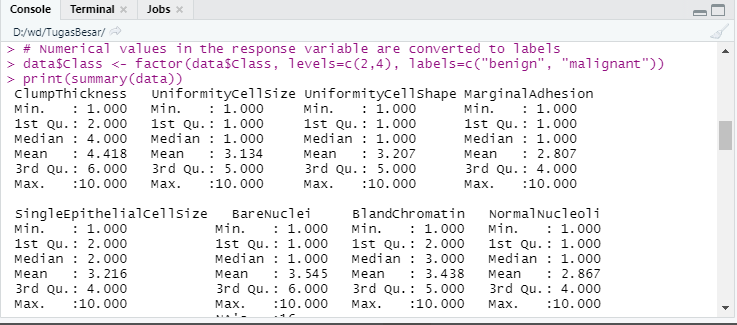
1. Menghapus kolom ID



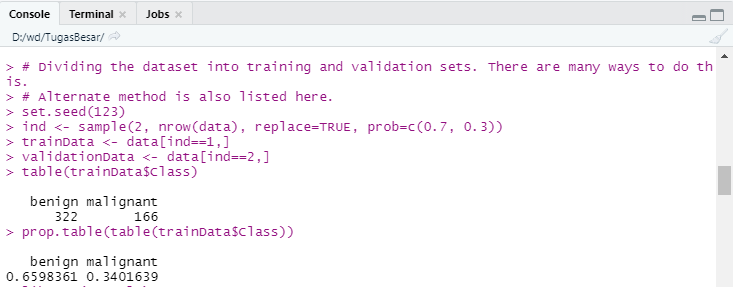
1. Mengganti nama kolom



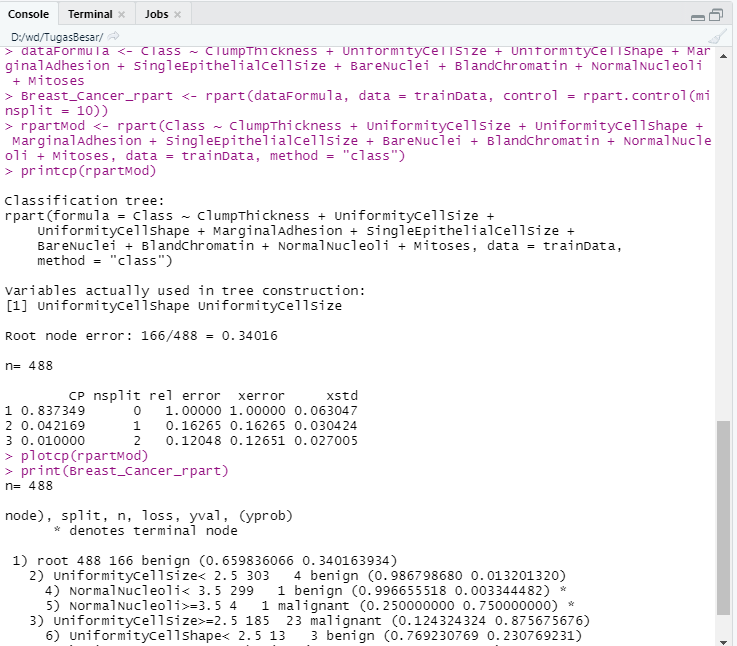
1. Mengkonversi nilai numerik dalam variabel respon menjadi label



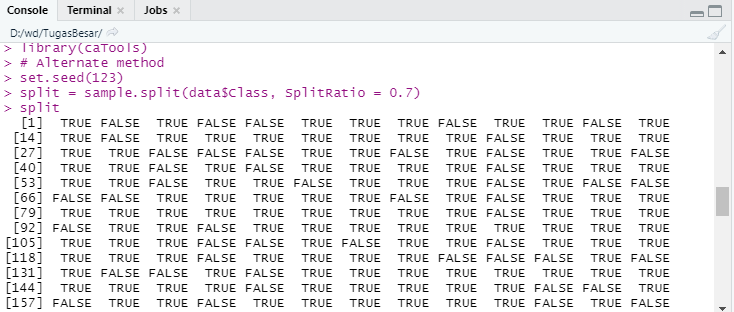
1. Membagi dataset kedalam training dan validation



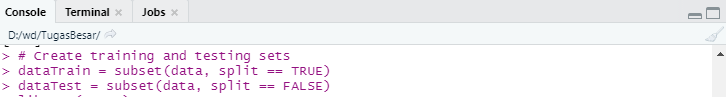
1. Memprediksi Benign Breast Cancer, Malignant Breast Cancer. Root Node Error dan Tingkat Akurasi



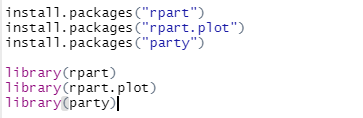
1. Menggunakan sample.split() untuk membuat vektor dengan dua nilai yaitu TRUE dan FALSE. Dengan mengatur SplitRatio menjadi 70% data training dan 30% data testing.



1. Membuat data training dan data testing



1. Install library rpart, rpart.plot dan party

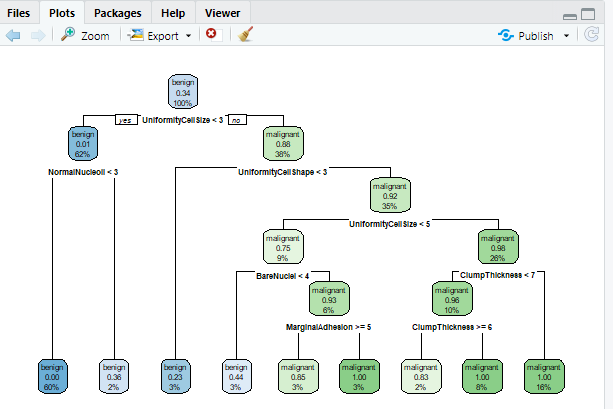


1. Membangun model



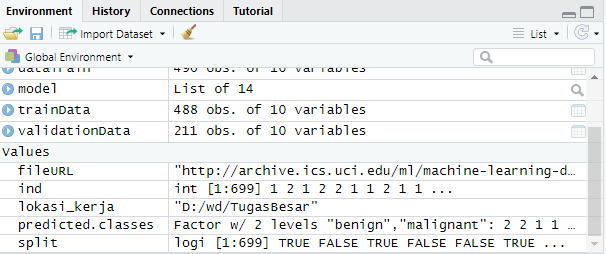
1. Menampilkan pohon model



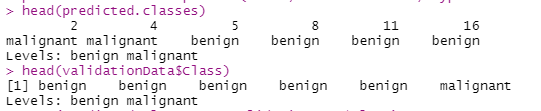


1. Memprediksi dan mengevaluasi kinerja pada trained tree model





1. Menampilkan beberapa nilai dari object yang telah diprediksi



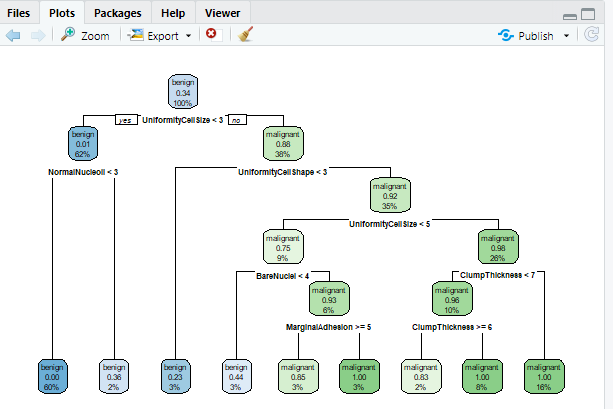
1. Membandingkan nilai yang diprediksi dengan nilai yang sebenarnya, lalu dihitung rata-ratanya



**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
* Nilai dari Membandingkan nilai yang diprediksi dengan nilai yang sebenarnya, lalu dihitung rata-ratanya adalah **0.9099**

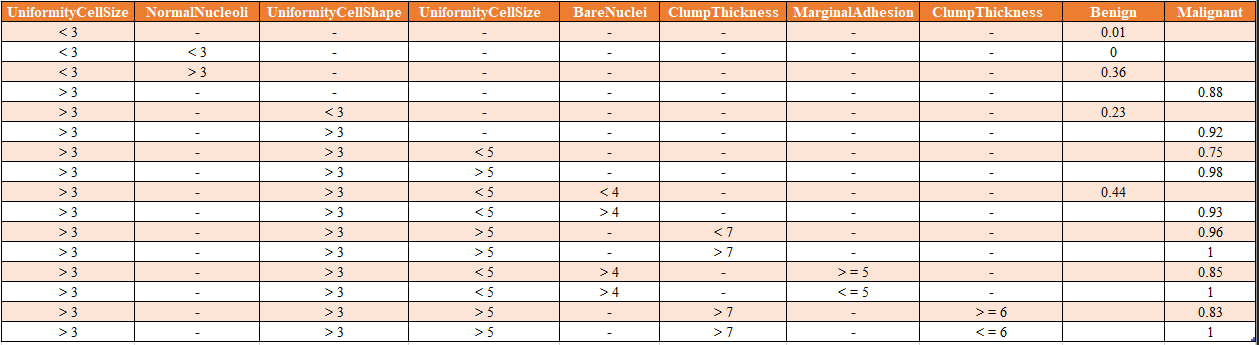
**Hasil analisa data mining :**



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 ;



**Referensi :**

* [UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29)[RPubs - Decision Tree modeling with Breast cancer Dataset](https://rpubs.com/mshaek/dt_breast_cancer)
* [Breast w - Dataset - DataHub - Frictionless Data](https://datahub.io/machine-learning/breast-w)
* [rpart.plot function | R Documentation](https://www.rdocumentation.org/packages/rpart.plot/versions/3.0.9/topics/rpart.plot)
* [BreastCancer function | R Documentation](https://www.rdocumentation.org/packages/mlbench/versions/2.1-1/topics/BreastCancer)
* [Splitting a data frame into training and testing sets in R – Stories Data Speak (duttashi.github.io)](https://duttashi.github.io/blog/splitting-a-data-frame-into-training-and-testing-sets-in-r/)
* [Breast Cancer Wisconsin (Original) dataset – ODDS (stonybrook.edu)](http://odds.cs.stonybrook.edu/breast-cancer-wisconsin-original-dataset/#:~:text=The%20Breast%20Cancer%20Wisconsin%20(Original,measurements%20for%20breast%20cancer%20cases.&text=The%20malignant%20class%20of%20this,benign%20class%20are%20considered%20inliers.)
* [Breast Cancer Analysis (rstudio-pubs-static.s3.amazonaws.com)](https://rstudio-pubs-static.s3.amazonaws.com/279492_f23b13028af747a695f79573de548dbb.html)