

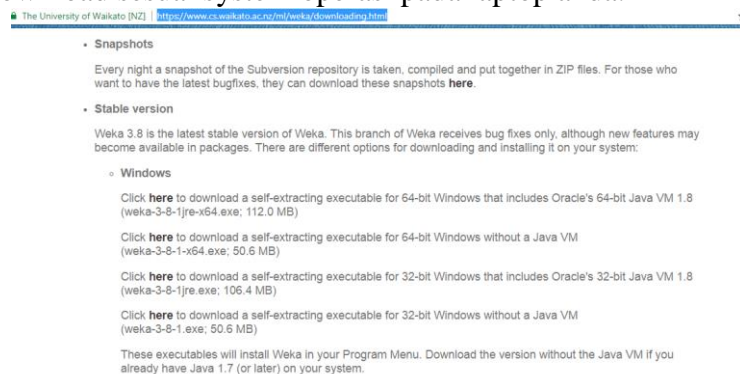
## Question

1. Get the "Abalone" dataset from UCI ML Repo
2. Use a decision tree (C4.5/J48) learner in WEKA to predict the number of rings Abalone will have
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?
4. What were your results? Show what decision trees you found.

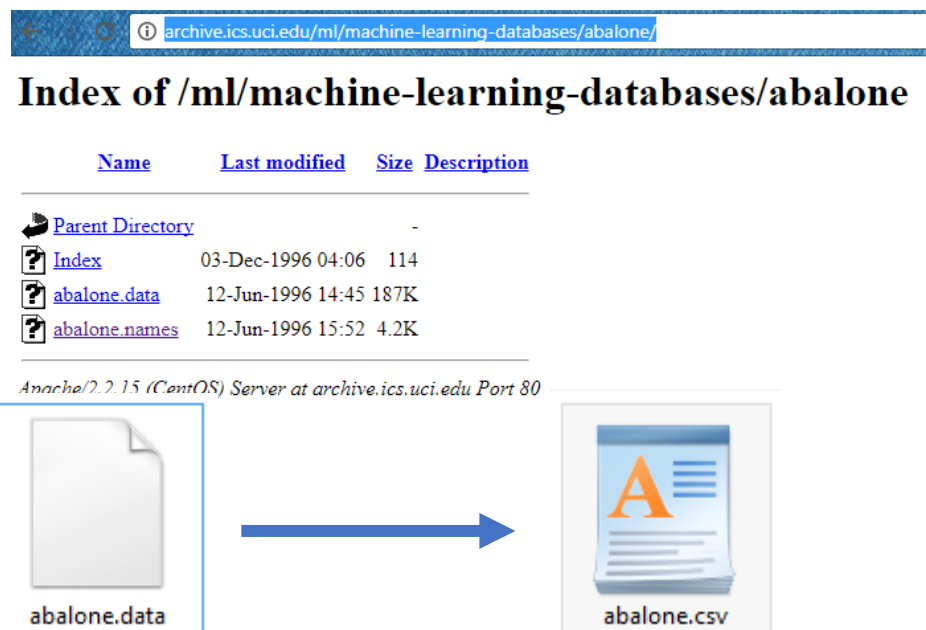
## Answer

### A. TUTORIAL PADA ABALONE DATA

1. Download WEKA dari alamat: <https://www.cs.waikato.ac.nz/ml/weka/downloading.html>  
Download sesuai system operasi pada laptop anda.



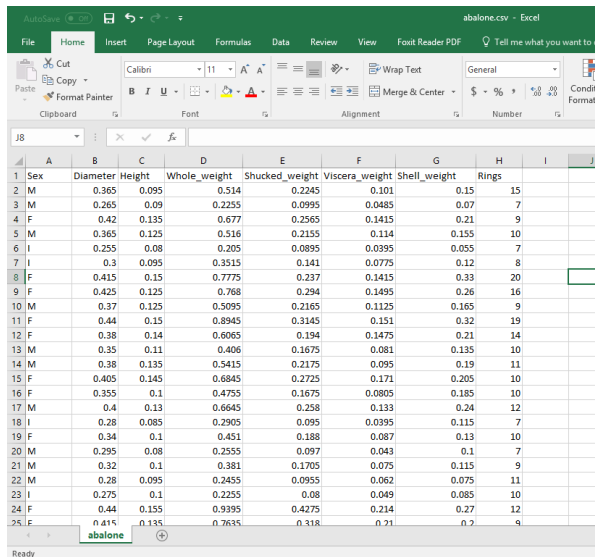
2. Download data Abalon di UCI Machine Learning Repository(abalone.data) dari alamat: <http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/>
3. Setelah terdownload, rename abalone.data menjadi abalone.csv
- 4.



5. Langkah selanjutnya adalah buka file abalone.scv dengan Microsoft excel

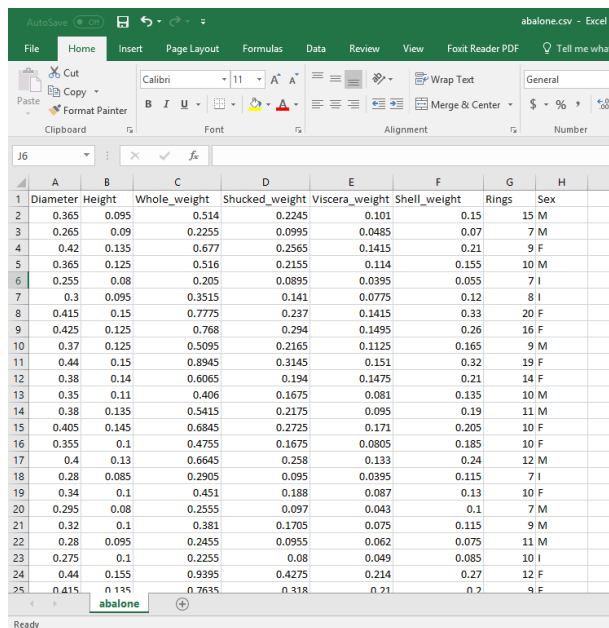
Fatimatus Zulfa  
05111540000073  
KK F

6. Berilah nama tiap kolom attribute dan class nya sesuai dengan yang ada pada abalone.name pada database UCI.



	A	B	C	D	E	F	G	H	I
1	Sex	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings	
2	M	0.365	0.095	0.514	0.2245	0.101	0.15	15	
3	M	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	
4	F	0.42	0.135	0.677	0.2565	0.1415	0.21	9	
5	M	0.365	0.125	0.516	0.2155	0.114	0.155	10	
6	I	0.255	0.08	0.205	0.0895	0.0395	0.055	7	
7	I	0.3	0.095	0.3515	0.141	0.0775	0.12	8	
8	F	0.415	0.15	0.7775	0.237	0.1415	0.33	20	
9	F	0.425	0.125	0.768	0.294	0.1495	0.26	16	
10	M	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	
11	F	0.44	0.15	0.8945	0.3145	0.151	0.32	19	
12	F	0.38	0.14	0.6065	0.194	0.1475	0.21	14	
13	M	0.35	0.11	0.406	0.1675	0.081	0.135	10	
14	M	0.38	0.135	0.5415	0.2175	0.095	0.19	11	
15	F	0.405	0.145	0.6845	0.2725	0.171	0.205	10	
16	F	0.355	0.1	0.4755	0.1675	0.0805	0.185	10	
17	M	0.4	0.13	0.6645	0.258	0.133	0.24	12	
18	I	0.28	0.085	0.2905	0.095	0.0395	0.115	7	
19	F	0.34	0.1	0.451	0.188	0.087	0.13	10	
20	M	0.295	0.08	0.2555	0.097	0.043	0.1	7	
21	M	0.32	0.1	0.381	0.1705	0.075	0.115	9	
22	M	0.28	0.095	0.2455	0.0955	0.062	0.075	11	
23	I	0.275	0.1	0.2255	0.08	0.049	0.085	10	
24	F	0.44	0.155	0.9395	0.4275	0.214	0.27	12	
25	F	0.415	0.135	0.7635	0.318	0.21	0.2	9	

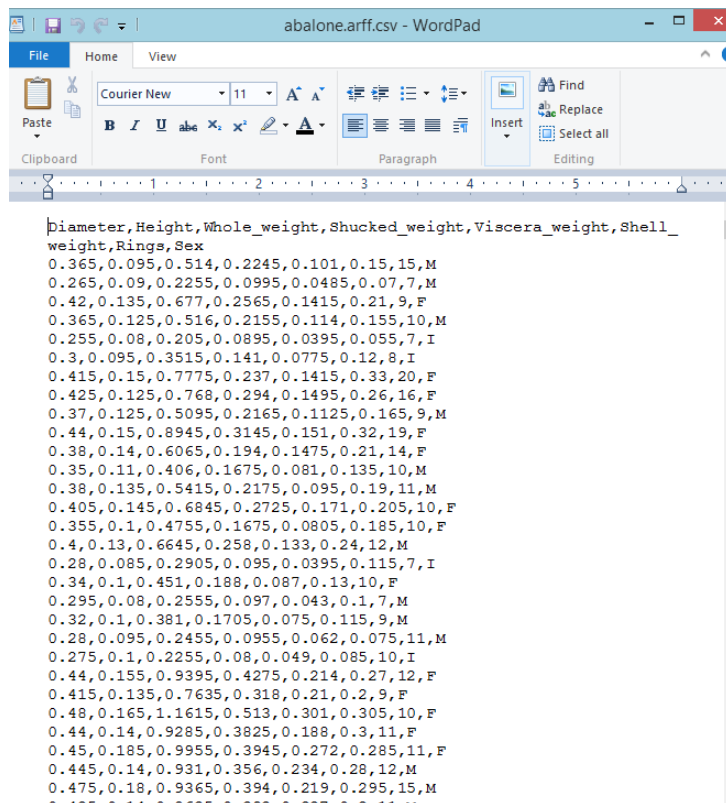
7. Pindahlah kolom Sex dari paling kiri ke kolom paling kanan(setelah Rings)



	A	B	C	D	E	F	G	H
1	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings	Sex
2	0.365	0.095	0.514	0.2245	0.101	0.15	15	M
3	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	M
4	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
5	0.365	0.125	0.516	0.2155	0.114	0.155	10	M
6	0.255	0.08	0.205	0.0895	0.0395	0.055	7	I
7	0.3	0.095	0.3515	0.141	0.0775	0.12	8	I
8	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
9	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
10	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	M
11	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F
12	0.38	0.14	0.6065	0.194	0.1475	0.21	14	F
13	0.35	0.11	0.406	0.1675	0.081	0.135	10	M
14	0.38	0.135	0.5415	0.2175	0.095	0.19	11	M
15	0.405	0.145	0.6845	0.2725	0.171	0.205	10	F
16	0.355	0.1	0.4755	0.1675	0.0805	0.185	10	F
17	0.4	0.13	0.6645	0.258	0.133	0.24	12	M
18	0.28	0.085	0.2905	0.095	0.0395	0.115	7	I
19	0.34	0.1	0.451	0.188	0.087	0.13	10	F
20	0.295	0.08	0.2555	0.097	0.043	0.1	7	M
21	0.32	0.1	0.381	0.1705	0.075	0.115	9	M
22	0.28	0.095	0.2455	0.0955	0.062	0.075	11	M
23	0.275	0.1	0.2255	0.08	0.049	0.085	10	I
24	0.44	0.155	0.9395	0.4275	0.214	0.27	12	F
25	0.415	0.135	0.7635	0.318	0.21	0.2	9	F

Fatimatus Zulfa  
05111540000073  
KK F

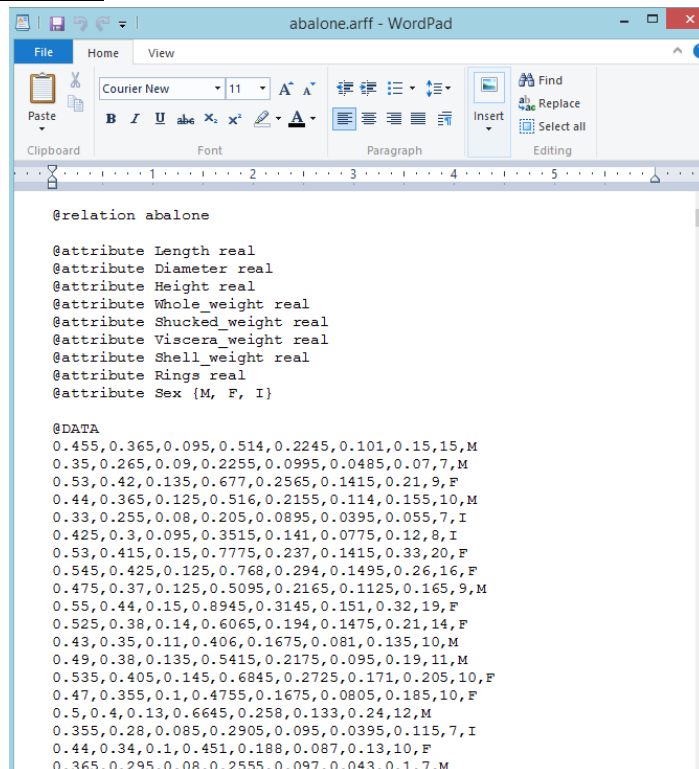
8. Langkah selanjutnya adalah save as data dengan nama abalone.arff dan selanjutnya buka file dengan Wordpad dan tambahkan @attribute dan real sebagai relation sebagai berikut:



The screenshot shows a WordPad window titled 'abalone.arff.csv - WordPad'. The text inside is a CSV file with the following structure:

```
Diameter,Height,Whole_weight,Shucked_weight,Viscera_weight,Shell_weight,Rings,Sex
0.365,0.095,0.514,0.2245,0.101,0.15,15,M
0.265,0.09,0.2255,0.0995,0.0485,0.07,7,M
0.42,0.135,0.677,0.2565,0.1415,0.21,9,F
0.365,0.125,0.516,0.2155,0.114,0.155,10,M
0.255,0.08,0.205,0.0895,0.0395,0.055,7,I
0.3,0.095,0.3515,0.141,0.0775,0.12,8,I
0.415,0.15,0.7775,0.237,0.1415,0.33,20,F
0.425,0.125,0.768,0.294,0.1495,0.26,16,F
0.37,0.125,0.5095,0.2165,0.1125,0.165,9,M
0.44,0.15,0.8945,0.3145,0.151,0.32,19,F
0.38,0.14,0.6065,0.194,0.1475,0.21,14,F
0.35,0.11,0.406,0.1675,0.081,0.135,10,M
0.38,0.135,0.5415,0.2175,0.095,0.19,11,M
0.405,0.145,0.6845,0.2725,0.171,0.205,10,F
0.355,0.1,0.4755,0.1675,0.0805,0.185,10,F
0.4,0.13,0.6645,0.258,0.133,0.24,12,M
0.28,0.085,0.2905,0.095,0.0395,0.115,7,I
0.34,0.1,0.451,0.188,0.087,0.13,10,F
0.295,0.08,0.2555,0.097,0.043,0.1,7,M
0.32,0.1,0.381,0.1705,0.075,0.115,9,M
0.28,0.095,0.2455,0.0955,0.062,0.075,11,M
0.275,0.1,0.2255,0.08,0.049,0.085,10,I
0.44,0.155,0.9395,0.4275,0.214,0.27,12,F
0.415,0.135,0.7635,0.318,0.21,0.2,9,F
0.48,0.165,1.1615,0.513,0.301,0.305,10,F
0.44,0.14,0.9285,0.3825,0.188,0.3,11,F
0.45,0.185,0.9955,0.3945,0.272,0.285,11,F
0.445,0.14,0.931,0.356,0.234,0.28,12,M
0.475,0.18,0.9365,0.394,0.219,0.295,15,M
0.425,0.14,0.8635,0.382,0.227,0.2,11,M
```

### MENJADI



The screenshot shows a WordPad window titled 'abalone.arff - WordPad'. The text inside is the ARFF file format for the abalone dataset:

```
@relation abalone

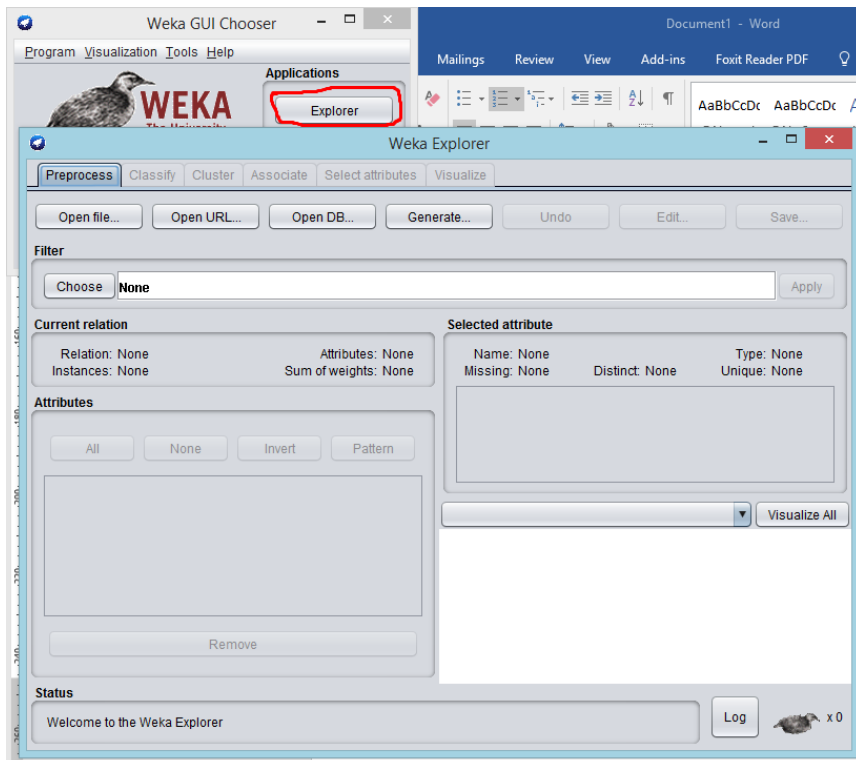
@attribute Length real
@attribute Diameter real
@attribute Height real
@attribute Whole_weight real
@attribute Shucked_weight real
@attribute Viscera_weight real
@attribute Shell_weight real
@attribute Rings real
@attribute Sex {M, F, I}

@DATA
0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15,M
0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7,M
0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9,F
0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10,M
0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7,I
0.425,0.3,0.095,0.3515,0.141,0.0775,0.12,8,I
0.53,0.415,0.15,0.7775,0.237,0.1415,0.33,20,F
0.545,0.425,0.125,0.768,0.294,0.1495,0.26,16,F
0.475,0.37,0.125,0.5095,0.2165,0.1125,0.165,9,M
0.55,0.44,0.15,0.8945,0.3145,0.151,0.32,19,F
0.525,0.38,0.14,0.6065,0.194,0.1475,0.21,14,F
0.43,0.35,0.11,0.406,0.1675,0.081,0.135,10,M
0.49,0.38,0.135,0.5415,0.2175,0.095,0.19,11,M
0.535,0.405,0.145,0.6845,0.2725,0.171,0.205,10,F
0.47,0.355,0.1,0.4755,0.1675,0.0805,0.185,10,F
0.5,0.4,0.13,0.6645,0.258,0.133,0.24,12,M
0.355,0.28,0.085,0.2905,0.095,0.0395,0.115,7,I
0.44,0.34,0.1,0.451,0.188,0.087,0.13,10,F
0.365,0.295,0.08,0.2555,0.097,0.043,0.1,7,M
```

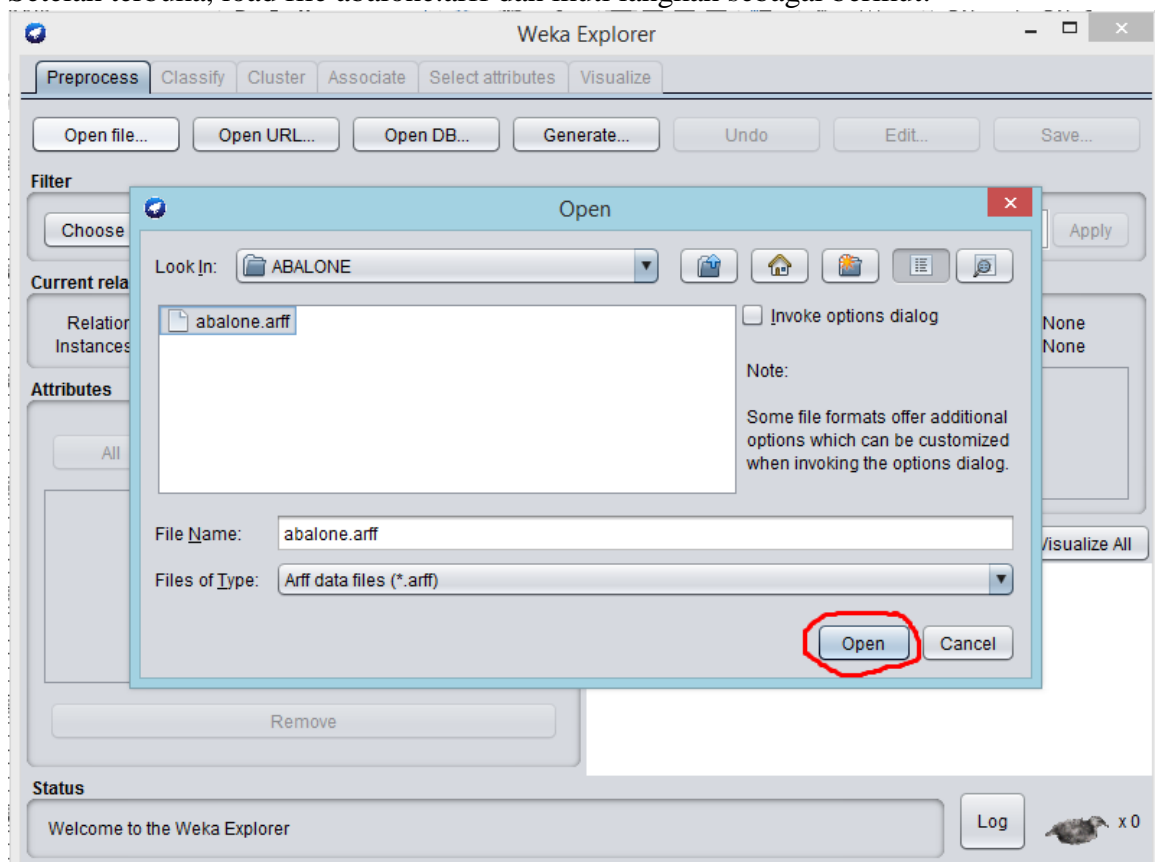
Dan simpan file abalone.arff

Fatimatus Zulfa  
05111540000073  
KK F

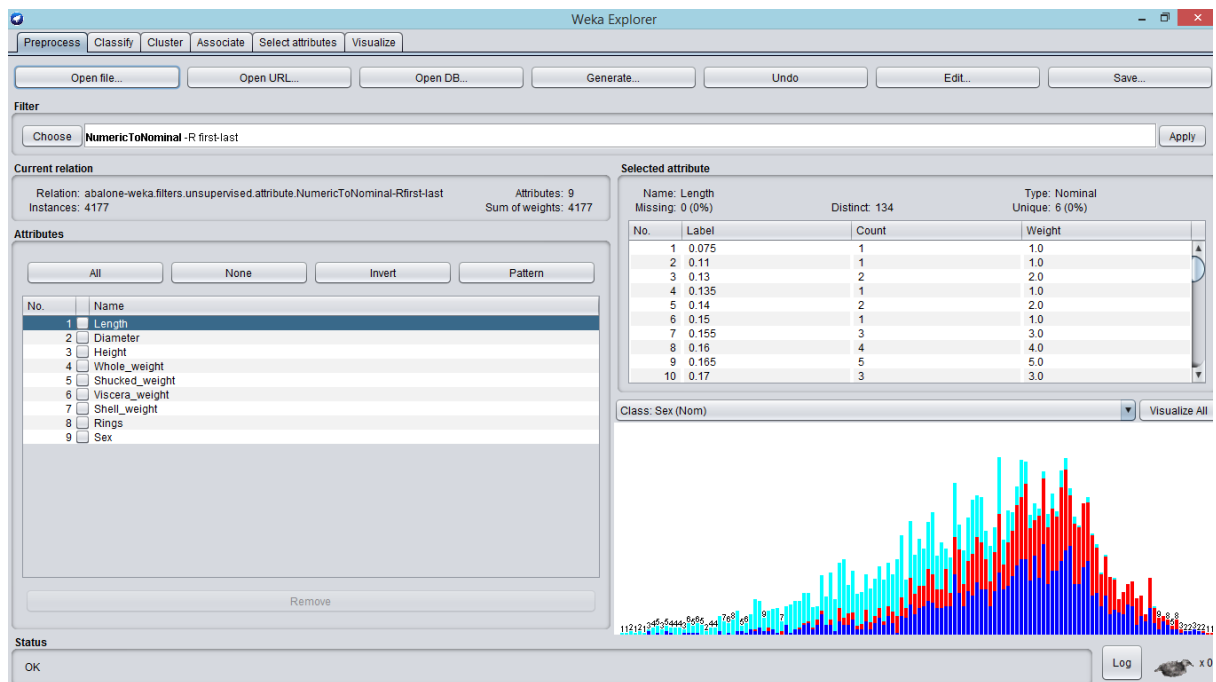
9. Langkah selanjutnya buka WEKA



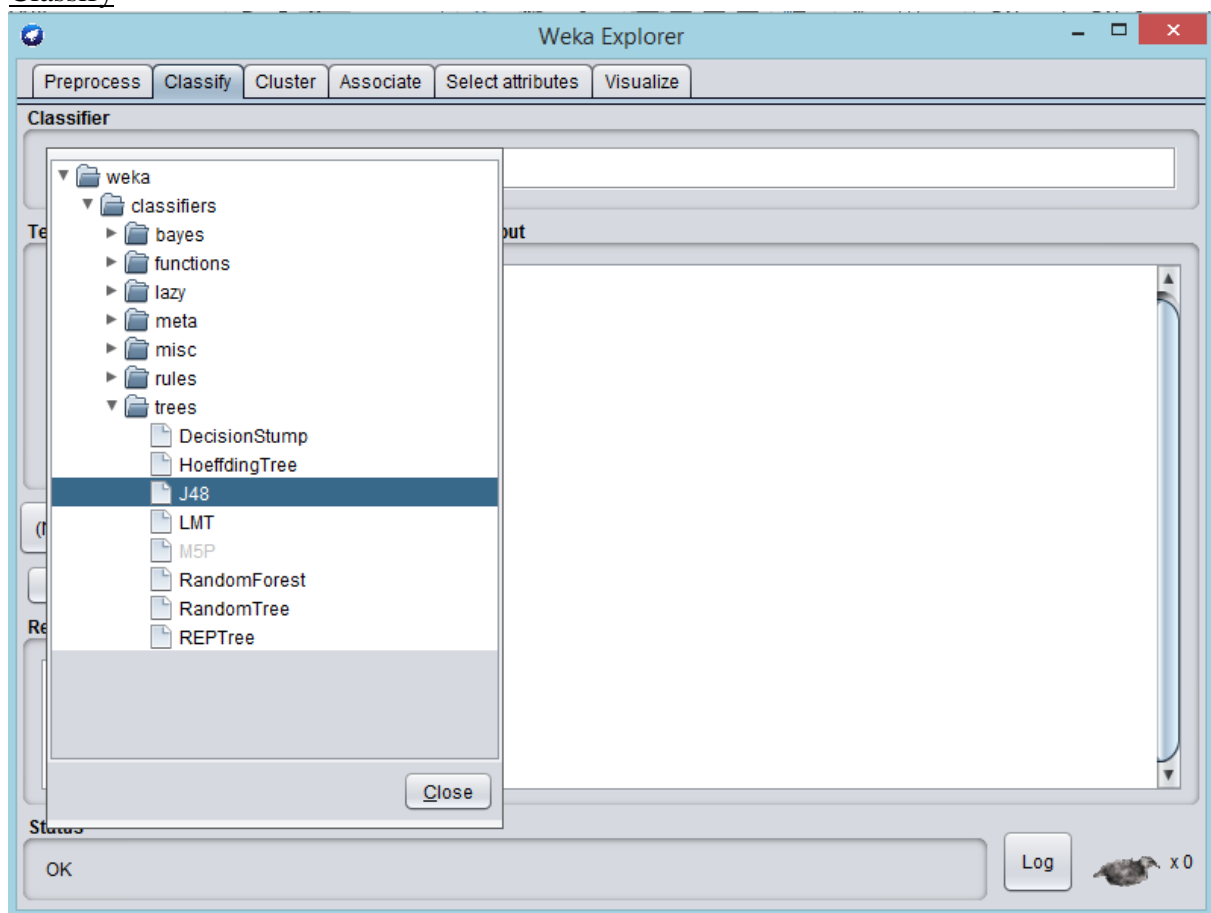
10. Setelah terbuka, load file abalone.arff dan ikuti langkah sebagai berikut:



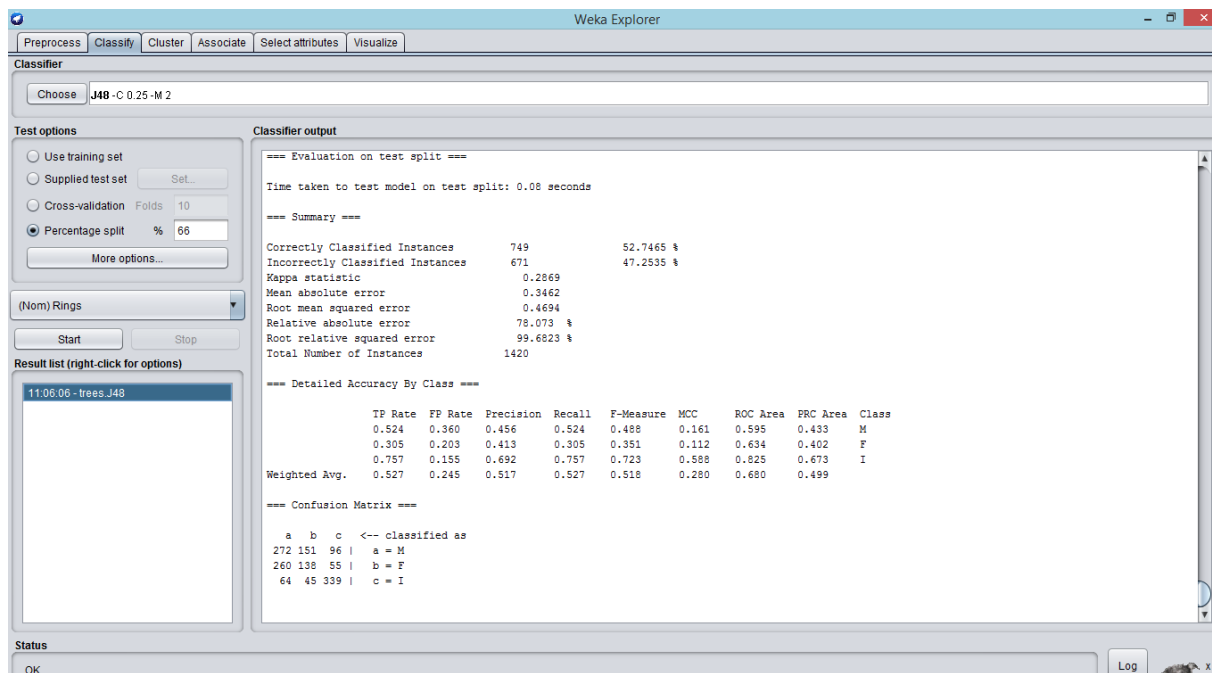
Fatimatus Zulfa  
05111540000073  
KK F



## Classify



Fatimatus Zulfa  
05111540000073  
KK F



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25-M 2

Test options

☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 66

More options...

(Nom) Rings

Start Stop

Result list (right-click for options)

11:06:06 - trees\_j48

Classifier output

=== Evaluation on test split ===

Time taken to test model on test split: 0.08 seconds

=== Summary ===

Correctly Classified Instances	749	52.7465 %
Incorrectly Classified Instances	671	47.2535 %
Kappa statistic	0.2869	
Mean absolute error	0.3462	
Root mean squared error	0.4694	
Relative absolute error	78.073 %	
Root relative squared error	99.6823 %	
Total Number of Instances	1420	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.524	0.360	0.456	0.524	0.488	0.161	0.595	0.433	M
	0.305	0.203	0.413	0.305	0.351	0.112	0.634	0.402	F
	0.757	0.155	0.692	0.757	0.723	0.588	0.825	0.673	I
Weighted Avg.	0.527	0.245	0.517	0.527	0.518	0.280	0.680	0.499	

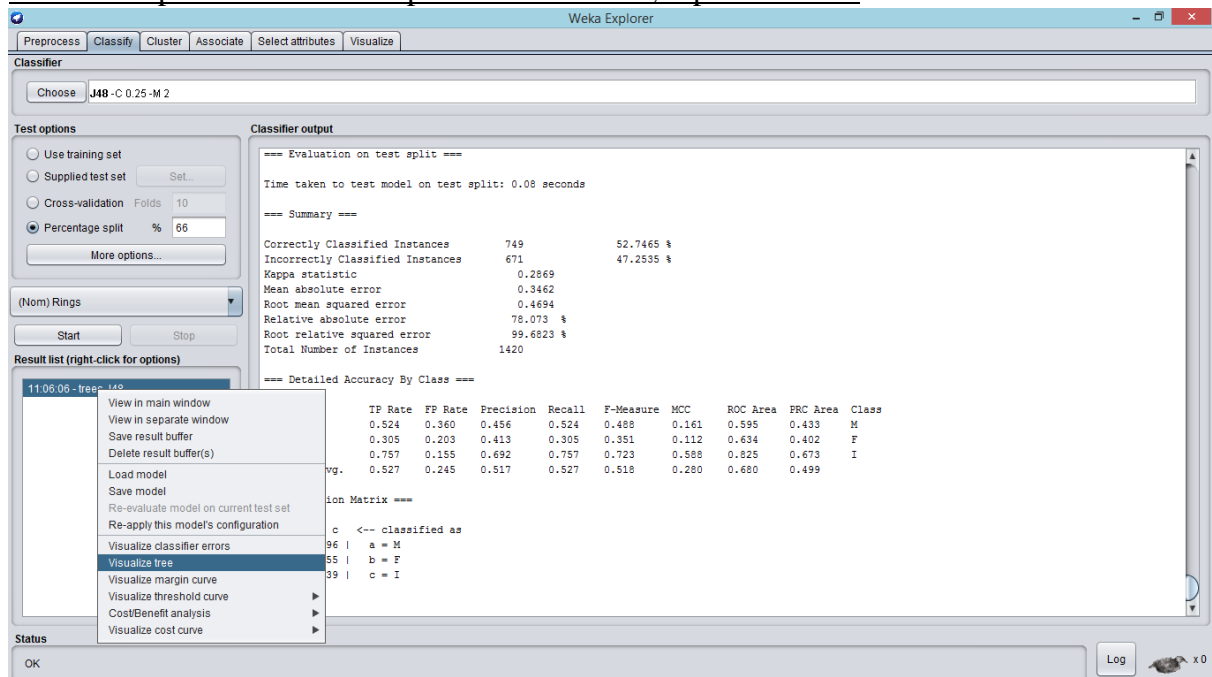
=== Confusion Matrix ===

	a	b	c	<-- classified as
272	151	96	I	a = M
260	138	55	I	b = F
64	45	339	I	c = I

Status

OK Log

Klik kanan pada result list dan pilih Visualize tree, seperti berikut:



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25-M 2

Test options

☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 66

More options...

(Nom) Rings

Start Stop

Result list (right-click for options)

11:06:06 - trees\_j48

Classifier output

=== Evaluation on test split ===

Time taken to test model on test split: 0.08 seconds

=== Summary ===

Correctly Classified Instances	749	52.7465 %
Incorrectly Classified Instances	671	47.2535 %
Kappa statistic	0.2869	
Mean absolute error	0.3462	
Root mean squared error	0.4694	
Relative absolute error	78.073 %	
Root relative squared error	99.6823 %	
Total Number of Instances	1420	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.524	0.360	0.456	0.524	0.488	0.161	0.595	0.433	M
	0.305	0.203	0.413	0.305	0.351	0.112	0.634	0.402	F
	0.757	0.155	0.692	0.757	0.723	0.588	0.825	0.673	I
Weighted Avg.	0.527	0.245	0.517	0.527	0.518	0.280	0.680	0.499	

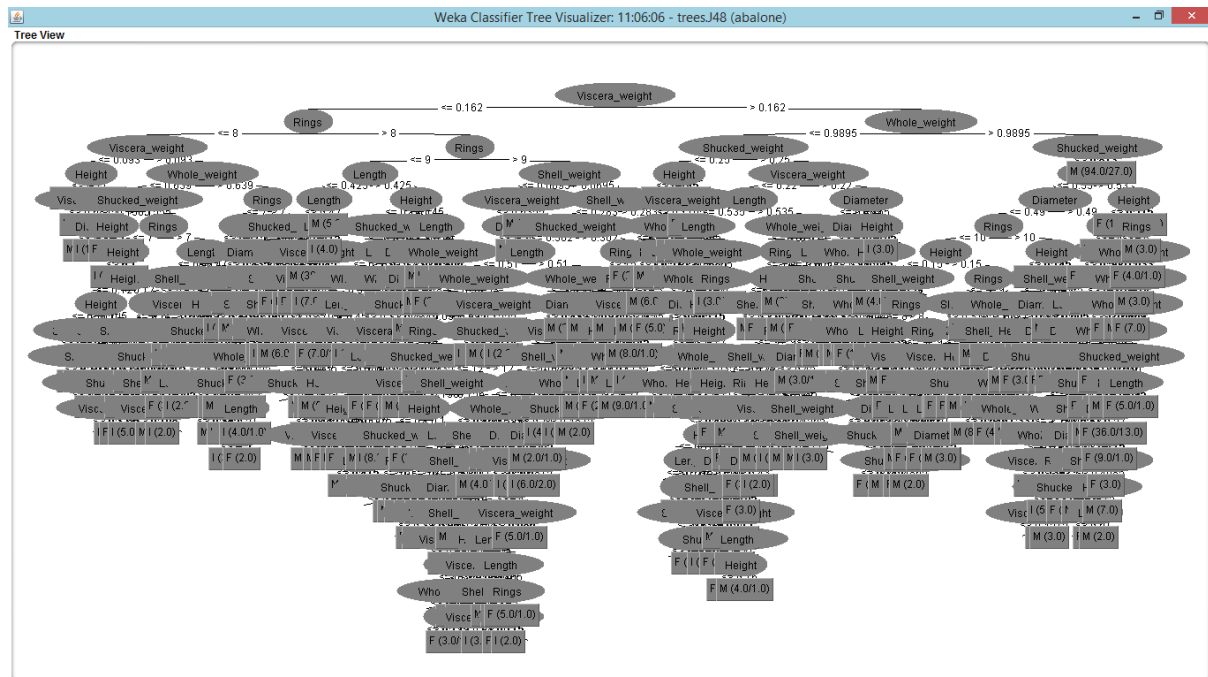
=== Confusion Matrix ===

	a	b	c	<-- classified as
272	151	96	I	a = M
260	138	55	I	b = F
64	45	339	I	c = I

Status

OK Log

Fatimatus Zulfa  
05111540000073  
KK F



## Parameter

weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4. More Capabilities

batchSize 100

binarySplits False

collapseTree True

confidenceFactor 0.25

debug False

doNotCheckCapabilities False

doNotMakeSplitPointActualValue False

minNumObj 2

numDecimalPlaces 2

numFolds 3

reducedErrorPruning False

saveInstanceData False

seed 1

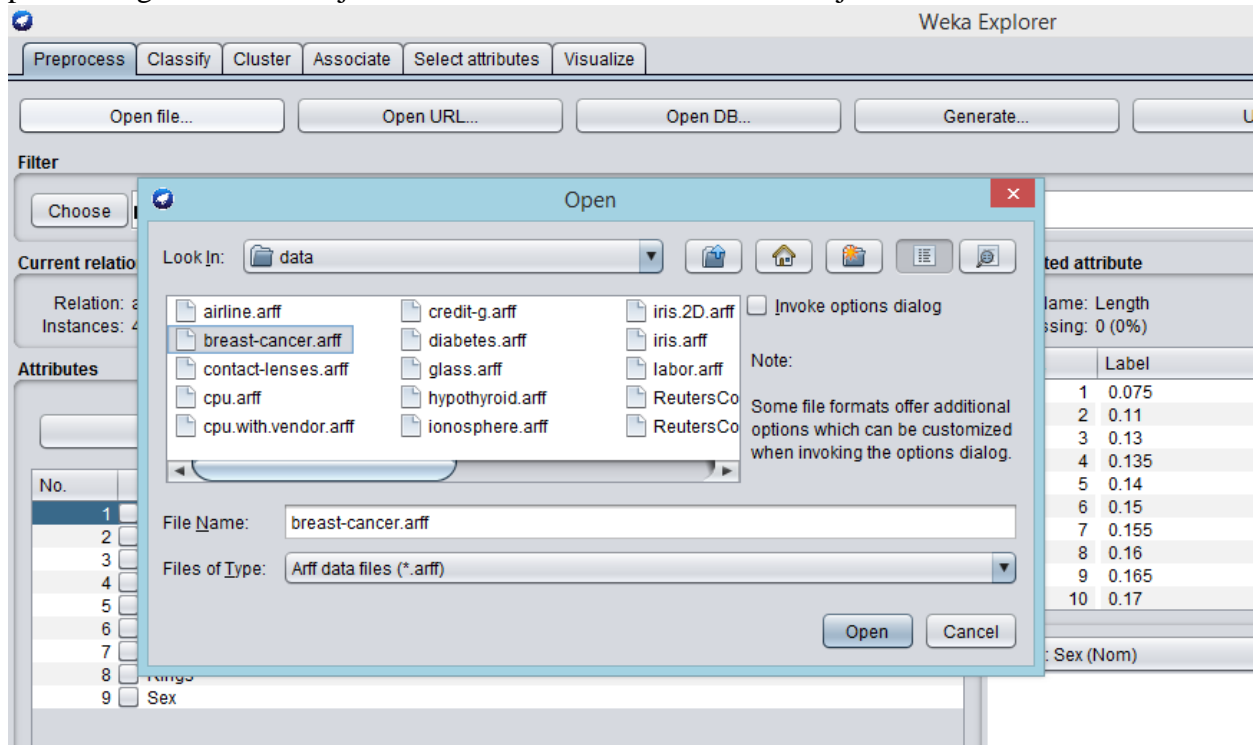
subtreeRaisino True

Open... Save... OK Cancel

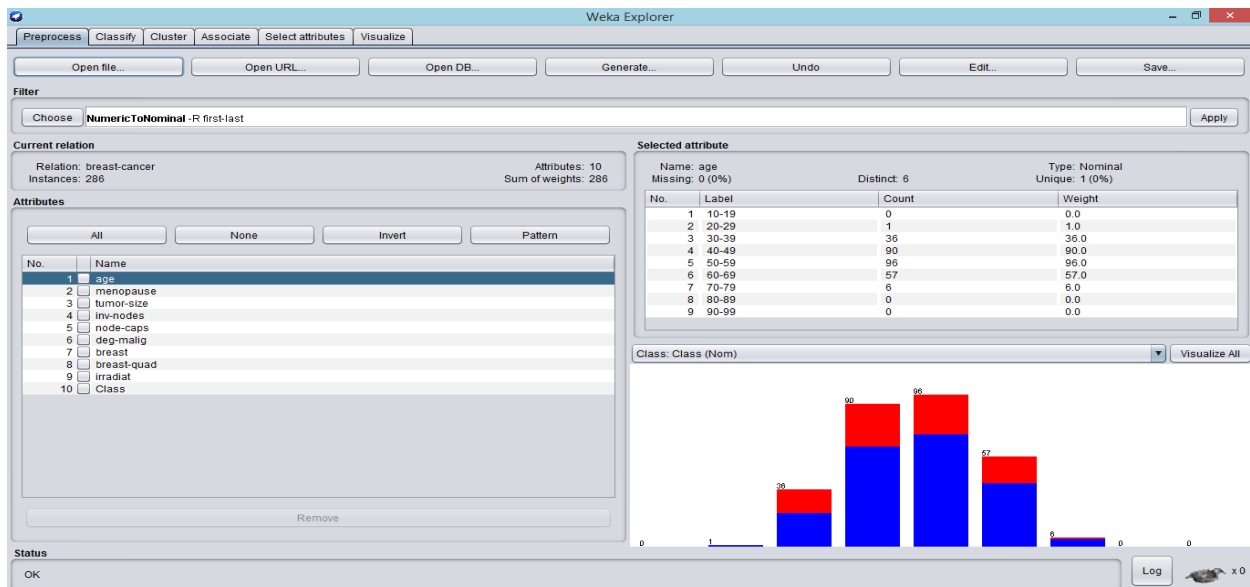
## B. TUTORIAL DATA BREAST CANCER

Fatimatus Zulfa  
05111540000073  
KK F

1. Untuk data breast cancer, datanya sudah tersedia pada weka jadi langkah-langkahnya tidak perlu mengubah data menjadi .arff karena dalam weka udah menjadi breast-cancer.arff



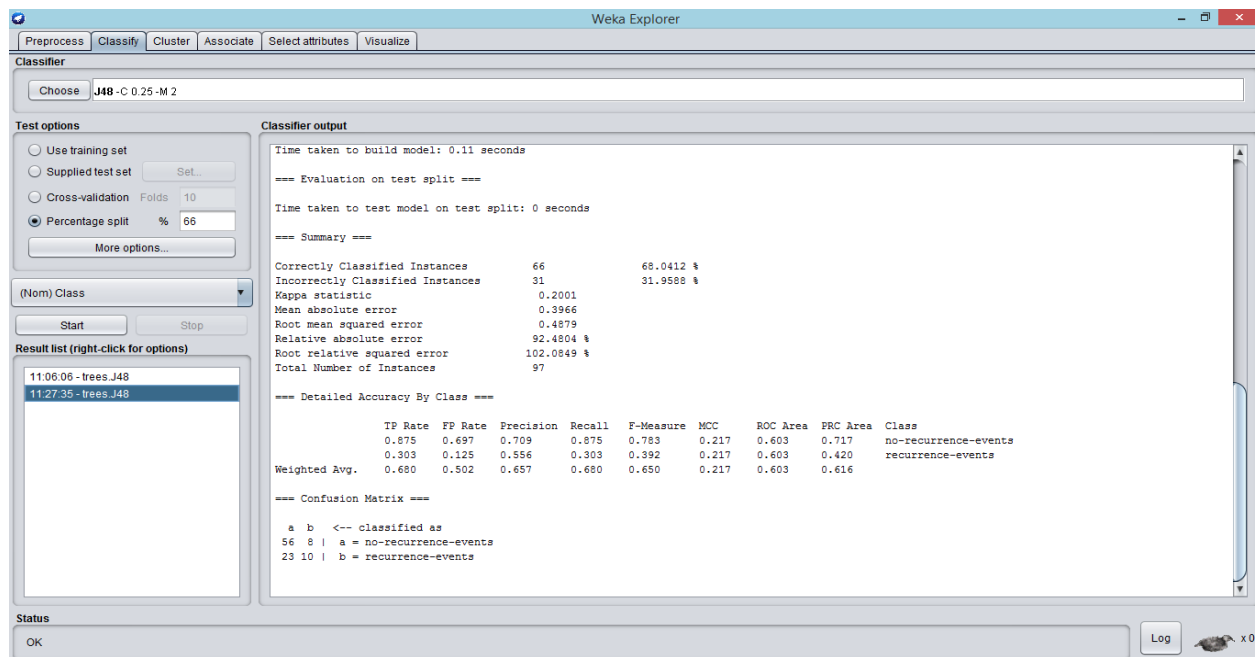
2. Setelah open file akan muncul gambar seperti di bawah ini:



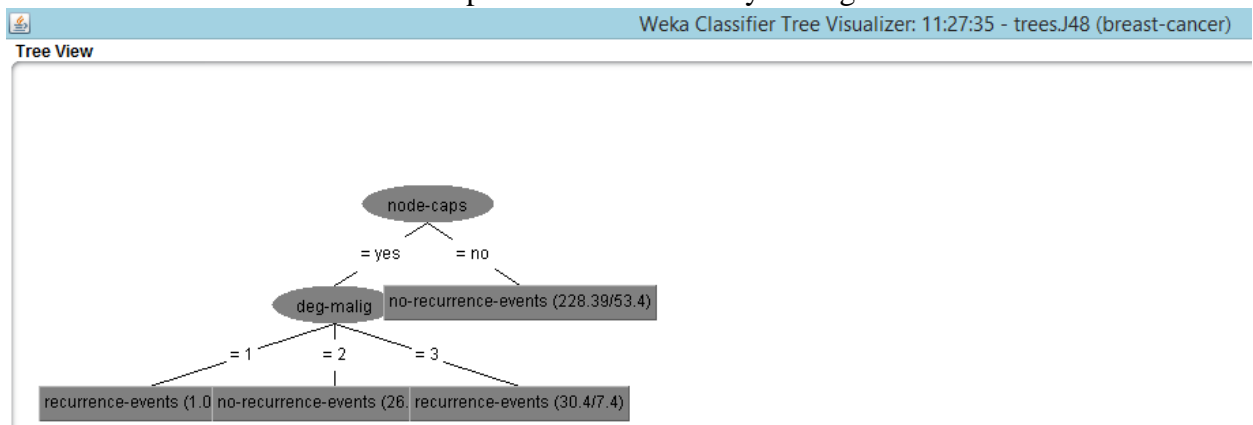
3. Langkah selanjutnya klik Classify dan ikuti langkah berikut:



Fatimatus Zulfa  
05111540000073  
KK F



Klik Visualize Tree dari klik kanan pada result dan hasilnya sebagai berikut:



## Question

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

## Answer

- Use Training Set

Correctly Classified Instances	2829	67.728 %
Incorrectly Classified Instances	1348	32.272 %
Kappa statistic	0.6358	
Mean absolute error	0.0244	
Root mean squared error	0.1105	
Relative absolute error	38.1689 %	
Root relative squared error	61.7967 %	
Total Number of Instances	4177	

- Cross Validation Folds: 10

Correctly Classified Instances	214	15.0704 %
Incorrectly Classified Instances	1206	84.9296 %
Kappa statistic	0.0222	
Mean absolute error	0.0622	
Root mean squared error	0.2108	
Relative absolute error	97.0872 %	
Root relative squared error	117.7184 %	
Total Number of Instances	1420	

- Percentage Splits 66%

Correctly Classified Instances	779	18.6497 %
Incorrectly Classified Instances	3398	81.3503 %
Kappa statistic	0.0658	
Mean absolute error	0.0609	
Root mean squared error	0.2042	
Relative absolute error	95.0928 %	
Root relative squared error	114.1805 %	
Total Number of Instances	4177	

Classifier output									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	0.000	0.000	0.000	0.000	?	?	1
	0.000	0.000	0.000	0.000	0.000	0.000	?	?	2
	0.000	0.000	0.000	0.000	0.000	0.000	0.774	0.003	3
	0.103	0.002	0.500	0.103	0.171	0.221	0.623	0.075	4
	0.115	0.010	0.316	0.115	0.169	0.173	0.651	0.082	5
	0.193	0.034	0.274	0.193	0.227	0.188	0.573	0.120	6
	0.172	0.059	0.224	0.172	0.195	0.128	0.586	0.120	7
	0.142	0.095	0.175	0.142	0.157	0.052	0.504	0.130	8
	0.471	0.576	0.134	0.471	0.208	-0.077	0.465	0.153	9
	0.088	0.069	0.196	0.088	0.122	0.027	0.539	0.170	10
	0.070	0.045	0.176	0.070	0.100	0.039	0.559	0.136	11
	0.023	0.026	0.054	0.023	0.032	-0.005	0.531	0.066	12
	0.014	0.023	0.031	0.014	0.019	-0.014	0.523	0.055	13
	0.000	0.009	0.000	0.000	0.000	-0.018	0.464	0.031	14
	0.000	0.009	0.000	0.000	0.000	-0.016	0.519	0.027	15
	0.000	0.006	0.000	0.000	0.000	-0.010	0.504	0.018	16
	0.000	0.006	0.000	0.000	0.000	-0.009	0.547	0.013	17
	0.000	0.004	0.000	0.000	0.000	-0.006	0.468	0.010	18
	0.000	0.004	0.000	0.000	0.000	-0.005	0.514	0.017	19
	0.000	0.001	0.000	0.000	0.000	-0.001	0.766	0.002	20
	0.000	0.001	0.000	0.000	0.000	-0.002	0.397	0.002	21
	0.000	0.001	0.000	0.000	0.000	-0.001	0.273	0.001	22
	0.000	0.000	0.000	0.000	0.000	0.000	0.438	0.002	23
	0.000	0.000	0.000	0.000	0.000	0.000	0.273	0.001	24
	0.000	0.000	0.000	0.000	0.000	0.000	?	?	25
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.001	26
	0.000	0.000	0.000	0.000	0.000	0.000	0.273	0.001	27
	0.000	0.000	0.000	0.000	0.000	0.000	?	?	29
Weighted Avg.	0.151	0.131	0.159	0.151	0.128	0.035	0.531	0.116	

Dilihat dari hasil akurasi yang telah diujikan terdapat perbedaan di masing-masing parameter. Hal tersebut terjadi disebabkan karena *misclassified*.