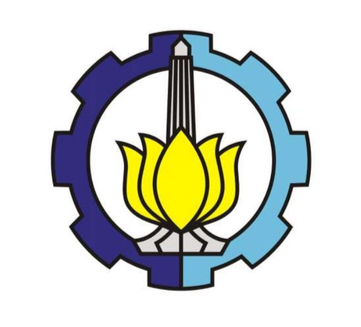
## TUGAS 2

**UJI COBA ALGORITMA J48 WEKA DENGAN ABALONE & BREAST CANCER DATASET**

****

### **Disusun oleh :**

Dzaky Zakiyal Fawwaz 5114100067

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**SURABAYA**

**DATASET ABALONE**

### **Penjelasan Dataset**

* 1. Jumlah data : 4117
  2. Jumlah atribut : 9 ditambah atribut kelas
  3. Nilai atibut yang hilang : Tidak ada
  4. Atribut :

|  |  |
| --- | --- |
| **Nama** | **Tipe data** |
| Sex | Nominal |
| Length | Kontinu |
| Diameter | Kontinu |
| Height | Kontinu |
| Whole weight | Kontinu |
| Shucked weight | Kontinu |
| Viscera weight | Kontinu |
| Shell weight | Kontinu |
| Rings | Integer |

* 1. Kelas : Rings
  2. Distribusi kelas (Rings) :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Kelas** | **Jumlah** | **Kelas** | **Jumlah** | **Kelas** | **Jumlah** |
| 1 | 1 | 11 | 487 | 21 | 14 |
| 2 | 1 | 12 | 267 | 22 | 6 |
| 3 | 15 | 13 | 203 | 23 | 9 |
| 4 | 57 | 14 | 126 | 24 | 2 |
| 5 | 115 | 15 | 103 | 25 | 1 |
| 6 | 259 | 16 | 67 | 26 | 1 |
| 7 | 391 | 17 | 58 | 27 | 2 |
| 8 | 568 | 18 | 42 | 29 | 1 |
| 9 | 689 | 19 | 32 |  |  |
| 10 | 634 | 20 | 26 |  |  |

### **Langkah uji coba**

* 1. Download data pada <https://archive.ics.uci.edu/ml/datasets/abalone>, terutama “abalone.data”
  2. Rename ekstensi “abalone.data” menjadi “abalone.arff”
  3. Kemudian ubah isi datanya sesuai Attribute-Relation File Format (ARFF) <https://www.cs.waikato.ac.nz/ml/weka/arff.html> yaitu dengan menambah header diatas data

@RELATION abalone

@ATTRIBUTE Sex {M,F,I}

@ATTRIBUTE Length numeric

@ATTRIBUTE Diameter numeric

@ATTRIBUTE Height numeric

@ATTRIBUTE Whole\_weight numeric

@ATTRIBUTE Shucked\_weight numeric

@ATTRIBUTE Viscera\_weight numeric

@ATTRIBUTE Shell\_weight numeric

@ATTRIBUTE Rings {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,29}

@DATA

........

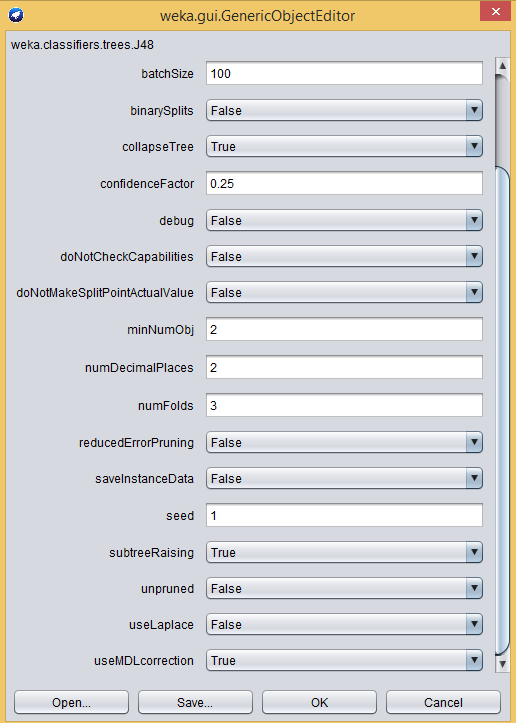
* 1. Hal diatas dilakukan karena J48 pada wekaa hanya dapat berjalan jika kelas data berformat nominal
  2. Buka aplikasi Weka
  3. Pilih Explorer



* 1. Pada tab Prepocess open file “abalone.arff”
  2. Kemudian pilih tab Classify
  3. Pilih choose lalu pilih algoritma J48 pad folder trees
  4. Klik dua kali pada J48 – C 0.25 – M2, akan muncul parameter-parameter yang bisa kita atur. Untuk penjelasan masing-masing parameter, pilih tombol “more”.

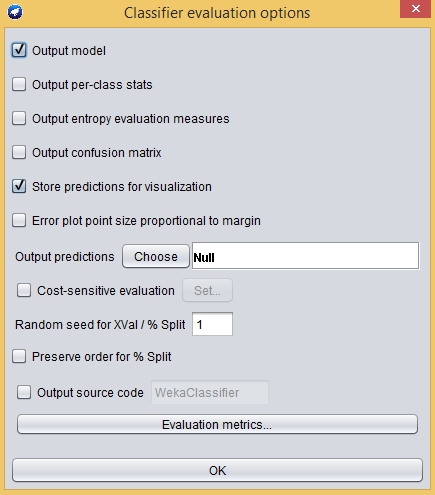


* 1. Berikut adalah pengaturan parameter yang digunakan pada uji coba

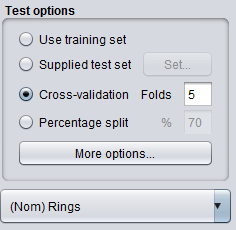


Penjelasan :

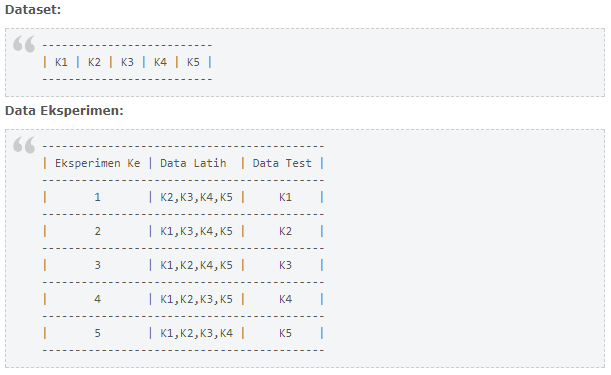
* seed – digunakan untuk randomisasi data ketika reduced-error pruning digunakan
* unpruned – pengaturan apakah pruning digunakan
* confidenceFactor -- confidence factor digunakan dalam pruning (semakin kecil semakin banyak proses pruning)
* numFolds – menentukan jumlah data yang digunakan untuk reduced-error pruning. sekalinya fold digunakan untuk pruning, sisanya digunakan untuk membangun tree
* numDecimalPlaces – jumlah desimal dibelakang koma yang digunakan pada ouput di model
* batchSize – ukuran batch data yang diproses yang diinginkan
* reducedErrorPruning – pengatuan apakah reduced-error pruning digunakan
* useLaplace – pengaturan apakah menggunakan laplace, yang dipakai dalam menghitung leaf lebih mudah
* doNotMakeSplitPointActualValue – Jika aktif, split point tidak akan ditampung dalam nilai data sesungguhnya. Fitur ini dapat mempercepat dataset besar dengan atribut numerik
* debug – Jika aktif, cclassifier akan memberi informasi output tambahan pada konsol
* subtreeRaising – pengaturan apakah menggunakan operasi subtree raising saat pruning.
* saveInstanceData – pengaturan apakah menyimpan training data untuk visualisasi
* binarySplits – pengaturan apakah menggunakan binary splits pada atribut nominal
* doNotCheckCapabilities – jika aktif, kapabilitas classifier tidak akan dicek sebelum classifier dibentuk
* minNumObj – minimal jumlah data per leaf
* useMDLcorrection – pengaturan apakah menggunakan MDL correction ketika mencari split pada atribut numerik
* collapseTree – pengaturan apakah jika bagian-bagian terbuang akan tidak mempengaruhi reduce training error
  1. Kemudian klik OK
  2. Pada tombol more option Anda bisa mengatur apa saja hasil keluaran pada konsol output, agar lebih mudah dibaca gunakan pengaturan berikut



* 1. Lalu atur test options, pada uji coba kali ini menggunakan Cross validation dengan fold 5. Lalu dropdown dibawahnya “(Nom) Rings” adalah target klasifikasi



* 1. Berikut gambaran 5-Fold Cross validation, yaitu dataset di bagi menjadi 5 kelompok secara acak.

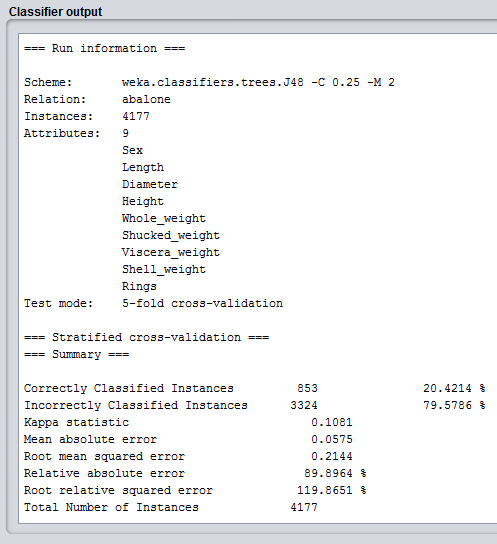


Jadi seluruh dataset akan dijadikan training set sekaligus test set, total data abalone ialah 4177

* 1. Kemudian pilih Start untuk memproses
  2. Hasil akan dibahas dibab selanjutnya

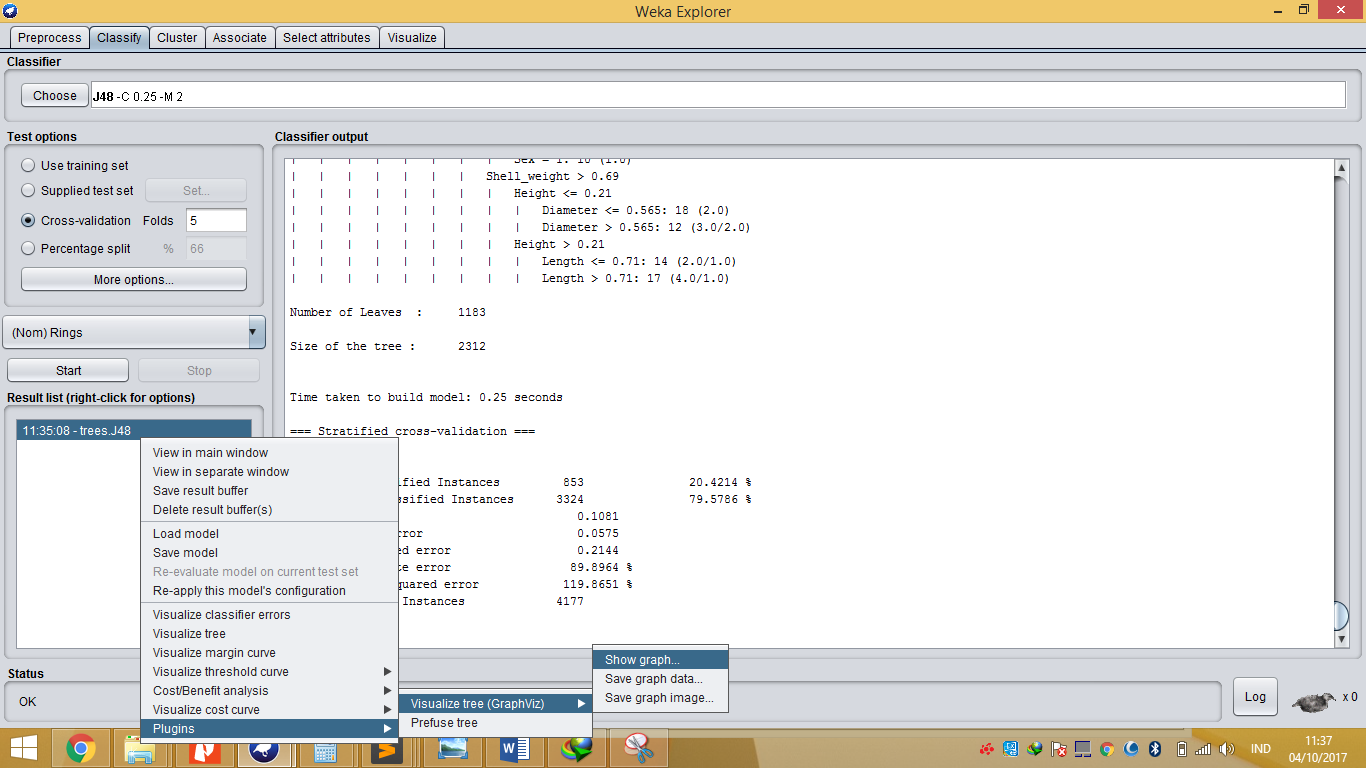
### **Hasil uji coba**

* 1. Berikut hasil dari uji boba yang dilakukan



Yaitu model j48 yang dibentuk berhasil mengklasifikasi 853 data dan gagal 3324 data. Akurasi dari model ini adalah 20.4214%

* 1. Pilih visualize tree untuk menampilkan tree dengan tool bawaan, tapi jika pada uji coba kali ini digunakan Graphviz, sehingga gambar bisa di save <https://github.com/fracpete/graphviz-treevisualize-weka-package/>



Karena tree terlalu besar, gambar bisa dilihat melalui <https://drive.google.com/file/d/0B977-wY19vMaaG10d3k1R3hUUFk/view>

### **Pembahasan hasil**

Meningkatkan akurasi

* 1. Menjadikan true pada binarySplits –apakah menggunakan binary splits pada atribut nominal. Pada uji coba, akurasi naik menjadi 21.355%
  2. Menurunkan confidenceFactor -- confidence factor digunakan dalam pruning (semakin kecil semakin banyak proses pruning). Pada uji coba nilai 0.05 akurasi 21.8339%, pada nilai 0.001 akurasi 26.2868%
  3. Menaikan minNumObj – minimal jumlah data per leaf. Pada uji coba nilai 5 akurasi 23.1027%, pada nilai 15 akurasi 25.2574%
  4. Menjadikan true pada reducedErrorPruning – apakah reduced-error pruning digunakan. Pada ujicoba, akurasu naik menjadi 25.018%
  5. Gunakan test options “use training set” dimana seluruh training set akan menjadi test set. Akurasi akan meningkat menjadi 75.7242%

**DATASET BREAST CANCER**

### **Penjelasan Dataset**

* 1. Jumlah data : 699
  2. Jumlah atribut : 10 ditambah atribut kelas
  3. Nilai atibut yang hilang : 16
  4. Atribut :

|  |  |
| --- | --- |
| **Nama** | **Tipe data** |
| Sample code number | Ordinal |
| Clump thickness | Nominal (1-10) |
| Uniformity of cell size | Nominal (1-10) |
| Uniformity of cell shape | Nominal (1-10) |
| Marginal adhesion | Nominal (1-10) |
| Single epithelial cell size | Nominal (1-10) |
| Bare nuclei | Nominal (1-10) |
| Bland chromatin | Nominal (1-10) |
| Normal nucleoli | Nominal (1-10) |
| Mitoses | Nominal (1-10) |
| Class | Nominal (2,4) |

* 1. Kelas : Class
  2. Distribusi kelas (Class) : 2 – Benign ada 458 (65.5%) dan 4 – Malignant ada 241 (34.5%)

### **Langkah uji coba**

* 1. Lakukan langkah-langkah seperti uji coba dataset abalone
  2. Berikut header diatas data asli setelah merubah “breast-cancer-wisconsin.data” menjadi “breast-cancer-wisconsin.arff”

@RELATION breast\_cancer

@ATTRIBUTE ID\_number numeric

@ATTRIBUTE Clump\_Thickness numeric

@ATTRIBUTE Cell\_Size numeric

@ATTRIBUTE Cell\_Shape numeric

@ATTRIBUTE Marginal\_Adhesion numeric

@ATTRIBUTE Epithelial\_Size numeric

@ATTRIBUTE Bare\_Nuclei numeric

@ATTRIBUTE Bland\_Chromatin numeric

@ATTRIBUTE Normal\_Nucleoli numeric

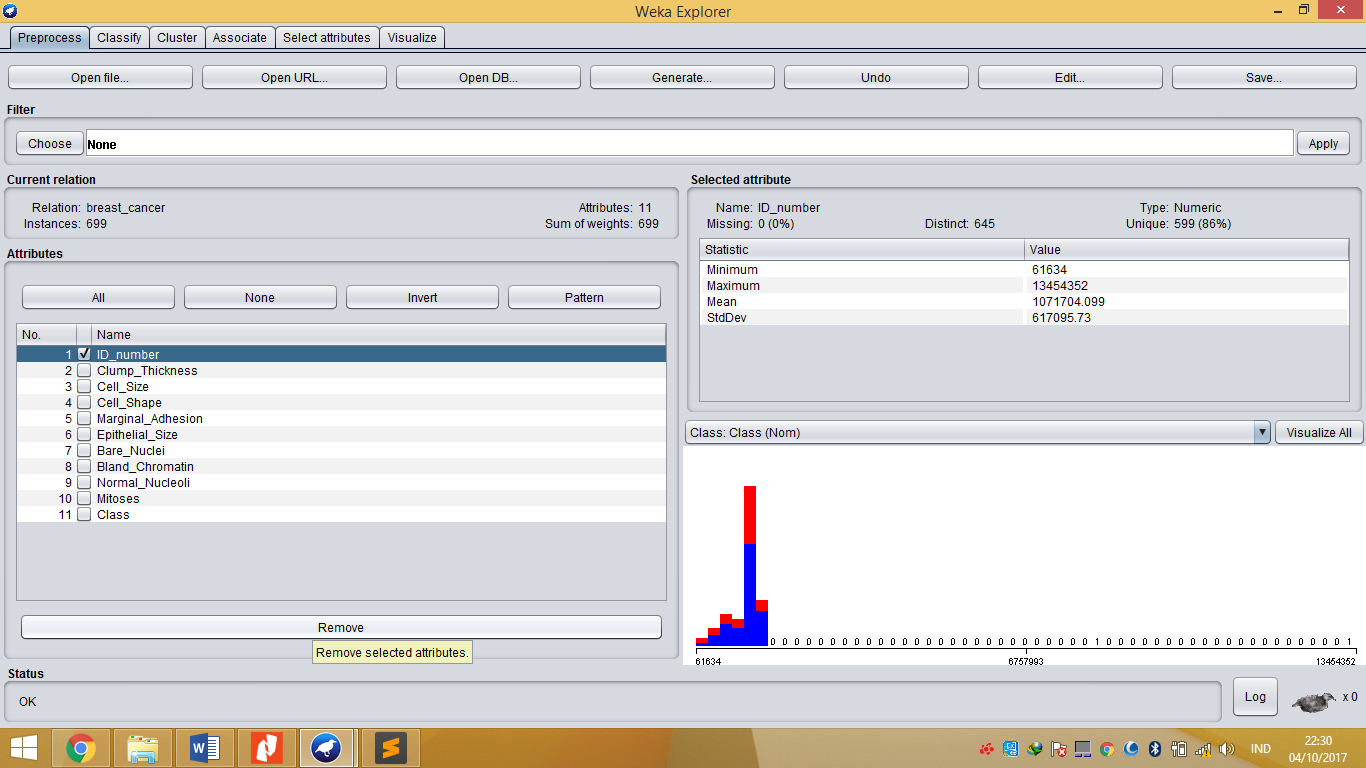
@ATTRIBUTE Mitoses numeric

@ATTRIBUTE Class {2,4}

@DATA

........

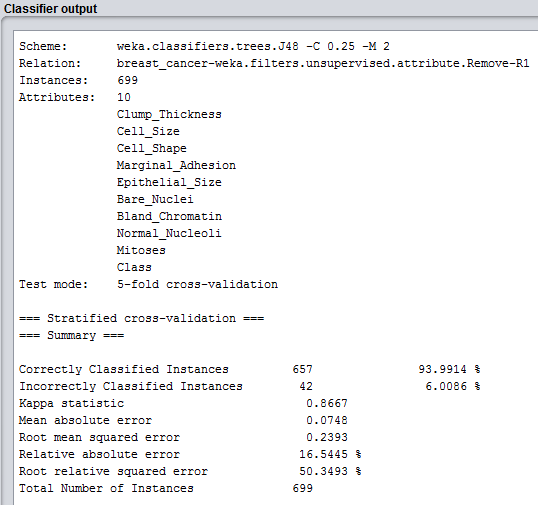
* 1. Pada “Preprocess” check ID\_number dan remove, karena atribut tersebut tidak diperlukan dalam klasifikasi, jika diikutkan justru hasil model tidak akan bagus.



* 1. Hasil akan dibahas dibab selanjutnya

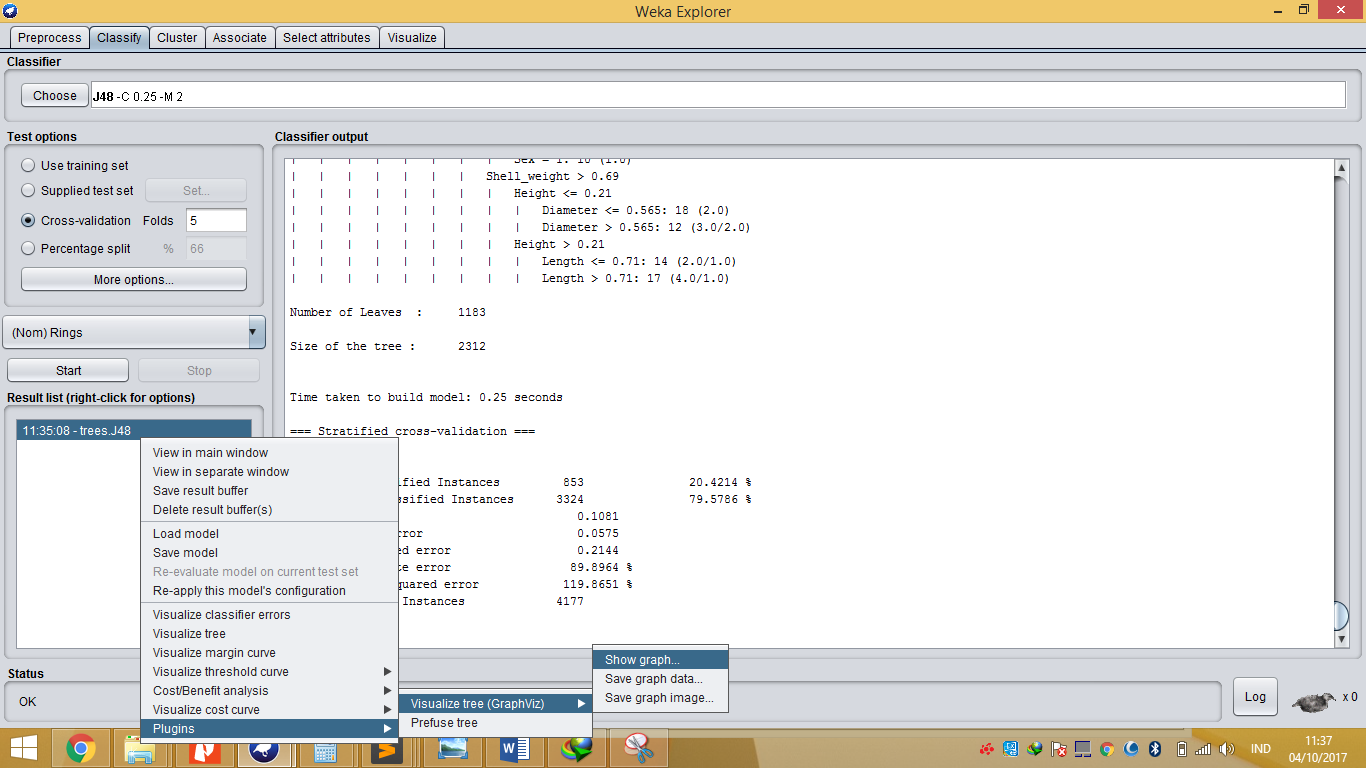
### **Hasil uji coba**

* 1. Berikut hasil dari uji boba yang dilakukan

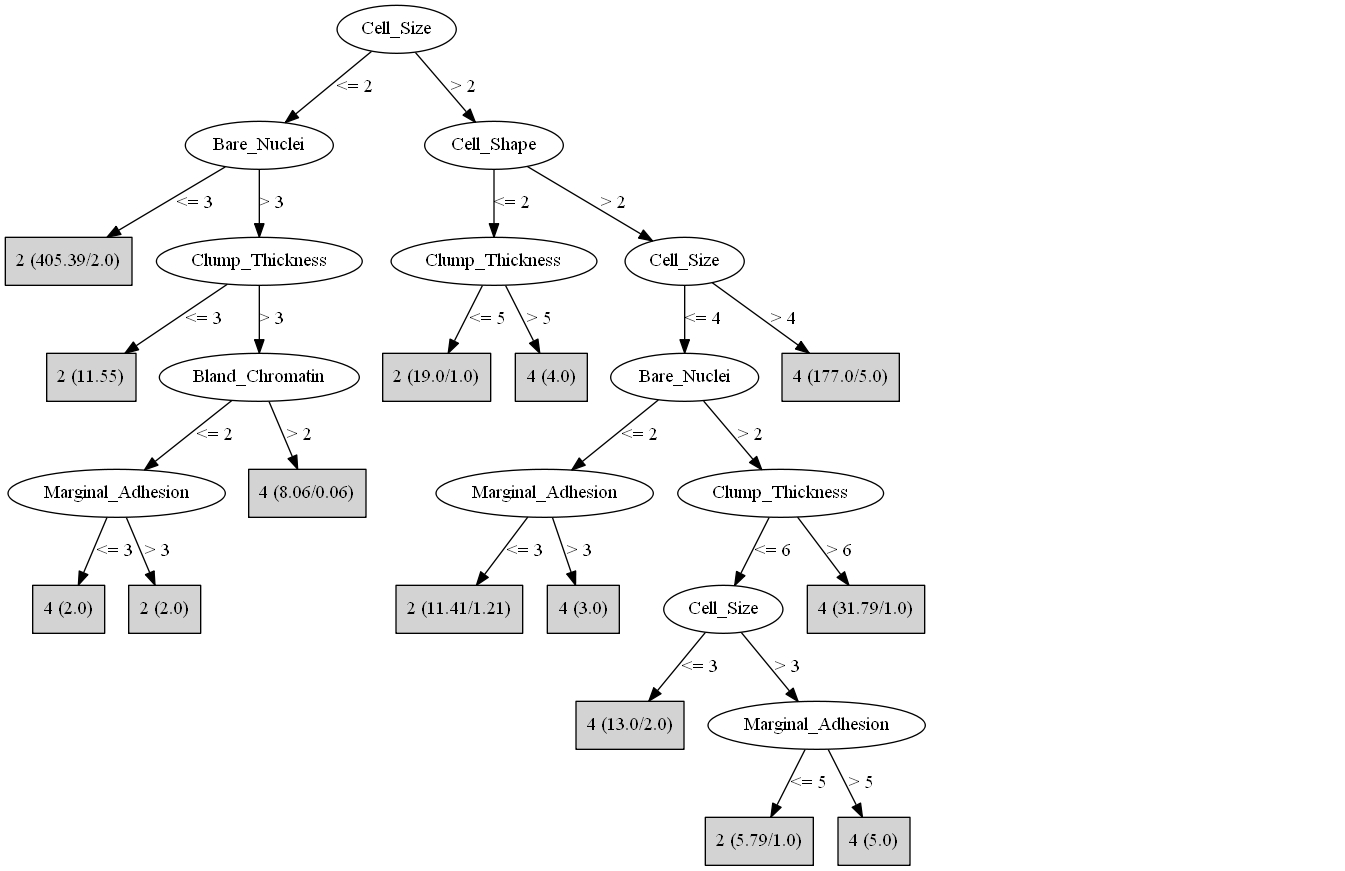


Yaitu model j48 yang dibentuk berhasil mengklasifikasi 657 data dan gagal 42 data. Akurasi dari model ini adalah 93.9914%

* 1. Pilih visualize tree untuk menampilkan tree dengan tool bawaan, tapi jika pada uji coba kali ini digunakan Graphviz, sehingga gambar bisa di save <https://github.com/fracpete/graphviz-treevisualize-weka-package/>



Berikut model tree yang terbentuk



### **Pembahasan hasil**

Meningkatkan akurasi

* 1. Menurunkan confidenceFactor -- confidence factor digunakan dalam pruning (semakin kecil semakin banyak proses pruning). Pada uji coba nilai 0.05 akurasi 94.4206%, pada nilai 0.001 akurasi 94.5637%
  2. Menaikan minNumObj – minimal jumlah data per leaf. Pada uji coba nilai 5 akurasi 94.2775%, pada nilai 7 akurasi 94.7067%
  3. Menjadikan true pada reducedErrorPruning – apakah reduced-error pruning digunakan. Pada ujicoba, akurasu naik menjadi 94.4206%
  4. Gunakan test options “use training set” dimana seluruh training set akan menjadi test set. Akurasi akan meningkat menjadi 98.1402%