Tugas Kelompok Data Mining

# Laporan Data Mining



Disusun oleh :

* Novitasari / 3311801012
* Karniah Hasanah / 3311601006

- Henni Hendrani Nasution / 3311801020

Kelas : Informatika 3 A

## **Program Studi D3 Teknik Informatika**

## **Jurusan Teknik Informatika**

## **Politeknik Negeri Batam**

## **2019**

# HALAMAN PENGESAHAN

**DATA MINING DARI DATASET “*Breast Cancer Coimbra”***

**DENGAN MENGGUNAKAN ALGORITMA C5.0**

**Disusun oleh:**

**Novitasari / 3311801012**

**Link Github : github.com/25Novitasari**

**Karniah Hasanah / 3311601006**

**Link Github : github.com/karniahhasanah**

**Henni Hendrani Nasution / 3311801020**

**Link Github : github.com/hennynst**

**Link Github Kelompok 6 : https://github.com/karniahhasanah/3311601006\_3311801012\_3311801020\_datamining\_polibatam**

Batam, …………….. 2019

Disetujui dan disahkan oleh:

|  |
| --- |
| Dosen pengajar,  **Muhammad Nashrullah**  **NIK/NIP.** |
|  |

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311801012

Nama : Novitasari

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET “*Breast Cancer Coimbra”*

DENGAN MENGGUNAKAN ALGORITMA C5.0

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**Novitasari**

**NIM. 3311801012**

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311601006

Nama : Karniah Hasanah

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET “*Breast Cancer Coimbra”*

DENGAN MENGGUNAKAN ALGORITMA C5.0

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**Karniah Hasanah**

**NIM. 3311601006**

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311801020

Nama : Henni Hendrani Nasution

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET “*Breast Cancer Coimbra”*

DENGAN MENGGUNAKAN ALGORITMA C5.0

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

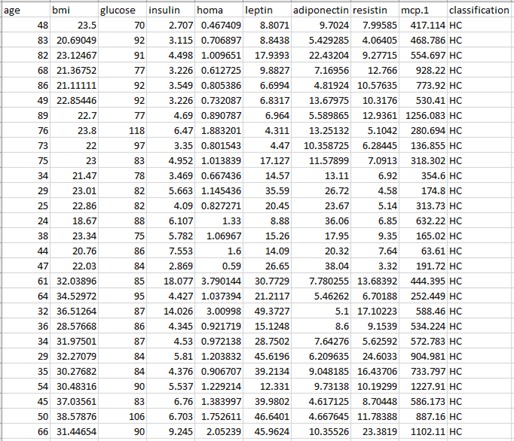
**Henni Hendrani Nasution**

**NIM. 3311801020**

# Dataset Yang Digunakan

Dataset yang kami gunakan adalah dataset yang bernama “*Breast Cancer Coimbra”*.Kanker payudara atau breast cancer merupakan kanker kedua yang paling banyak diderita serta menjadi penyebab kelima kematian kanker diseluruh dunia dengan presentse sebesar 6.4%. Guna bertahan hidup, diagnosis kanker payudara harus dilakukan sedini mungkin sebelum sel kanker menyebar ke bagian tubuh lainnya.

* Tujuan Dataset : Dataset ini bertujuan untuk menentukan apakahn seseorang Menderita kanker payudara atau hanya melakukan pemeriksaan kesehatan.
* Dataset Breast Cancer Coimbra Memiliki 116 data. Dengan isi sebagai berikut :



* Informasi Kumpulan Data:

Ada 10 prediktor, semuanya kuantitatif, dan variabel dependen biner, yang menunjukkan ada tidaknya kanker payudara.

Prediktor adalah data antropometrik dan parameter yang dapat dikumpulkan dalam analisis darah rutin.

Model prediksi berdasarkan prediksi ini, jika akurat, berpotensi dapat digunakan sebagai biomarker kanker payudara.

* DatasetBreast Cancer Coimbra ini memiliki 10 attribut diantaranya :

1. Age (Umur ) : Usia responden
2. BMI ( Kg/m2) : *Body Mass Index*
3. Glukosa ( Mg/dl) : Kandungan Glukosa dalam darah
4. Insulin (ML) : Kandungan Insulin dalam darah
5. Homa : *Homeostasis Model Assessment*
6. Leptin (ML) : Kandungan Leptin dalam darah
7. Adiponectin (ML) : Kandungan Adiponectin dalam darah
8. Resistin (ML) : Kandungan Resistin dalam darah
9. MCP-1 (PG/DL) : Kandungan *Monocyte Chemoattractan Protein-1*
10. Classification : Status pasien (HC = Health Control, P = Patients)

# Algoritma Yang Digunakan (*C5.0*)

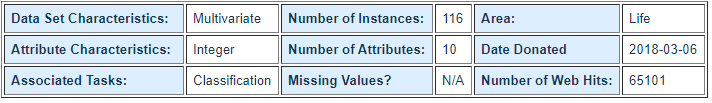
Algoritma C5.0 adalah salah satu algoritma data mining yang khususnya diterapkan pada teknik decision tree. C5.0 merupakan penyempurnaan algoritma sebelumnya yang dibentuk oleh Ross Quinlan pada tahun 1987, yaitu ID3 dan C4.5. Dalam algoritma ini pemilihan atribut yang akan diproses menggunakan information gain. Dalam memilih atribut untuk pemecah obyek dalam beberapa kelas harus dipilih atribut yang menghasilkan information gain paling besar. Atribut dengan nilai information gain tertinggi akan dipilih sebagai parent bagi node selanjutnya {4} C5.0 adalah versi komersial dari C4.5 yang secara luas digunakan di banyak pemaketan data mining seperti Clementine and RuleQuest. Tidak seperti C4.5, penggunaan algoritma yang tepat untuk C5.0 belum terungkap. Hasil menunjukkan bahwa C5.0 meningkatkan pada penggunaaan memori sekitar 90%, lebih cepat daripada C4.5.

**Alat:**

* R dan R Studio

**Deskripsi:**

* Gambaran klinis yang diamati dari 64 pasien dengan kanker payudara dan 52 orang yang kontrol kesehatan.



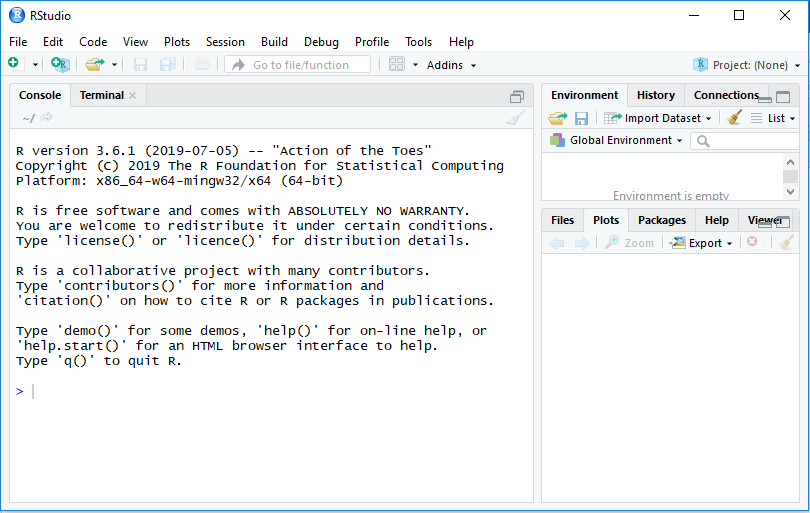
* Dataset ini memiliki 10 attribut. Attribut 1-9 adalah masukan sedangkan attribut 10 adalah keluaran. Keluaran dari data ini yang menunjukkan apakah seseorang menderita kanker payudara atau tidak. Dataset ini merupakan data yang diambil dari pemeriksaan darah rutin. Model prediksi berdasarkan data ini, jika akurat, berpotensi dapat digunakan sebagai biomarker kanker payudara sehingga dapat menentukan langkah apa yang harus diambil selanjutnya.

**Labels:**

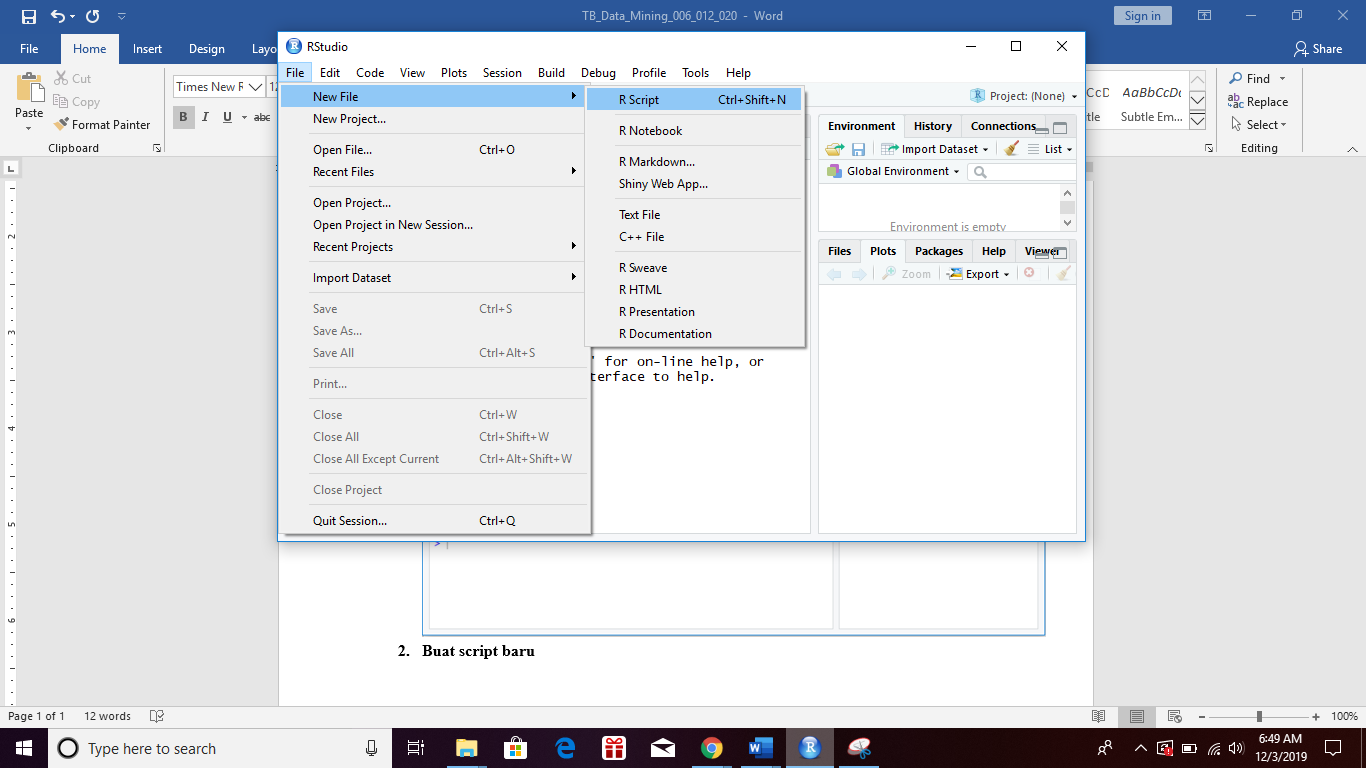
HC : Health Controls

P : Patients

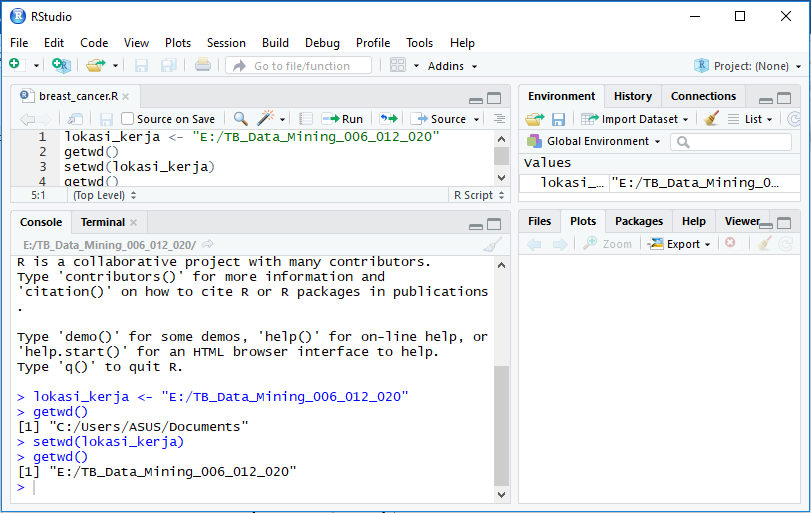
1. **Implementasi**
2. **Buka Rstudio**



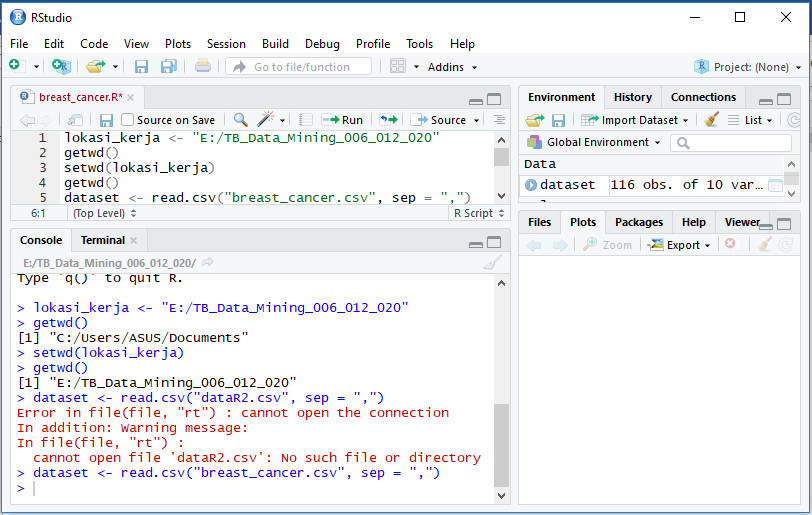
1. **Buat script baru**



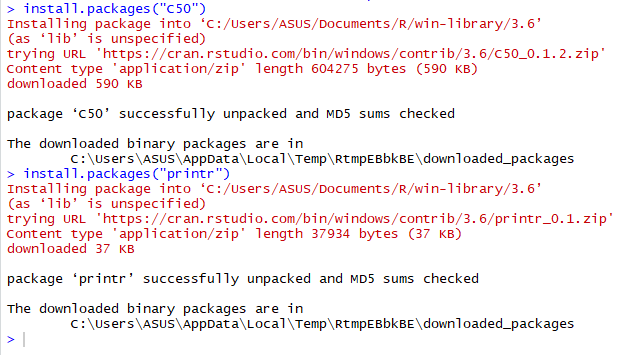
1. **Pengaturan lokasi directory**



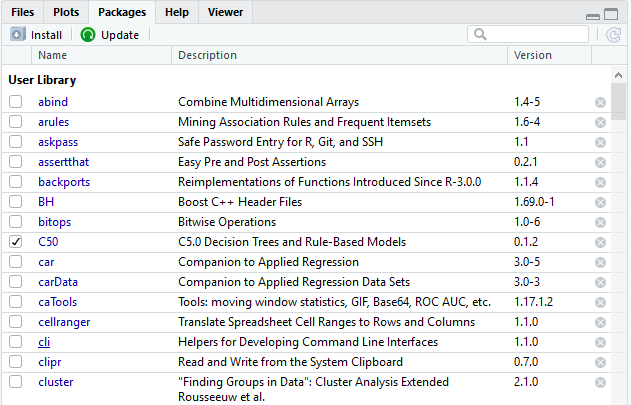
1. **Membaca data pada file breast\_cancer.csv**



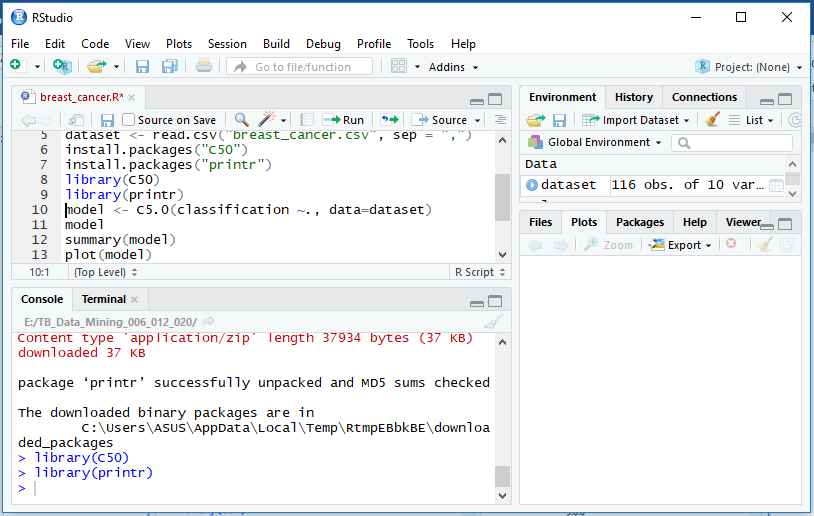
1. **Instalasi Package**



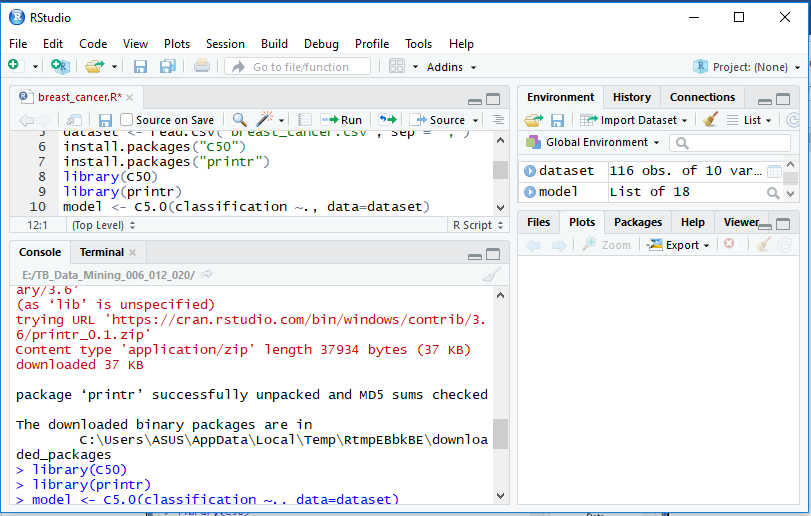
1. **Untuk melihat package yang sudah terinstal pada computer anda, perhatikan pada bagian kanan bawah**



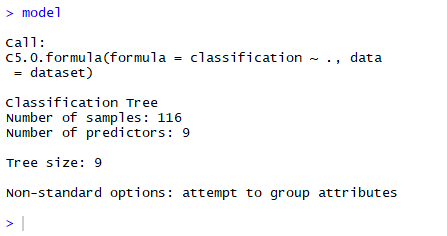
1. **Gunakan package**

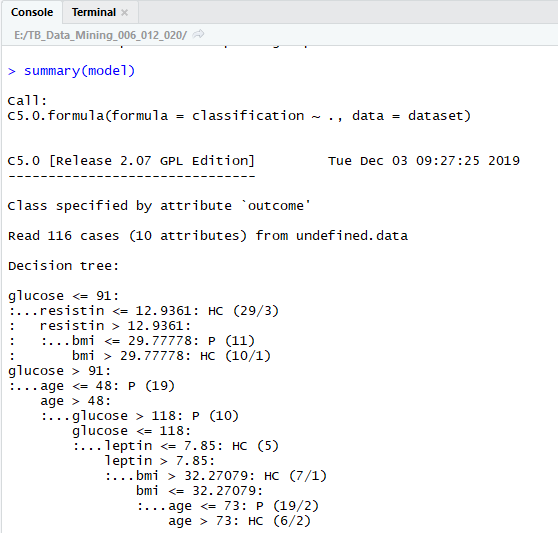


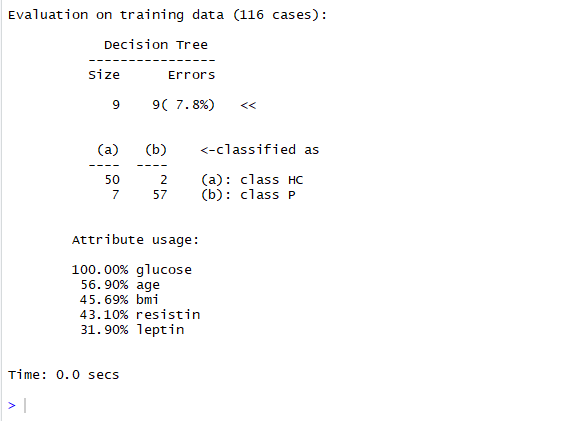
1. **Pembuatan model decision tree menggunakan algoritman C5.0**



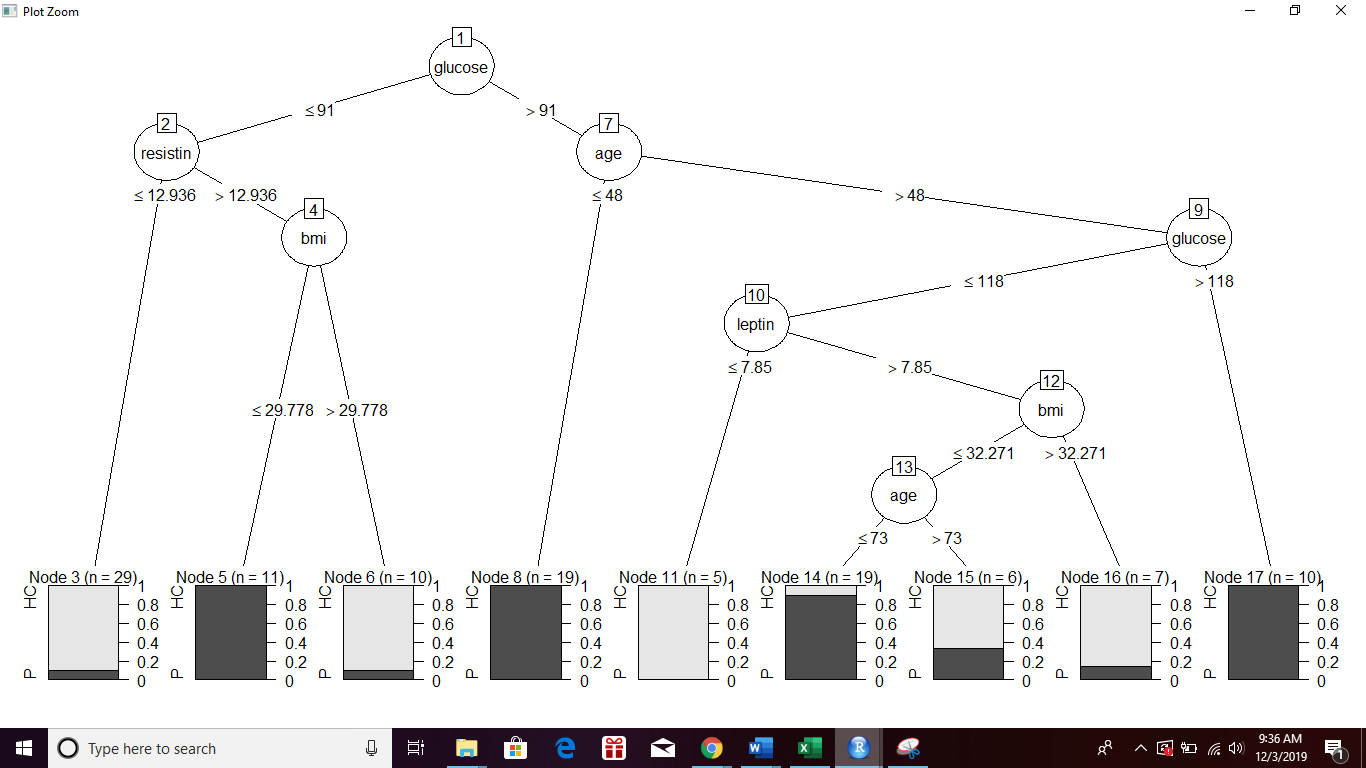
1. **Melihat model**



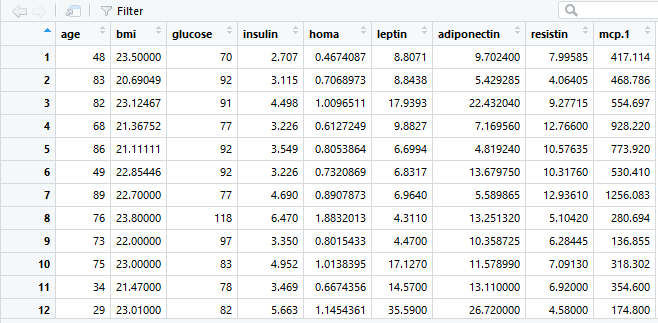




1. **Menampilkan pohon yang sudah dibangun**



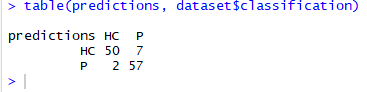
1. **Menjadikan dataset, sebagai data testing. Namun hanya 9 kolom saja dan tanpa label.**



1. **Prediksi**



1. **Membandingkan hasil prediksi dengan dataset**



1. **Rule yang dihasilkan dari proses klasifikasi**

**Rule 1:** if (glucose <=91) and (resistin <=12,9361) then classification = HC

**Rule 2:** if (glucose <=91) and (resistin > 12,9361) and (bmi <= 29,77778) then classification = P

**Rule 3:** if (glucose <=91) and (resistin > 12,9361) and (bmi > 29,77778) then classification = HC

**Rule 4:** if (glucose > 91) and (age <= 48) then classification = P

**Rule 5:** if (age > 48) and (glucose > 118) then classification = P

**Rule 6:** if (glucose <= 118) and (leptin <= 7,85) then classification = HC

**Rule 7:** if (leptin > 7,85) and (bmi > 32,27079) then classification = HC

**Rule 8:** if (bmi <= 32,27079) and (age <= 73) then classification = P

**Rule 9:** if (bmi <= 32,27079) and (age > 73) then classification = HC

1. **Kesimpulan**

Breast Cancer Coimbra Dataset adalah sekumpulan data pemeriksaan darah rutin yang dikumpulkan oleh patricio, dkk di University Hospital Centre of Coimbra. Dataset ini digunakan untuk memprediksikan seseorang menderita kanker payudara atau tidak dengan menggunakan metode klasifikasi dan algoritma C5.0. Kesimpulan dari hasil klasifikasi dataset ini adalah:

1. Memiliki glucose kurang dari atau sama dengan 91 dan resistin lebih dari 12,9361 memiliki 2 kemungkinan yaitu dengan melihat hasil data bmi. Jika data bmi kurang dari atau sama dengan 29,77778 maka orang itu menderita kanker payudara.
2. Memiliki glucose lebih dari 91 dan berumur dibawah atau sama dengan 48 juga menderita kanker payudara.
3. Memiliki glucose lebih dari 118 dan berumur lebih dari 48 juga menderita kanker payudara.
4. Memiliki bmi kurang dari atau sama dengan 32,27079 dan berumur kurang dari atau sama dengan 73 juga menderita kanker payudara.

# Link

* Link Referensi
* <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>
* Link Github Kelompok
* <https://github.com/karniahhasanah/3311601006_3311801012_3311801020_datamining_polibatam>
* Link Akun Github Anggota
* Kelompok 6

- Henni (github.com/hennynst)

- Kak Karniah (github.com/karniahhasanah)

- Novita (github.com/25Novitasari)