

DATA MINING TOOL *WEKA*

Weka adalah sebuah perangkat lunak yang memiliki banyak algoritma *machine learning* untuk keperluan *data mining*. Weka juga memiliki banyak *tools* untuk pengolahan data, mulai dari *pre-processing*, *classification*, *regression*, *clustering*, *association rules*, dan *visualization*.

ABALONE DATASET

- number of instances: 4177
- number of attribute: 8
- attribute information

Name of attribute	Data Type	Measure	Description
Sex	Nominal	M, F, I	
Length	Continuous	Mm	ukuran panjang tempurung/kerang/kulit
Diameter	Continuous	Mm	ukuran diameter secara tegak lurus
Height	Continuous	Mm	
Whole weight	Continuous	Gram	
Shucked weight	Continuous	Gram	
Viscera weight	Continuous	Gram	
Shell weight	Continuous	Gram	
Rings	Integer	+1.5	





- missing attribute values: none

PROSES DATA MINING MENGGUNAKAN SOFTWARE WEKA

a. Preprocess

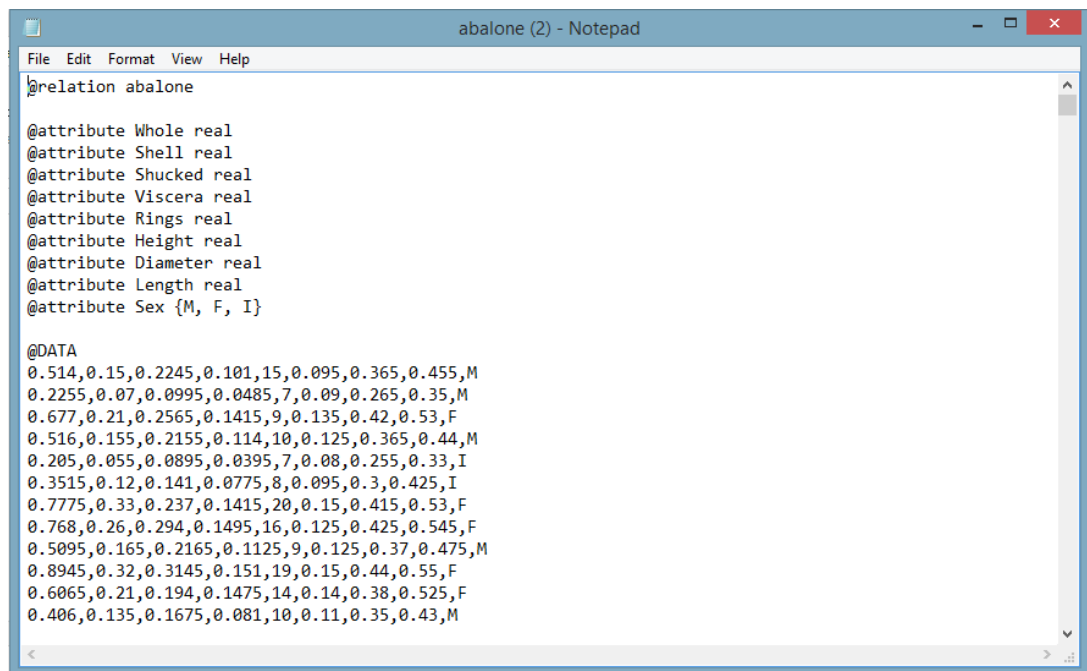
Memasukkan file dataset abalone yang sudah diubah ekstensinya menjadi .arff. Tahapan preprocess adalah sebagai berikut :

1. Download terlebih dahulu file abalone.data

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 Index	03-Dec-1996 04:06	114	
 abalone.data	12-Jun-1996 14:45	187K	
 abalone.names	12-Jun-1996 15:52	4.2K	

Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443

2. Kemudian ubah ekstensinya menjadi .arff seperti gambar berikut



```
@relation abalone

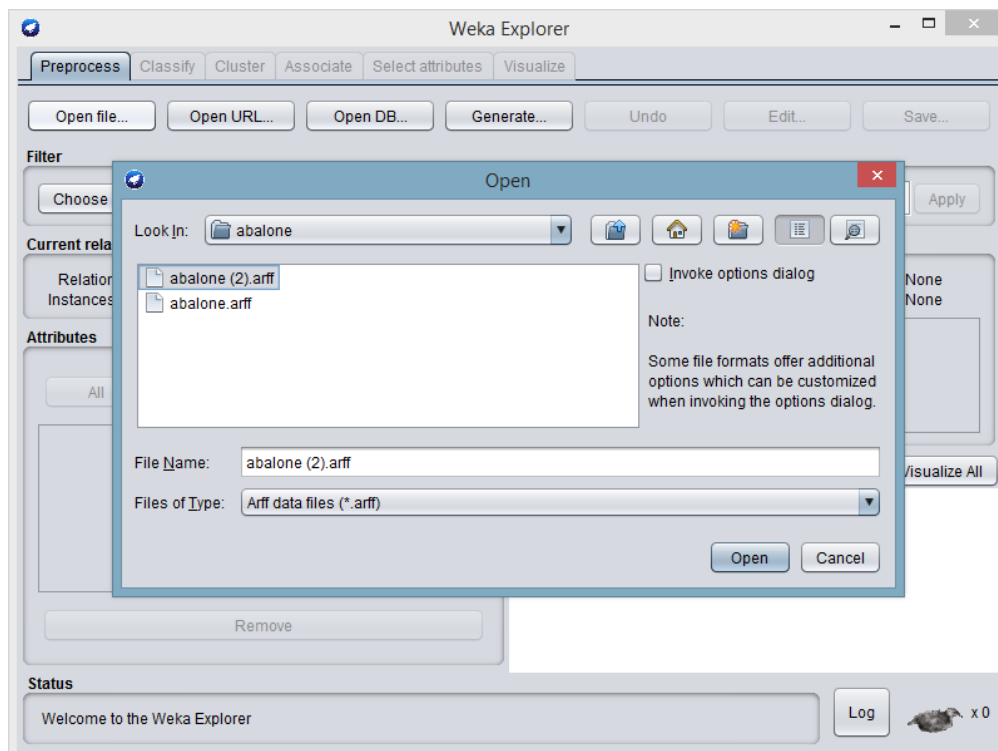
@attribute Whole real
@attribute Shell real
@attribute Shucked real
@attribute Viscera real
@attribute Rings real
@attribute Height real
@attribute Diameter real
@attribute Length real
@attribute Sex {M, F, I}

@DATA
0.514,0.15,0.2245,0.101,15,0.095,0.365,0.455,M
0.2255,0.07,0.0995,0.0485,7,0.09,0.265,0.35,M
0.677,0.21,0.2565,0.1415,9,0.135,0.42,0.53,F
0.516,0.155,0.2155,0.114,10,0.125,0.365,0.44,M
0.205,0.055,0.0895,0.0395,7,0.08,0.255,0.33,I
0.3515,0.12,0.141,0.0775,8,0.095,0.3,0.425,I
0.7775,0.33,0.237,0.1415,20,0.15,0.415,0.53,F
0.768,0.26,0.294,0.1495,16,0.125,0.425,0.545,F
0.5095,0.165,0.2165,0.1125,9,0.125,0.37,0.475,M
0.8945,0.32,0.3145,0.151,19,0.15,0.44,0.55,F
0.6065,0.21,0.194,0.1475,14,0.14,0.38,0.525,F
0.406,0.135,0.1675,0.081,10,0.11,0.35,0.43,M
```

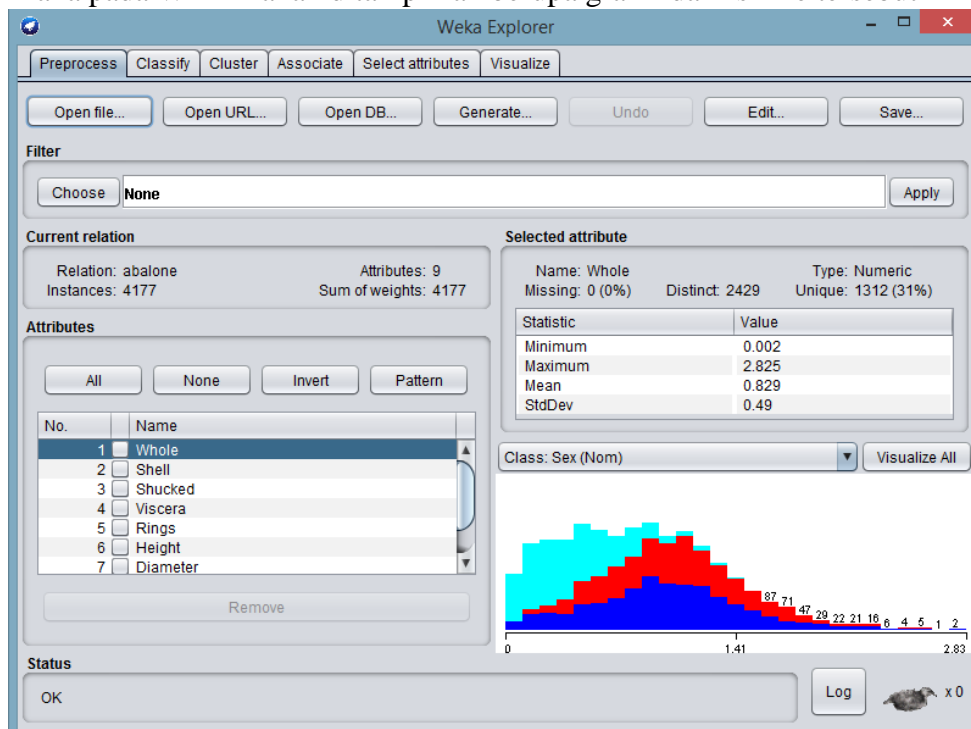
3. Kemudian buka weka dan pilih explorer



4. Kemudian akan muncul preproses, kemudian masukkan file abalone yang sudah diubah ekstensinya tadi

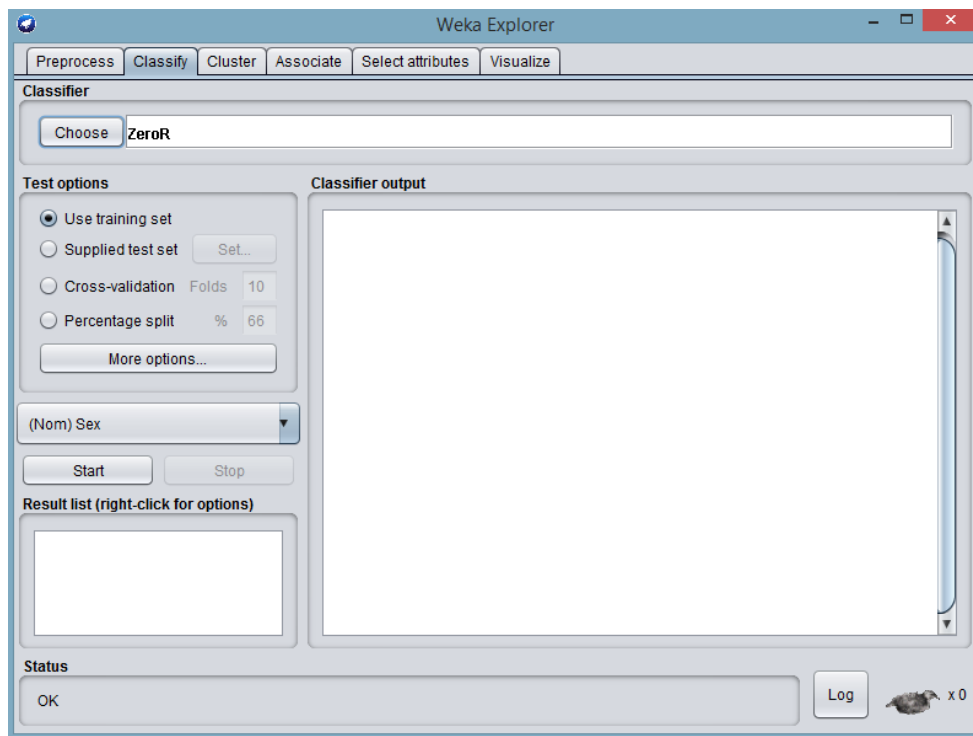


5. Maka pada WEKA akan ditampilkan berupa grafik dari isi file tersebut

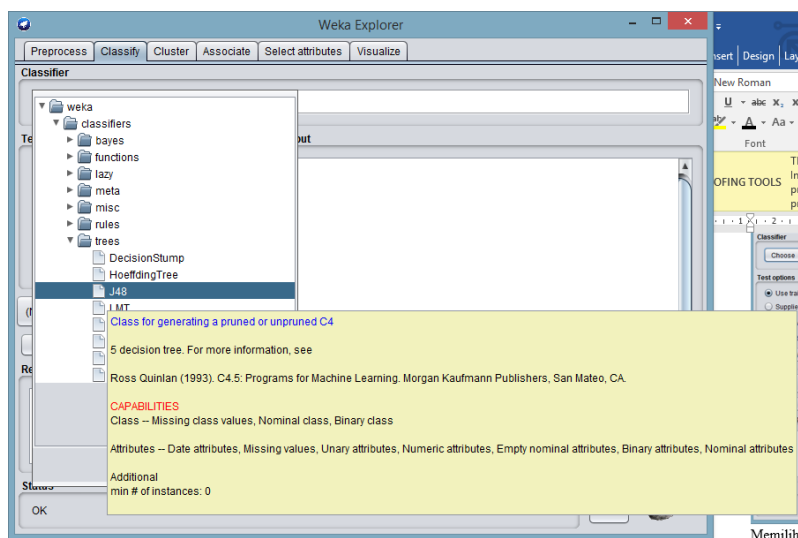


b. Classify

Pada classify ini memungkinkan untuk memilih salah satu dari pengklasifikasi yang tersedia di WEKA seperti yang ditunjukkan gambar berikut



Memilih metode Classify yang akan digunakan



Memilih algoritma yang digunakan sebagai *Classifier*

WEKA menyediakan penggunaan teknik klasifikasi menggunakan pohon keputusan dengan algoritma J48.

Kemudian klik start. Hasil menerapkan *classifier* yang dipilih akan diuji sesuai dengan pilihan yang ditetapkan dengan mengklik pada kotak Test Option

Ada empat option :

1. Use Training Set
Pengetesan dilakukan dengan menggunakan data training itu sendiri
2. Supplied Test Set

Pengetesan dilakukan menggunakan data lain. Dengan menggunakan option inilah, bisa dilakukan prediksi terhadap data tes

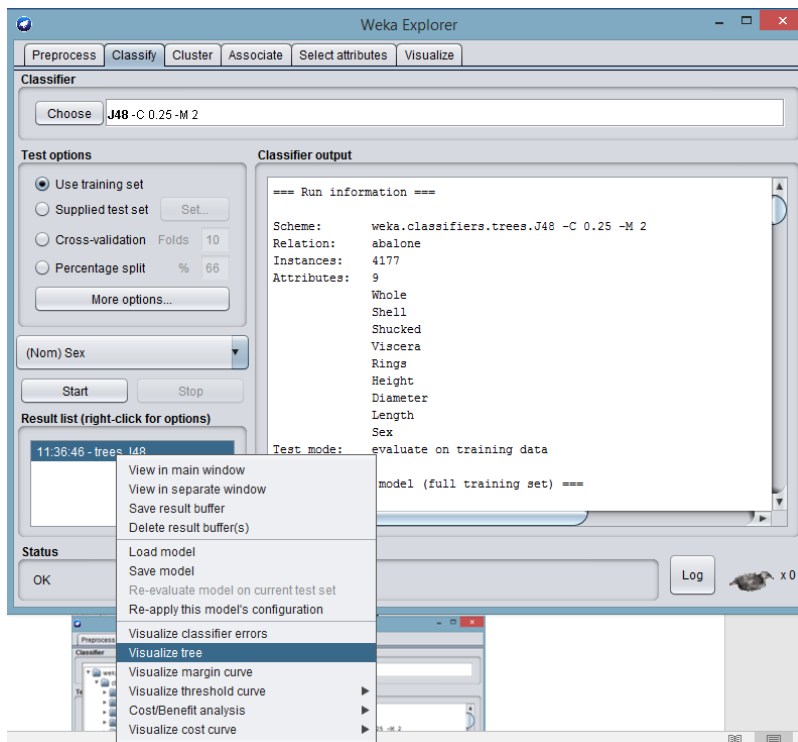
3. Cross-validation

Ada beberapa pilihan beberapa *fold* yang akan digunakan. Nilai *default*-nya adalah 10. Mekanismenya adalah data training dibagi menjadi k buah subset, dimana k adalah nilai dari *fold*. Selanjutnya untuk tiap subset akan dijadikan data tes dari hasil klasifikasi yang dihasilkan $k-1$ subset lainnya. Jadi akan ada 10 kali tes, dimana setiap *datum* akan menjadi data tes sebanyak 1x dan menjadi data training sebanyak $k-1$ x. Kemudian, error dari k test tersebut akan dihitung rata-ratanya.

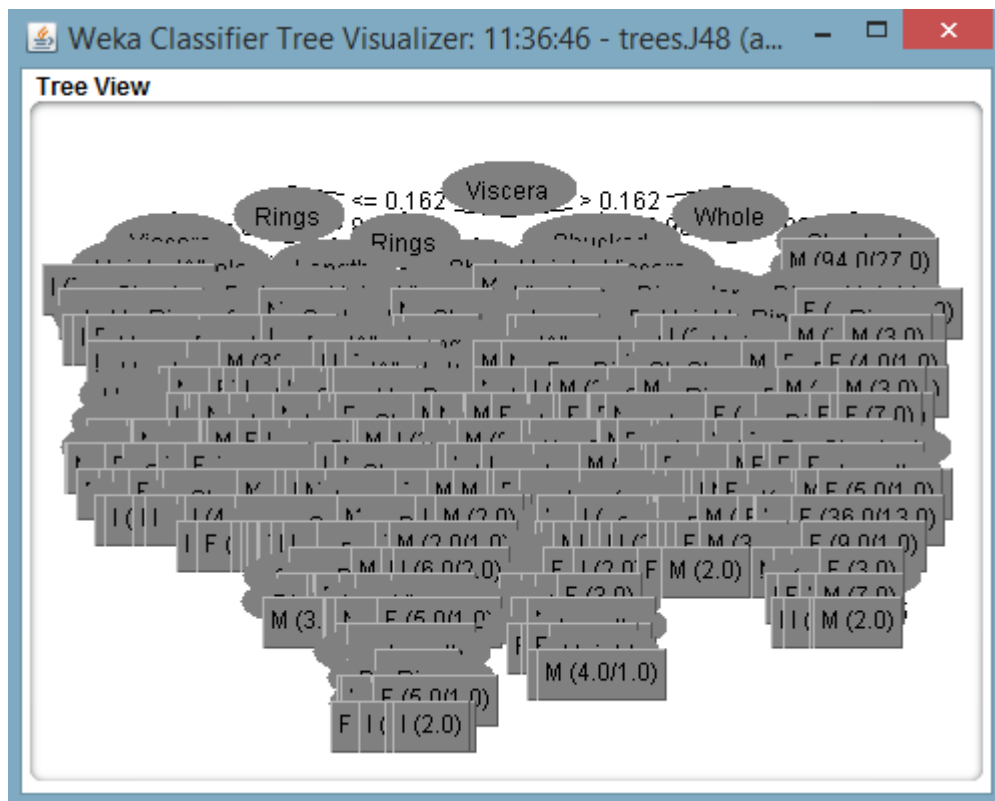
4. Percentage Split

Hasil klasifikasi akan dites dengan menggunakan $k\%$ dari data tersebut. k merupakan masukan dari user.

Kemudian klik start, maka akan ada informasi, kemudian klik kanan seperti gambar berikut dan klik visualize tree



Setelah di klik visualize tree akan muncul tree seperti berikut



Soal

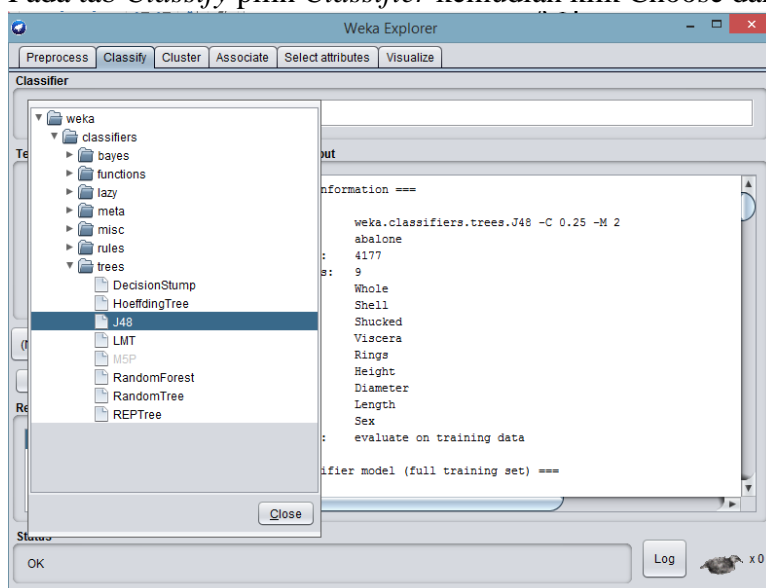
1. Get the “Abalone” dataset from UCI ML Repo

Jawab: sudah dijelaskan di atas 😊

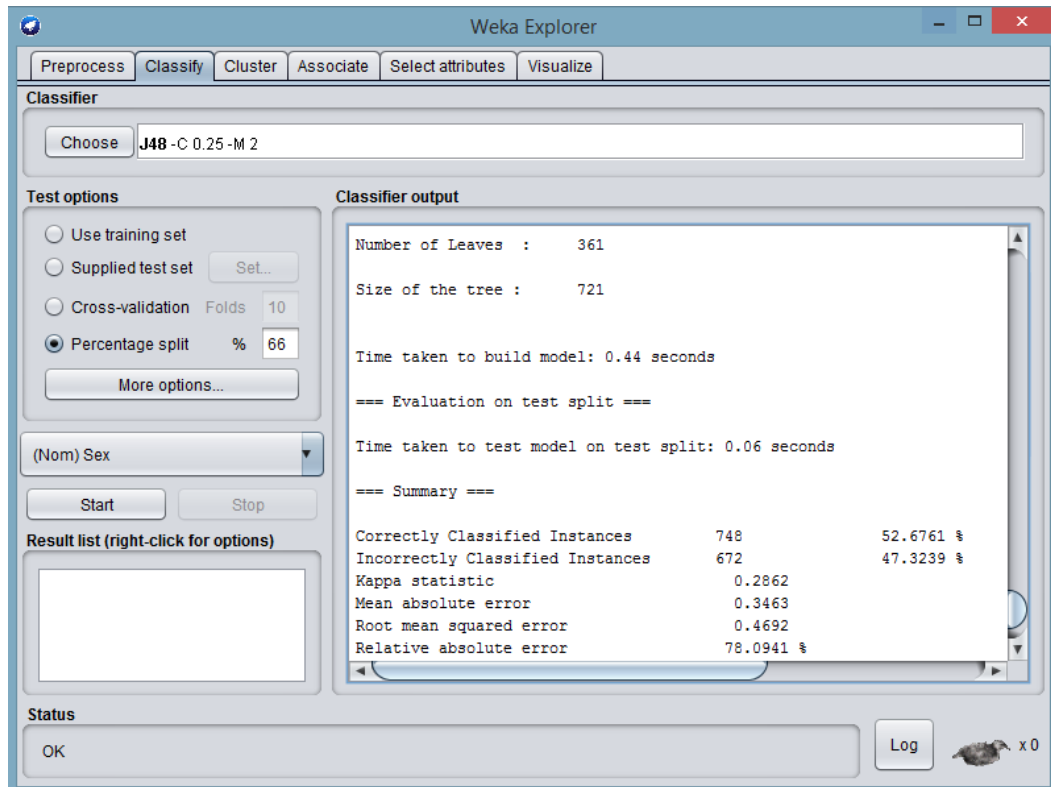
2. Use decision tree (C4.5/J48) learner in WEKA to predict the number of rings Abalone will have

Jawab:

Pada tab *Classify* pilih *Classifier* kemudian klik Choose dan pilih trees → J48



Kemudian klik Start, maka akan ada informasi pada *Classifier Output*, meliputi nilai keakurasian dan confusion matrix (nilai benar/salahnya prediksi)

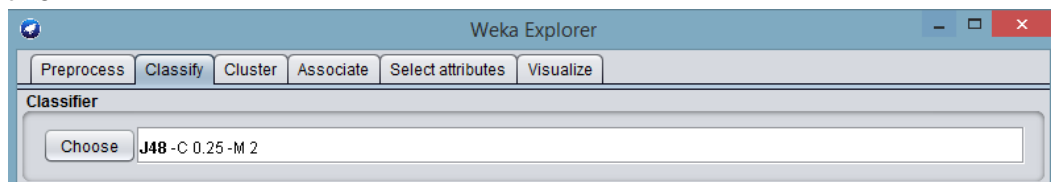


3. Write the description of your process and methods. What parameters (testing/training size, classification target, subtree raising etc.) did you use? Did you preprocess the dataset? Why did you select those parameters/preprocessing?

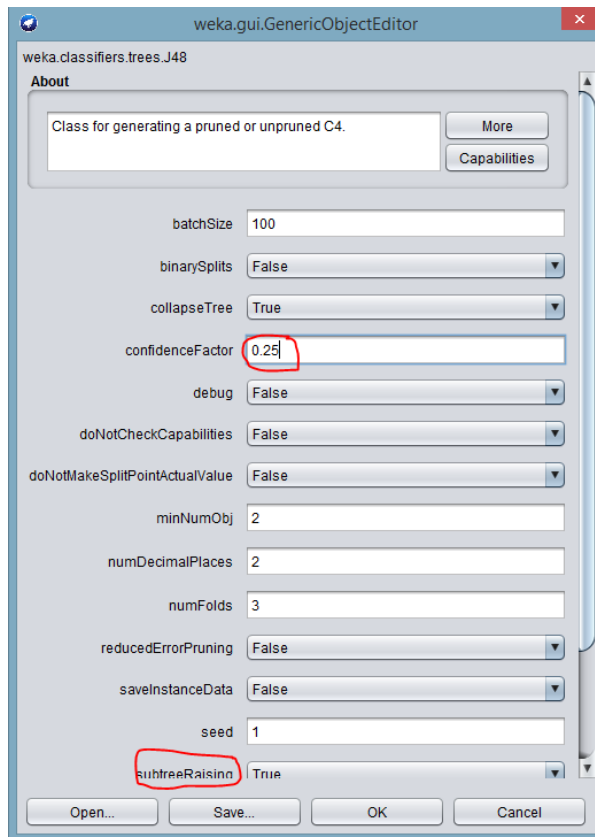
Jawab :

Proses dan metode serta attribute sudah dijelaskan di atas.

Iya, menggunakan decision tree yang ada di WEKA, yaitu dengan memilih *classifier* J48



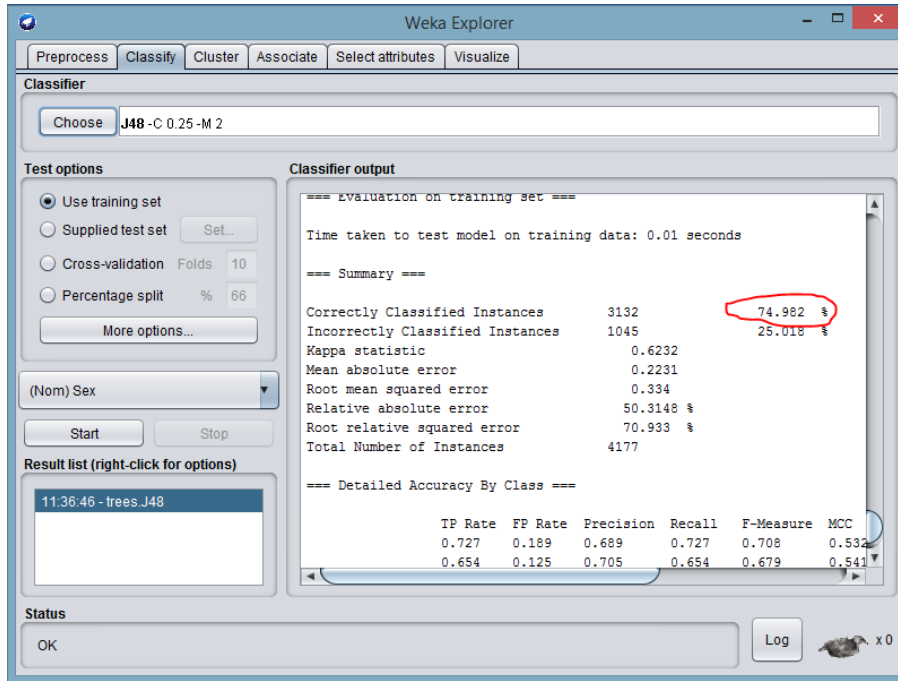
Disini saya menggunakan parameter default dari WEKA, seperti pada gambar dibawah ini



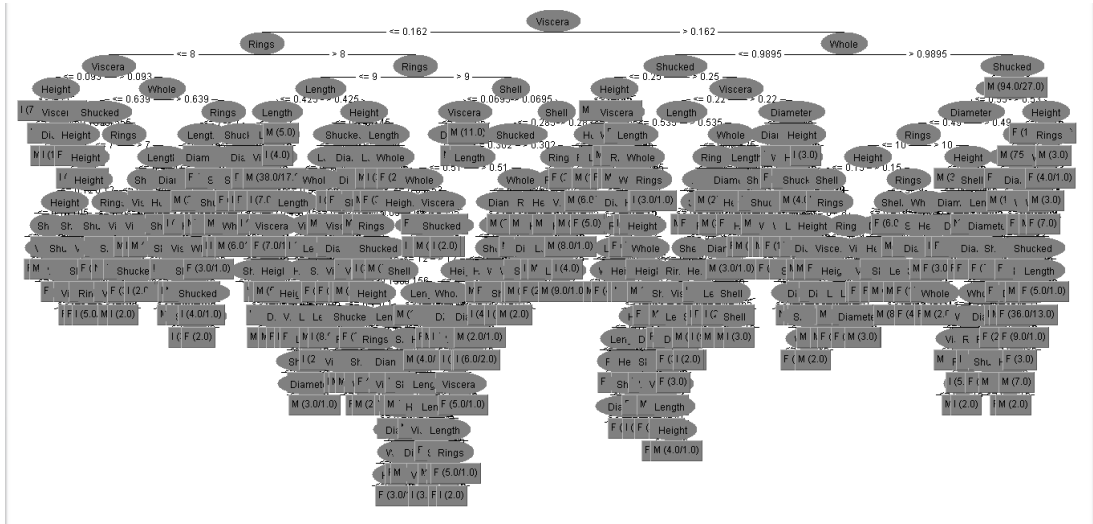
Jika kita mengubah confidenceFactor, aka nada perubahan nilai pada akurasinya, seperti terlihat dari gambar berikut ini

0.25			0.5		
Correctly Classified Instances	3132	74.982 %	Correctly Classified Instances	3206	76.7537 %
Incorrectly Classified Instances	1045	25.018 %	Incorrectly Classified Instances	971	23.2463 %
Kappa statistic	0.6232		Kappa statistic	0.6497	
Mean absolute error	0.2231		Mean absolute error	0.2019	
Root mean squared error	0.334		Root mean squared error	0.3177	
Relative absolute error	50.3148 %		Relative absolute error	45.5279 %	
Root relative squared error	70.933 %		Root relative squared error	67.4744 %	
Total Number of Instances	4177		Total Number of Instances	4177	

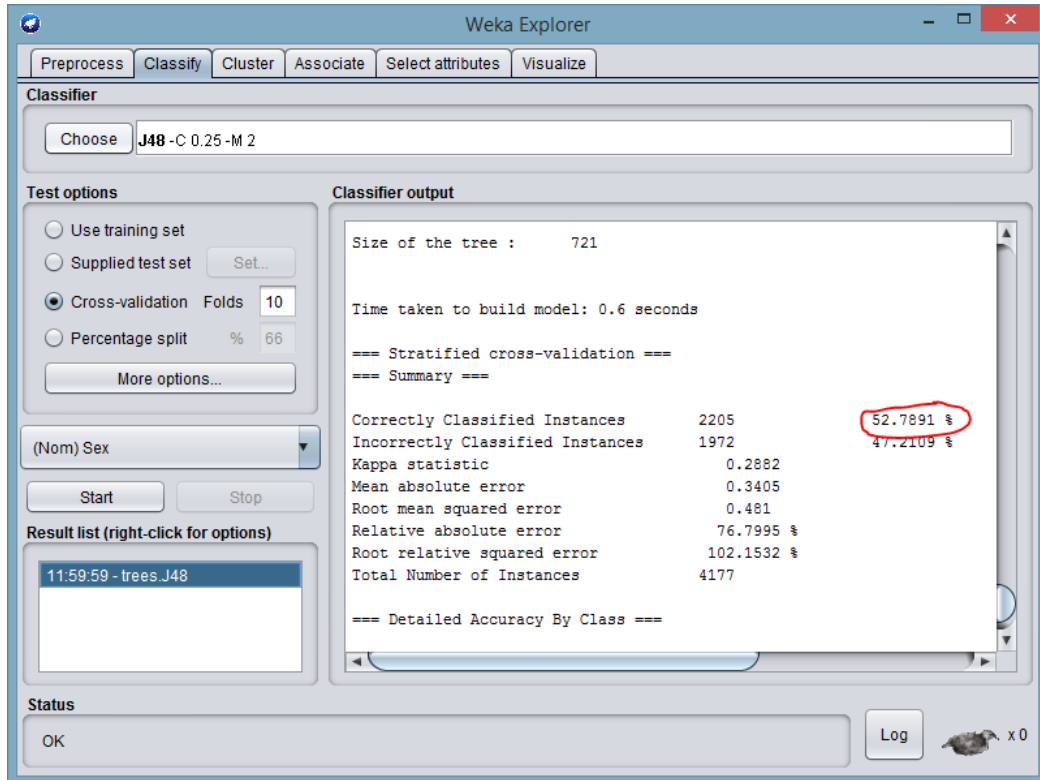
4. What were your results? Show what decision trees you found.
 - a. Menggunakan Use Training Set, didapatkan akuransi 74,982%



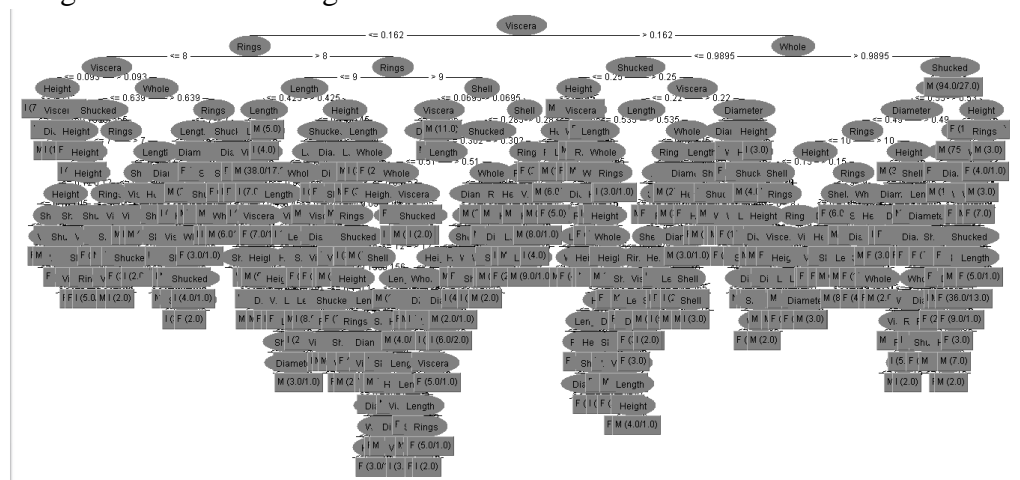
Dengan bentuk treenya sebagai berikut



- b. Menggunakan Cross-validation
Didapatkan akurasi yaitu

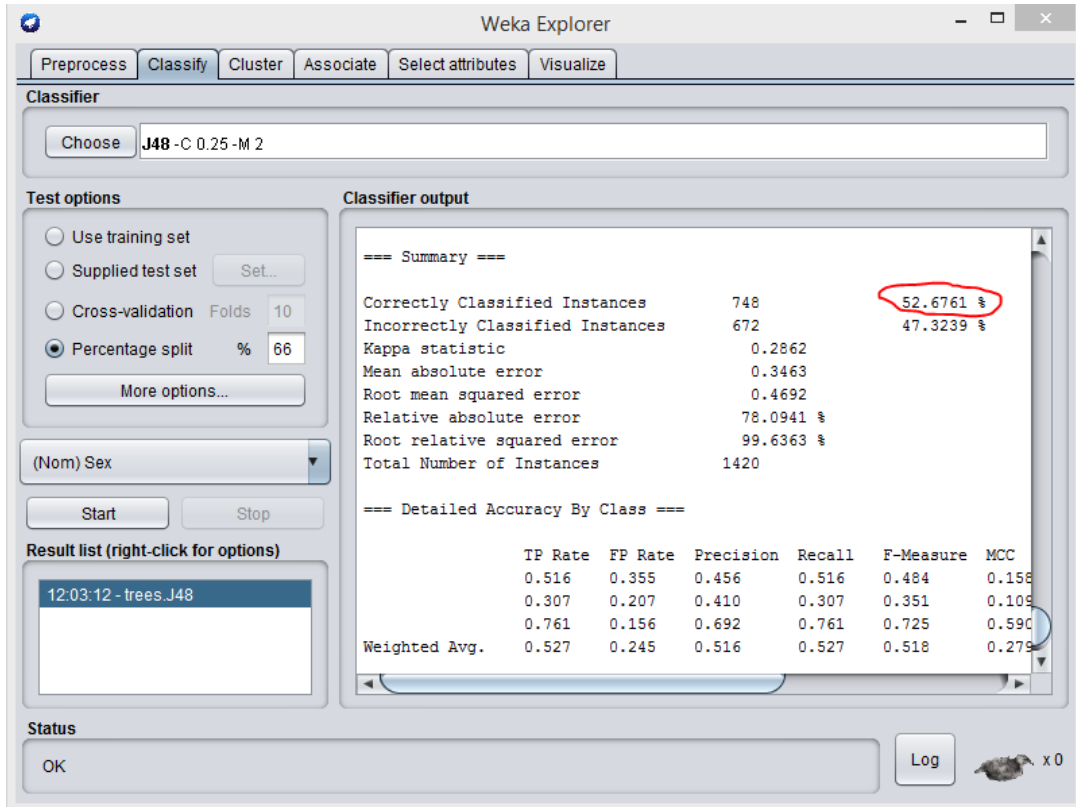


Dengan bentuk tree sebagai berikut

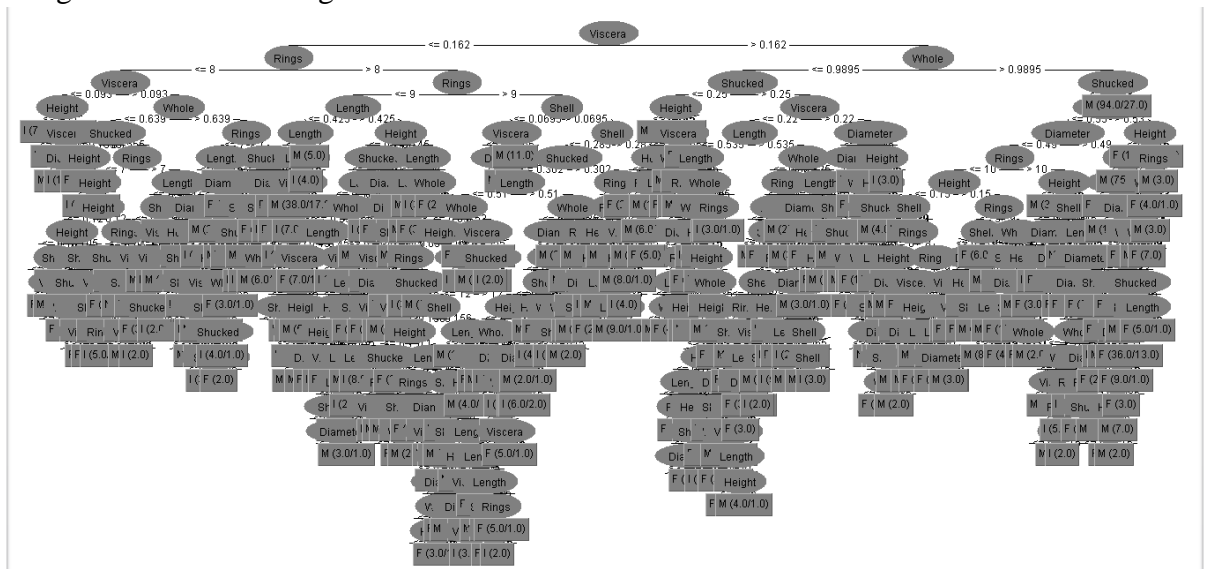


c. Percentage Split

Didapatkan akuransi yaitu 52,6761%



Dengan bentuk tree sebagai berikut :



5. What do the results tell us? Why are the results (in)accurate? Why did changing parameter(s) improve/degrade accuracy?

Jawab:

Berdasarkan hasil dari masing-masing parameter memiliki nilai akurasi yang berbeda-beda. Akurasi terbesar yaitu ketika menggunakan Use Training Set, sehingga Use Training Set memiliki keakurasian yang tinggi, tetapi tidak memberikan estimasi

NAFINGATUN NGALIAH_5115100032
KK F

akurasi yang sebenarnya terhadap data yang lain (yang dipakai untuk training). Tiap parameter berbeda akurasinya karena kemungkinan ada misclassified.