Materi
Praktikum Data Mining

Decision Tree

Program Studi Informatika / Matematika

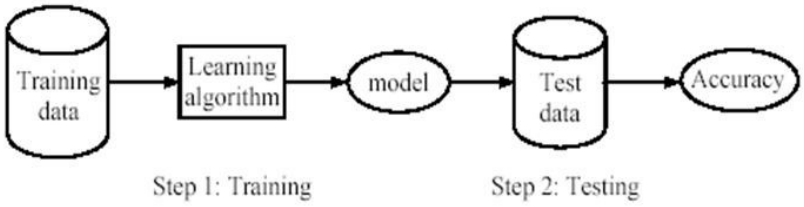
FMIPA Universitas Syiah Kuala

Dosen Pengasuh
Dr. Taufik Fuadi Abidin, M.Tech
Dr. Muhammad Subianto, M.Si

{tfa,subianto}@informatika.unsyiah.ac.id

PENDAHULUAN

Ada dua proses penting yang dilakukan saat melakukan klasifikasi. Proses yang pertama adalah **learning (training)** yaitu proses pembelajaran menggunakan training set. Untuk kasus Naïve Bayesian Classifier, perhitungan probabilitas dari data berdasarkan data pembelajaran dilakukan. Proses yang kedua adalah proses **testing** yaitu menguji model menggunakan data testing. Gambar berikut memperlihatkan alur dari kedua proses tersebut.



Gambar 1. Tahapan Proses Klasifikasi

Materi praktikum ini berkaitan dengan metode ***Decision Tree***. Ada dua tahapan yang harus dilakukan bila klasifikasi dilakukan menggunakan metode ini. Pertama adalah membangun pohon keputusan (decision tree) dan kedua, membangun aturan (rule) dari pohon keputusan yang dibangun. Tree dibangun secara ***top-down recursive divide-and-conquer*** dan data dipartisi secara rekursif berdasarkan atribut yang dipilih secdara heuristics menggunakan pengukuran statistik *information gain*. Partisi data berhenti jika tidak ada lagi data sampel yang tersisa, tidak ada lagi atribut yang dapat dipartisi atau semua data masuk ke dalam kelas label yang sama.

Perhitungan information gain untuk menentukan *atribut split* dilakukan sebagai berikut:

- Pilih atribut dengan nilai *information gain* tertinggi
- Jika S mengandung s_i sample dari class C_i untuk dimana $i = \{1, \ldots, m\}$
- Perhitungan besar informasi yang dibutuhkan untuk melakukan proses klasifikasi adalah:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} \frac{s_i}{S} \log_2 \frac{s_i}{S}$$

- Entropy dari atribut *A* dengan nilai {a_1, a_2,…,a_v}

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{S} I(s_{1j}, \ldots, s_{mj})$$

- Information gain dihitung sebagai berikut:

$$\text{Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

TUJUAN

Tujuan dari praktikum ini adalah a) meningkatkan pemahaman mahasiswa terhadap metode Decision Tree (yang dalam perangkat lunak disebut dengan metode J48), b) memahami cara melakukan klasifikasi dengan metode Decision Tree (membangun tree) menggunakan perangkat lunak Weka dan c) Memahami hasil klasifikasi.

Gambar 2. Tampilan GUI Weka

KEGIATAN PRAKTIKUM

1. Pahami data ***contact-lense.arff*** yang disimpan dalam direktori data dimana aplikasi Weka di install. Data ***contact-lense.arff*** memiliki 4 atribut nominal (categorical) dan 1 klas label yang menerangkan apakah seseorang menggunakan **hard contact lenses, soft contact lenses, atau tidak perlu menggunakan contact lenses.**

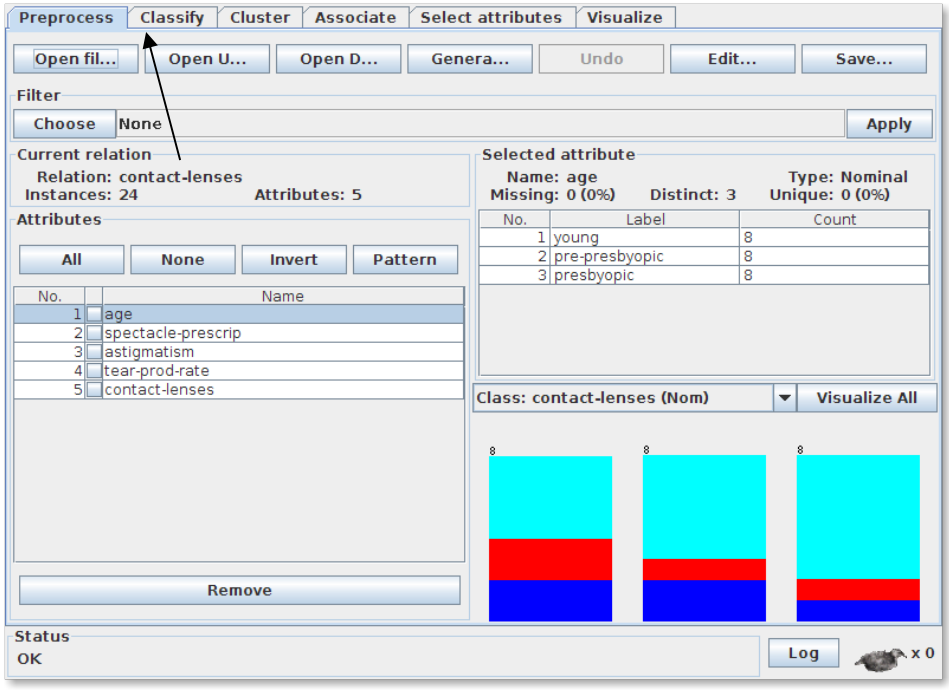
   Buka file arff tersebut dengan menggunakan text editor (*gedit, textpad atau vim*) dan pelajari bagaimana data dalam format arff disusun. Perhatikan penjelasan di bagian atas file tersebut dan perhatikan pula bagian deklarasi seperti potongan pernyataan berikut:

```
@relation contact-lenses

@attribute age              {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip   {myope, hypermetrope}
@attribute astigmatism          {no, yes}
@attribute tear-prod-rate       {reduced, normal}
@attribute contact-lenses       {soft, hard, none}
```

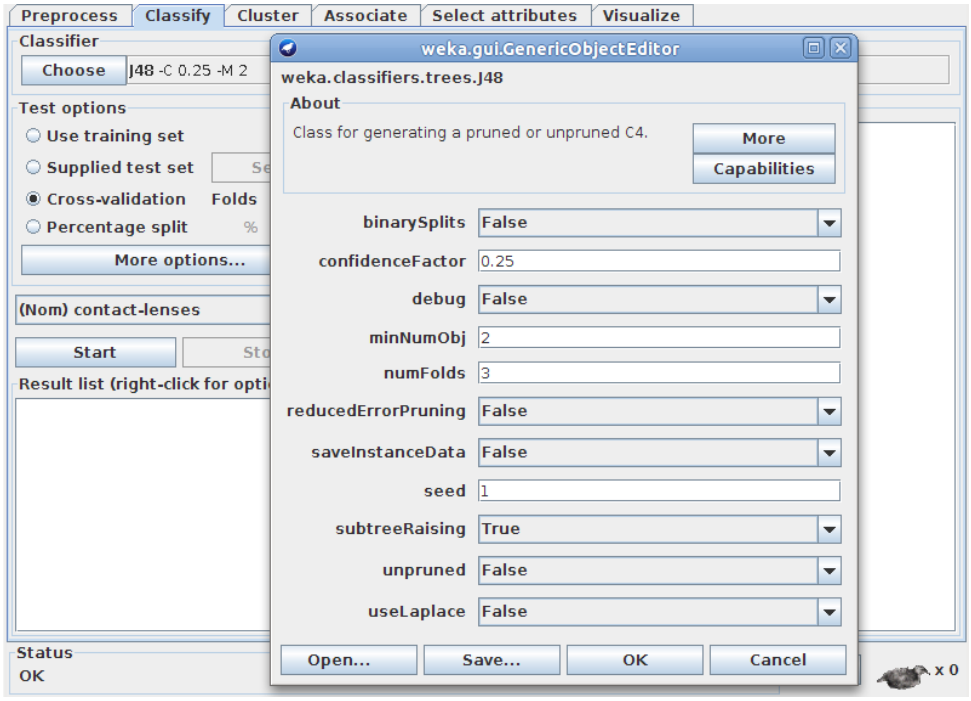
```
@data
%
% 24 instances
%
young,myope,no,reduced,none
young,myope,no,normal,soft
young,myope,yes,reduced,none
:
:
```

2. Setelah anda memahami bagaimana data dalam format arff dari data **contact-lense** disusun, lakukan proses klasifikasi menggunakan metode **J48**. Sebagai acuan, berikut snapshot dari Weka setelah file ***contact-lense .arff*** dibuka.



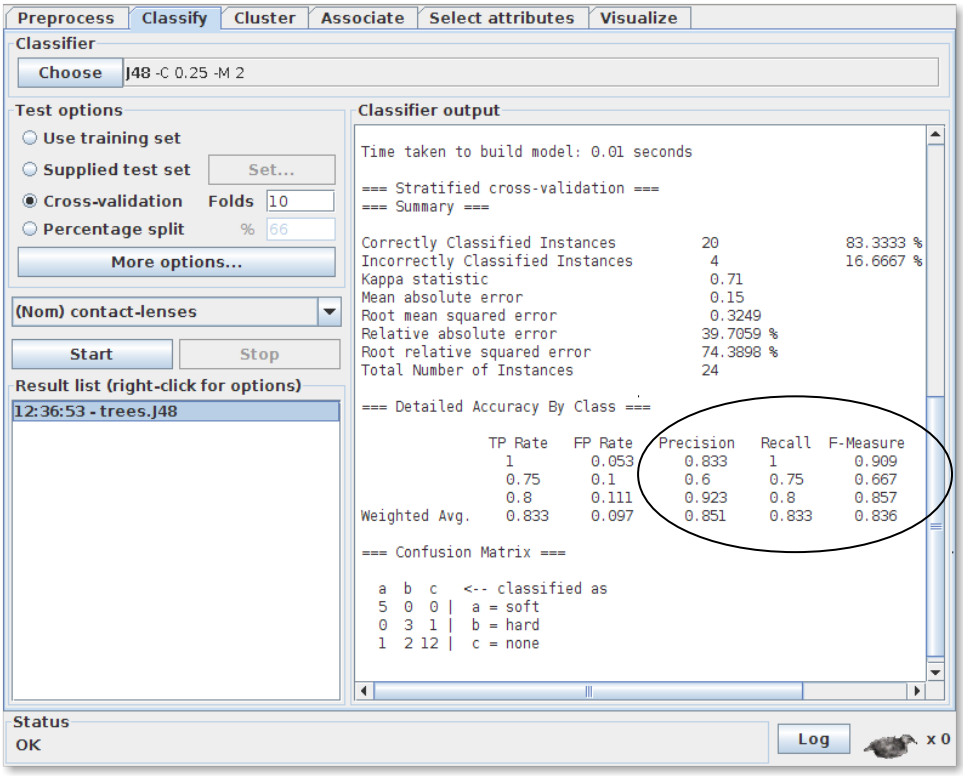
Gambar 3. Tampilan setelah dataset dibuka dalam Weka

- Lakukan observasi dan lanjutkan proses klasifikasi (*classify*) menggunakan metode **Decision Tree** (dalam software Weka disebut sebagai metode ***J48***). Snapshot berikut memperlihatkan tampilan Weka setelah metode ***J48*** dipilih.



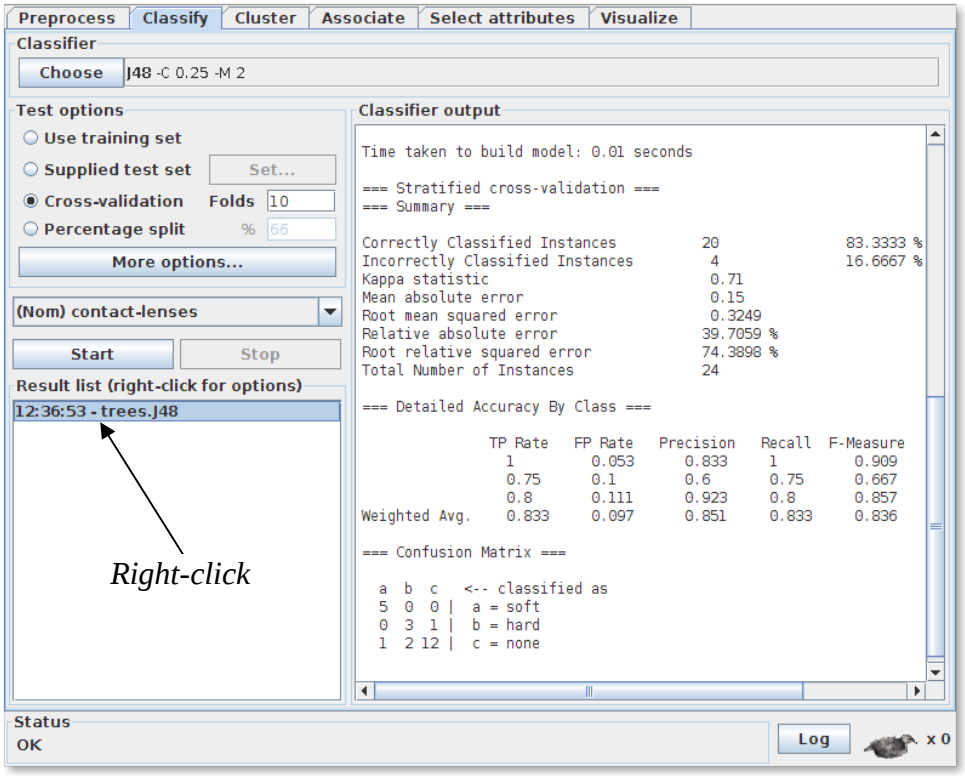
Gambar 4. Tampilan setelah metode J48 dipilih

- Lakukan proses testing menggunakan training set itu sendiri dengan memilih **Cross-validation** sebagai **Test Options**. Set parameter **folds = 10**, hal ini berarti sistem akan mengacak data training set dan mengambil sebagian dari datanya untuk dijadikan testing set. Proses ini dilakukan sebanyak 10 kali dan hasil akhir merupakan akurasi rata-rata dari kesepuluh percobaan tersebut. Diskusikan hasil yang diperoleh dengan teman dan asisten. Lakukan *tuning* (perubahan) pada parameter di atas dan perhatikan hasilnya.



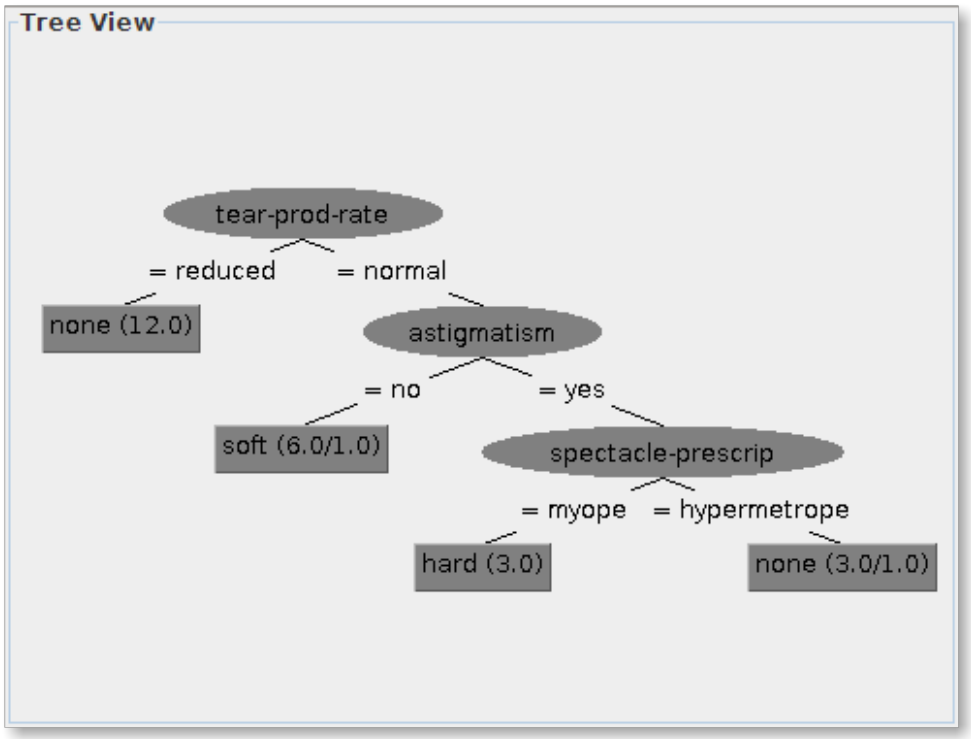
Gambar 5. Hasil klasifikasi menggunakan Decision Tree (J48 di Weka)

Klik tombol mouse-kanan pada bagian seperti yang ditunjukkan oleh gambar berikut dan pilih **Visualize Tree**, maka pohon yang terbentuk akan ditampilkan.



Gambar 6. Right-click dan pilih visualize tree

Pohon keputusan ditampilkan sebagai berikut (pahami arti dari pohon tersebut).



Gambar 7. Tree yang dibangun oleh metode Decision Tree (J48)

- Ulangi observasi menggunakan **Cross-validation** dengan parameter **folds = 5**. Apakah terjadi perubahan yang signifikan terhadap akurasi?

- Ulangi juga proses klasifikasi menggunakan data **weather.arff**. Perhatikan nilai *precision*, *recall*, dan *F-measurenya*.

- - - Happy Mining - - -