Nama : Ferza Reyaldi
NIM : 09021281924060
Kelas : TI REGULER
Mata Kuliah : Data Mining

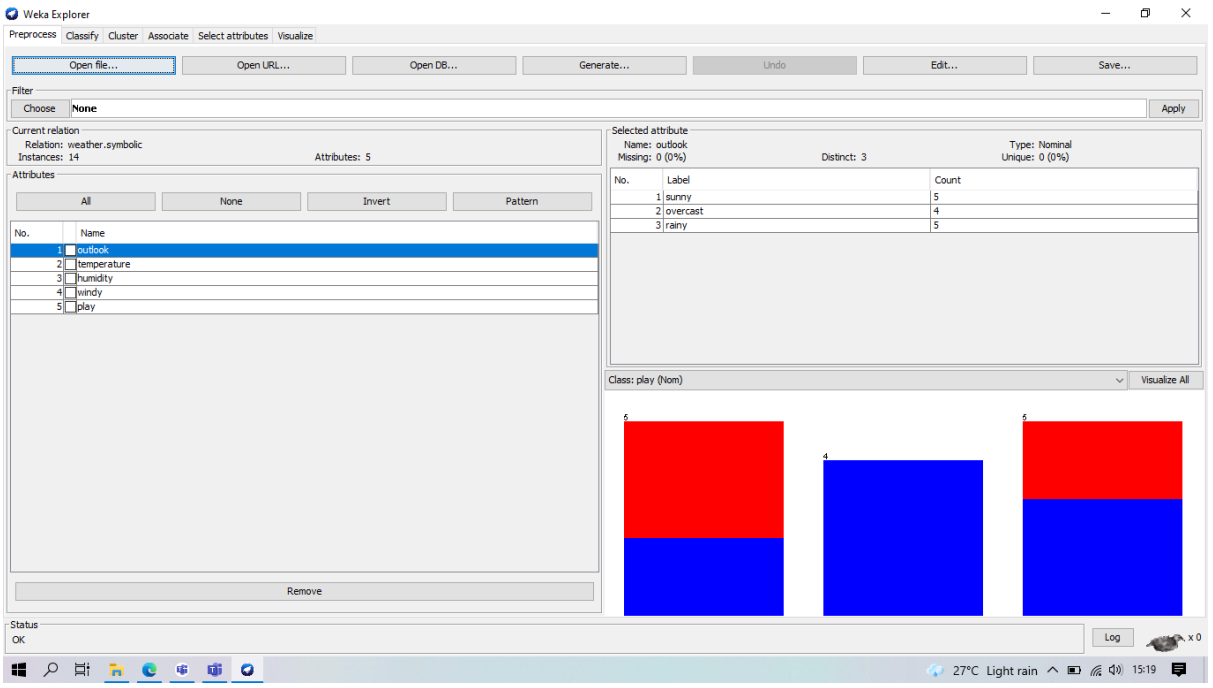
Data Mining with Weka (3.1-5.4)

3.1 Simplicity

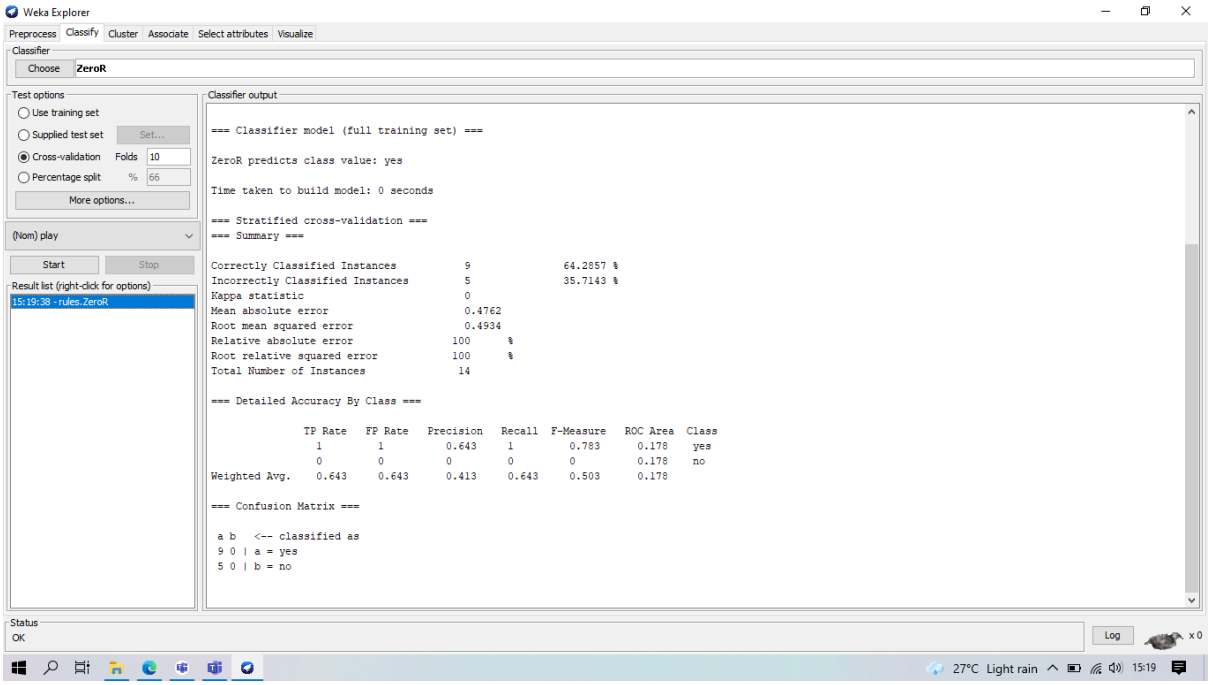
Dalam mencoba suatu algoritma dalam Data Mining diharuskan mencoba dari sesuatu yang paling sederhana terlebih dulu karena mungkin dengan metode paling sederhana telah menunjukkan hasil yang baik. Tidak ada algoritma yang terbaik secara umum, keberhasilan suatu algoritma ditentukan juga dengan domain dimana algoritma tersebut diterapkan.

Salah satu struktur model yang sederhana adalah satu atribut mengerjakan semua pekerjaan, contohnya adalah OneR.

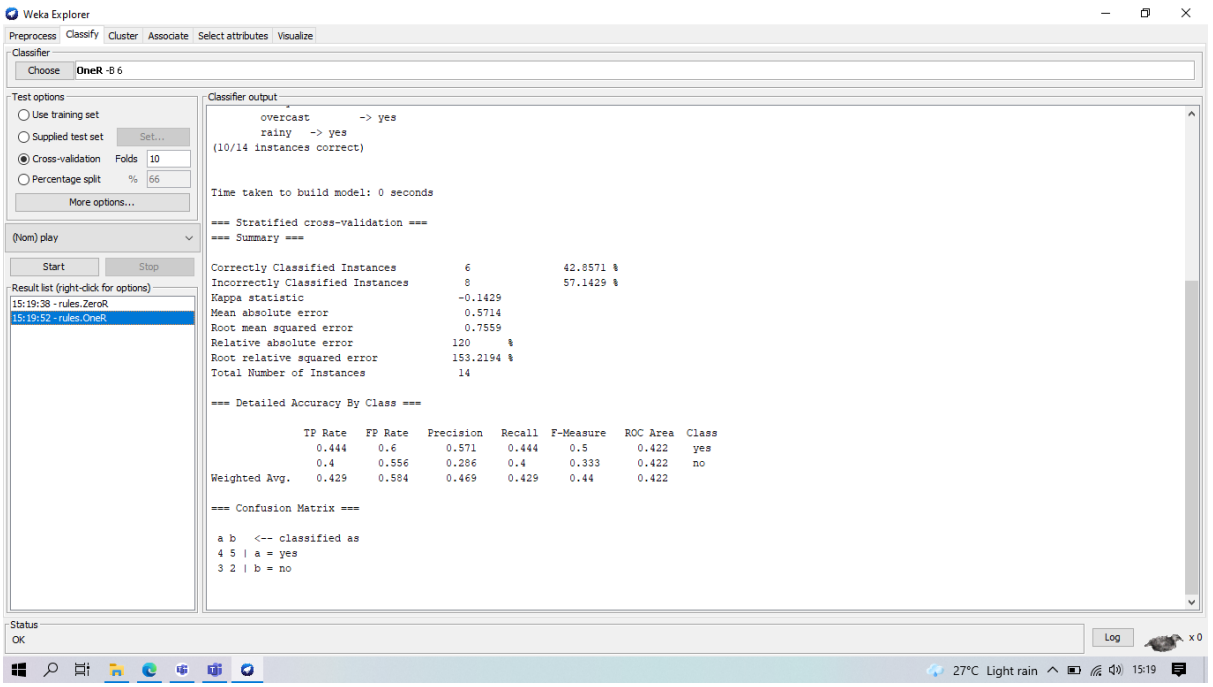
Load dataset weather.



Gunakan ZeroR untuk menentukan baseline accuracynya. Baseline accuracynya adalah 64,3%.



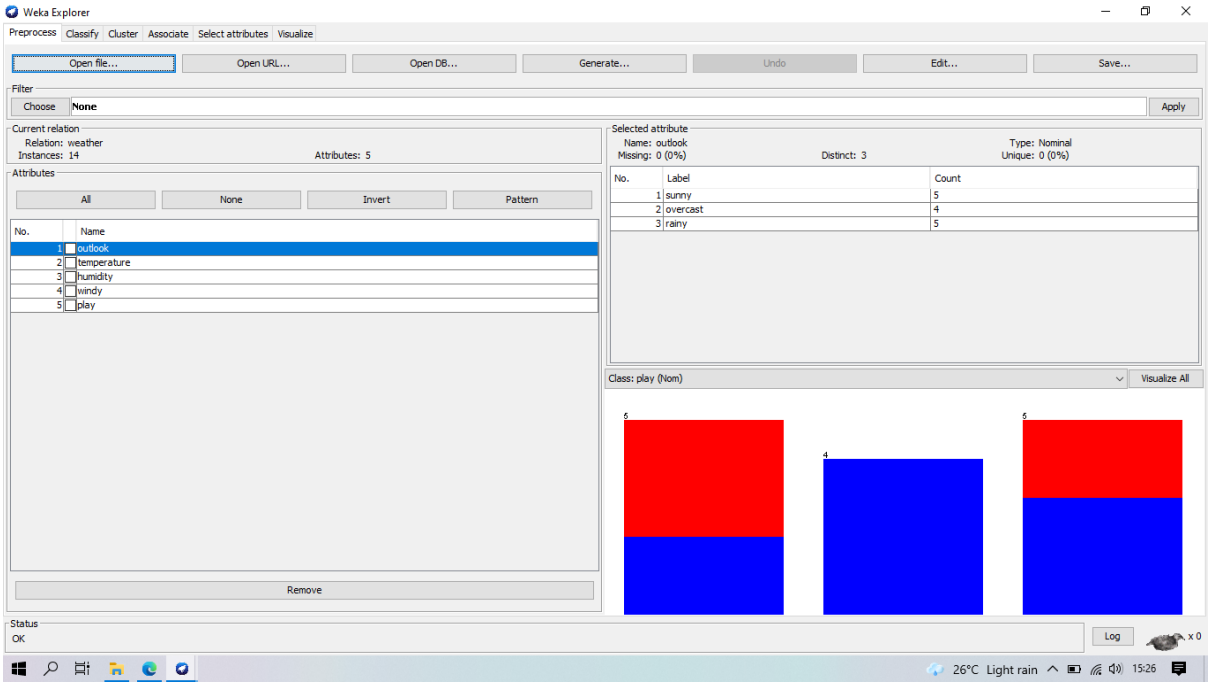
Berdasarkan prinsip simplicity, gunakan algoritma yang sederhana terlebih dulu, contohnya OneR. untuk akurasi prediksinya hanya 42,8% (lebih kecil dari baseline accuracy), sehingga bisa disimpulkan bahwa dataset tidak cocok dengan classifier OneR.



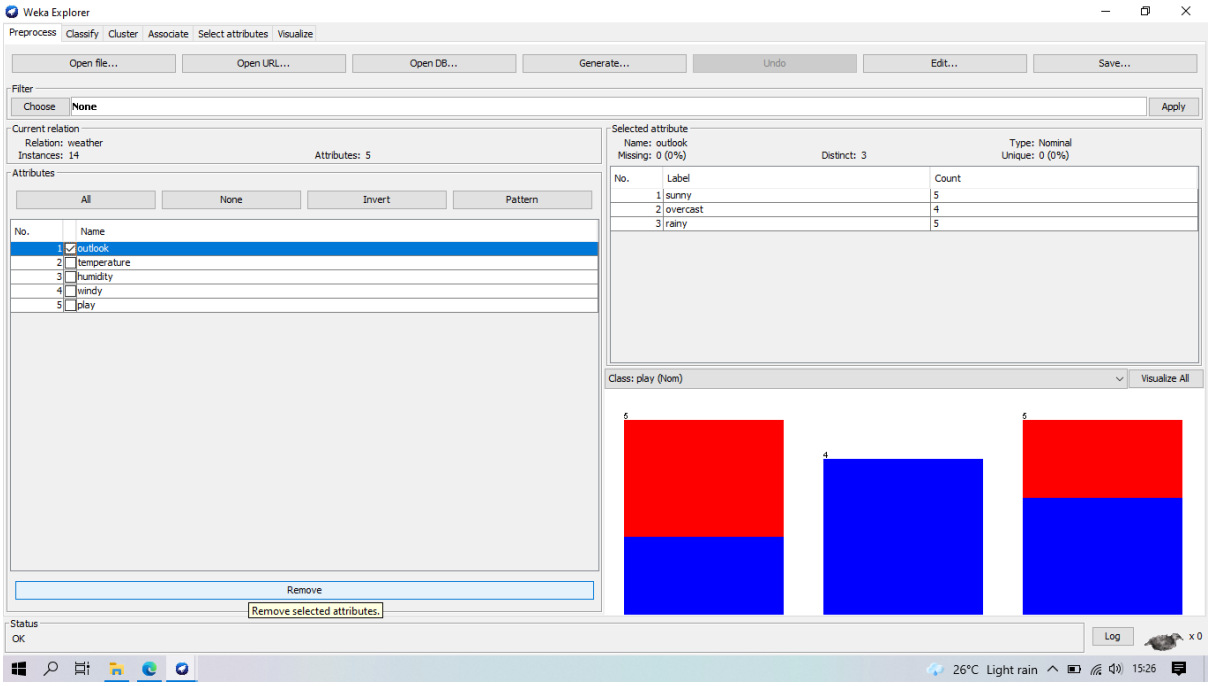
3.2 Overfitting

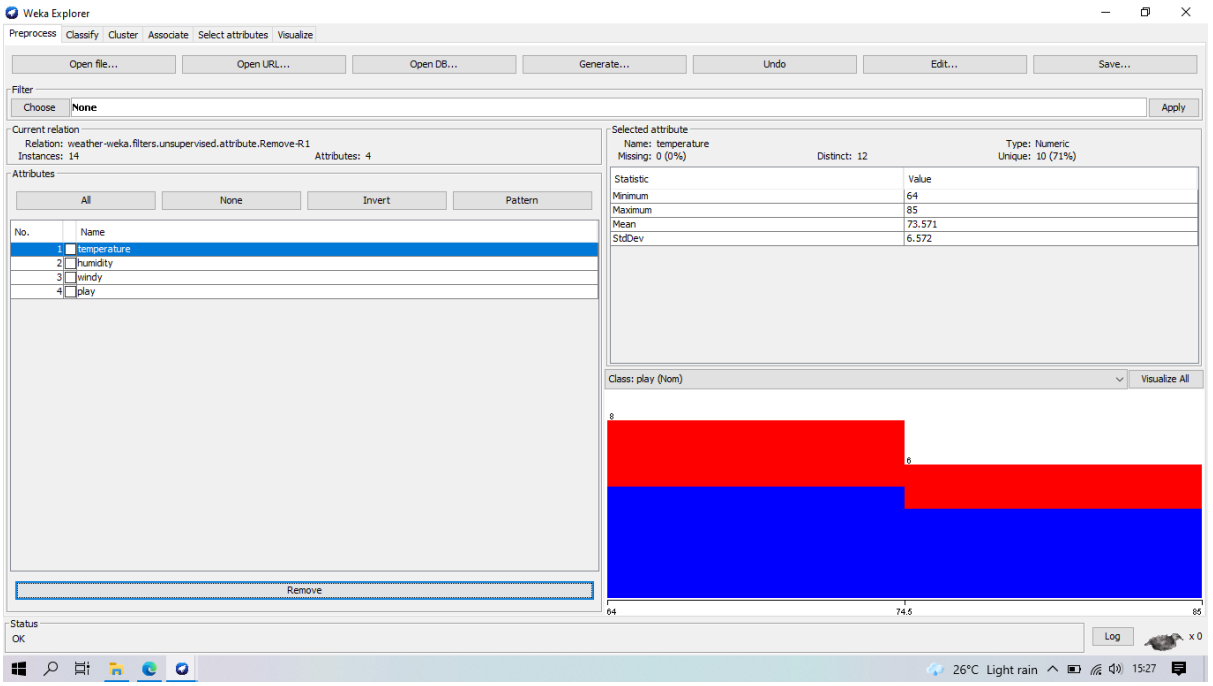
Overfitting sering terjadi karena model terlalu menyesuaikan dengan training data yang ada, sehingga ketika muncul data baru yang independen, hasil yang didapatkan sering kali tidak sesuai harapan.

Load dataset weather.

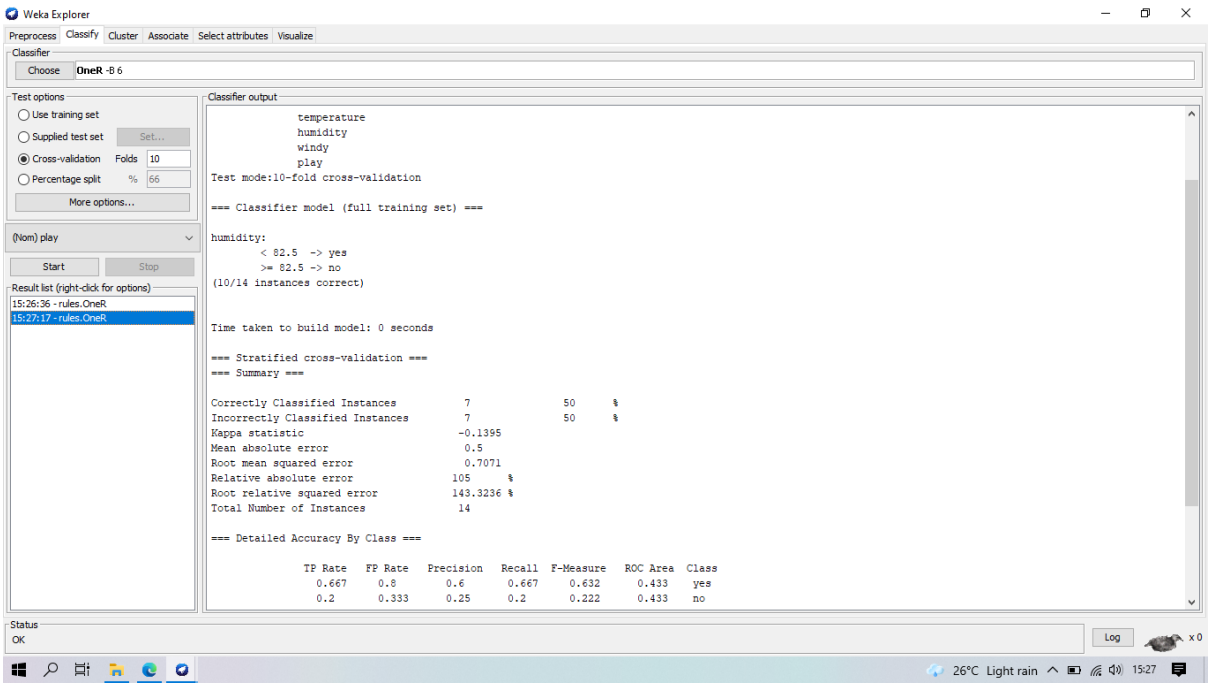


Hapus atribut outlook.

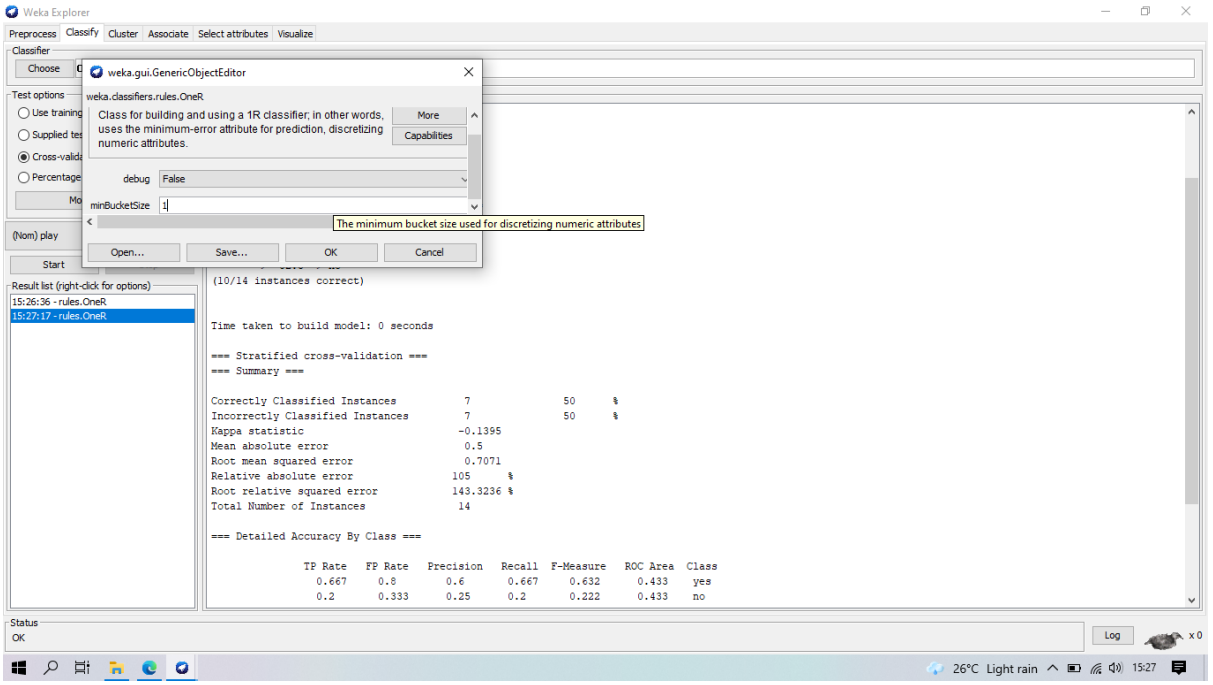


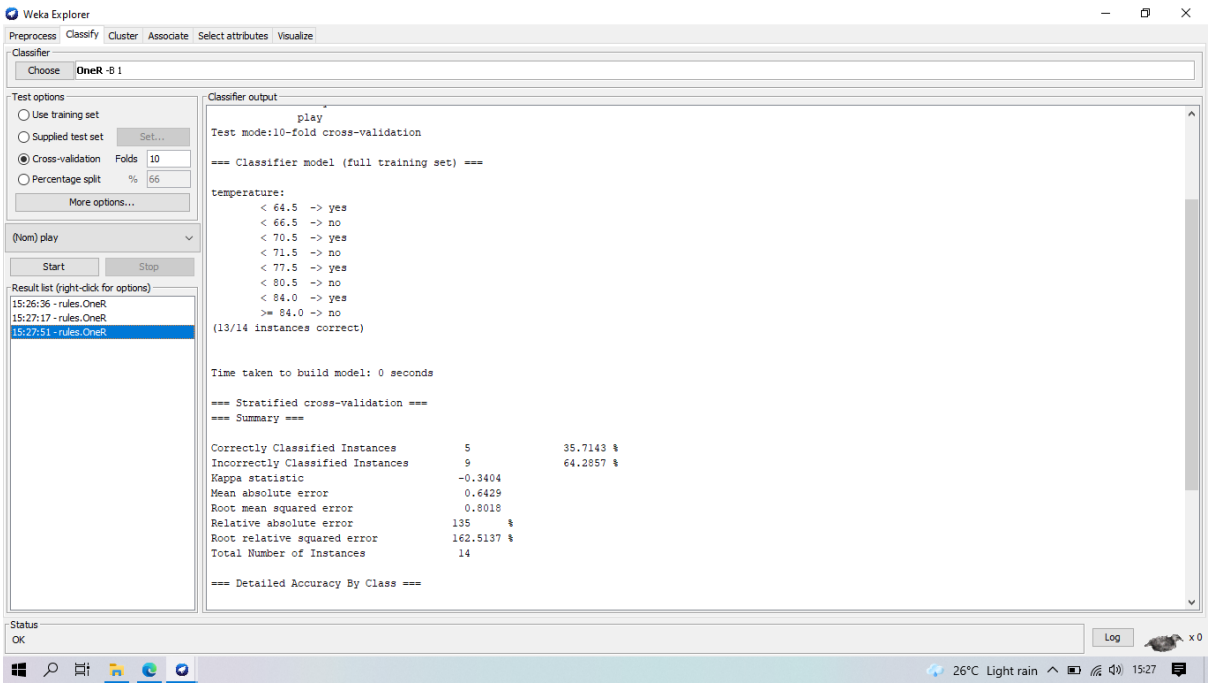


Gunakan classifier OneR untuk membuat model, hasilnya akurasinya 50%.



Klik nama classifier, kemudian terbuka jendela GenericObjectEditor. Ubah minBucketSize diubah menjadi 1. Ketika dijalankan kembali, akurasinya hanya bernilai 35,7%.

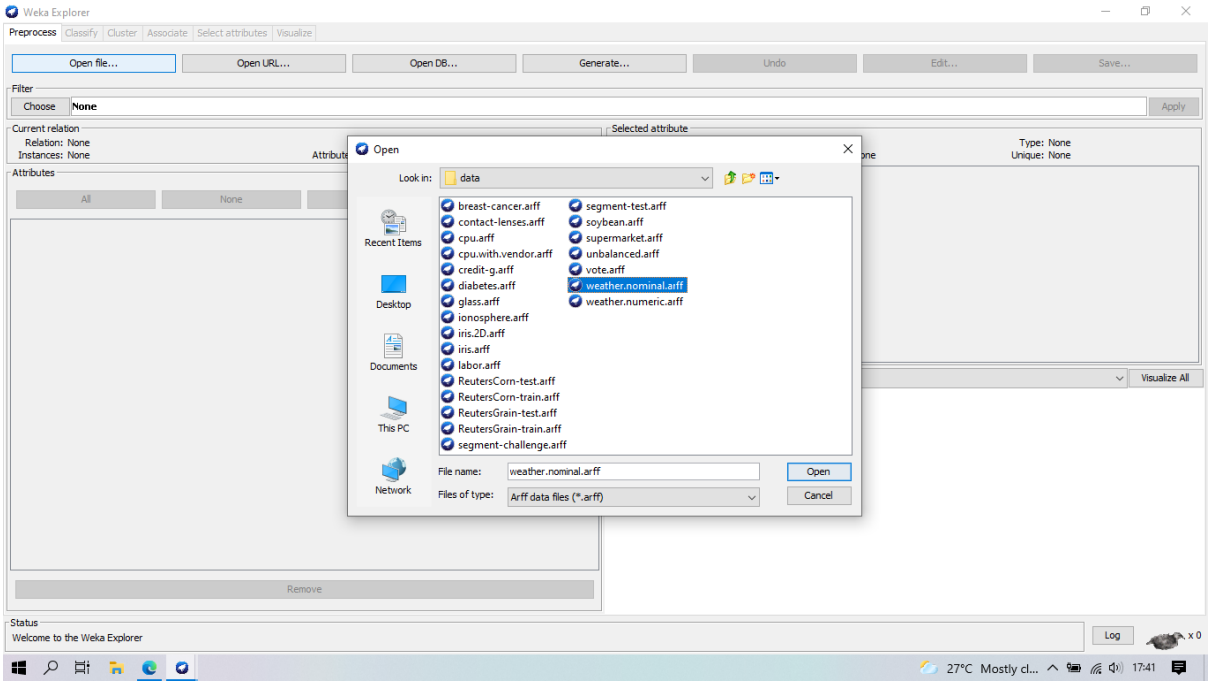


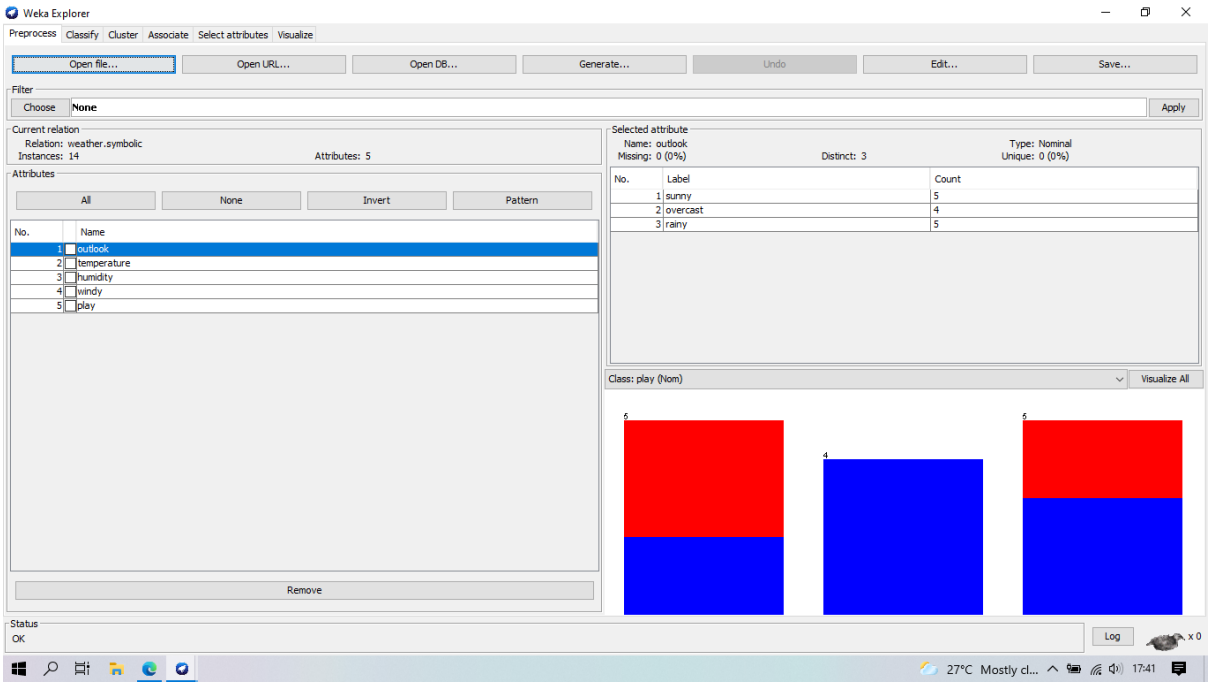


3.3 Using Probabilities

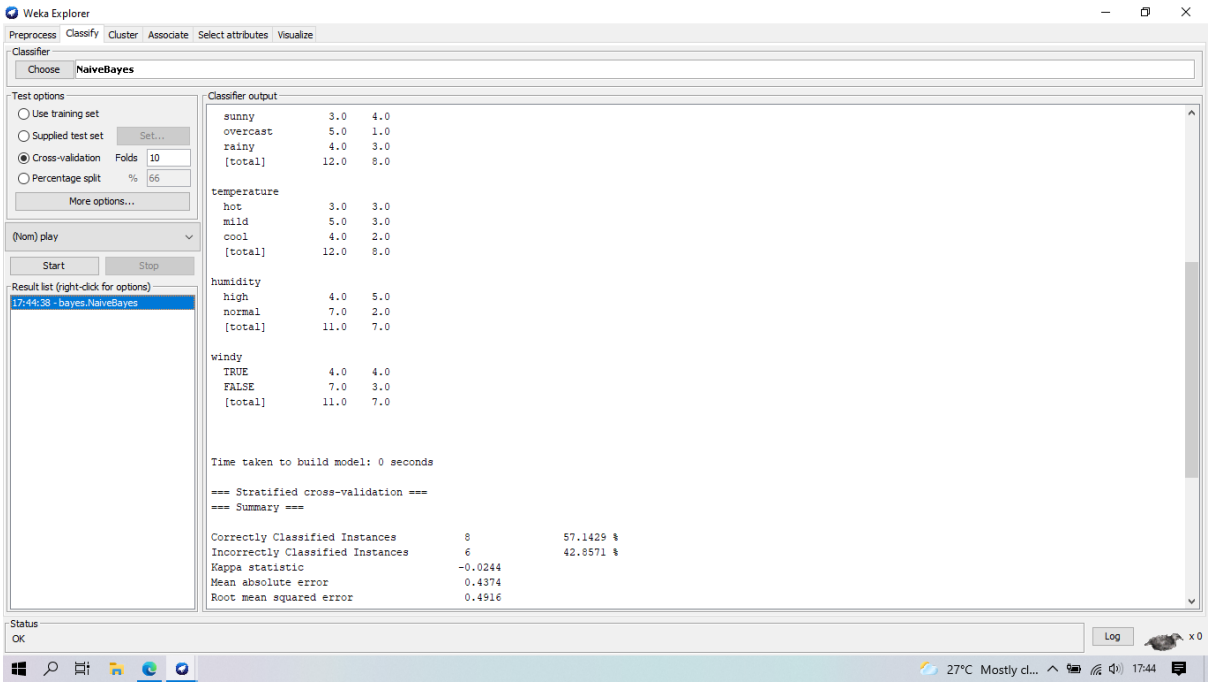
Salah satu bentuk model yang sederhana adalah menggunakan probabilitas, contohnya adalah algoritma naive bayes.

Load dataset weather nominal.





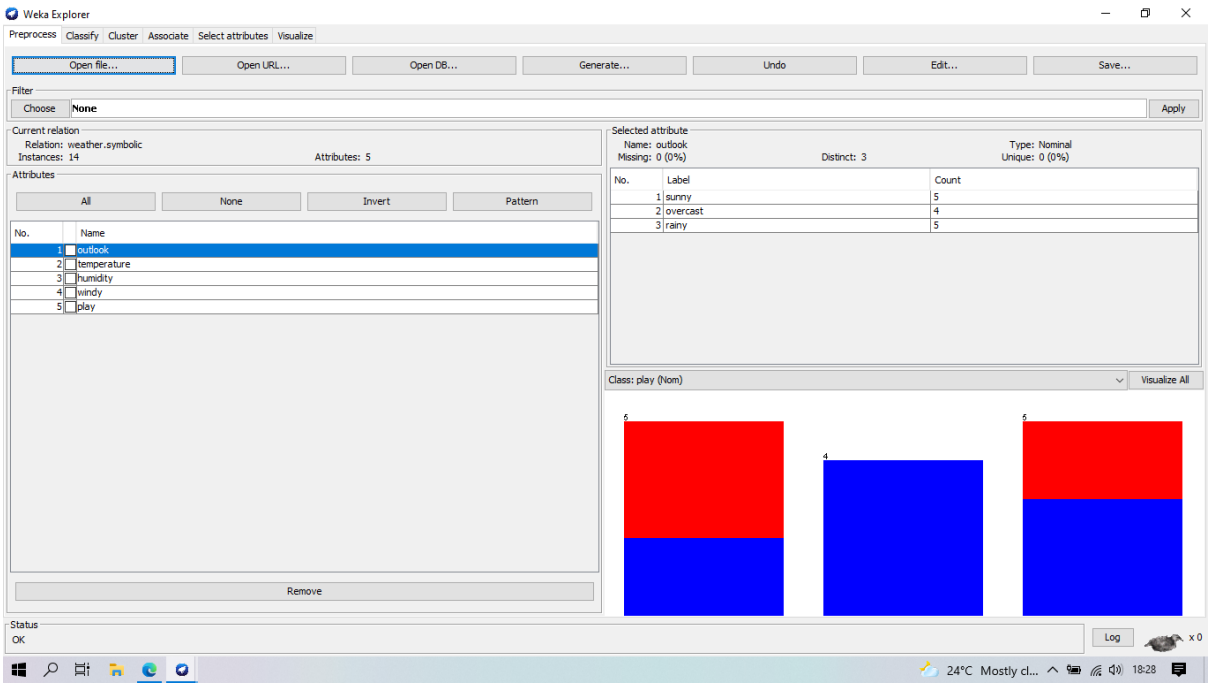
Pada tab classify, pilih classifier Naive Bayes. Akurasinya adalah 57%.



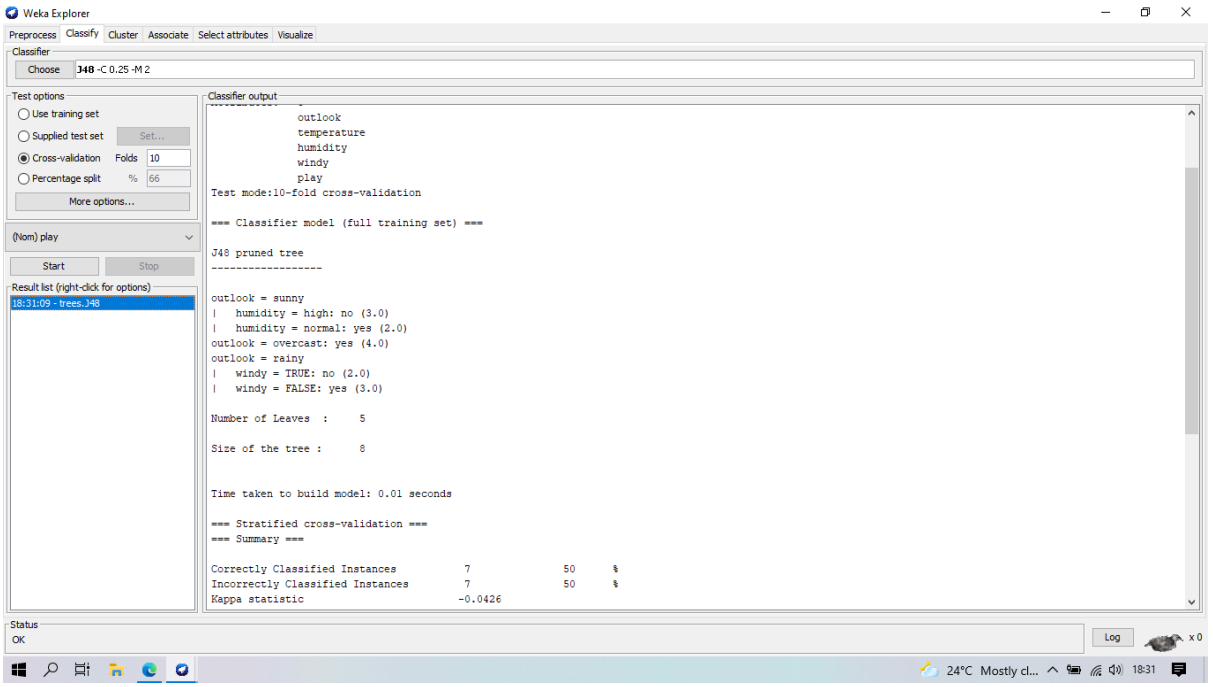
3.4 Decision Trees

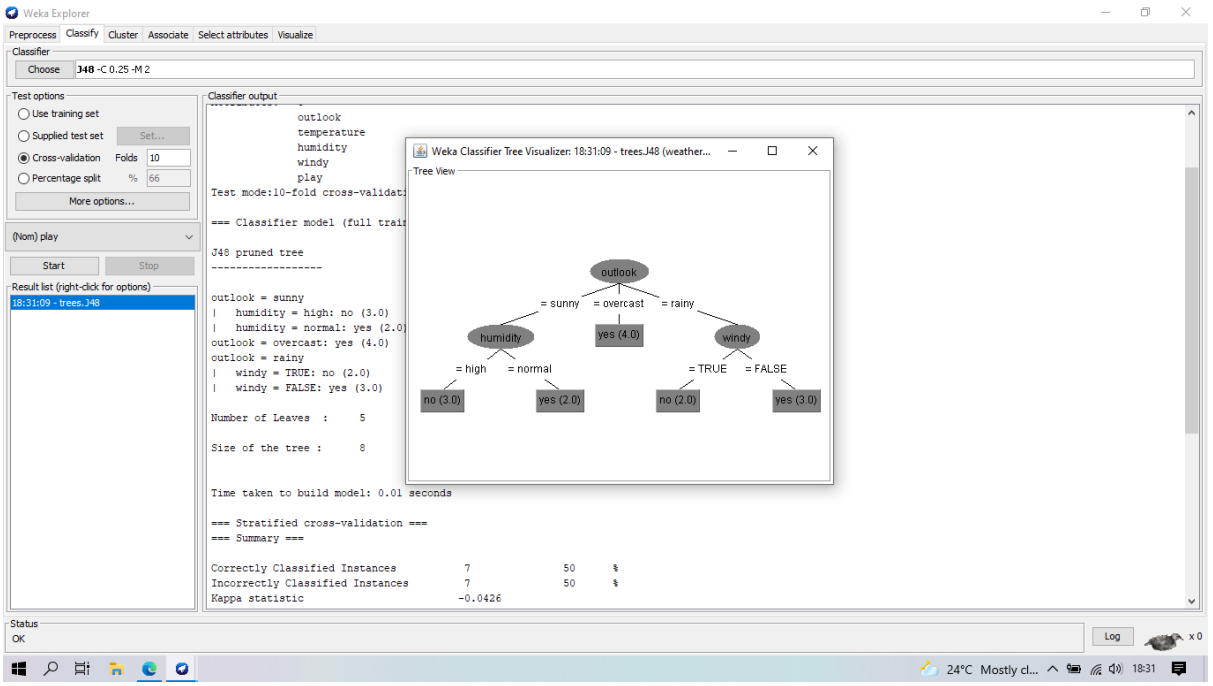
Decision Trees adalah satu satu bentuk model sederhana, yaitu menggunakan beberapa atribut dalam melakukan pembuatan model.

Load dataset Weather.



Pilih Classifier J48, hasil akurasinya 50%.

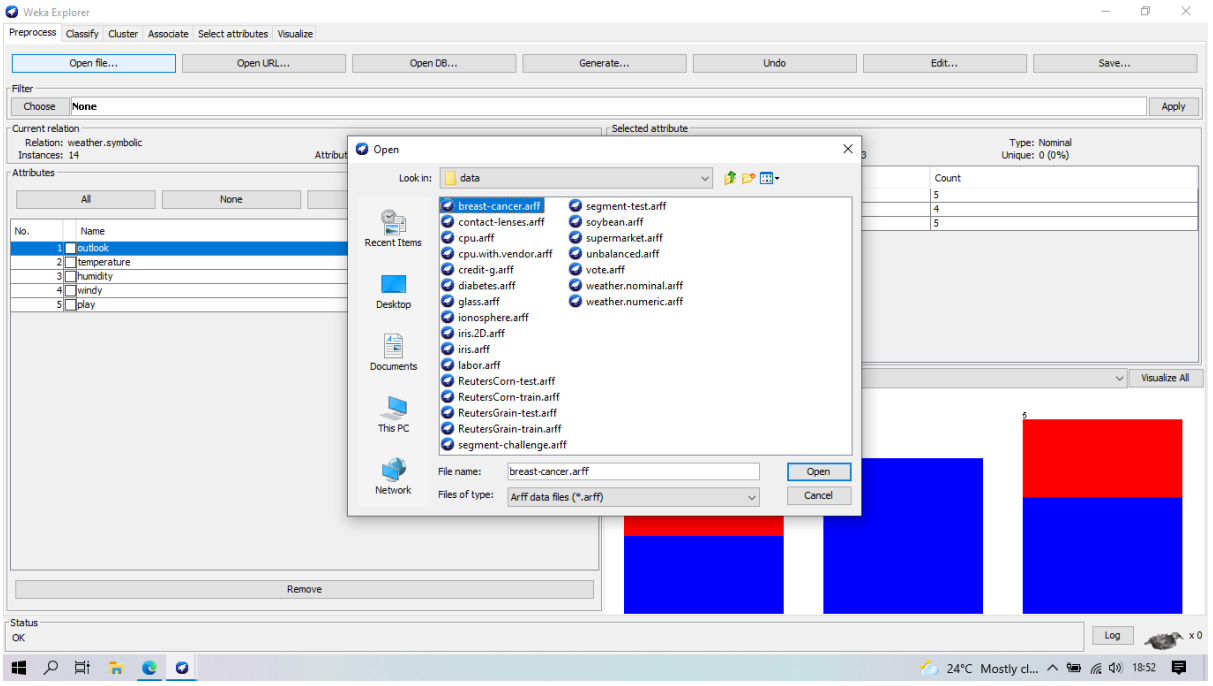


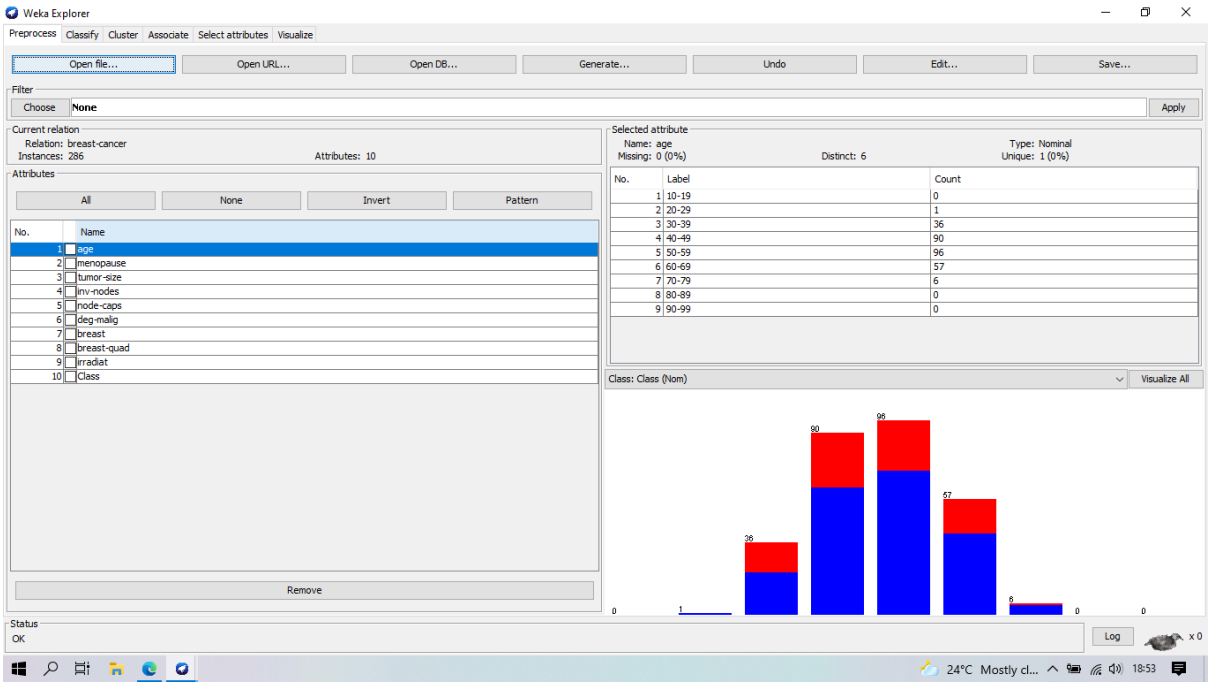


3.5 Pruning Decision Trees

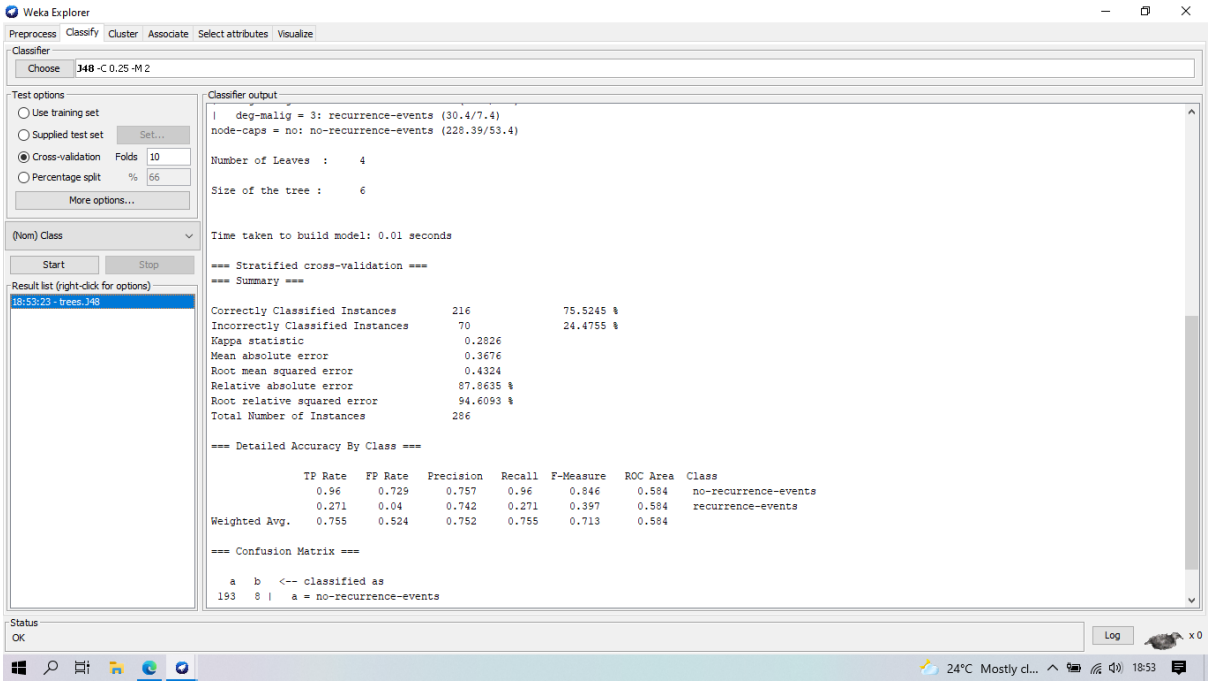
Pruning Decision Trees dimanfaatkan untuk pemangkasan Tree untuk mencegah terlalu banyak cabang pada suatu Tree dan mengoptimalkan akurasi model yang dibuat.

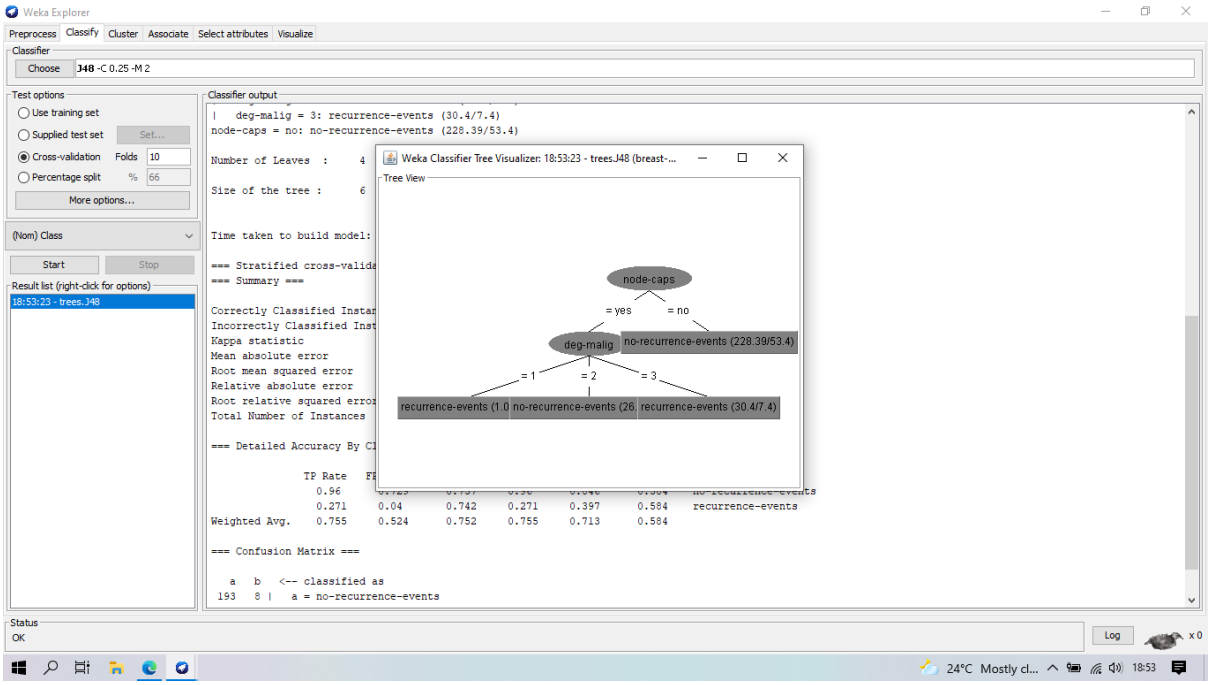
Load dataset breast cancer.

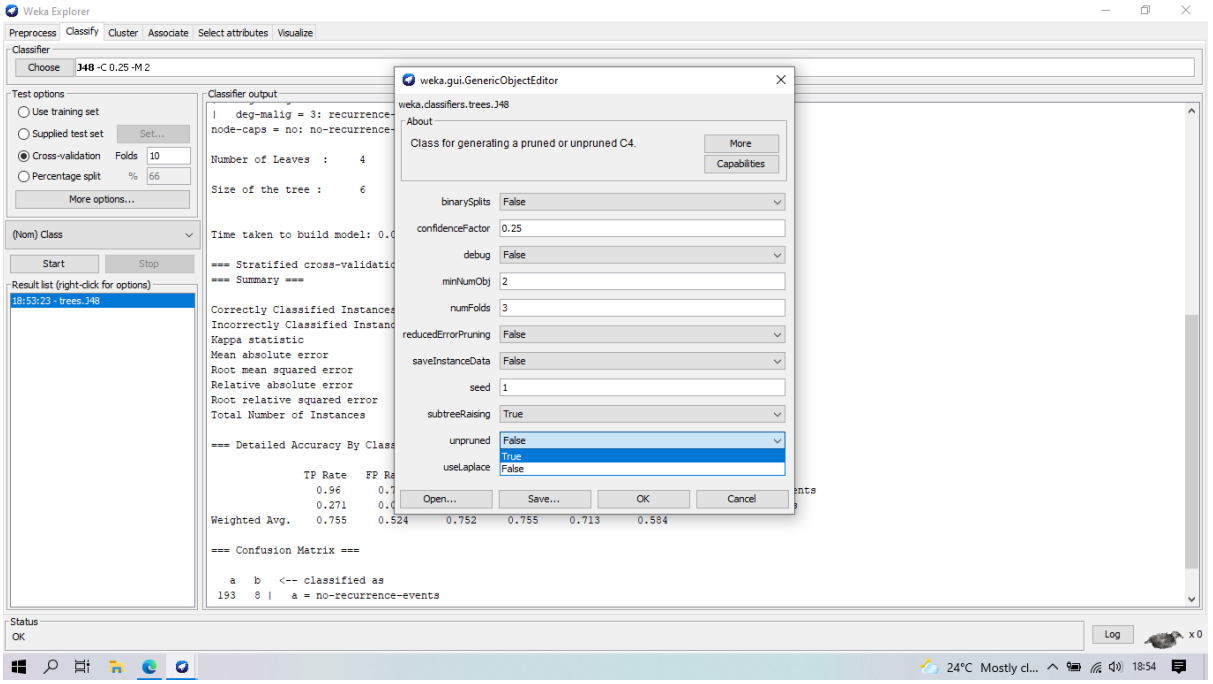




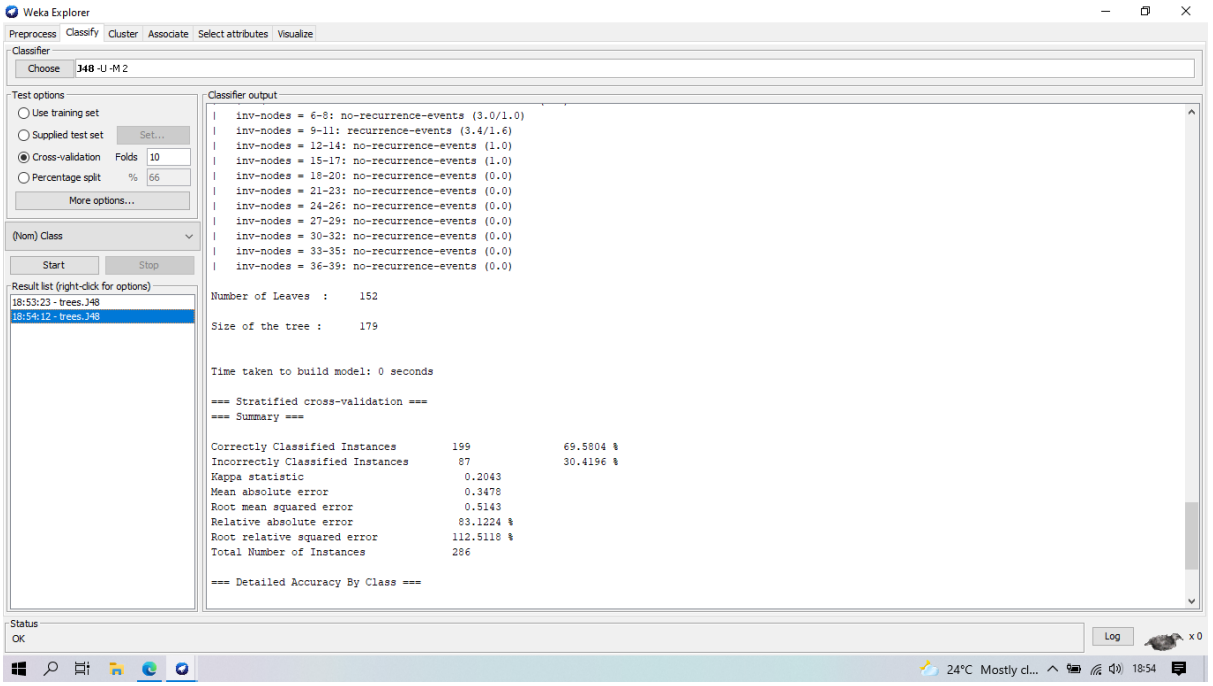
Gunakan classifier J48. Jalankan, akurasi 75,5% jika dilakukan pruning (secara default pruning digunakan dalam melatih model).

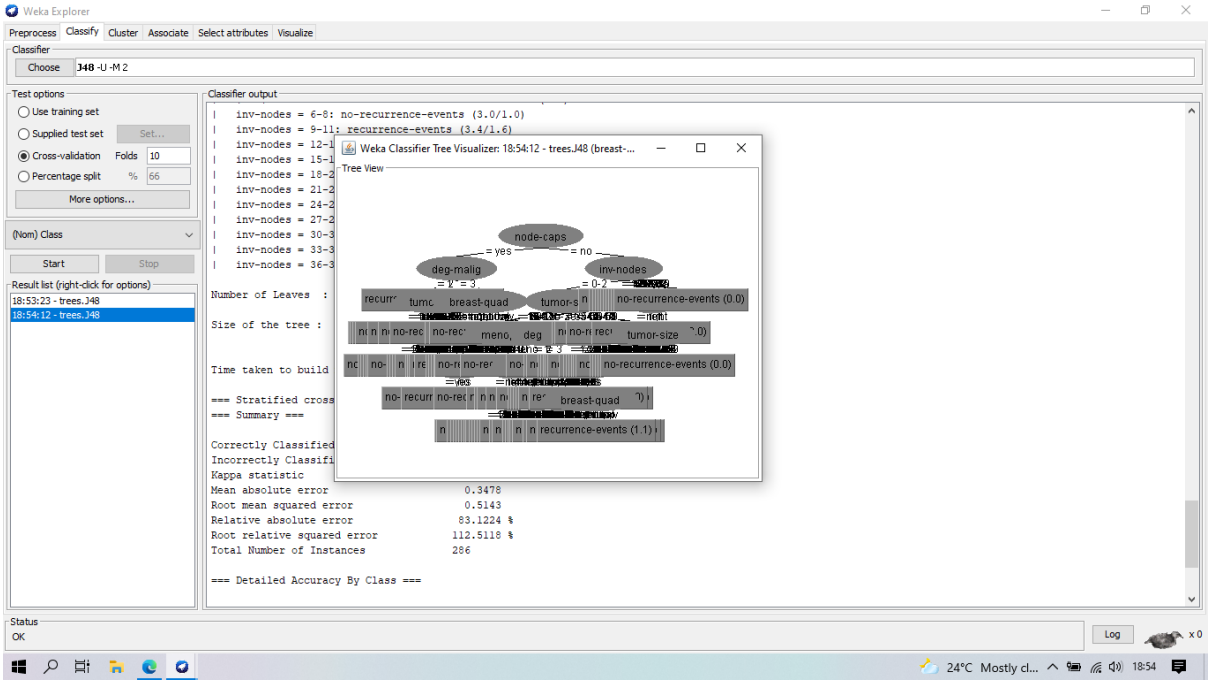






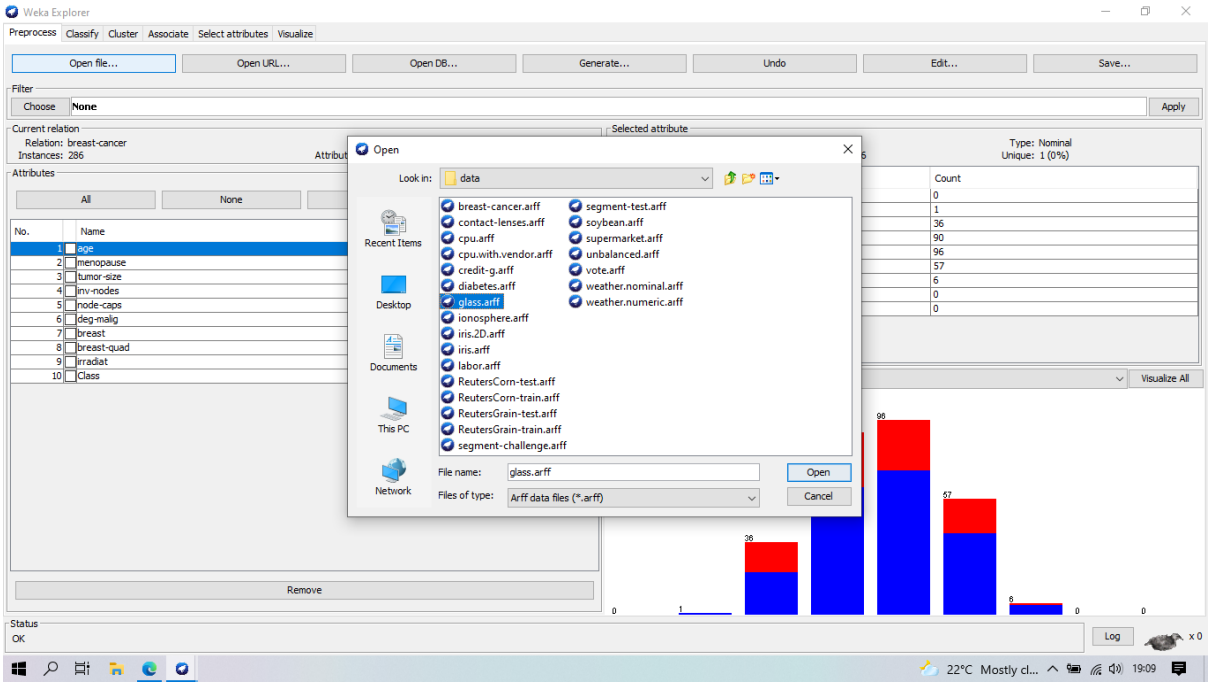
Ketika diatur pruningnya menjadi FALSE, hasil yang didapatkan seperti gambar dibawah. Terlalu banyak ranting dan daun, dan akurasinya pun menurun.

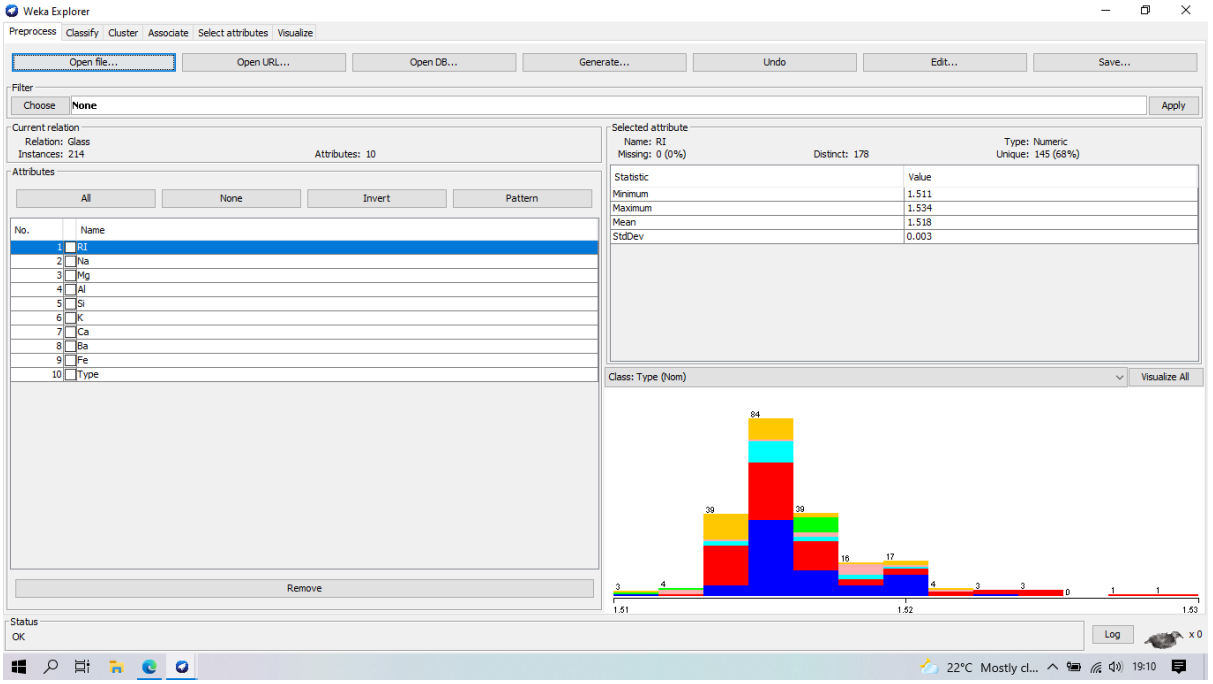




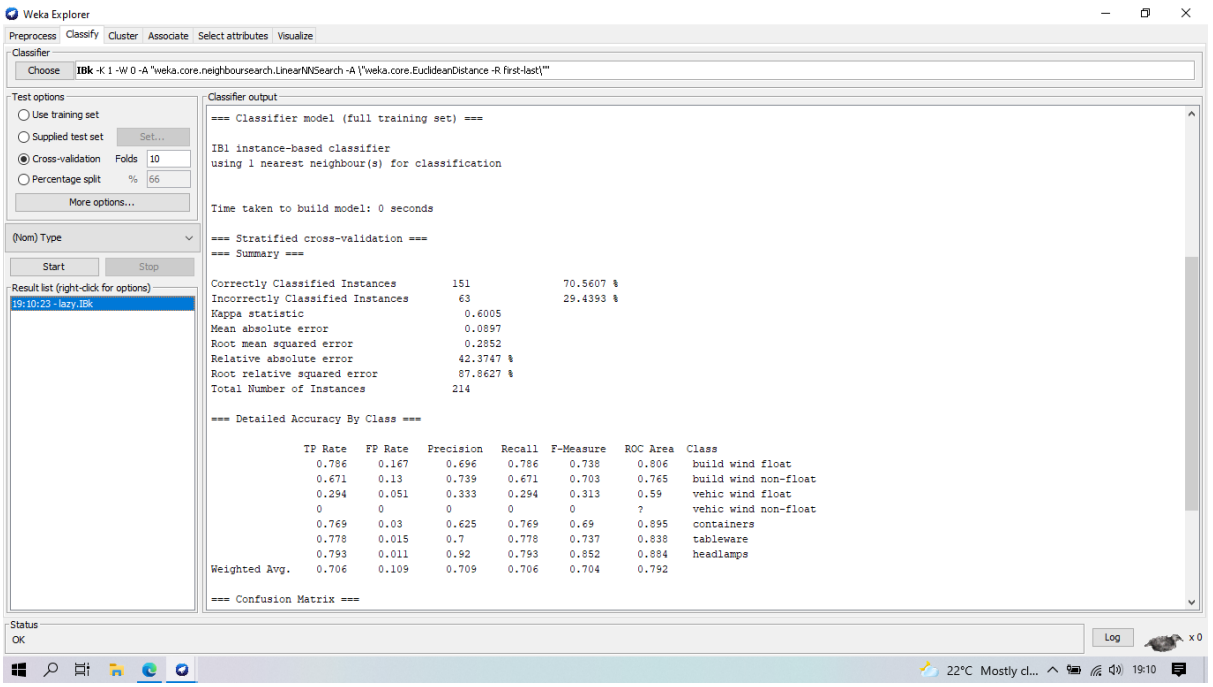
3.6 Nearest Neighbor

Load dataset glass.

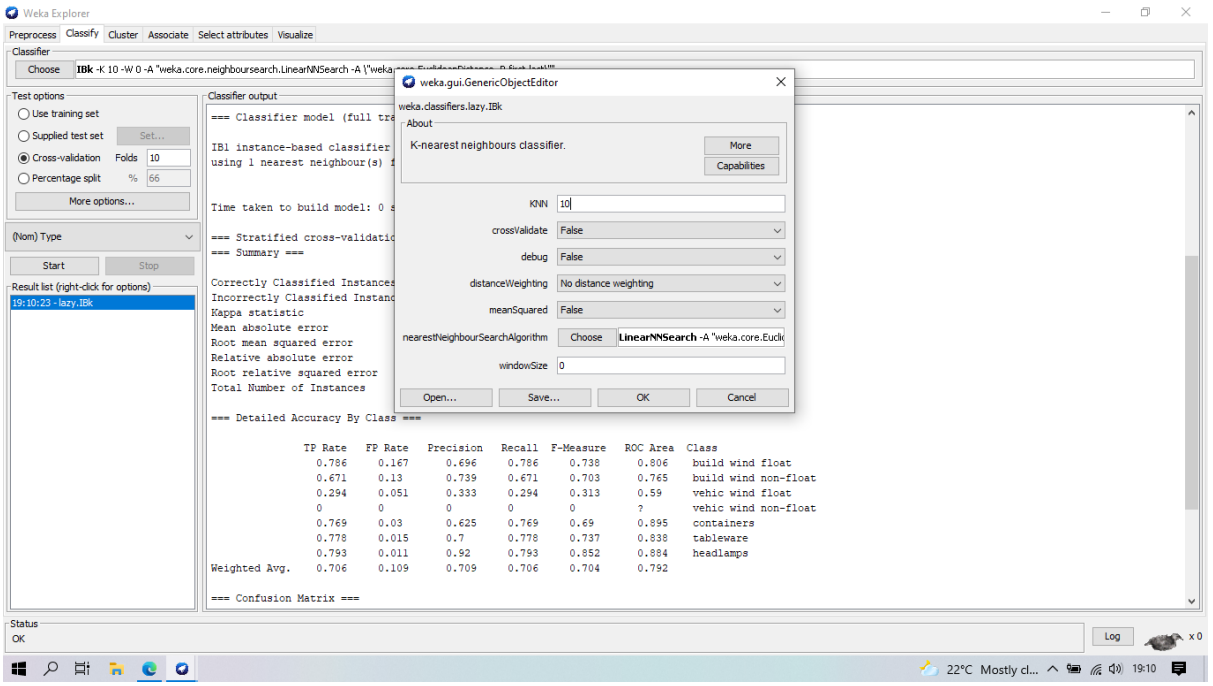


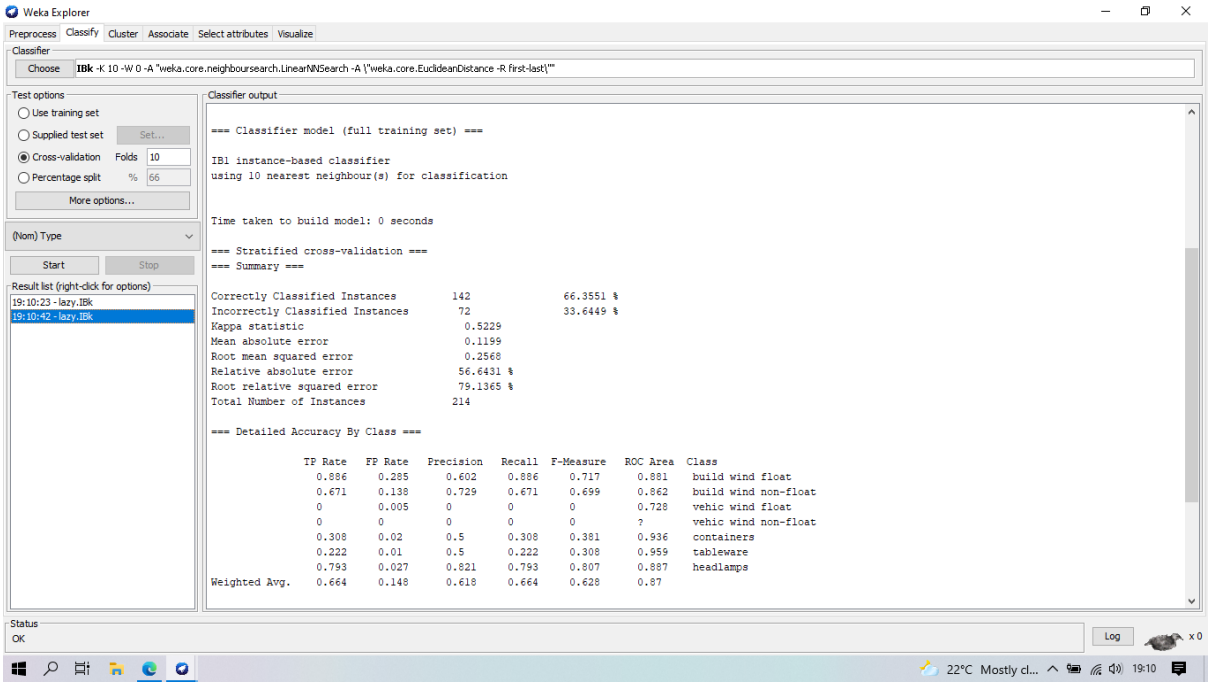


Gunakan classifier IBk. akurasi yang didapatkan 70,5%.

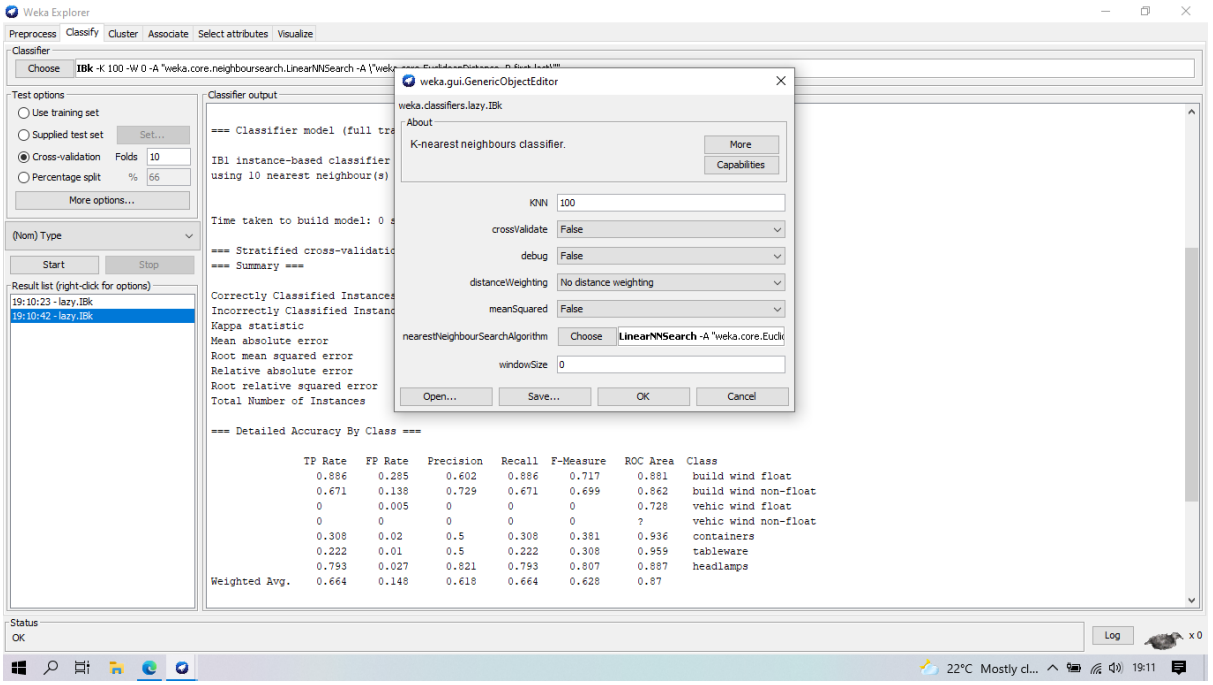


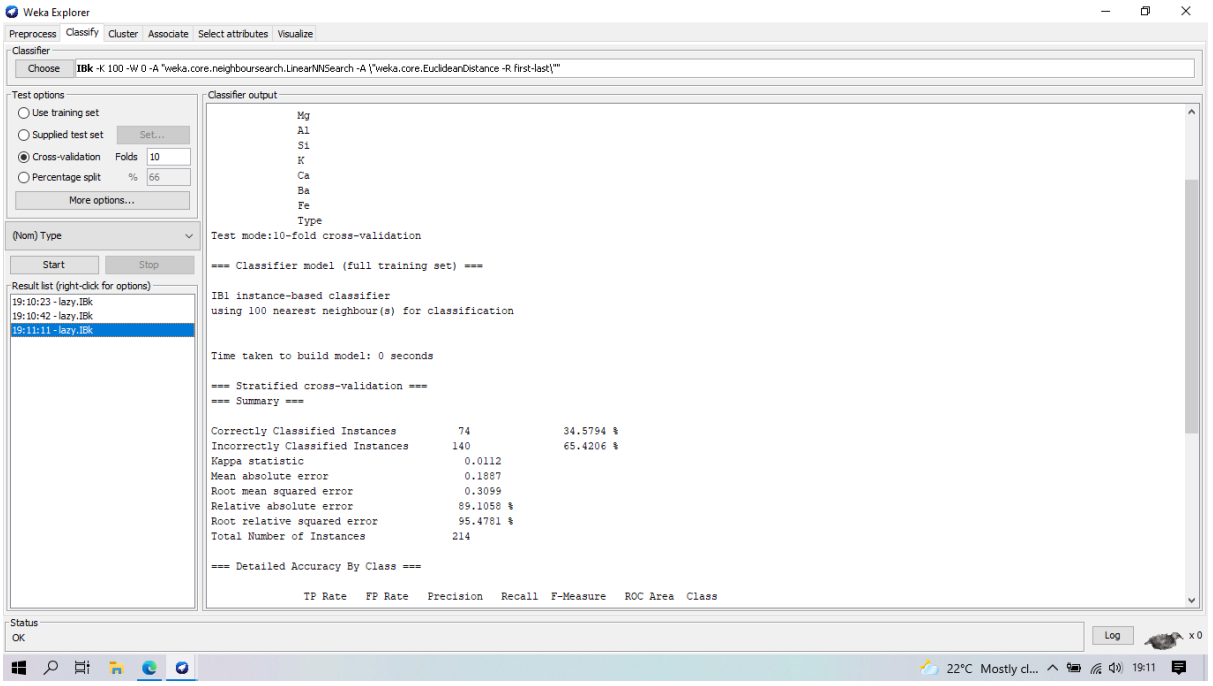
Ubah nilai KNN menjadi 10. (Hal ini berarti suatu instance baru yang ditambahkan, akan mencari 10 instance dengan jarak terdekat). Akurasi sekitar 66%, lebih kecil dibandingkan KNN bernilai 1.



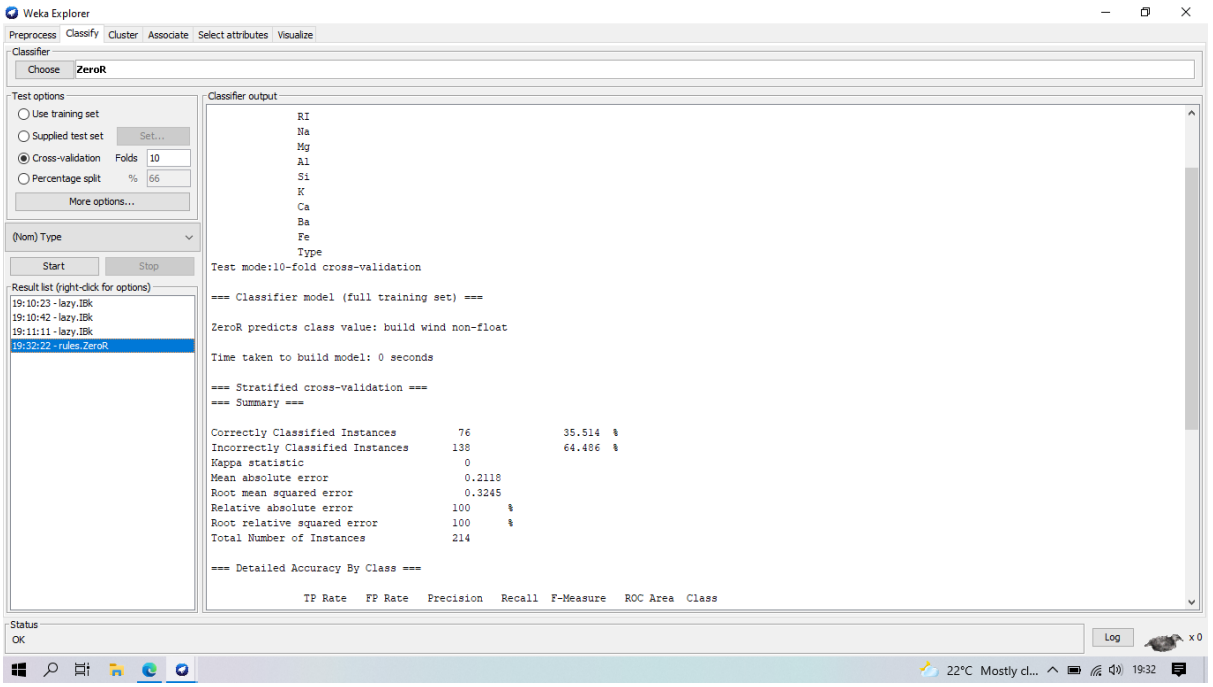


Ubah nilai KNN menjadi 100. (Hal ini berarti suatu instance baru yang ditambahkan, akan mencari 100 instance dengan jarak terdekat). Akurasi semakin jelek, hanya sekitar 34,6%,



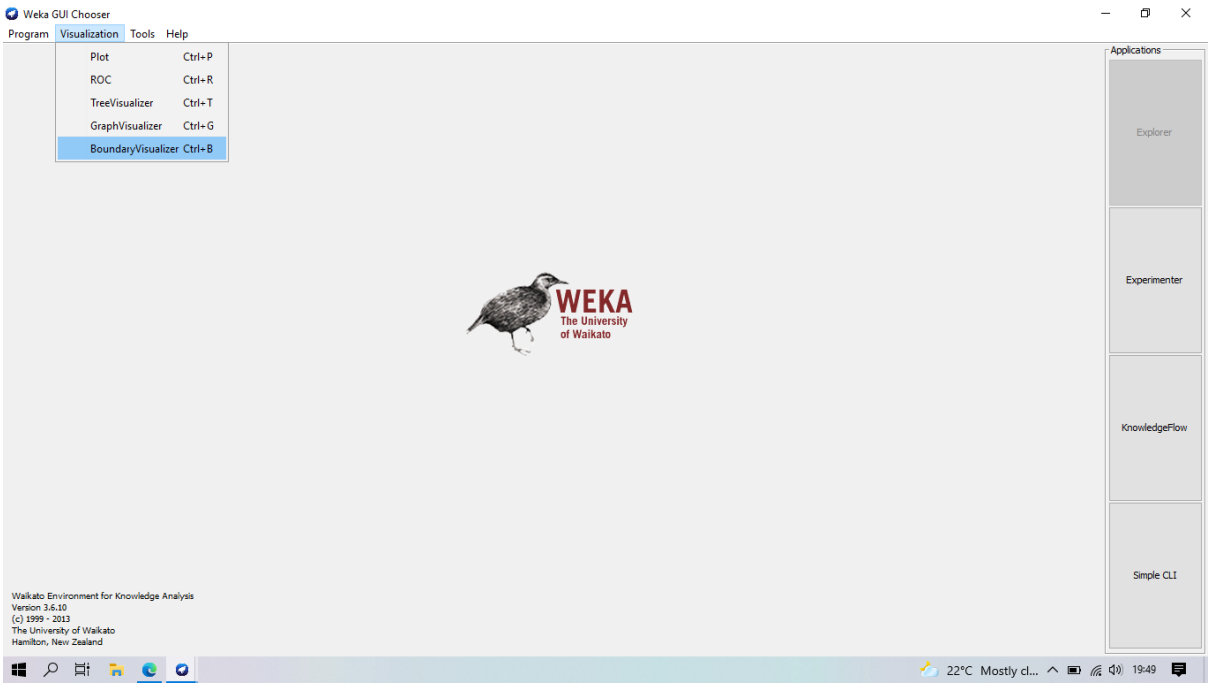


Setelah dianalisis, ketika nilai k = 100, nilai akurasinya mendekati baseline accuracy.

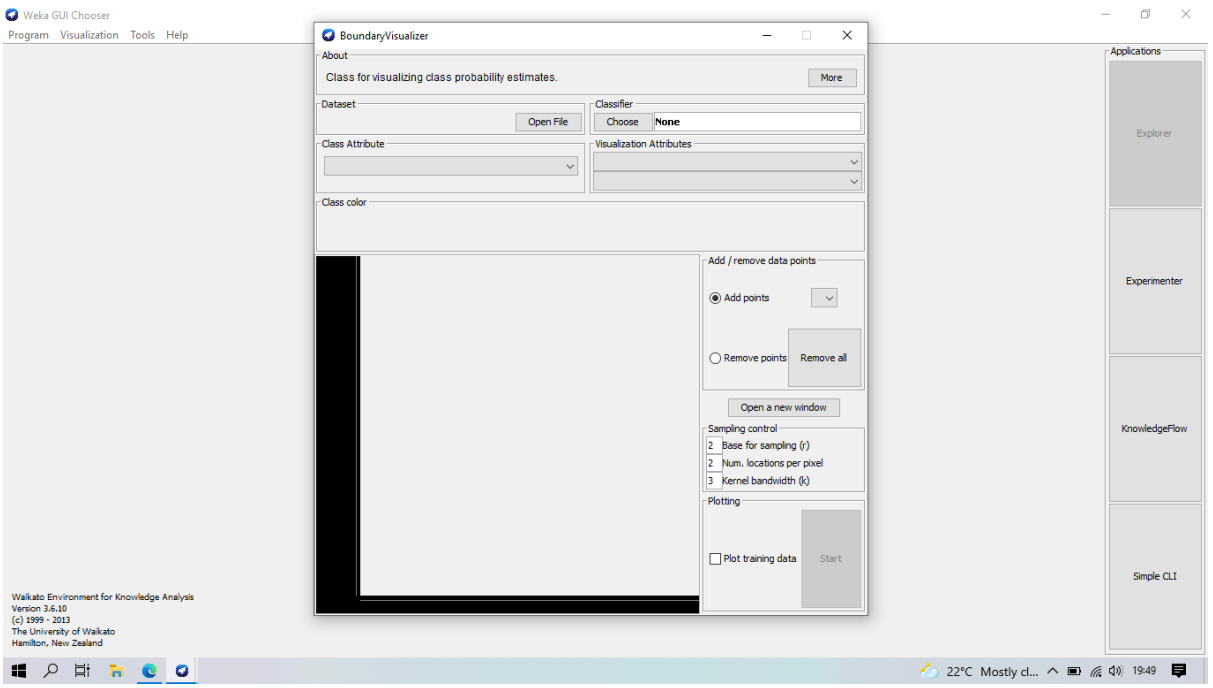


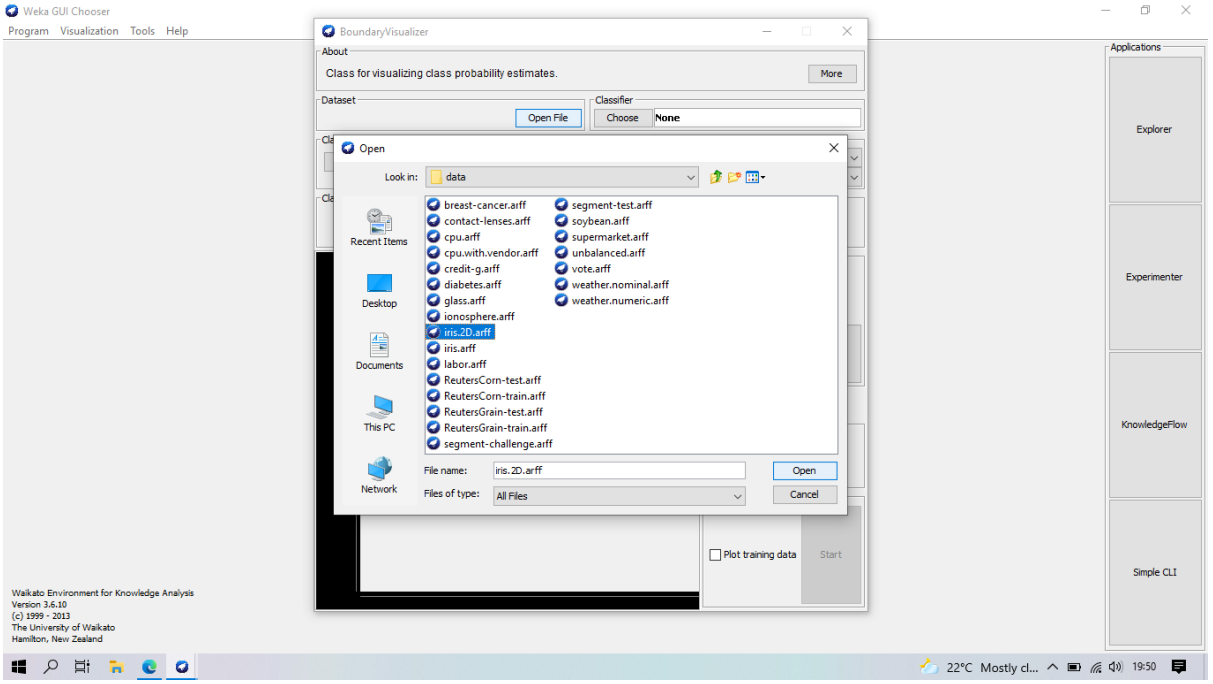
4.1 Classification Boundaries

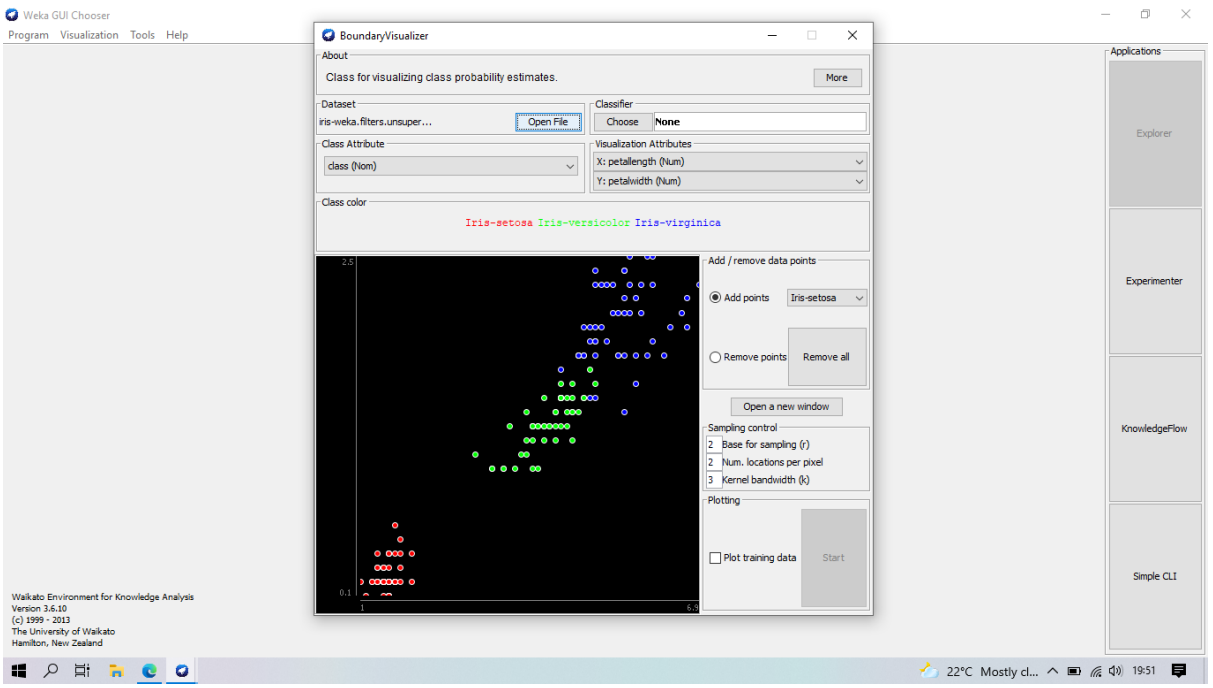
Buka Visualization > BoundaryVisualizer



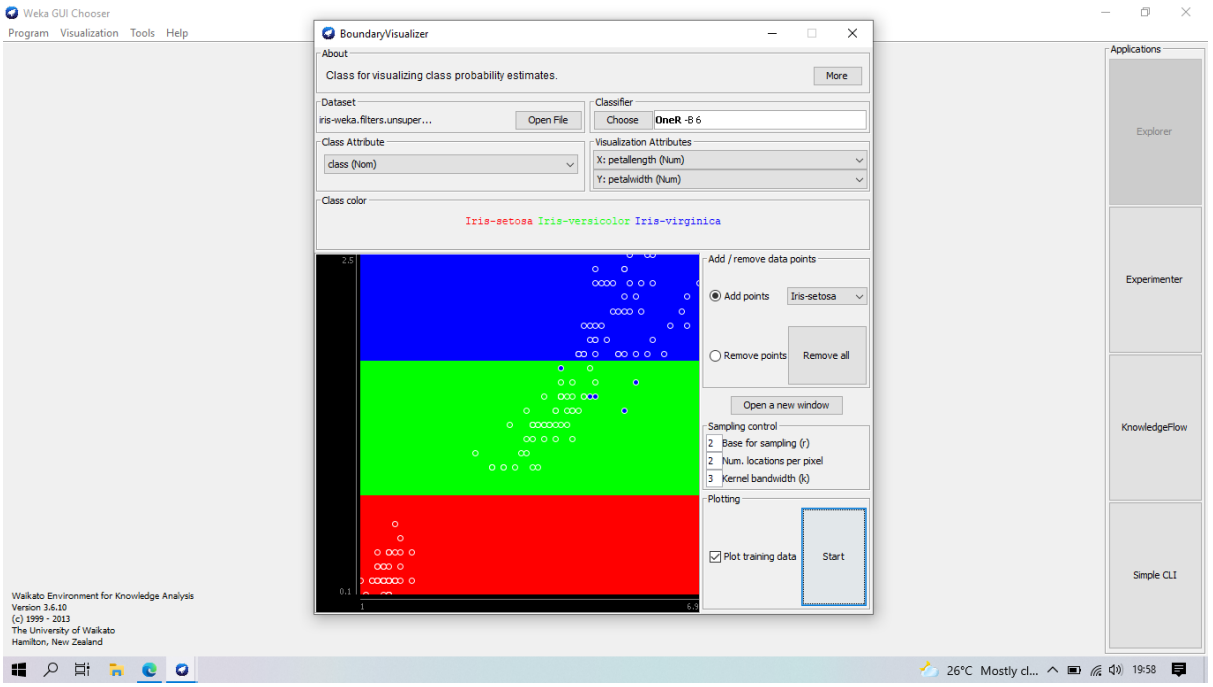
Buka dataset iris 2D.



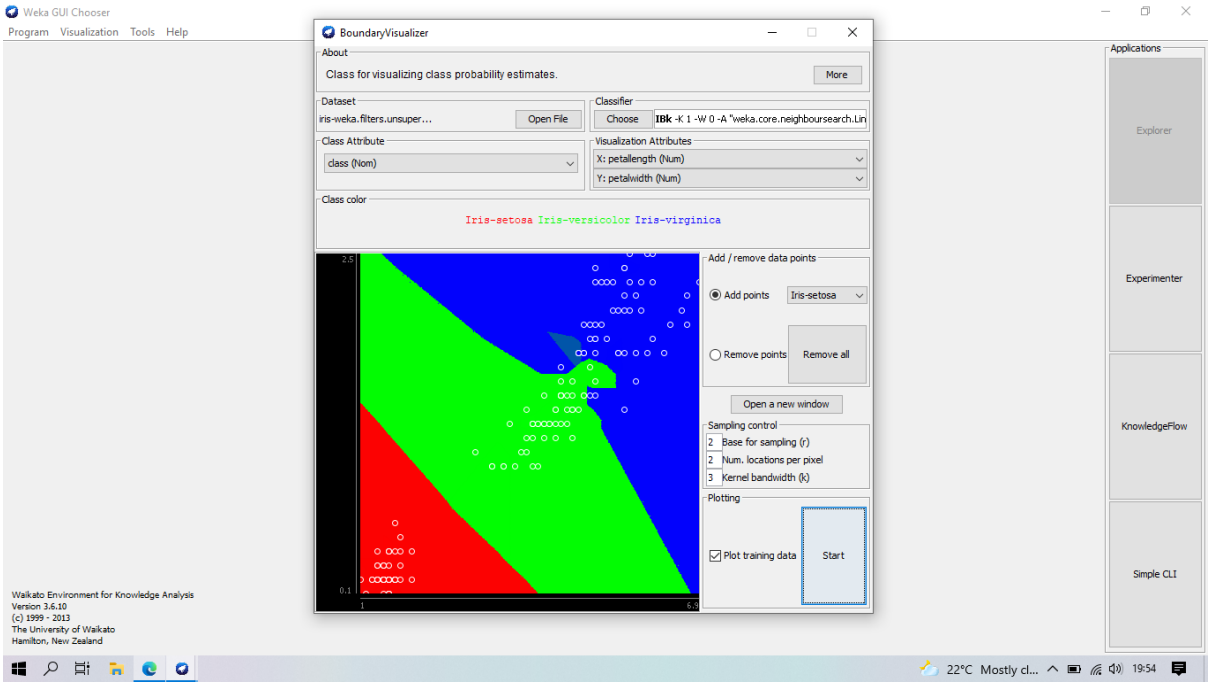

Weka GUI Chooser
Program Visualization Tools Help
BoundaryVisualizer
About
Class for visualizing class probability estimates.
More
Dataset
Open File
Classifier
Choose
None
Open
Look in:
data
breast-cancer.arff
contact-lenses.arff
cpu.arff
cpu.with.vendor.arff
credit-g.arff
diabetes.arff
glass.arff
ionosphere.arff
iris.2D.arff
iris.arff
labor.arff
ReutersCorn-test.arff
ReutersCorn-train.arff
ReutersGrain-test.arff
ReutersGrain-train.arff
segment-challenge.arff
segment-test.arff
soybean.arff
supermarket.arff
unbalanced.arff
vote.arff
weather.nominal.arff
weather.numeric.arff
Recent Items
Desktop
Documents
This PC
Network
File name:
iris.2D.arff
Files of type:
All Files
Open
Cancel
Plot training data
Start
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand
22°C Mostly cl...
19:50


Weka GUI Chooser
Program Visualization Tools Help
BoundaryVisualizer
About
Class for visualizing class probability estimates.
More
Dataset
iris-weka.filters.unsuper...
Open File
Classifier
Choose
None
Class Attribute
class (Nom)
Visualization Attributes
X: petallength (Num)
Y: petalwidth (Num)
Class color
Iris-setosa Iris-versicolor Iris-virginica
Add / remove data points
Add points
Iris-setosa
Remove points
Remove all
Open a new window
Sampling control
2 Base for sampling (r)
2 Num. locations per pixel
3 Kernel bandwidth (k)
Plotting
Plot training data
Start
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand
22°C Mostly cl...
19:51

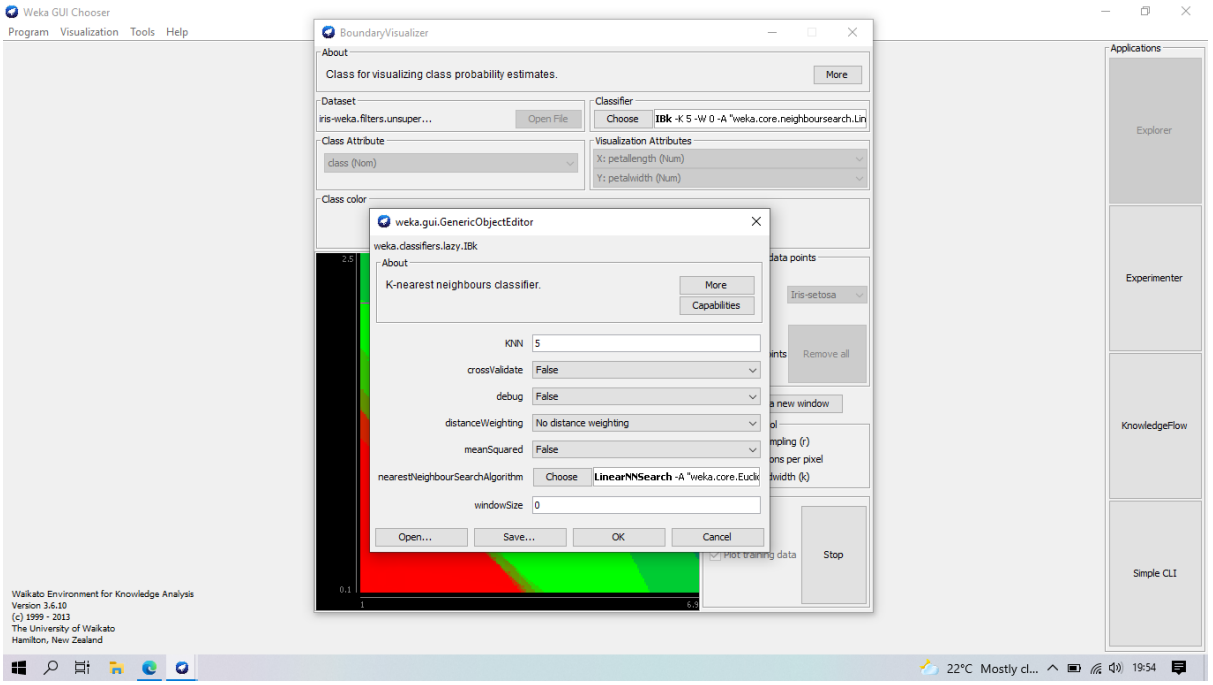
Gunakan classifier OneR. Berikut hasil visualisasinya:

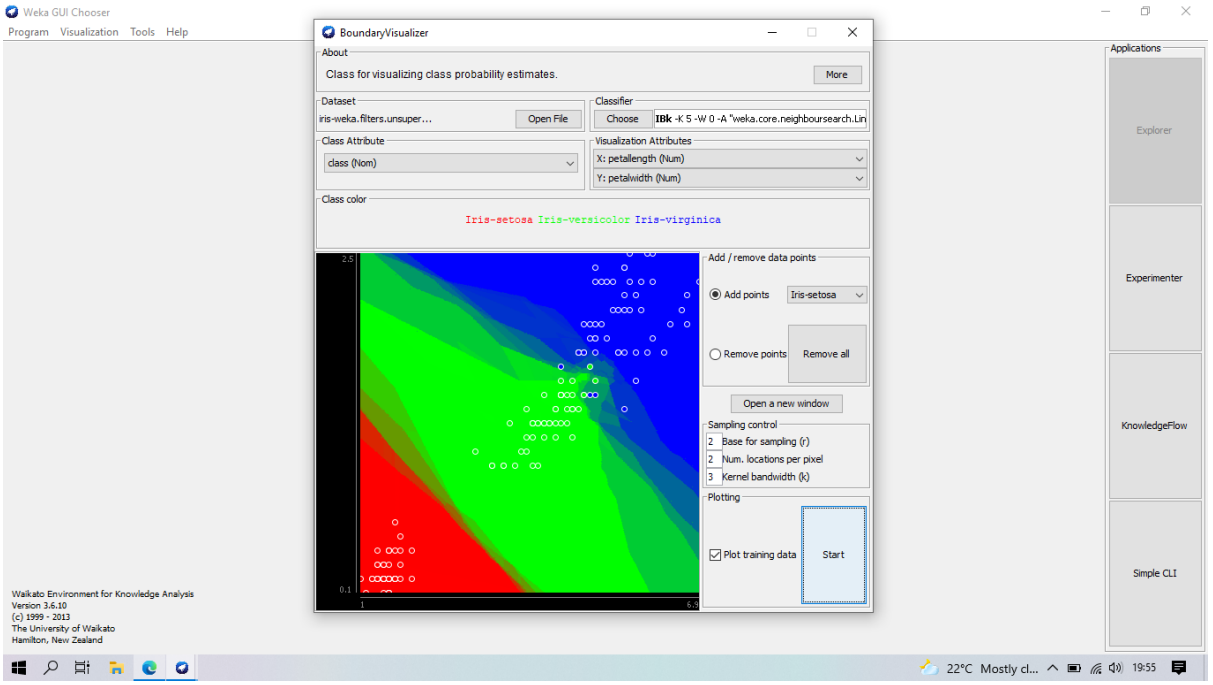


Gunakan classifier IBk. Berikut hasil visualisasinya:

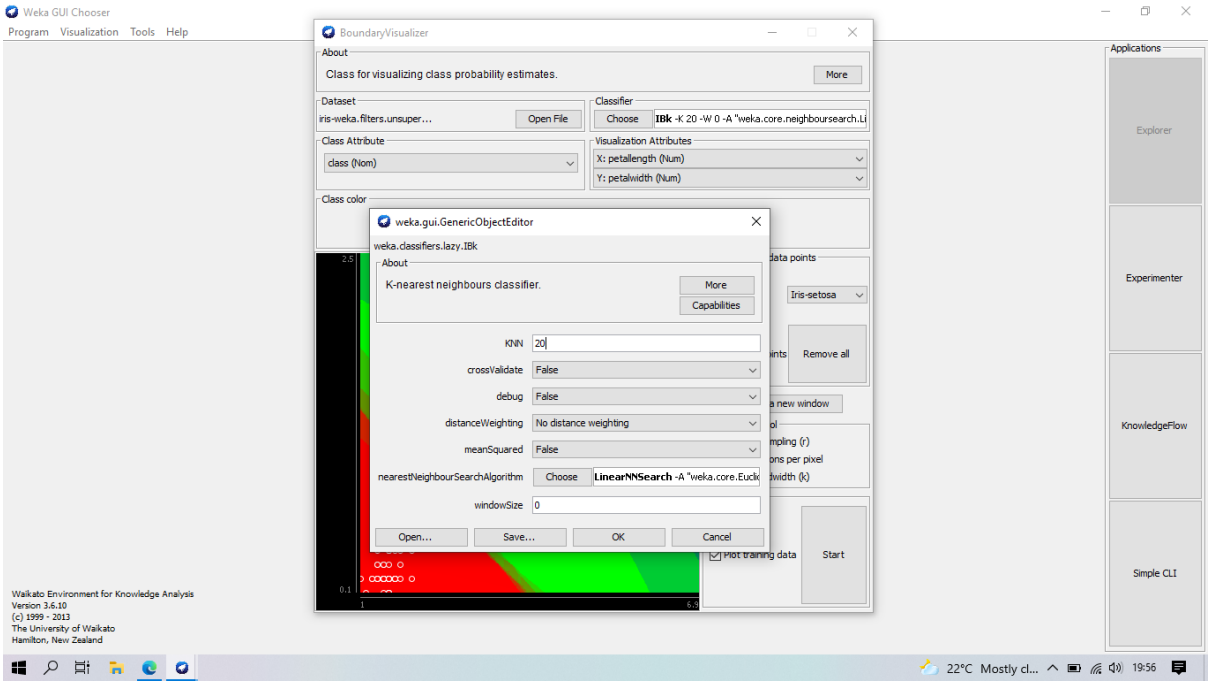


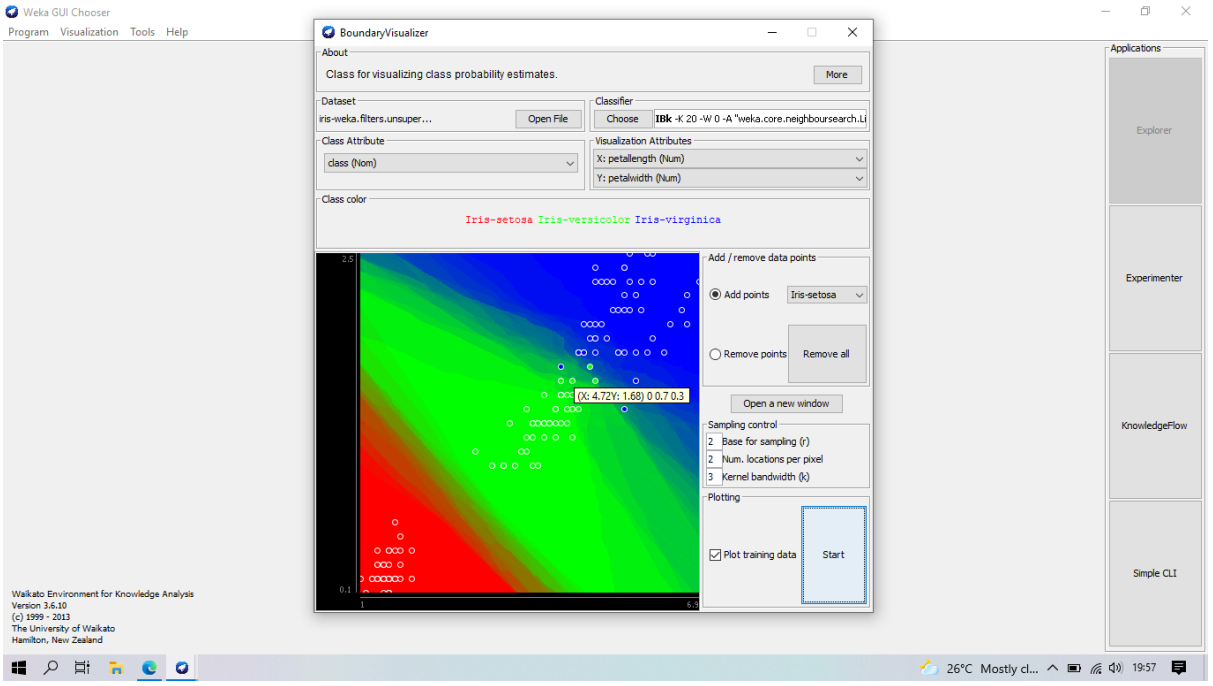
Ubah nilai KNN menjadi 5. Setelah nilai KNN berubah menjadi 5, ada daerah yang transisi antara merah dan hijau, atau warna lainnya. Berikut hasil visualisasinya:



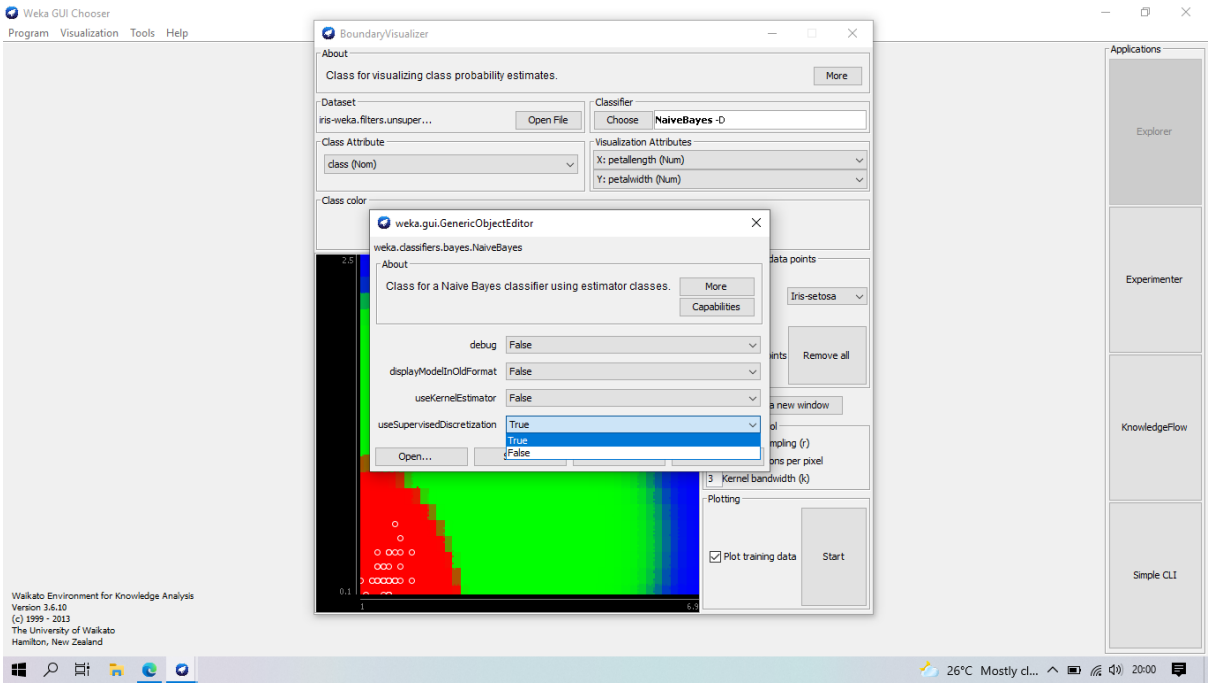


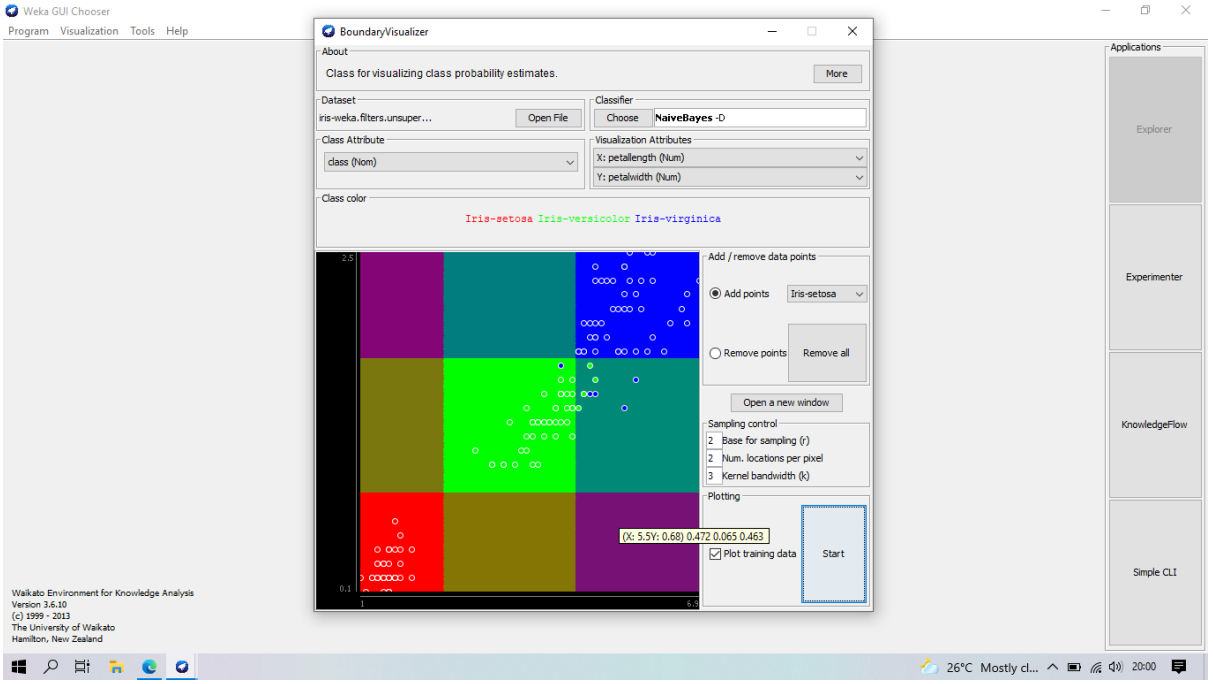
Ubah nilai KNN menjadi 20. Setelah nilai KNN berubah menjadi 20, daerah yang transisi antara merah dan hijau, atau warna lainnya menjadi lebih lebar. Berikut hasil visualisasinya:



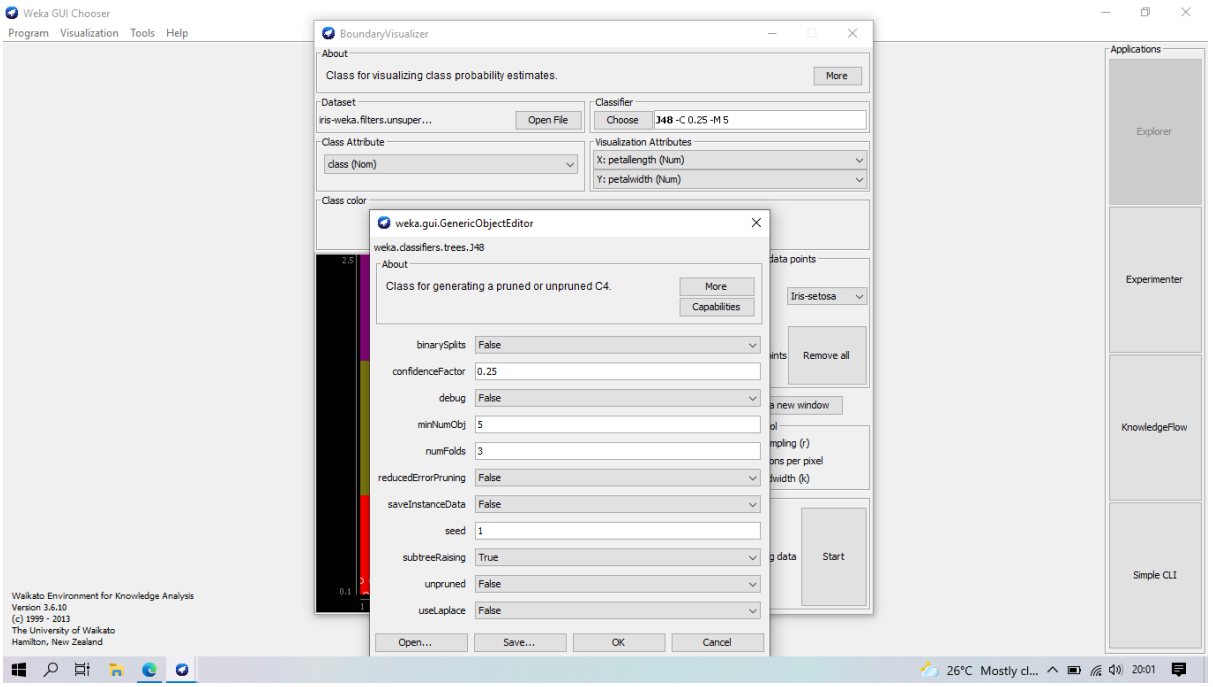


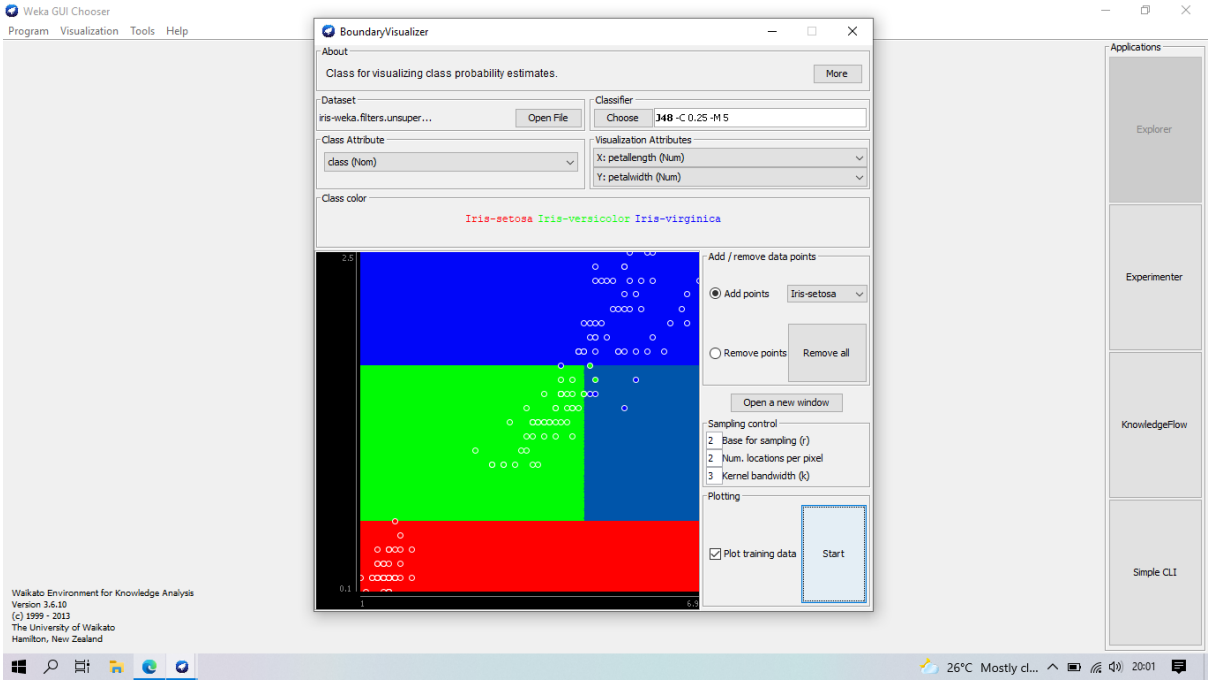
Gunakan Classifier NaiveBayes. Berikut hasil visualisasinya:



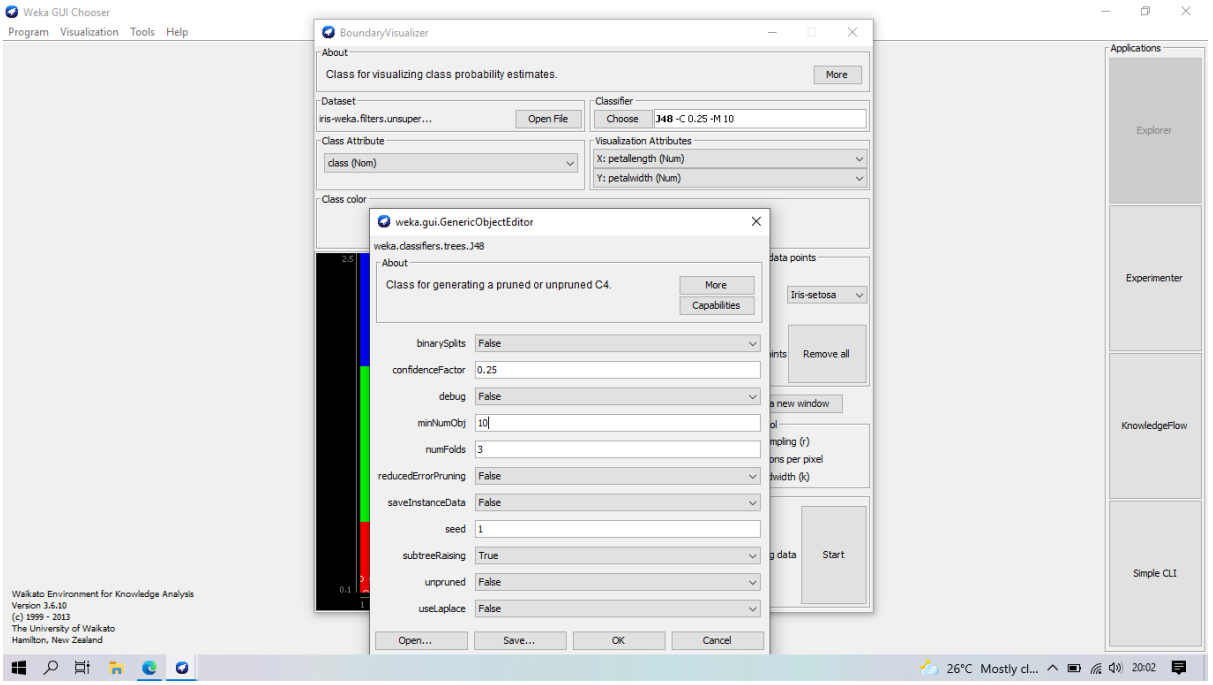


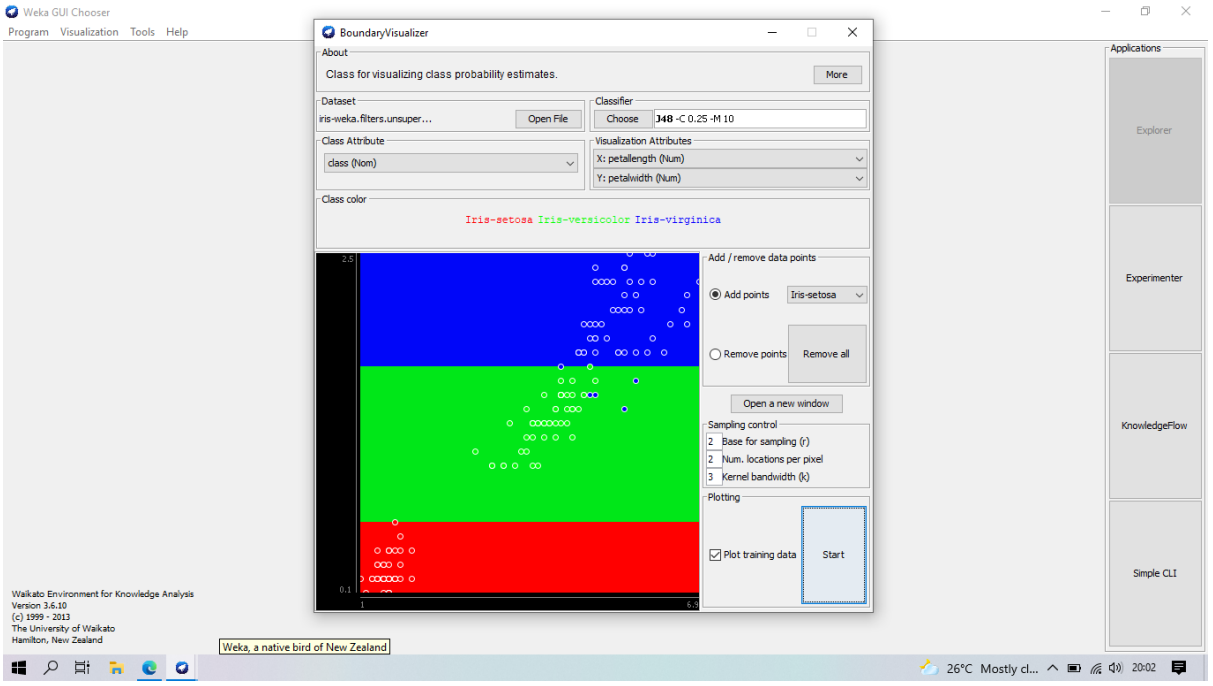
Gunakan Classifier J48. Berikut hasil visualisasinya:





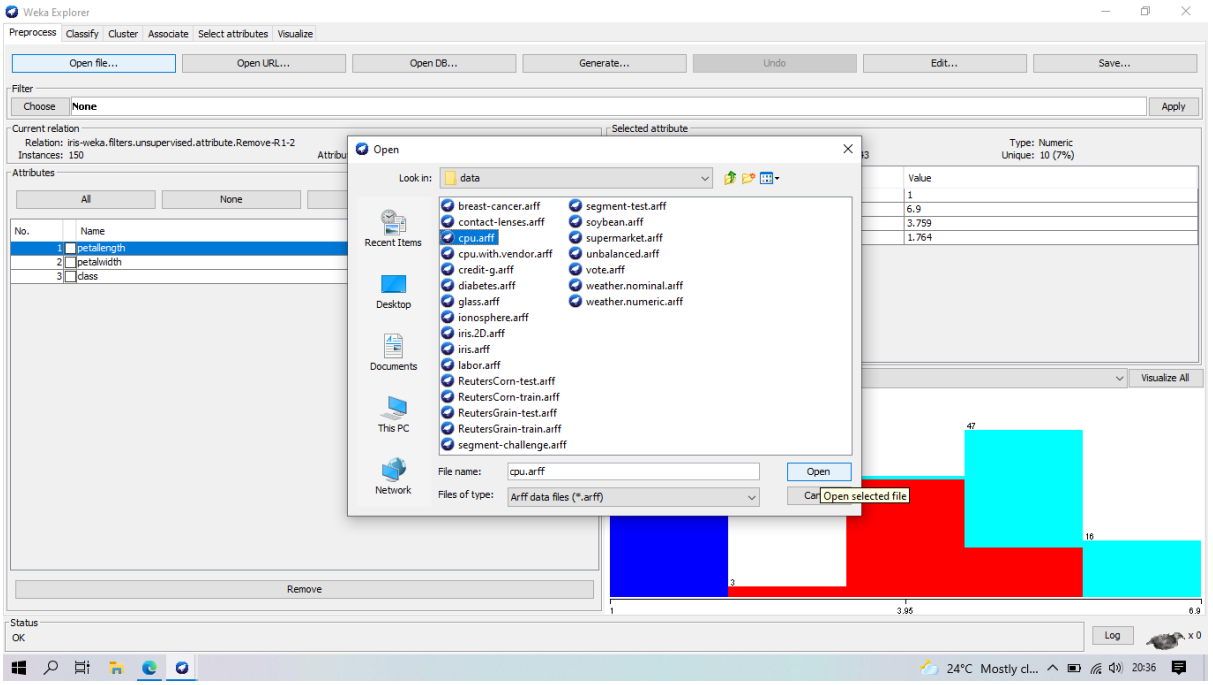
Klik classifier J48 untuk melakukan konfigurasi. Atur minNumObj menjadi 10. Berikut hasil visualisasinya:



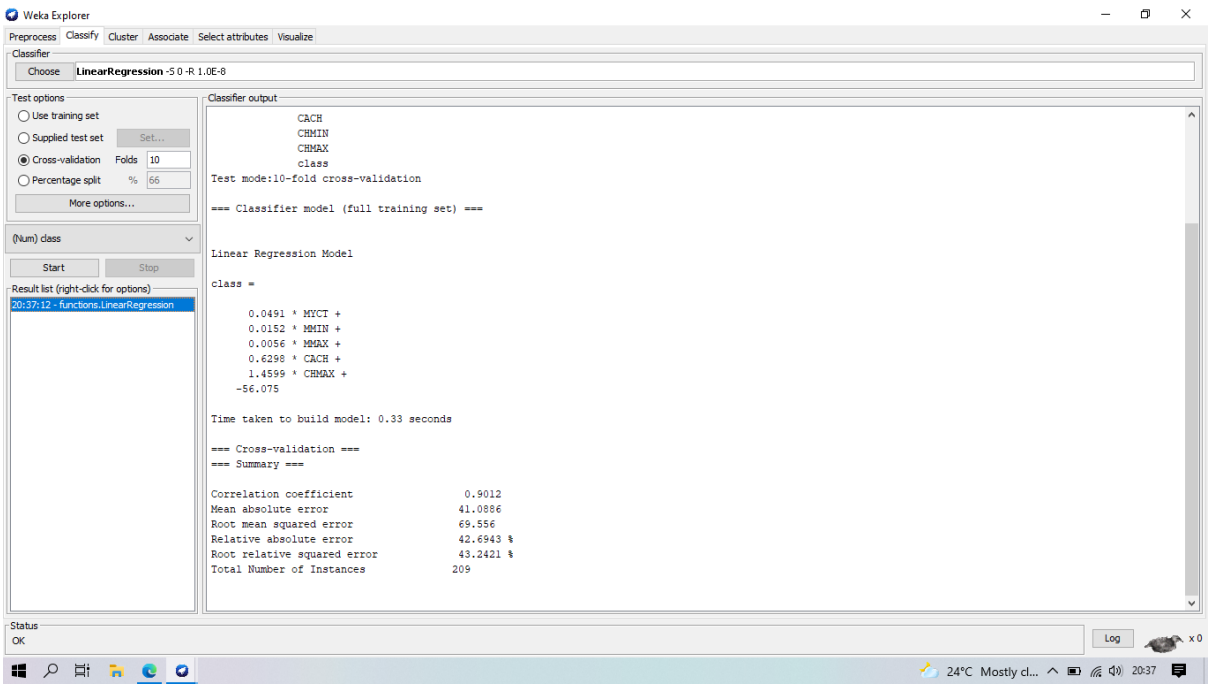


4.2 Linear Regression

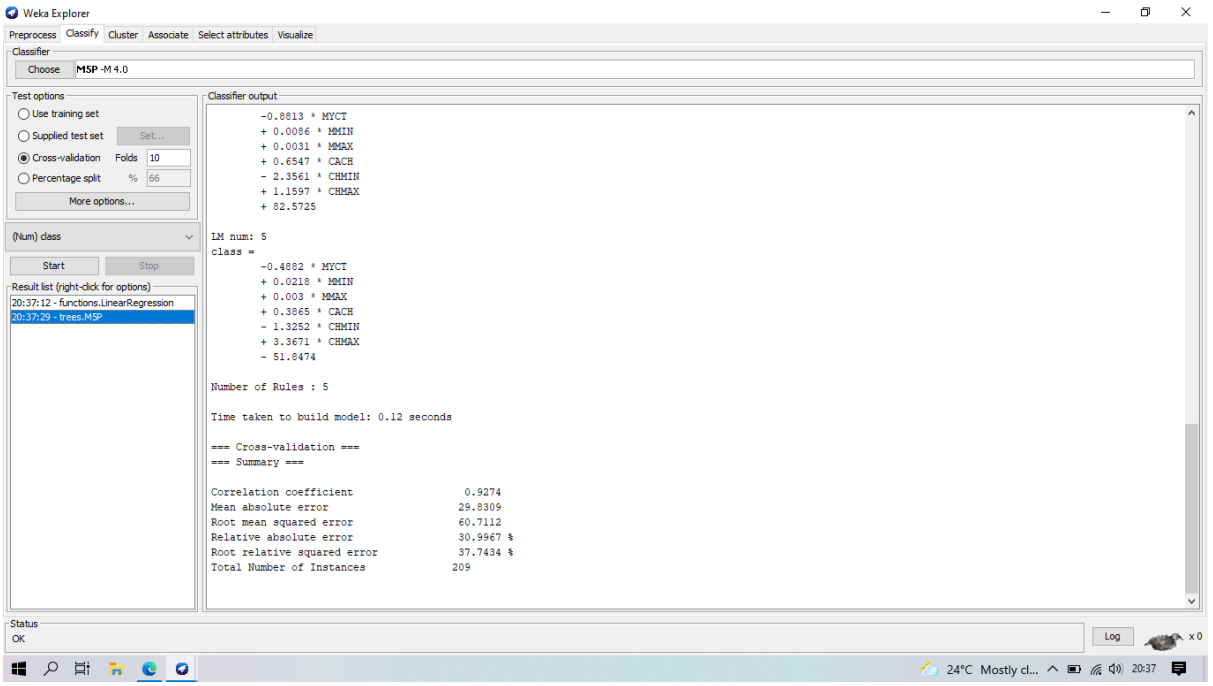
Buka dataset CPU.



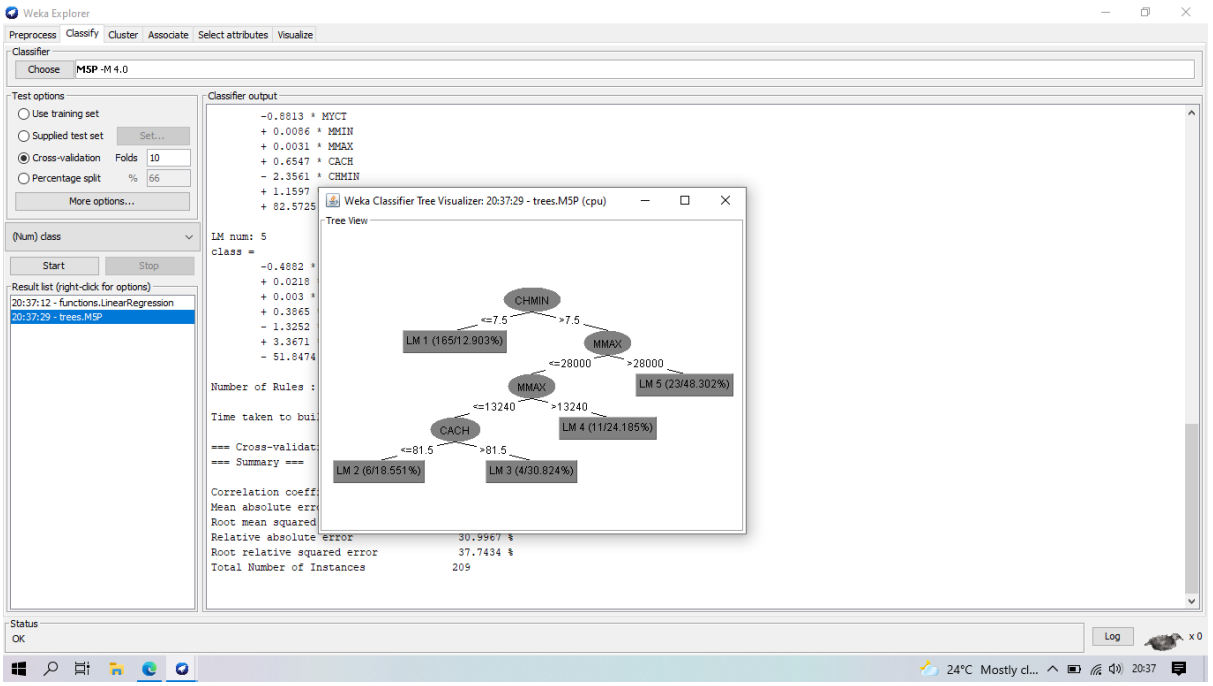
Gunakan Classifier Linear Regression. Berikut rincian model yang dihasilkan.



Berikut adalah classifier M5P, merupakan bentuk pohon dari model regresi. Berikut hasil dari penggunaan classifier M5P.

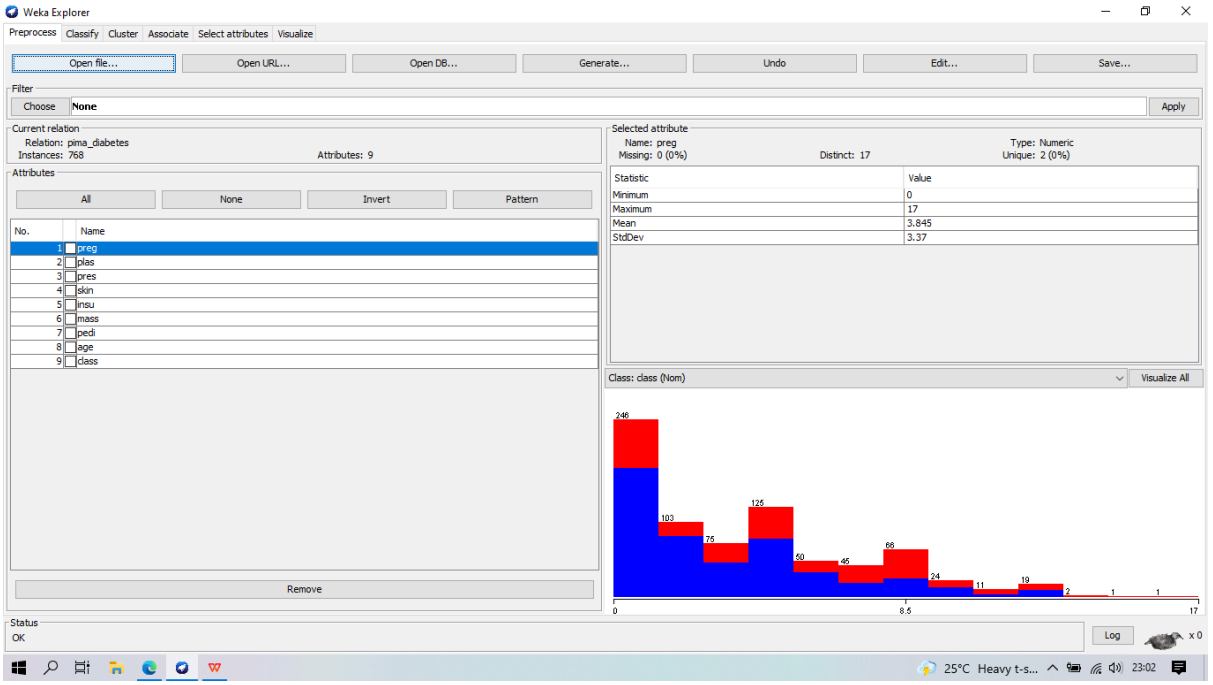


Berikut adalah visualisasi pohonnya.

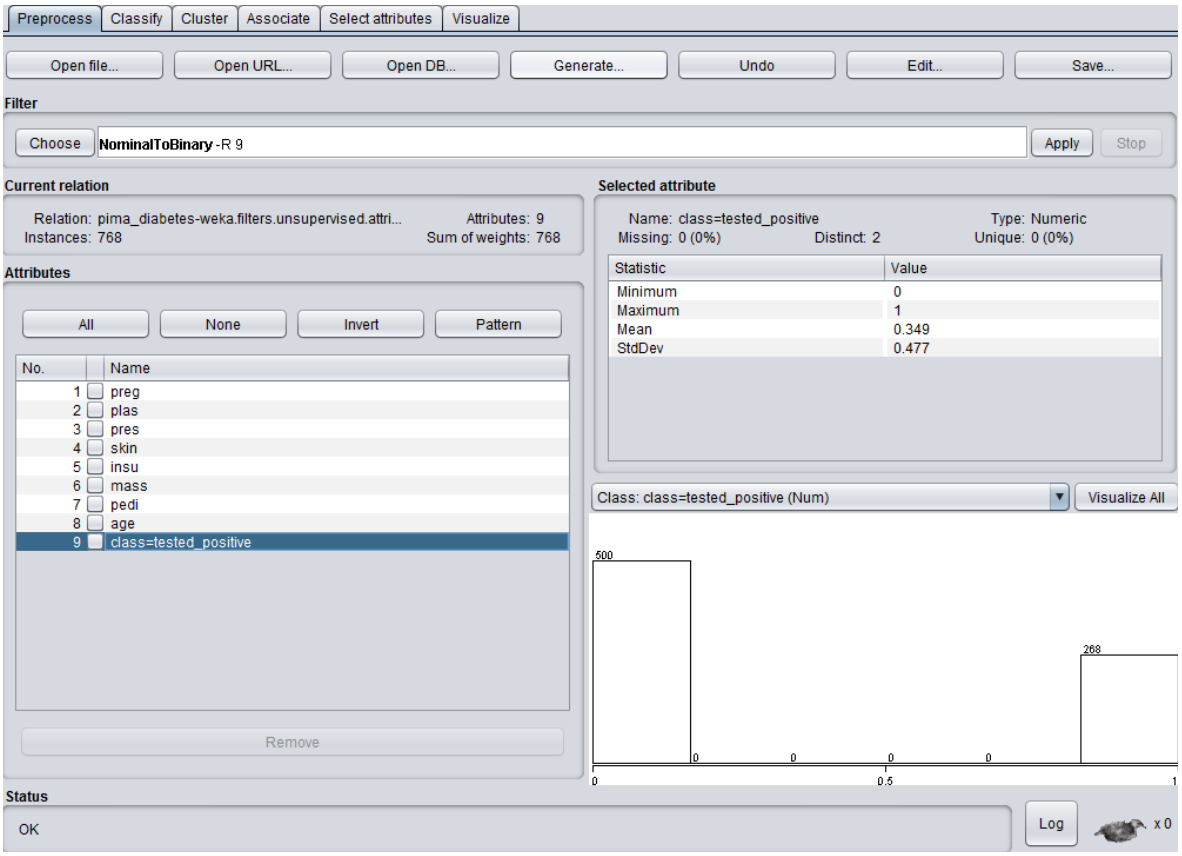


4.3 Classification by Regression

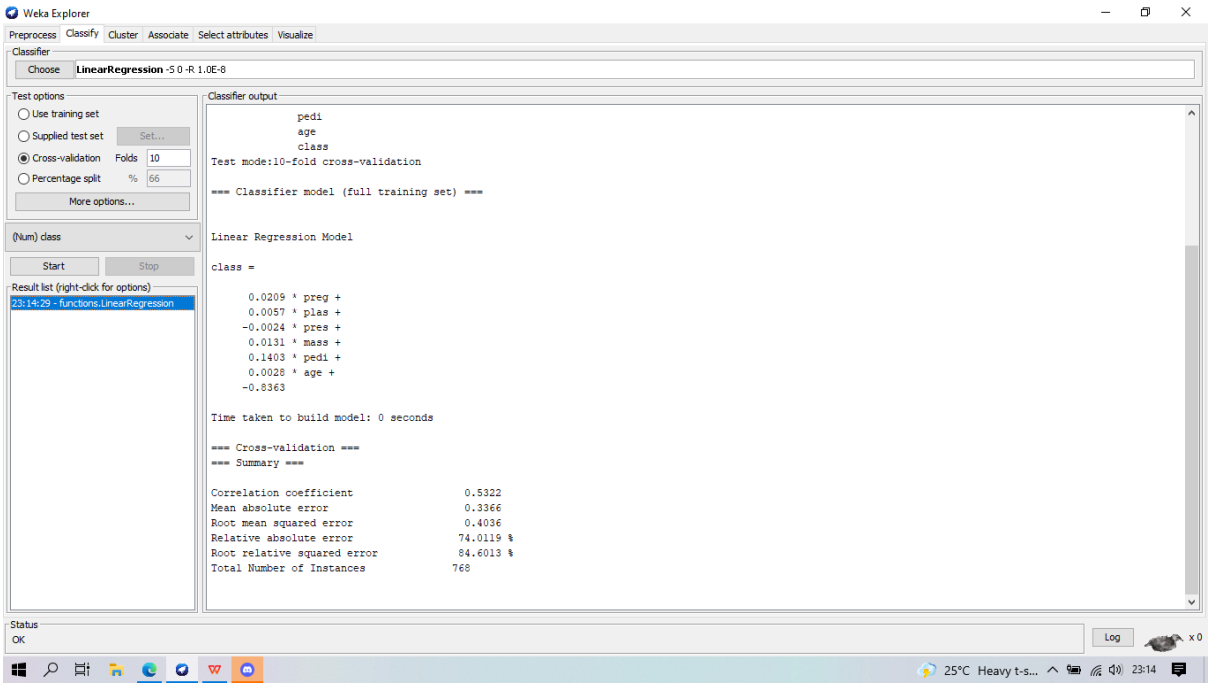
Buka dataset diabetes



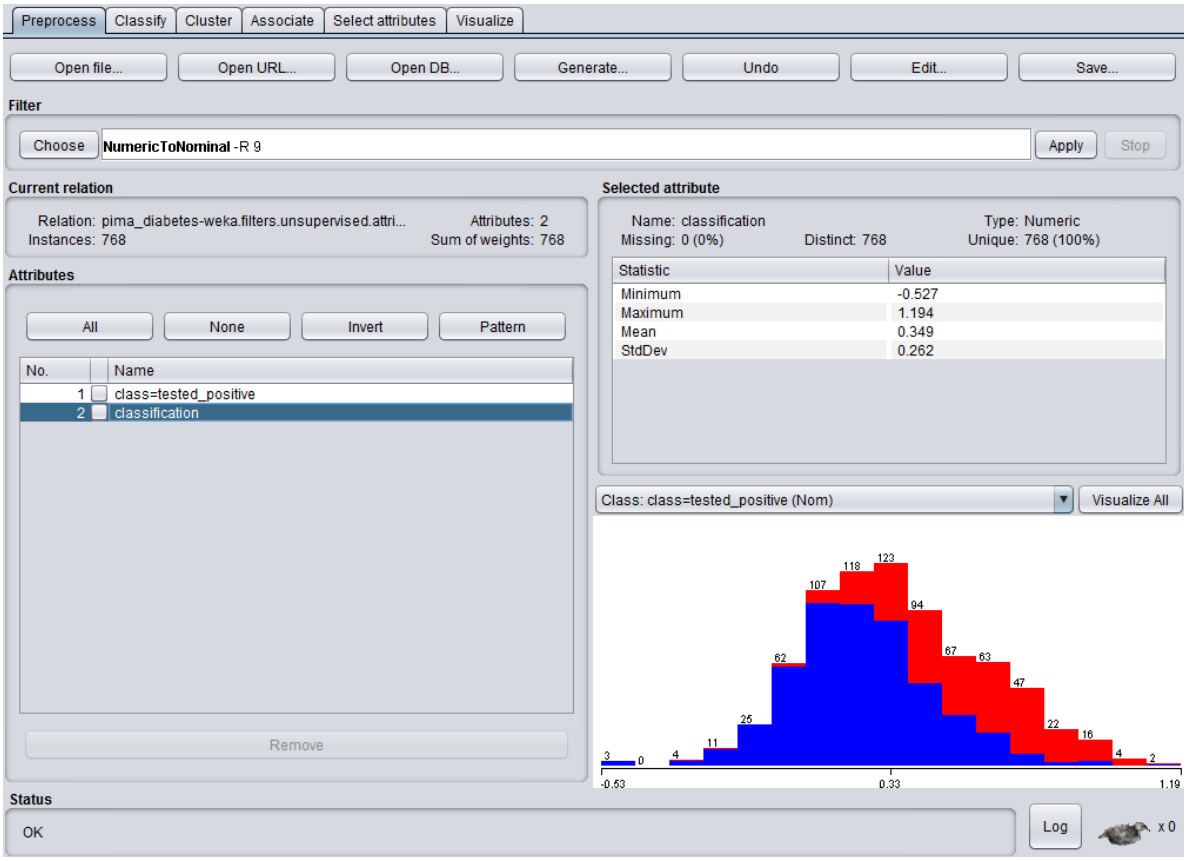
Ubah atribut kelas menjadi binary dengan filter NominalToBinary.



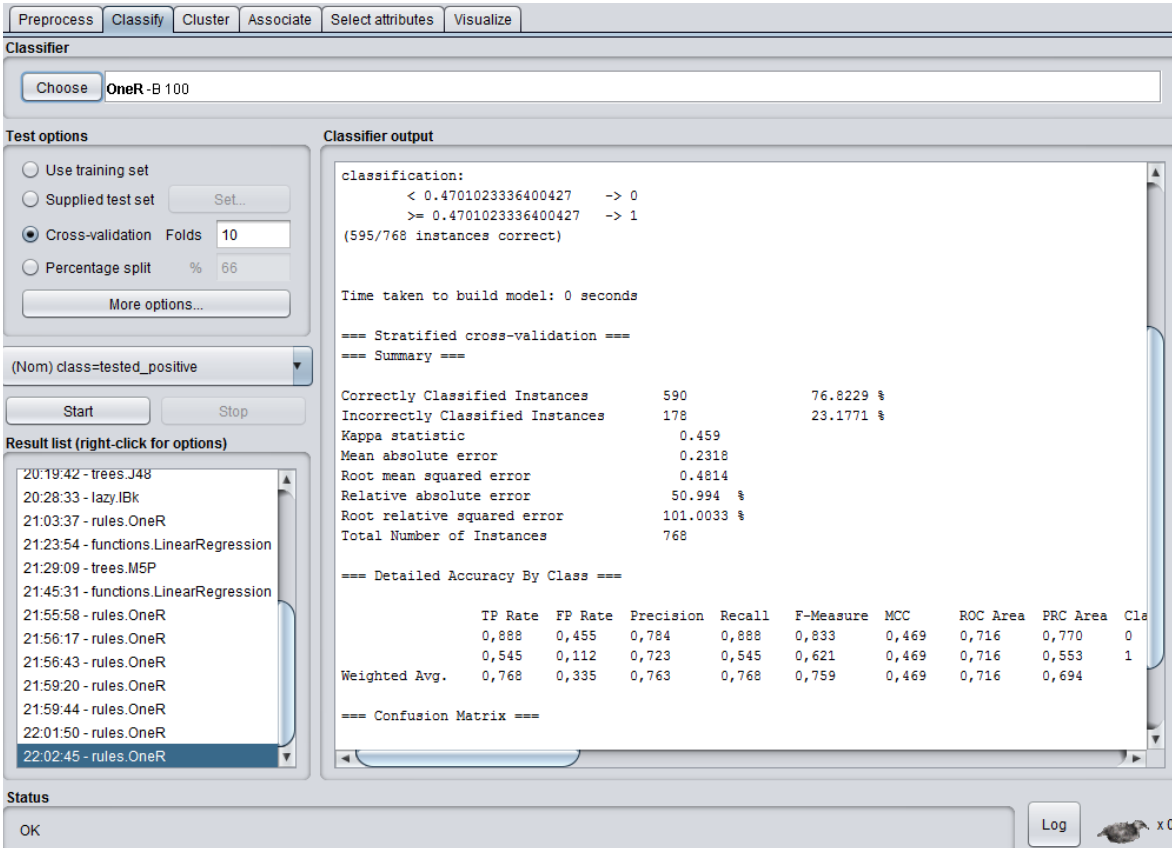
Lakukan Classifier menggunakan Linear Regression



Modifikasi dataset dengan ditambah atribut classification yang berasal dari filer addClassification dengan classifier Linear Regression dan parameter outputClassification=True, pengembalian atribut class menjadi nominal, dan menghapus attribut selain kedua atribut tersebut.

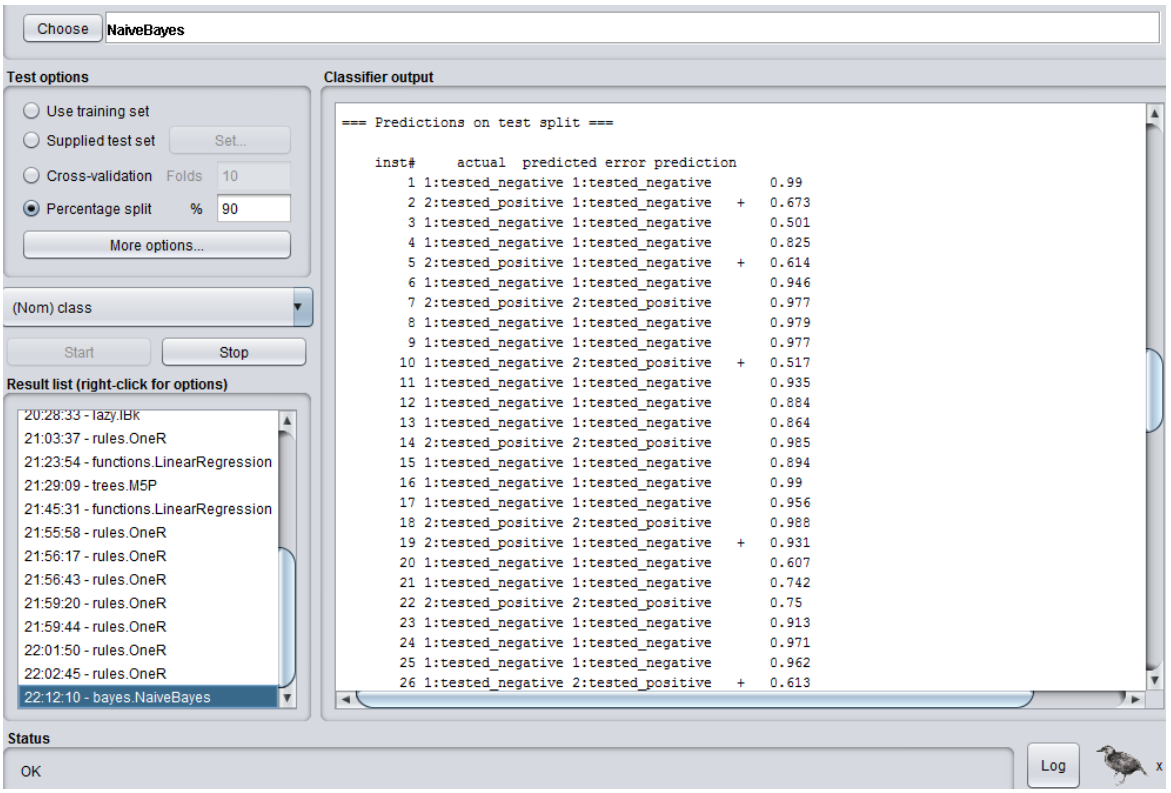


Lakukan classifier OneR dengan parameter minBucketSize=100.

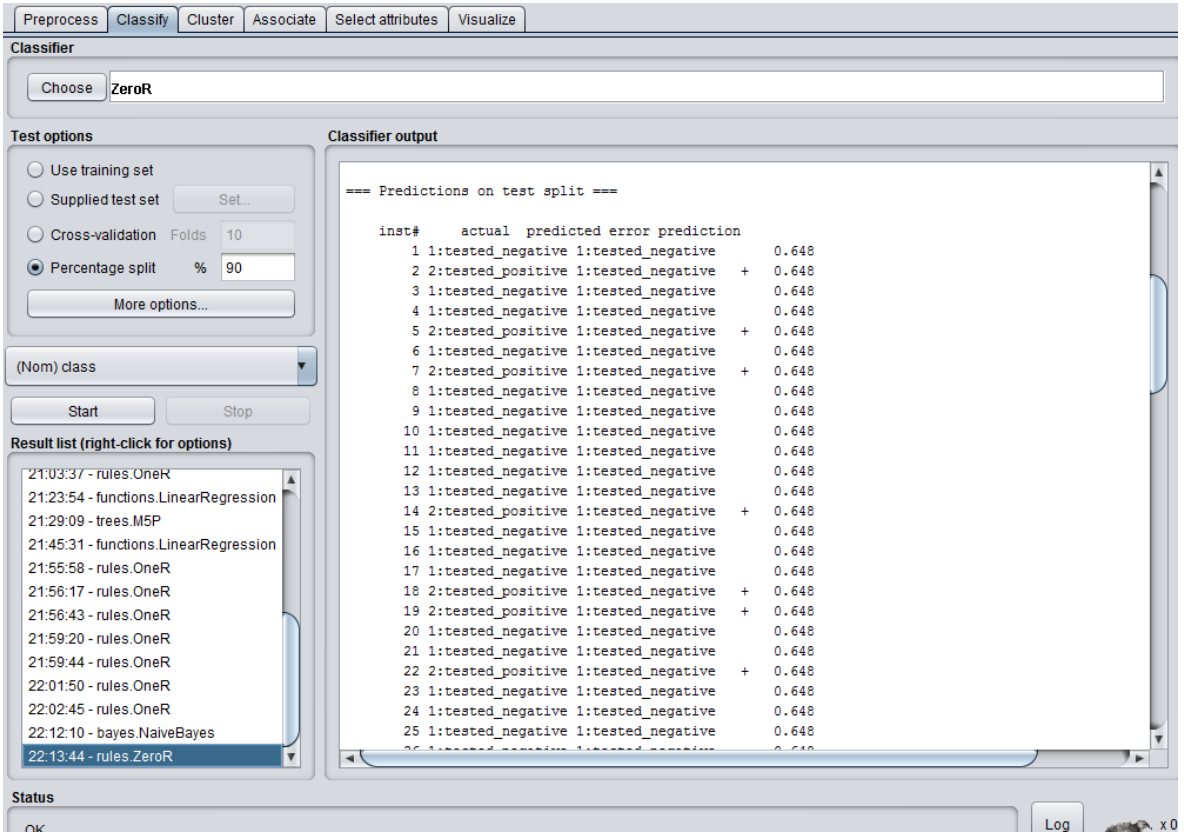


4.4 Logistic Regression

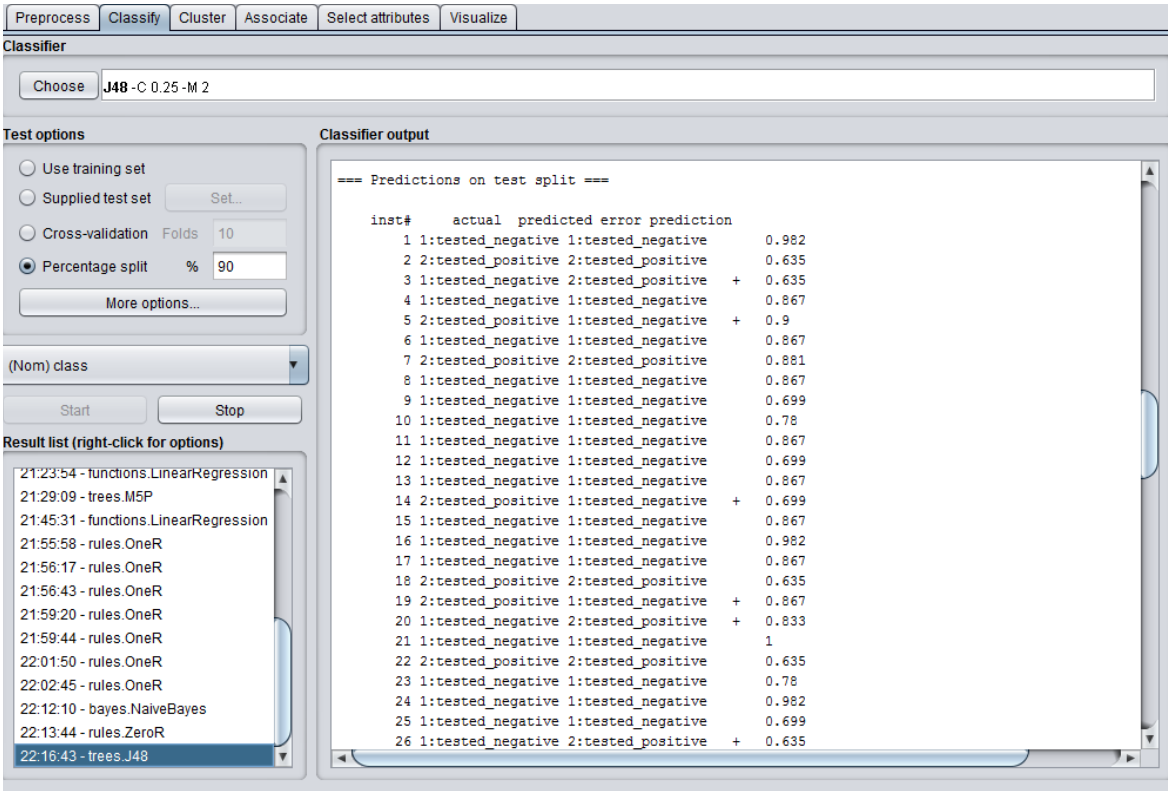
Buka Dataset Diabetes. Kemudian gunakan classifier NaiveBayes dengan metode 90% percentage split.



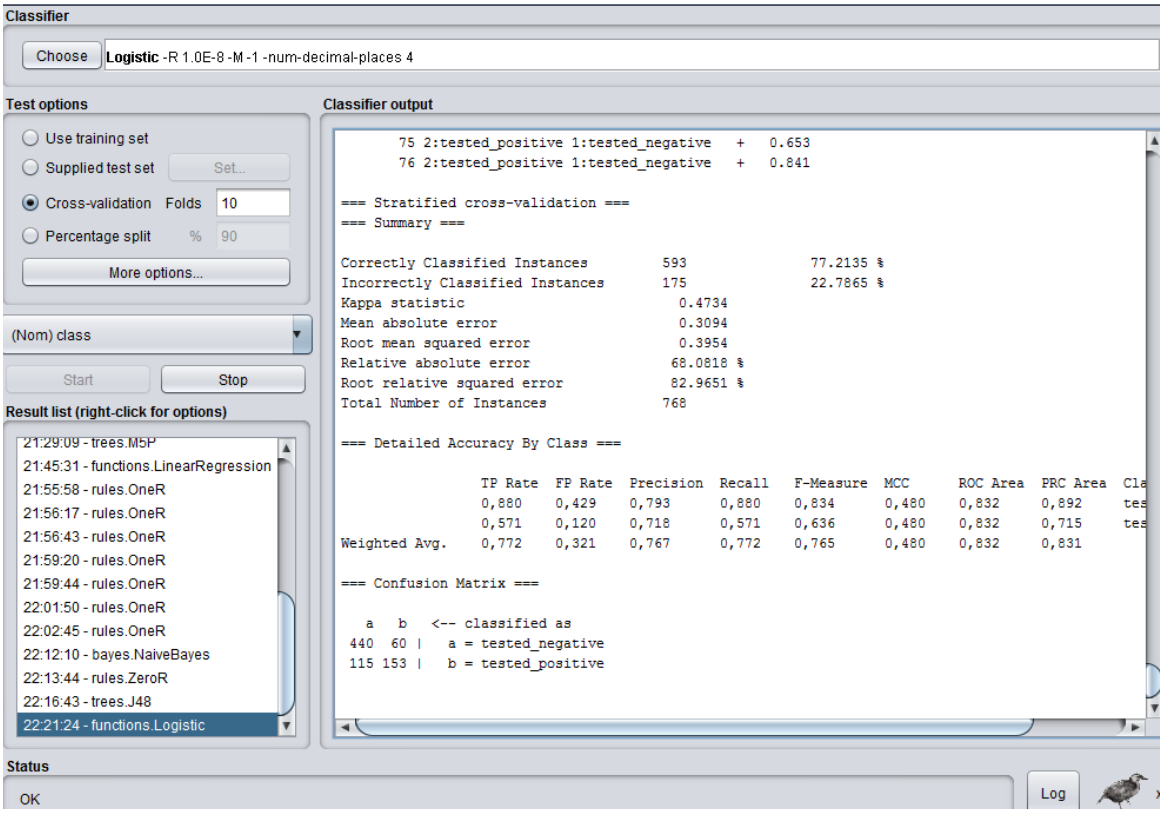
Gunakan classifier ZeroR dengan metode 90% percentage split.



Gunakan classifier J48 dengan metode 90% percentage split.

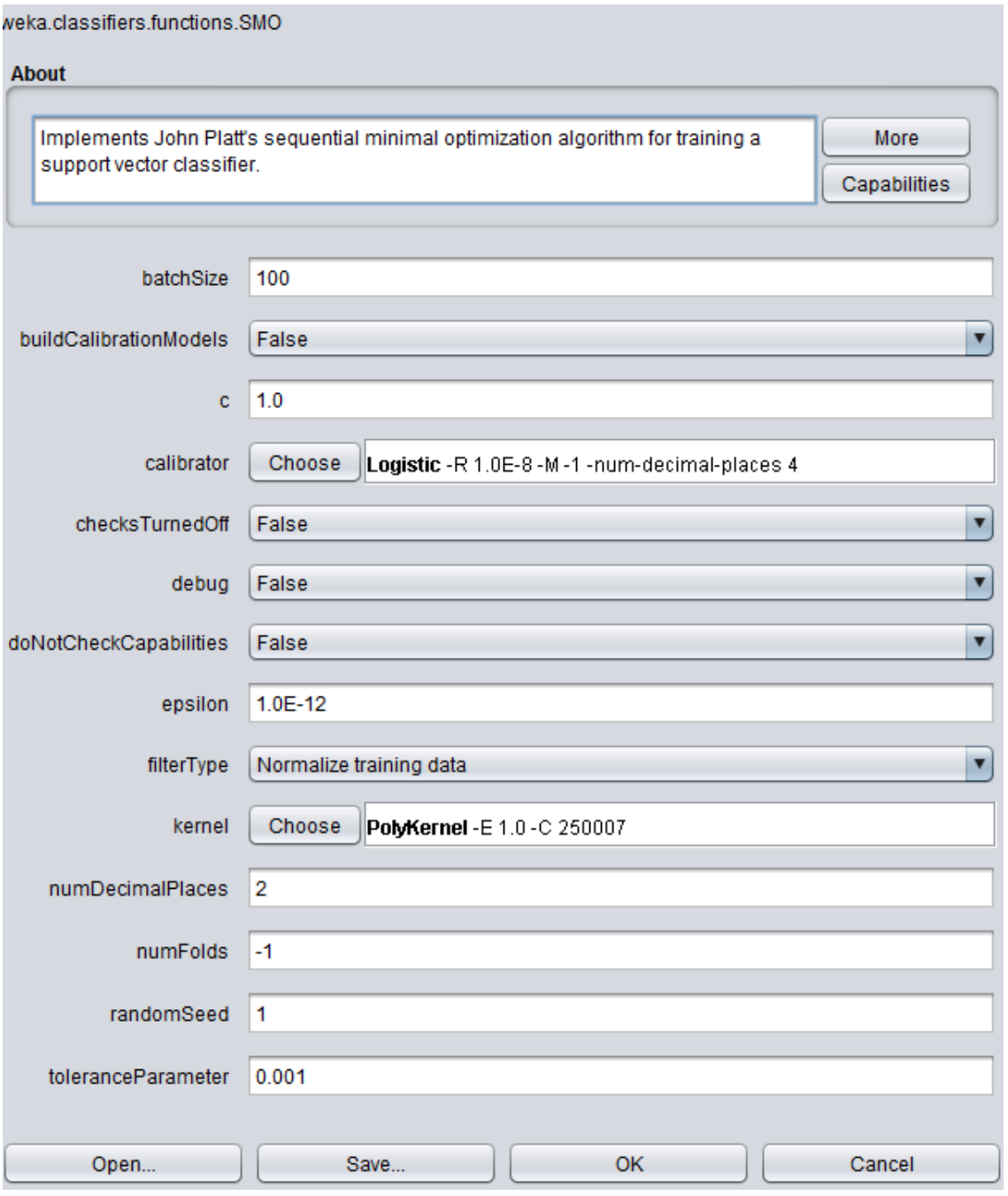


Gunakan classifier Logistic dengan metode 10 fold cross-validation



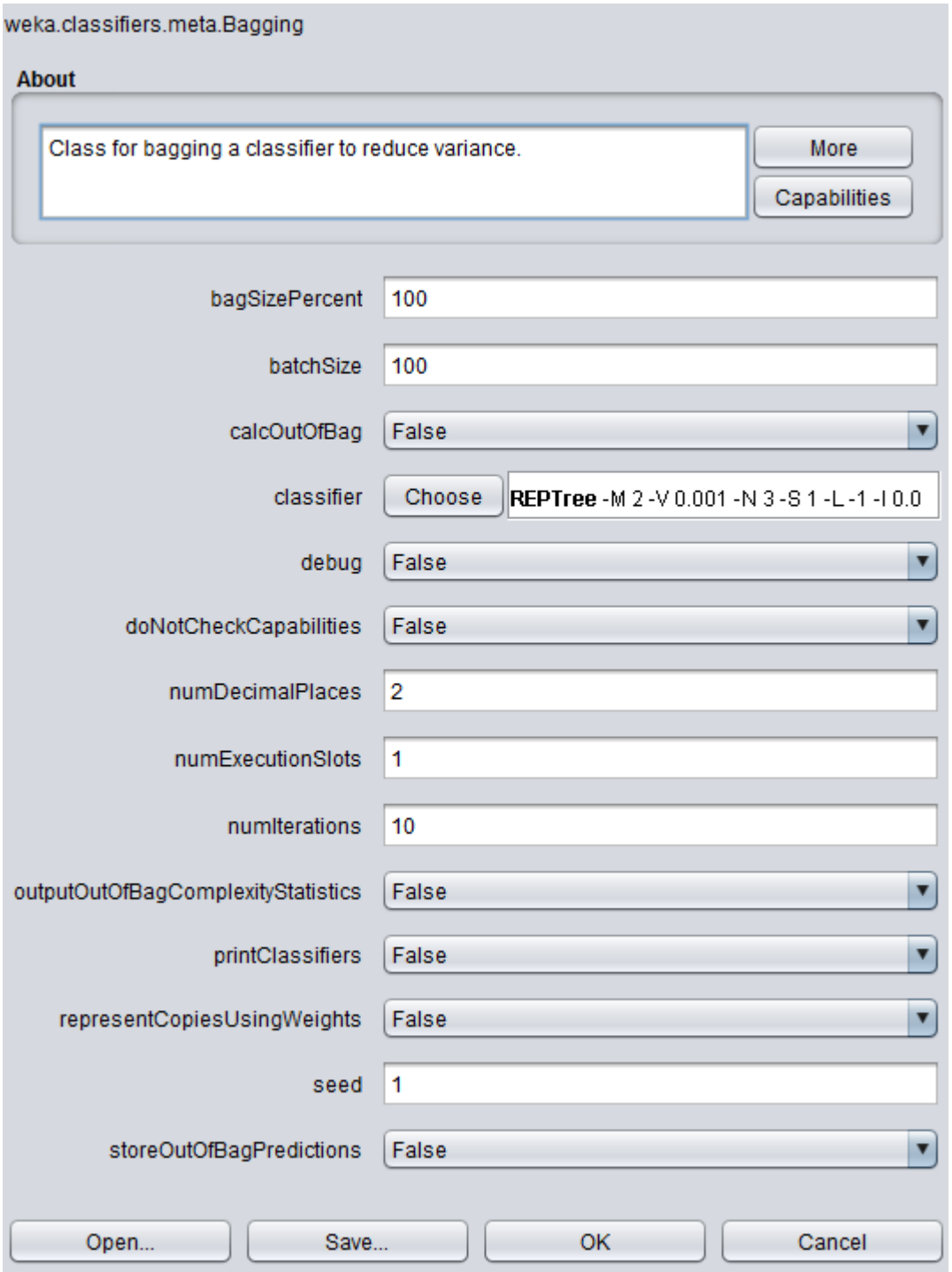
4.5 Support Vector Machines

Gunakan parameter pada SMO (Sequential Minimal Optimization) untuk mengimplementasi support vector machines.

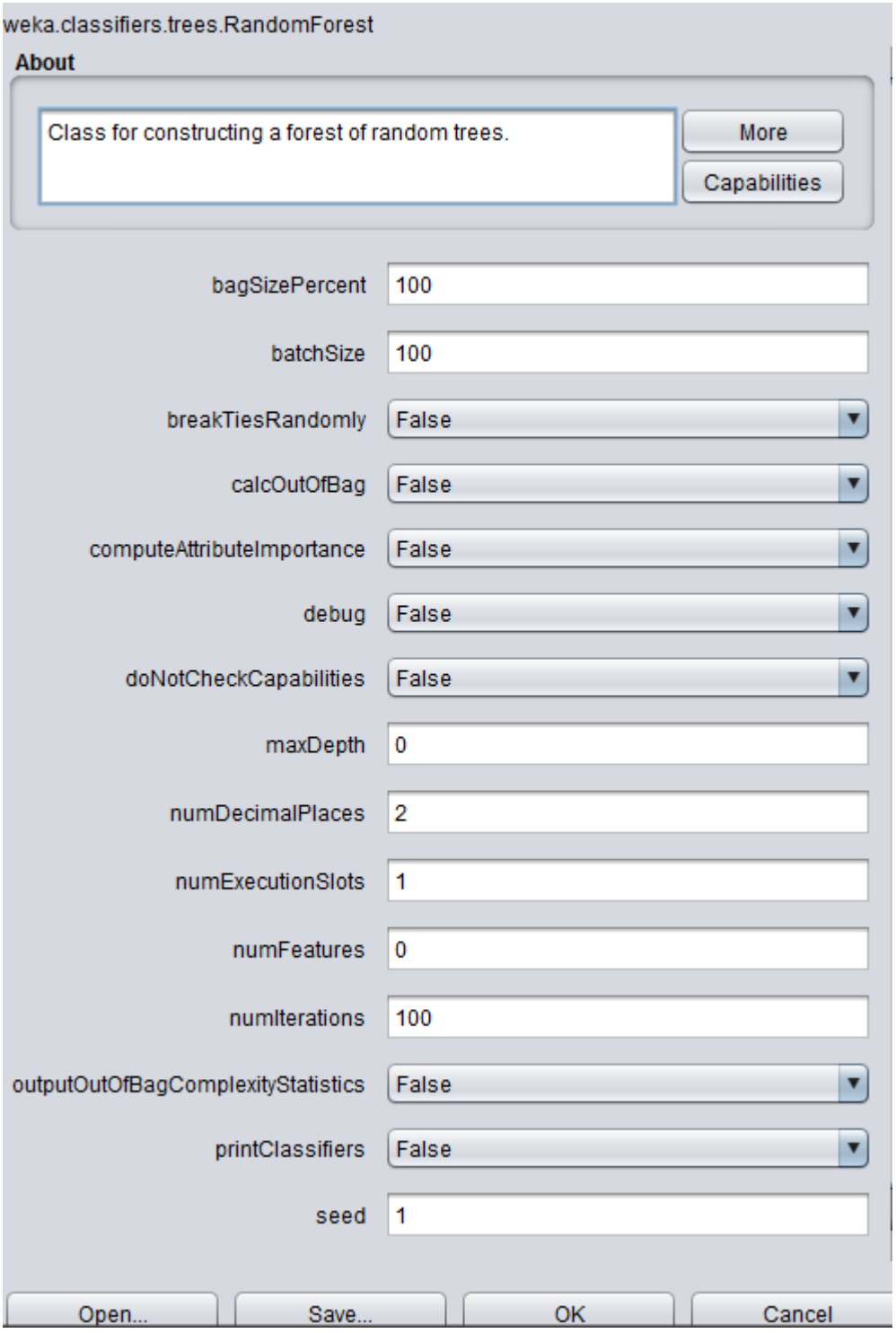


4.6 Ensemble Learning

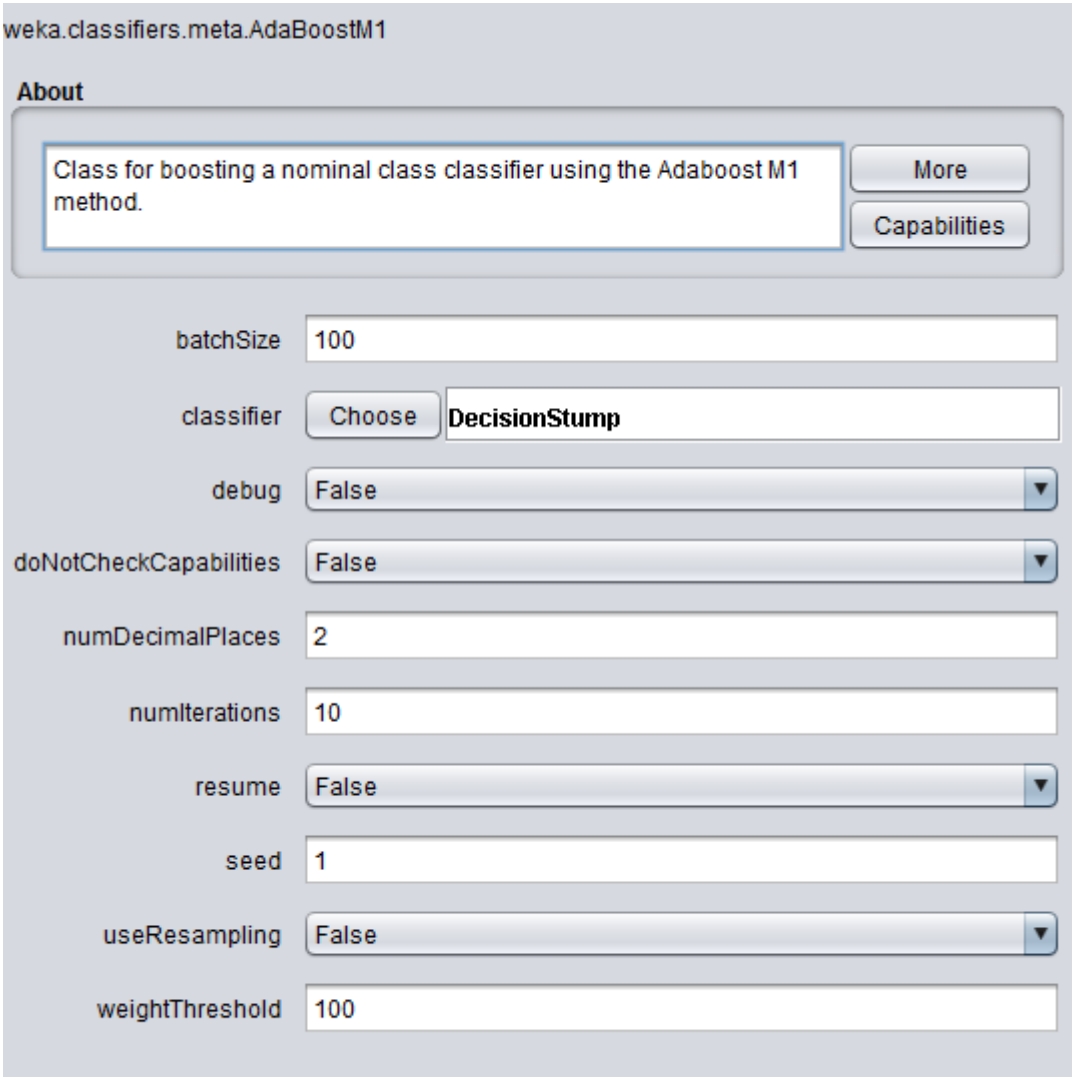
Gunakan parameter pada classifier Bagging yang merupakan training set di-sample sesuai batchSize dengan replacement yang kemudian dilakukan iterasi sejumlah numiterations dengan classifier tertentu.



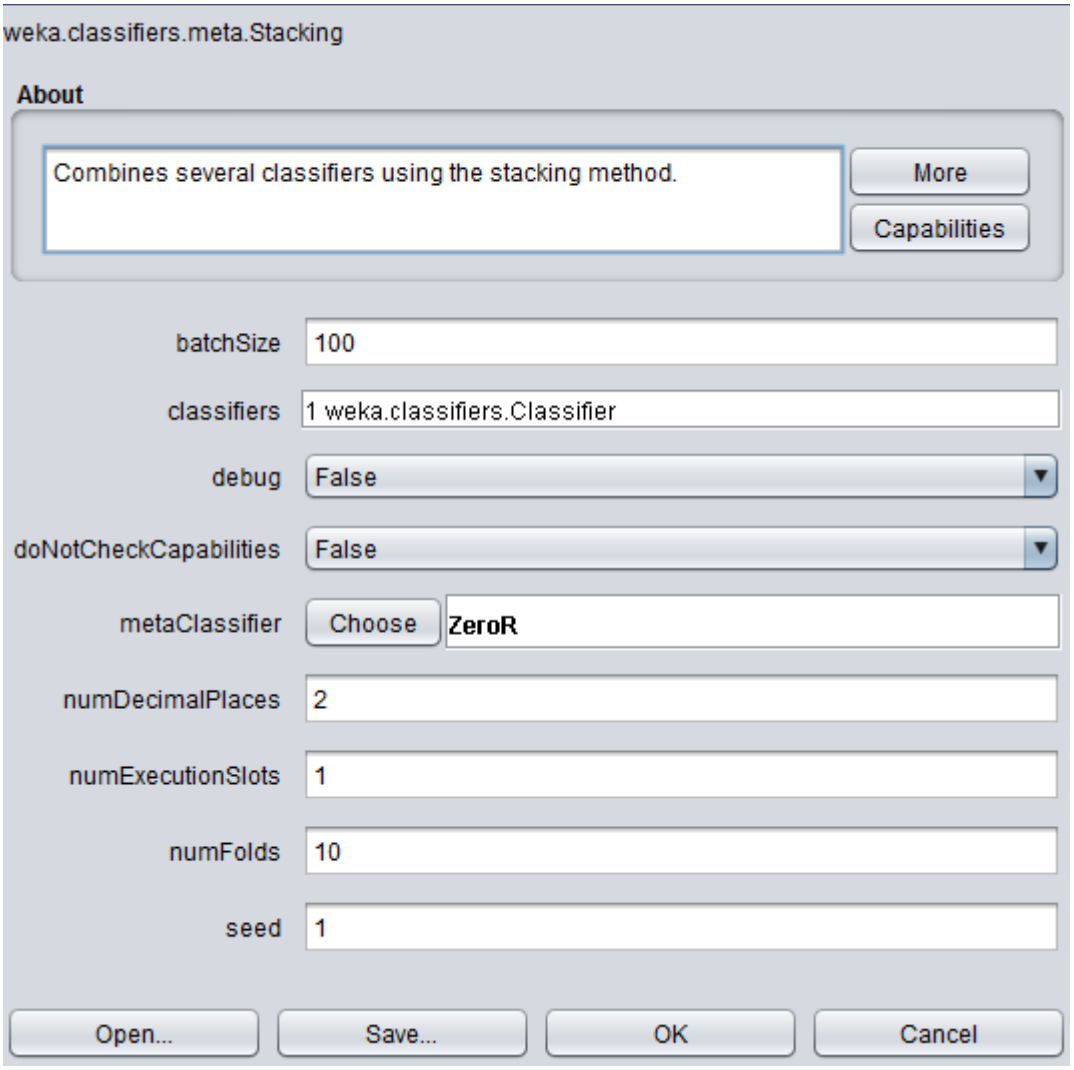
Gunakan parameter pada classifier Random Forest yang merupakan kebalikan dari Bagging yang merandom training set, tetapi me-random classifier.



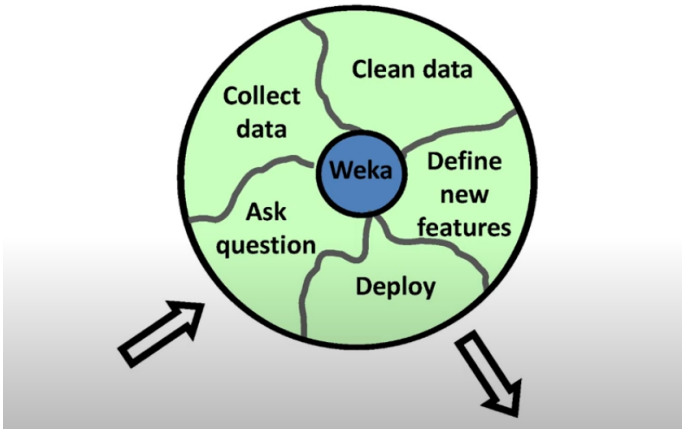
Gunakan parameter classifier AdaBoost M1 yang merupakan metode belajar Boost dimana model akan diperbaiki sejumlah numiterations.



Gunakan parameter classifier Stacking dimana output (model) dari classifier (level-0) akan menjadi input pada metaClassifier (level-1).



5.1 The Data Mining Process



Proses yang dilalui dalam Data Mining secara garis besar ada 5, yaitu:

1. *Ask a question*: Apa yang kita ingin ketahui dari data
2. *Collect data:* mengumpulkan beberapa dari beberapa sumber untuk memperkaya informasi yang ingin didapatkan,

3. *Clean data:* Melakukan pembersihan terhadap data yang dimiliki karena data sebenarnya sangatlah kotor dan tidak beraturan.
4. *Define a new features:* Menentukan fitur baru adalah kunci dalam data mining. Fitur tersebut disebut *feature engineering*.
5. *Deploy the result*: Setelah berhasil melalui semua tahapan yang ada, model prediksi atau informasi yang didapatkan di-*deploy* agar bisa digunakan.

Dalam proses data mining, WEKA hanya membantu dalam bagian kecil saja, terutama bagian teknikal.

5.2 Pitfalls and Pratfalls

Dalam proses data mining, sering kali ditemui *pitfalls* dan *pratfalls*. *Pitfalls* adalah kesulitan atau bahaya yang tidak terdeteksi dan *pratfalls* adalah aksi yang kurang baik atau ceroboh. Dalam data mining, sangat mudah sekali dilakukan kecurangan pada pemrosesan data, baik secara sadar maupun tidak sadar. Untuk melakukan tes yang *reliable*, gunakanlah data yang benar-benar baru dan belum dipakai sebelumnya.

Dalam proses pembuatan model, seringkali terjadi overfitting. Hal yang dapat dilakukan untuk menghindari hal tersebut adalah dengan tidak memakai training data dalam melakukan testing dan melakukan evaluasi terhadap model yang telah dibuat.

5.3 Data Mining and Ethics

Dalam melakukan data mining ada beberapa hukum privasi yang diperhatikan, di setiap negara/benua memiliki aturan yang tidak semuanya sama, namun secara garis besar memiliki makna yang sama.

Berikut contohnya hukum privasi yang ada di Eropa.

Information privacy laws (in Europe, but not US)

- A purpose must be stated for any personal information collected
- Such information must not be disclosed to others without consent
- Records kept on individuals must be accurate and up to date
- To ensure accuracy, individuals should be able to review data about themselves
- Data must be deleted when it is no longer needed for the stated purpose
- Personal information must not be transmitted to locations where equivalent data protection cannot be assured
- Some data is too sensitive to be collected, except in extreme circumstances

5.4 Summary

Dari video-video yang sudah dipelajari sebelumnya, dapat disimpulkan bahwa banyak teknik yang bisa dilakukan dalam melakukan data mining. Metode-metode yang ada di dunia tidak satupun cocok ke setiap kondisi data, sehingga perlu dievaluasi setiap algoritma yang berbeda dan performa yang ditunjukkan, dari algoritma yang telah dipakai.