**DT for Abalone Using WEKA**

**Weka** adalah aplikasi data mining open source berbasis Java. Aplikasi ini dikembangkan pertama kali oleh Universitas Waikato di Selandia Baru sebelum menjadi bagian dari Pentaho. Weka terdiri dari koleksi algoritma machine learning yang dapat digunakan untuk melakukan generalisasi / formulasi dari sekumpulan data sampling. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.

**Data Set Information:**

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 4177 | **Area:** | Life |
| **Attribute Characteristics:** | Categorical, Integer, Real | **Number of Attributes:** | 8 | **Date Donated** | 1995-12-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 497125 |

**Attribute Information:**

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name Data Type Meas. Description

---- --------- ----- -----------

Sex nominal M, F, and I (infant)

Length continuous mm Longest shell measurement

Diameter continuous mm perpendicular to length

Height continuous mm with meat in shell

Whole weight continuous grams whole abalone

Shucked weight continuous grams weight of meat

Viscera weight continuous grams gut weight (after bleeding)

Shell weight continuous grams after being dried

Rings integer +1.5 gives the age in years

Statistics for numeric domains:

Length Diam Height Whole Shucked Viscera Shell Rings

Min 0.075 0.055 0.000 0.002 0.001 0.001 0.002 1

Max 0.815 0.650 1.130 2.826 1.488 0.760 1.005 29

Mean 0.524 0.408 0.140 0.829 0.359 0.181 0.239 9.934

SD 0.120 0.099 0.042 0.490 0.222 0.110 0.139 3.224

Correl 0.557 0.575 0.557 0.540 0.421 0.504 0.628 1.0

Class Distribution:

Class Examples

----- --------

1 1

2 1

3 15

4 57

5 115

6 259

7 391

8 568

9 689

10 634

11 487

12 267

13 203

14 126

15 103

16 67

17 58

18 42

19 32

20 26

21 14

22 6

23 9

24 2

25 1

26 1

27 2

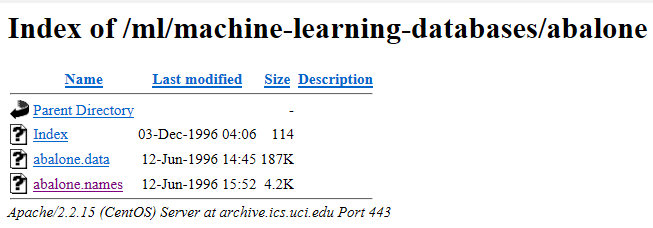
29 1

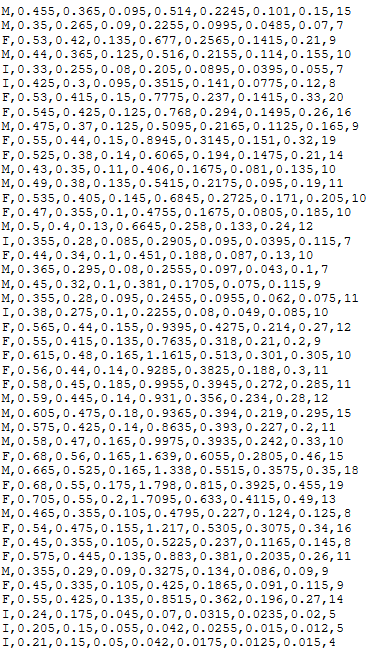
----- ----

Total 4177

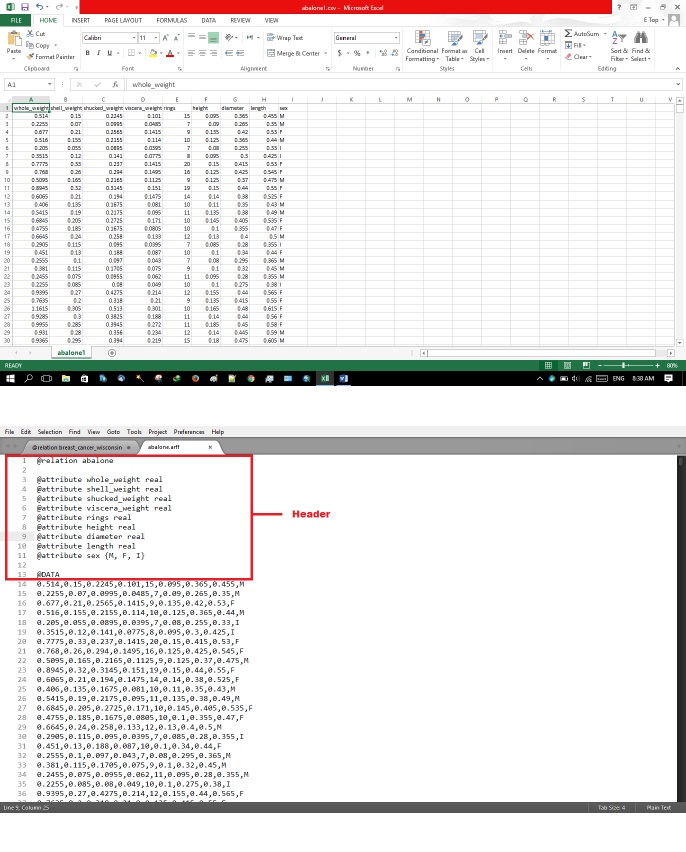
Langkah-langkah yang harus dilakukan:

1. Download data Abalone pada UCI (<https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/>)

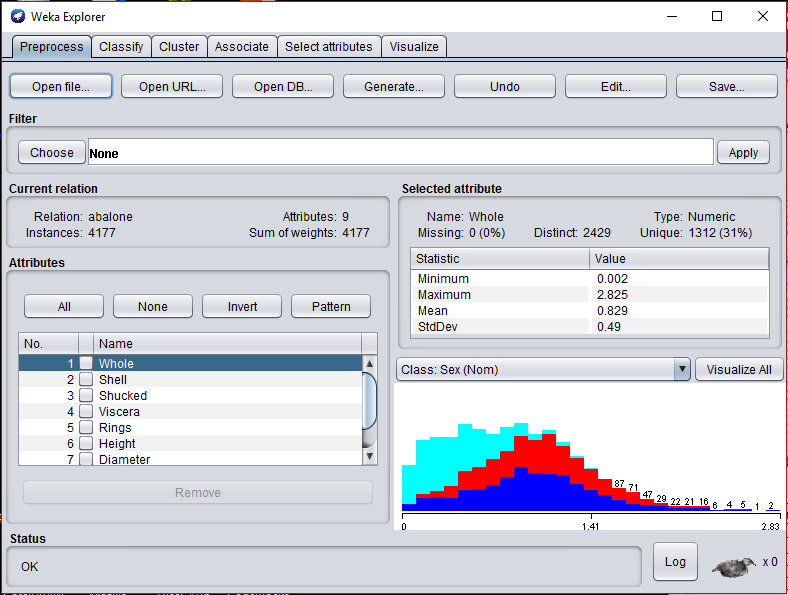




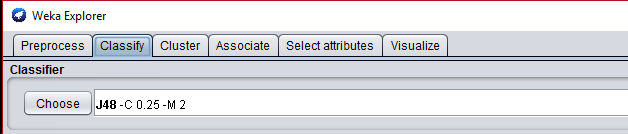
1. Ubah extensinya dari abalone.data menjadi abalone.csv
2. Buka file abalone.csv dan Urutan atribut file  adalah sebagai berikut: -Kolom 1 = Sex -Kolom 2 = Length -Kolom 3 = Diameter -Kolom 4 = Height -Kolom 5 = Whole weight -Kolom 6 = Shucked weight -Kolom 7 = Viscera weight -Kolom 8 = Shell weight -Kolom 9 = Rings
3. Letakkan atribut sex pada kolom terakhir lalu save data
4. Buka file .csv ke notepad. Lalu tambahkan bagian header. Setelah itu save as maenjadi extensi .arff



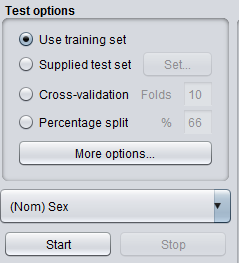
1. Jalankan WEKA GUI Chooser lalu pilih Explorer. Masukkan file Abalone.arff



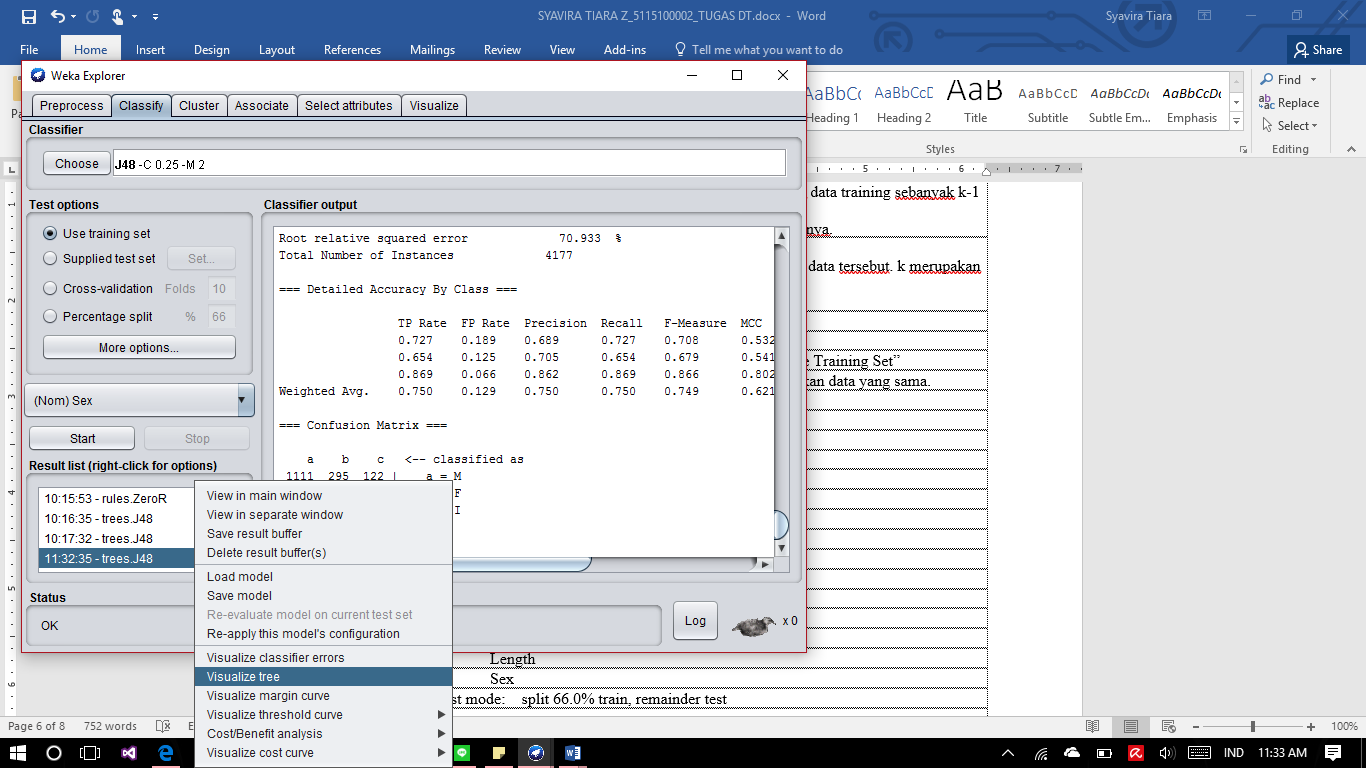
1. Pilih classiffy dan pilih tree tipe J48



1. Lalu pilih test option untuk pemilihan metode dan klik start



1. *Use training set*  
   Pengetesan dilakukan dengan menggunakan data training itu sendiri.
2. *Supplied test set*  
   Pengetesan dilakukan dengan menggunakan data lain. Dengan menggunakan option inilah, kita bisa melakukan prediksi terhadap data tes. Akan dijelaskan di postingan selanjutnya.
3. *Cross-validation*  
   Pada *cross-validation*, akan ada pilihan berapa *fold* yang akan digunakan. Nilai *default*-nya adalah 10. Mekanisme-nya adalah sebagai berikut :  
   Data training dibagi menjadi k buah subset (subhimpunan). Dimana k adalah nilai dari fold. Selanjutnya, untuk tiap dari subset, akan dijadikan data tes dari hasil klasifikasi yang dihasilkan dari k-1 subset lainnya. Jadi, akan ada 10 kali tes. Dimana, setiap datum akan menjadi data tes sebanyak 1 kali, dan menjadi data training sebanyak k-1 kali.  
   Kemudian, error dari k tes tersebut akan dihitung rata-ratanya.
4. *Percentage split*  
   Hasil klasifikasi akan dites dengan menggunakan k% dari data tersebut. k merupakan masukan dari *user*.
5. Lalu akan muncul output. Dan jika ingin menampilkan tree klik kanan pada result list

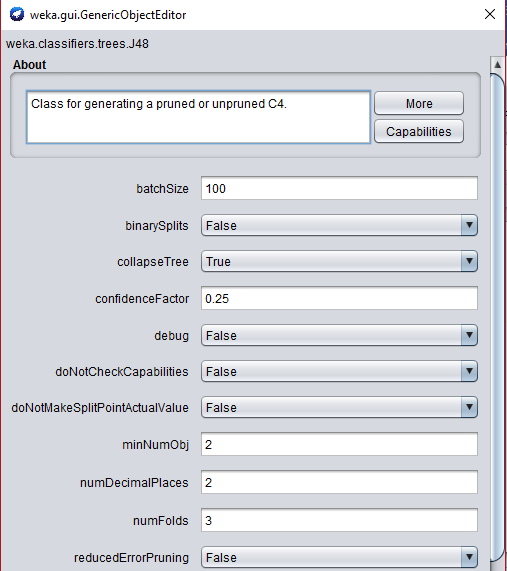


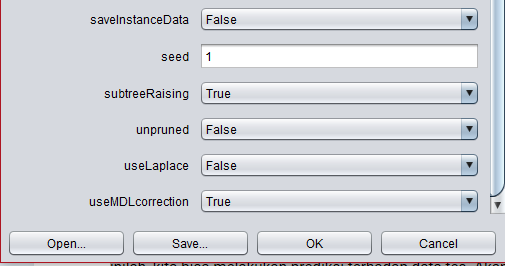
Soal:

1. Describe your process and methods

Langkah-langkah(proses dan metode) yang digunakan telah dijelaskah disebelumnya.

Dan menggunakan parameter default dari Weka





1. What were your result?
   1. Pertama kita akan mencoba metode “Use Training Set”

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: abalone

Instances: 4177

Attributes: 9

Whole

Shell

Shucked

Viscera

Rings

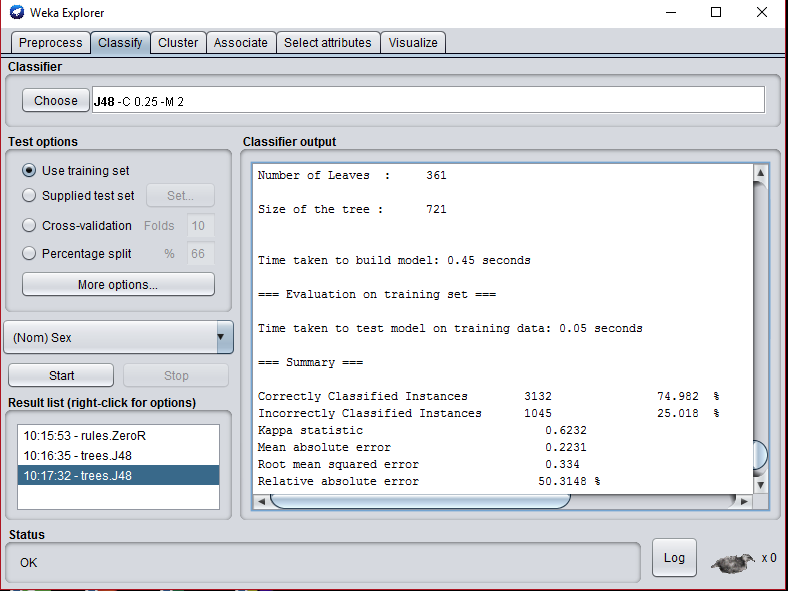
Height

Diameter

Length

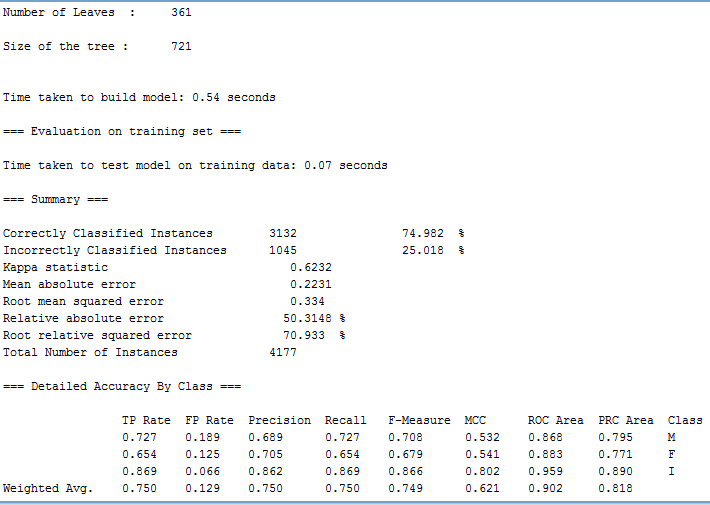
Sex

Test mode: split 66.0% train, remainder test

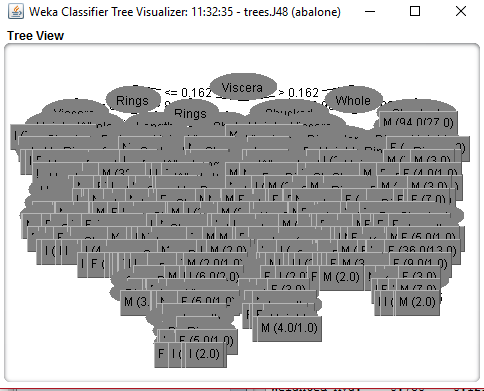


Number of Leaves : 361

Size of the tree : 721



Decision Tree:



* 1. Menggunakan Cross validation

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: abalone

Instances: 4177

Attributes: 9

Whole

Shell

Shucked

Viscera

Rings

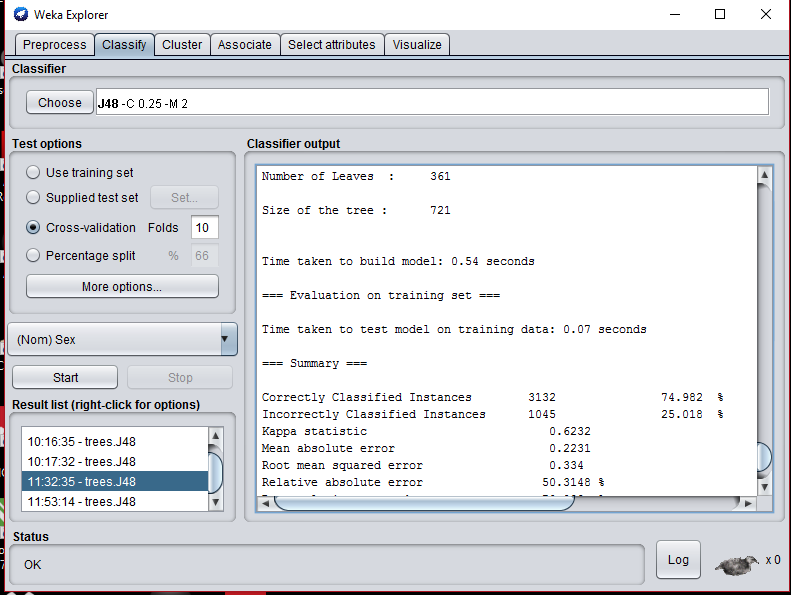
Height

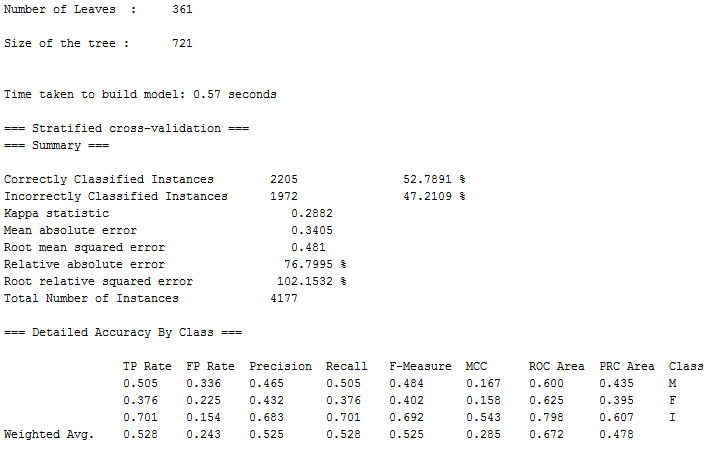
Diameter

Length

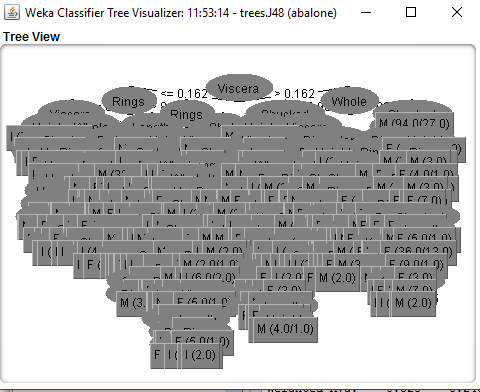
Sex

Test mode: 10-fold cross-validation





Decision Tree



* 1. Percentage split

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: abalone

Instances: 4177

Attributes: 9

Whole

Shell

Shucked

Viscera

Rings

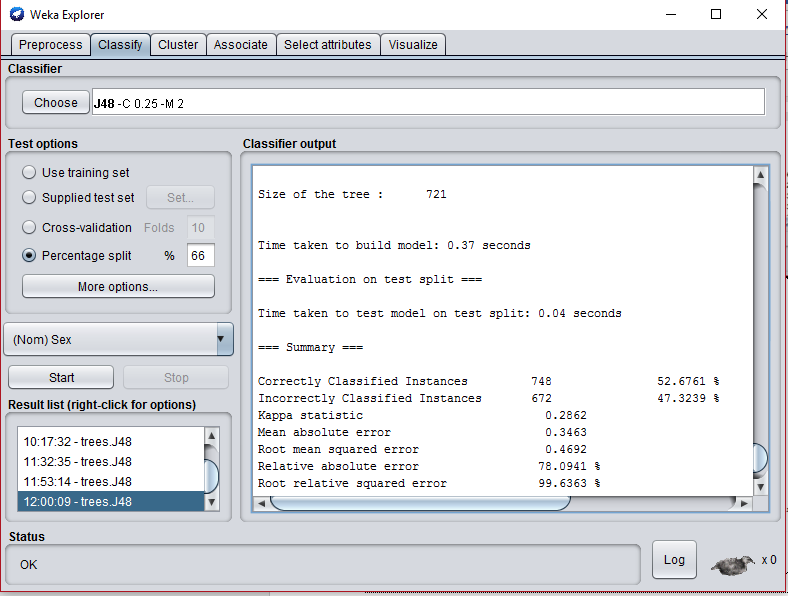
Height

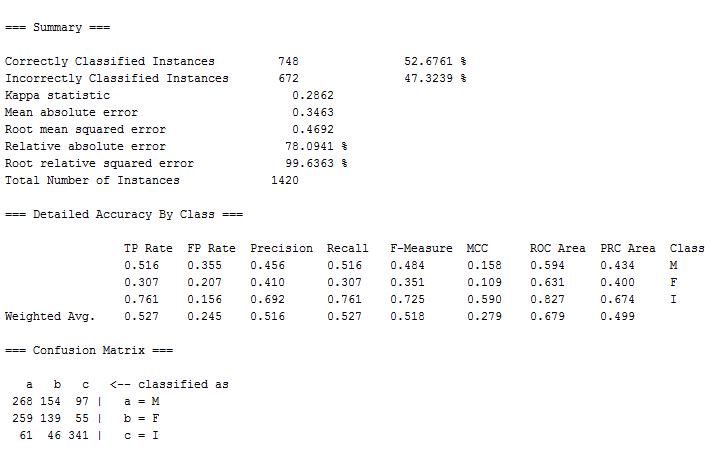
Diameter

Length

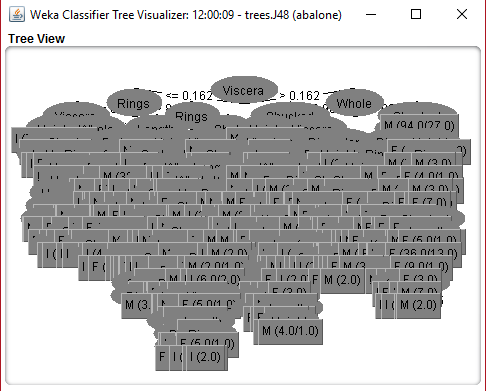
Sex

Test mode: split 66.0% train, remainder test





Decision Tree



1. a. Use Training Set

Correctly Classified Instances 3132 74.982 %

Incorrectly Classified Instances 1045 25.018 %

Number of Leaves : 361

Size of the tree : 721

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.727 0.189 0.689 0.727 0.708 0.532 0.868 0.795 M

0.654 0.125 0.705 0.654 0.679 0.541 0.883 0.771 F

0.869 0.066 0.862 0.869 0.866 0.802 0.959 0.890 I

Weighted Avg. 0.750 0.129 0.750 0.750 0.749 0.621 0.902 0.818

b. Cross Validation

Correctly Classified Instances 2205 52.7891 %

Incorrectly Classified Instances 1972 47.2109 %

Number of Leaves : 361

Size of the tree : 721

c. Percentage split

Correctly Classified Instances 748 52.6761 %

Incorrectly Classified Instances 672 47.3239 %

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

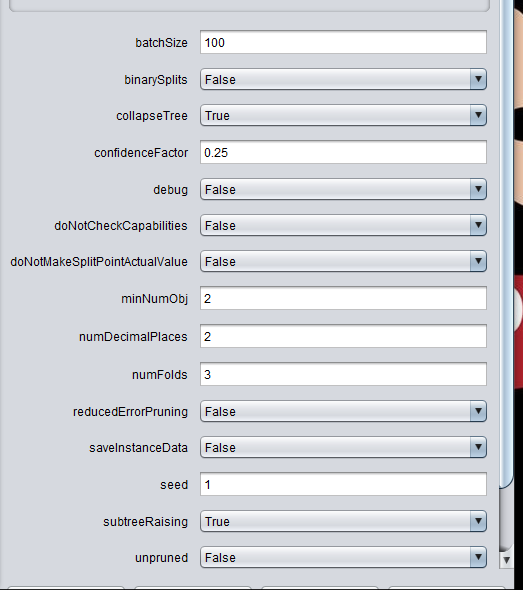
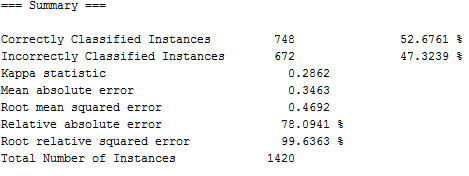
0.516 0.355 0.456 0.516 0.484 0.158 0.594 0.434 M

0.307 0.207 0.410 0.307 0.351 0.109 0.631 0.400 F

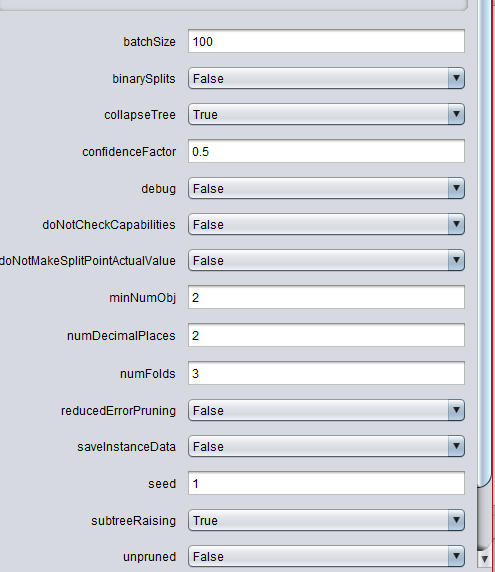
0.761 0.156 0.692 0.761 0.725 0.590 0.827 0.674 I

Weighted Avg. 0.527 0.245 0.516 0.527 0.518 0.279 0.679 0.499

use training set lebih tinggi nilai akurasinya dari cross-validation, kemungkinan, karena use training set diuji dengan dia sendiri, sementara cross validation data yang digunakan untuk membentuk hasil klasifikasi dan untuk mengetesnya berbeda.

Paramater default mengikuti di WEKA. Dengan Confidence factor 0.25

 Confidence factor 0.5 beserta hasilnya

