Tugas 5 Laporan Mata Kuliah Data Mining

LANGKAH – LANGKAH MENGUBAH HIMPUNAN DATA WDBC, ABALONE DAN BANK MARKETING SERTA MEMBAGINYA MENJADI DATA TRAINING DAN TESTING

disusun untuk memenuhi
tugas mata kuliah Data Mining

oleh :

Khairul Umam Albi
2008107010072

JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
2022

A. Dataset WDBC

- Tools yang digunakan adalah WEKA explorer

1. Langkah-langkah untuk memisahkan data WDBC

- Buka WEKA Explorer dan buka file WDBC.arff.
- Kemudian pada bagian filter pilih Randomize pada package weka => filters => unsupervised=>instance => Randomize.
- Tanpa melakukan konfigurasi tambahan maka dataset tersebut telah teracak.
- Selanjutnya Kembali ke bagian filter pilih RemovePercentage pada package weka => filters =>unsupervised => instance => RemovePercentage.
- Pada bagian tersebut mengantikan presentase menjadi 80.0 untuk data testing maka data terhapus sebanyak 80 persen.
- Selanjutnya tekan apply dan simpan data tersebut dengan menekan save as di WEKA menjadi wdbc-testing.arrf.
- Kembali ke halaman utama weka dan klik undo, selanjutnya ganti nilai invertSelection menjadi True, dengan percentage masih sama maka data tersimpan sebanyak 80 persen.

2. Uraikan hasil klasifikasi yang diperoleh.

Proses klasifikasi menggunakan Classifier IBk (Lazy/KNN) dengan k=3, k=5, k=7, k=9,dan k=11 serta pembobotan dalam voting penentuan class adalah distanceWeighting =Weight by 1/distance.

a. Supplied test set WDBC-testing.arff pada k = 3

```
=== Summary ===

Correctly Classified Instances         110               96.4912 %
Incorrectly Classified Instances         4                3.5088 %
Kappa statistic                          0.923
Mean absolute error                      0.0562
Root mean squared error                  0.1738
Relative absolute error                 12.1191 %
Root relative squared error             36.3571 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.027    0.950      0.950    0.950      0.923    0.983     0.975     M
                 0.973    0.050    0.973      0.973    0.973      0.923    0.983     0.982     B
Weighted Avg.    0.965    0.042    0.965      0.965    0.965      0.923    0.983     0.980

=== Confusion Matrix ===

  a  b   <-- classified as
 38  2 |  a = M
  2 72 |  b = B
```

Precision bernilai 0.965, Recall bernilai 0.965, dan F-Measure bernilai 0.965

b. Supplied test set WDBC-testing.arff pada k = 5

```
=== Summary ===

Correctly Classified Instances         111               97.3684 %
Incorrectly Classified Instances         3                2.6316 %
Kappa statistic                          0.9426
Mean absolute error                      0.0495
Root mean squared error                  0.1402
Relative absolute error                 10.6772 %
Root relative squared error             29.3291 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.027    0.951      0.975    0.963      0.943    0.997     0.993     M
                 0.973    0.025    0.986      0.973    0.980      0.943    0.997     0.998     B
Weighted Avg.    0.974    0.026    0.974      0.974    0.974      0.943    0.997     0.996

=== Confusion Matrix ===

  a  b   <-- classified as
 39  1 |  a = M
  2 72 |  b = B
```

Precision bernilai 0.974, Recall bernilai 0.974, dan F-Measure bernilai 0.974

c. Supplied test set WDBC-testing.arff pada k = 7

```
=== Summary ===

Correctly Classified Instances         113               99.1228 %
Incorrectly Classified Instances         1                0.8772 %
Kappa statistic                          0.9809
Mean absolute error                      0.0529
Root mean squared error                  0.1311
Relative absolute error                 11.4089 %
Root relative squared error             27.4322 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.014    0.976      1.000    0.988      0.981    0.999     0.998     M
                 0.986    0.000    1.000      0.986    0.993      0.981    0.999     0.999     B
Weighted Avg.    0.991    0.005    0.991      0.991    0.991      0.981    0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 40  0 |  a = M
  1 73 |  b = B
```

Precision bernilai 0.991, Recall bernilai 0.991, dan F-Measure bernilai 0.991.

d. Supplied test set WDBC-testing.arff pada k = 9

```
Correctly Classified Instances         113               99.1228 %
Incorrectly Classified Instances         1                0.8772 %
Kappa statistic                          0.9809
Mean absolute error                      0.0616
Root mean squared error                  0.1385
Relative absolute error                 13.286  %
Root relative squared error             28.9701 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.014    0.976      1.000    0.988      0.981    0.999     0.998     M
                 0.986    0.000    1.000      0.986    0.993      0.981    0.999     0.999     B
Weighted Avg.    0.991    0.005    0.991      0.991    0.991      0.981    0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 40  0 |  a = M
  1 73 |  b = B
```

Precision bernilai 0.991, Recall bernilai 0.991, dan F-Measure bernilai 0.991.

e. Supplied test set WDBC-testing.arff pada k = 11

```
=== Summary ===

Correctly Classified Instances         112               98.2456 %
Incorrectly Classified Instances         2                1.7544 %
Kappa statistic                          0.961
Mean absolute error                      0.068
Root mean squared error                  0.1473
Relative absolute error                 14.6526 %
Root relative squared error             30.8048 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.000    1.000      0.950    0.974      0.962    0.999     0.998     M
                 1.000    0.050    0.974      1.000    0.987      0.962    0.999     0.999     B
Weighted Avg.    0.982    0.032    0.983      0.982    0.982      0.962    0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 38  2 |  a = M
  0 74 |  b = B
```

Precision bernilai 0.982, Recall bernilai 0.982, dan F-Measure bernilai 0.982.

f. Cross-validation dengan file WDBC.arff pada k = 3

```
=== Summary ===

Correctly Classified Instances         551               96.8366 %
Incorrectly Classified Instances        18                3.1634 %
Kappa statistic                          0.9318
Mean absolute error                      0.0487
Root mean squared error                  0.1711
Relative absolute error                 10.4128 %
Root relative squared error             35.3883 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.939    0.014    0.975      0.939    0.957      0.932    0.980     0.973     M
                 0.986    0.061    0.964      0.986    0.975      0.932    0.980     0.978     B
Weighted Avg.    0.968    0.044    0.969      0.968    0.968      0.932    0.980     0.976

=== Confusion Matrix ===

   a   b   <-- classified as
 199  13 |   a = M
   5 352 |   b = B
```

Precision bernilai 0.969, Recall bernilai 0.968, dan F-Measure bernilai 0.968.

g. Cross-validation dengan file WDBC.arff pada k = 5

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         552               97.0123 %
Incorrectly Classified Instances        17                2.9877 %
Kappa statistic                          0.9355
Mean absolute error                      0.0503
Root mean squared error                  0.1621
Relative absolute error                 10.7568 %
Root relative squared error             33.5346 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.939    0.011    0.980      0.939    0.959      0.936    0.987     0.982     M
                 0.989    0.061    0.964      0.989    0.976      0.936    0.987     0.987     B
Weighted Avg.    0.970    0.043    0.970      0.970    0.970      0.936    0.987     0.985

=== Confusion Matrix ===

   a   b   <-- classified as
 199  13 |   a = M
   4 353 |   b = B
```

Precision bernilai 0.970, Recall bernilai 0.970, dan F-Measure bernilai 0.970

h. Cross-validation dengan file WDBC.arff pada k = 7

```
=== Summary ===

Correctly Classified Instances         553               97.188  %
Incorrectly Classified Instances        16                2.812  %
Kappa statistic                          0.9394
Mean absolute error                      0.0568
Root mean squared error                  0.1639
Relative absolute error                 12.1383 %
Root relative squared error             33.9003 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.011    0.980      0.943    0.962      0.940    0.989     0.983     M
                 0.989    0.057    0.967      0.989    0.978      0.940    0.989     0.989     B
Weighted Avg.    0.972    0.040    0.972      0.972    0.972      0.940    0.989     0.987

=== Confusion Matrix ===

   a   b   <-- classified as
 200  12 |   a = M
   4 353 |   b = B
```

Precision bernilai 0.927, Recall bernilai 0.972, dan F-Measure bernilai 0.972

i. Cross-validation dengan file WDBC.arff pada k = 9

```
=== Summary ===

Correctly Classified Instances         553               97.188  %
Incorrectly Classified Instances        16                2.812  %
Kappa statistic                          0.9392
Mean absolute error                      0.0611
Root mean squared error                  0.1643
Relative absolute error                 13.073  %
Root relative squared error             33.9814 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.934    0.006    0.990      0.934    0.961      0.940    0.991     0.989     M
                 0.994    0.066    0.962      0.994    0.978      0.940    0.991     0.991     B
Weighted Avg.    0.972    0.044    0.972      0.972    0.972      0.940    0.991     0.990

=== Confusion Matrix ===

   a   b   <-- classified as
 198  14 |   a = M
   2 355 |   b = B
```

Precision bernilai 0.927, Recall bernilai 0.972, dan F-Measure bernilai 0.972

j. Cross-validation dengan file WDBC.arff pada k = 11

```
=== Summary ===

Correctly Classified Instances         550               96.6608 %
Incorrectly Classified Instances        19                3.3392 %
Kappa statistic                          0.9278
Mean absolute error                      0.067
Root mean squared error                  0.1722
Relative absolute error                 14.3168 %
Root relative squared error             35.6207 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.011    0.980      0.929    0.954      0.929    0.990     0.987     M
                 0.989    0.071    0.959      0.989    0.974      0.929    0.990     0.990     B
Weighted Avg.    0.967    0.049    0.967      0.967    0.966      0.929    0.990     0.989

=== Confusion Matrix ===

   a   b   <-- classified as
 197  15 |   a = M
   4 353 |   b = B
```

Precision bernilai 0.967, Recall bernilai 0.967 dan F-Measure bernilai 0.966

3. Kesimpulan

Precision terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 7 dan k = 9 berada pada angka 0.991. Recall terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 7 dan k = 9 berada pada angka 0.991. F-Measure terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 7 dan k = 9 berada pada angka 0.991. Pada kasus dataset WDBC, menggunakan ke-5 k menunjukkan pengujian supplied test dapat menghasilkan akurasi yang lebih baik daripada cross-validation, selain itu dari ke-5 k menunjukkan penggunaan k = 7 dan k = 9 lebih baik digunakan untuk dataset ini.

B. Abalone

- Tools yang digunakan adalah WEKA explorer

1. Langkah-langkah untuk memisahkan data Abalone

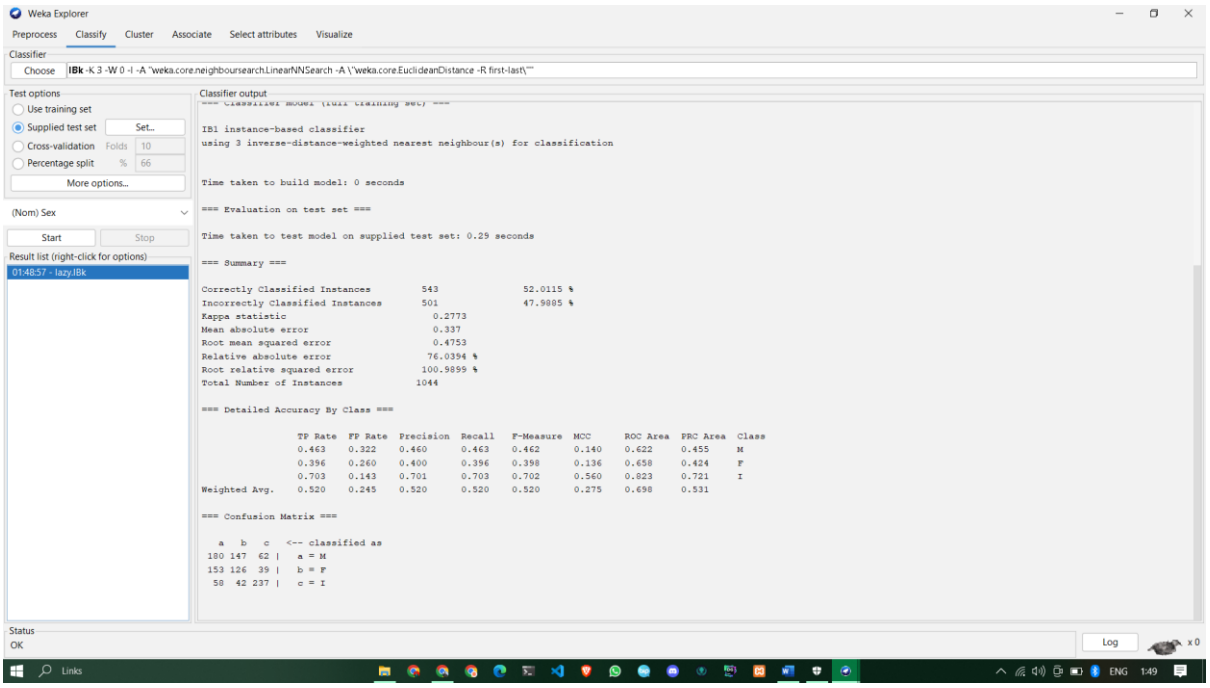
- Buka WEKA Explorer dan buka file Abalone.arff.
- Kemudian pada bagian filter pilih Randomize pada package weka => filters => unsupervised=>instance => Randomize.
- Tanpa melakukan konfigurasi tambahan maka dataset tersebut telah teracak.
- Selanjutnya Kembali ke bagian filter pilih RemovePercentage pada package weka => filters =>unsupervised => instance => RemovePercentage.
- Pada bagian tersebut mengantikan presentase menjadi 75.0 untuk data testing maka data terhapus sebanyak 75 persen.
- Selanjutnya tekan apply dan simpan data tersebut dengan menekan save as di WEKA menjadi Abalone-testing.arrf.

- Kembali ke halaman utama weka dan klik undo, selanjutnya ganti nilai invertSelection menjadi True, dengan percentage masih sama maka data tersimpan sebanyak 75 persen.

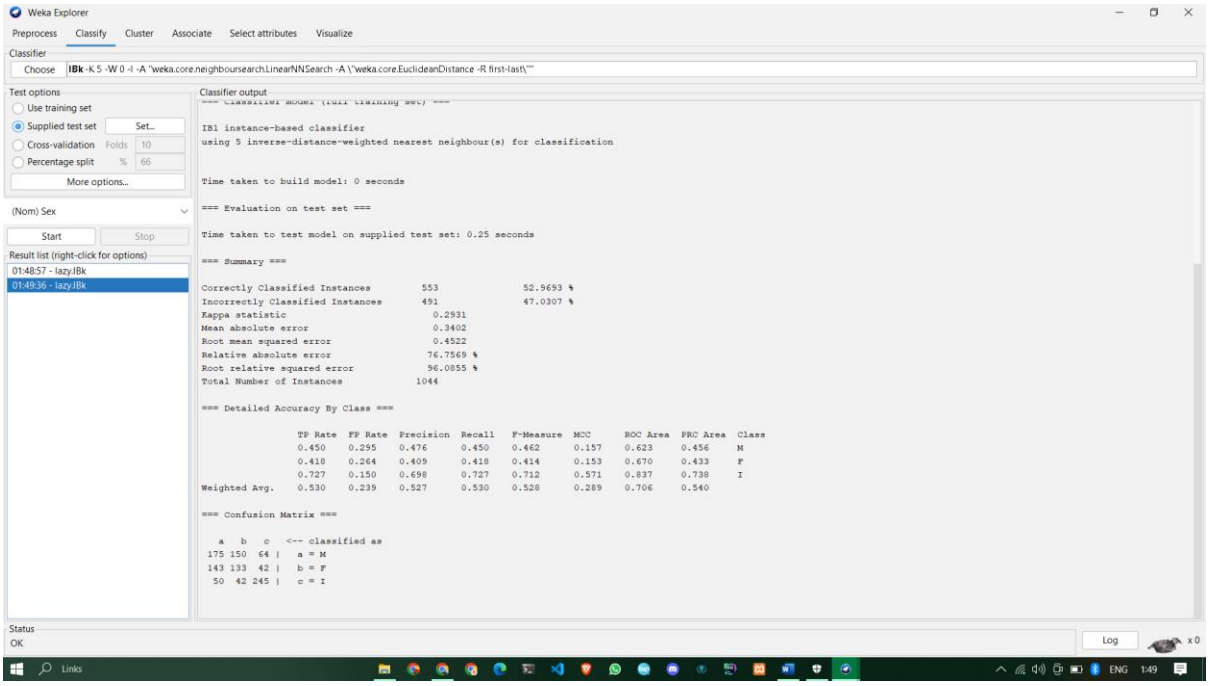
2. Uraikan hasil klasifikasi yang diperoleh.

Proses klasifikasi menggunakan Classifier IBk (Lazy/KNN) dengan k=3, k=5, k=7, k=9,dan k=11 serta pembobotan dalam voting penentuan class adalah distanceWeighting =Weight by 1/distance.

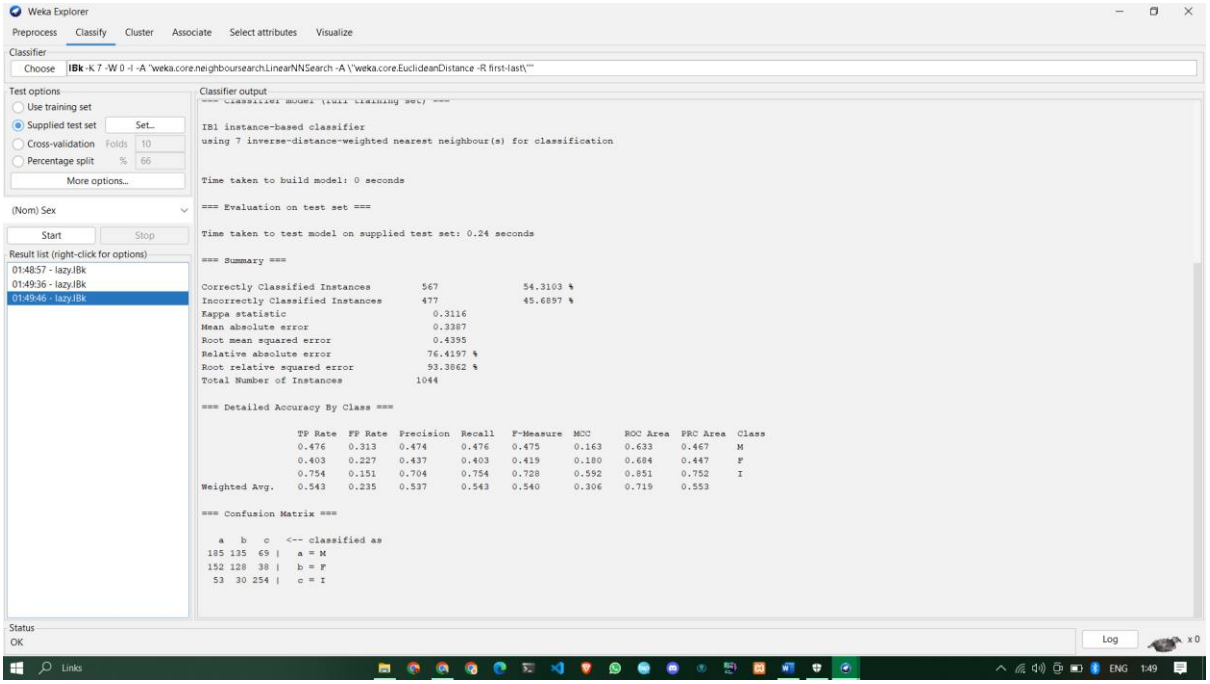
a. Supplied test set Abalone-testing.arff pada k = 3



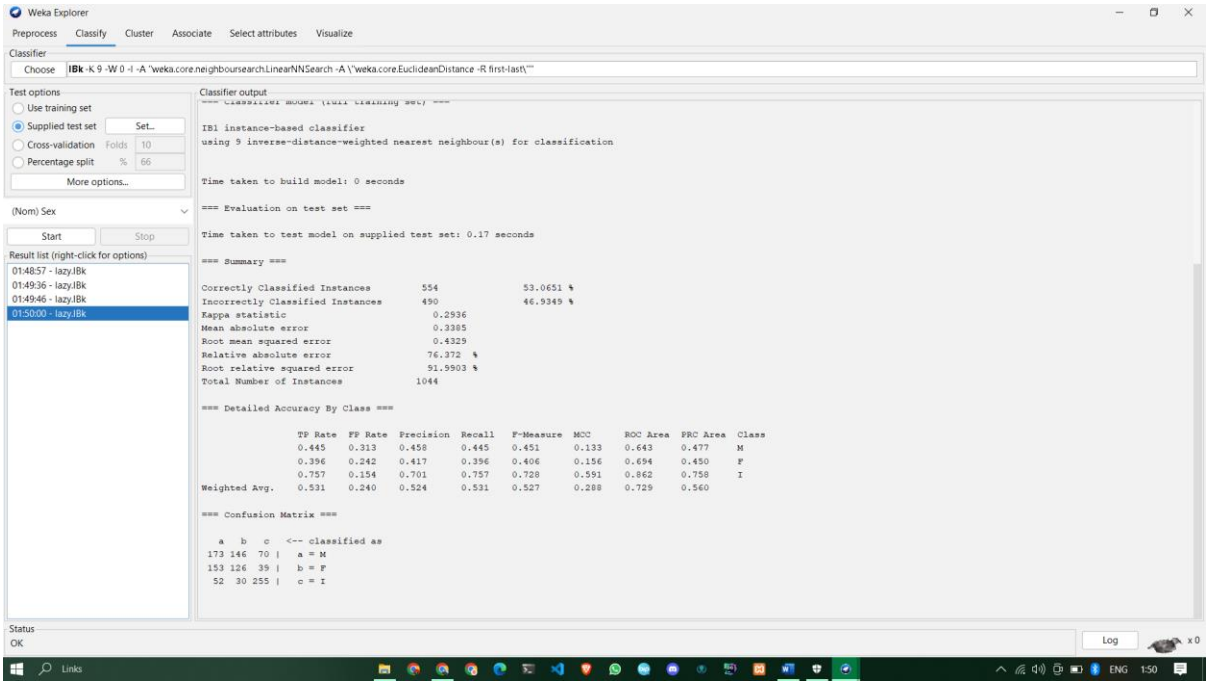
b. Supplied test set Abalone-testing.arff pada k = 5



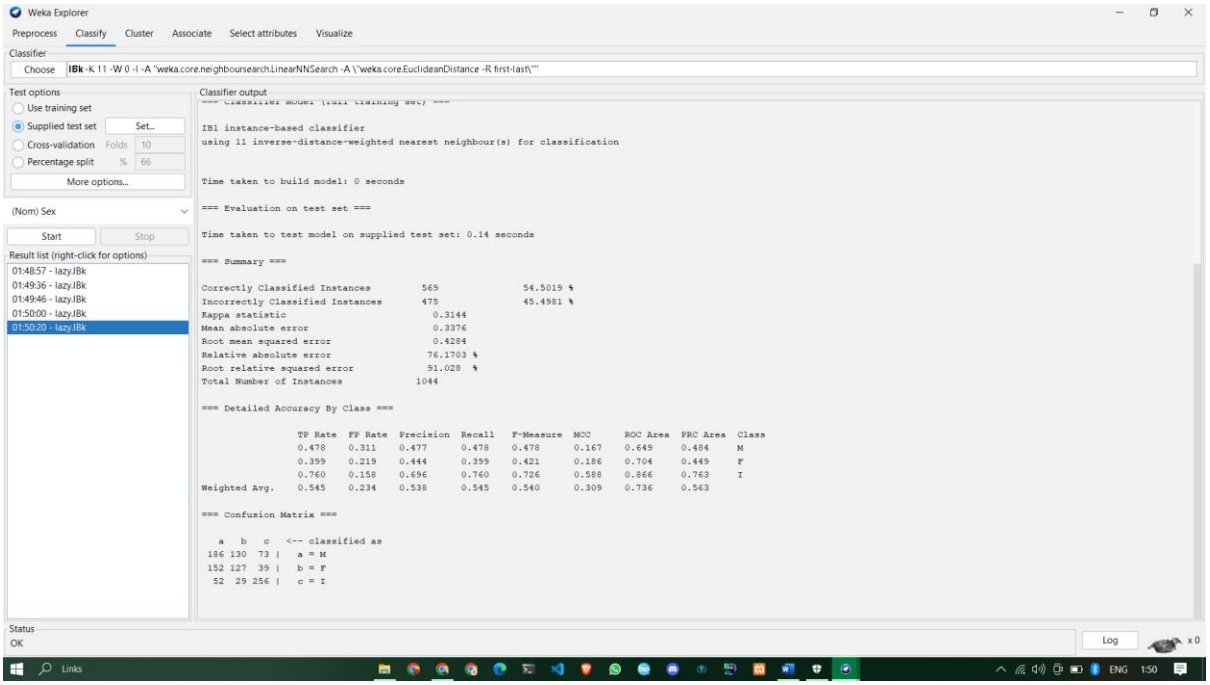
c. Supplied test set Abalone-testing.arff pada k = 7



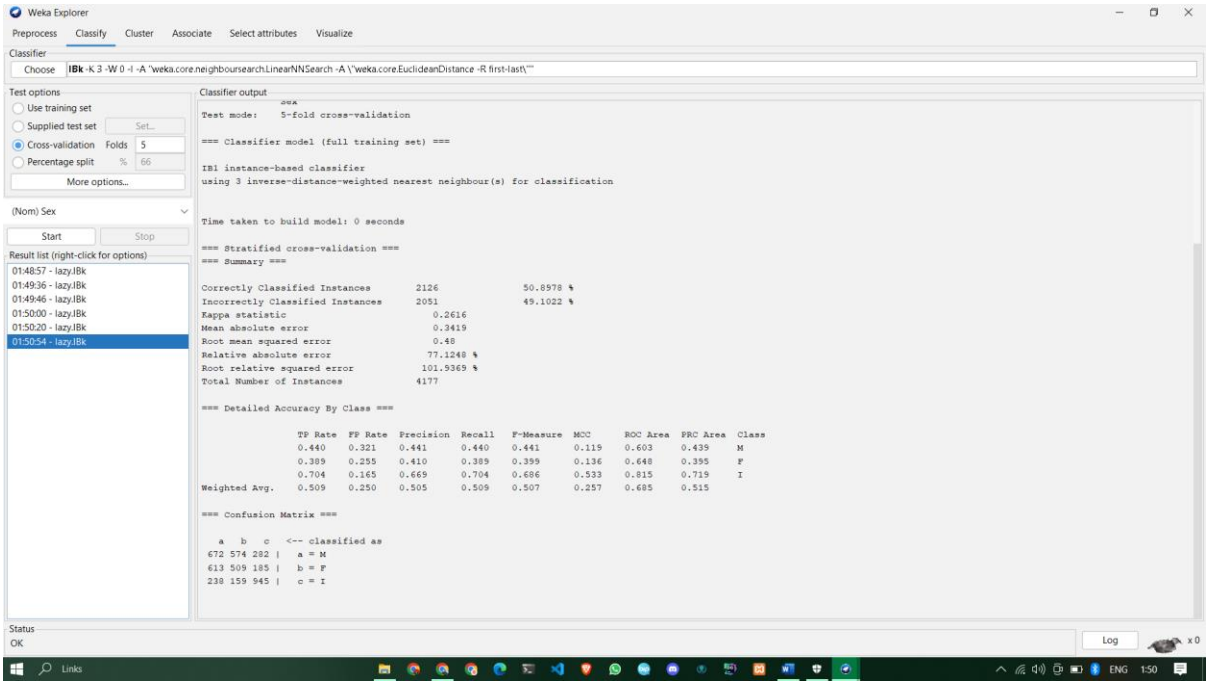
d. Supplied test set Abalone-testing.arff pada k = 9



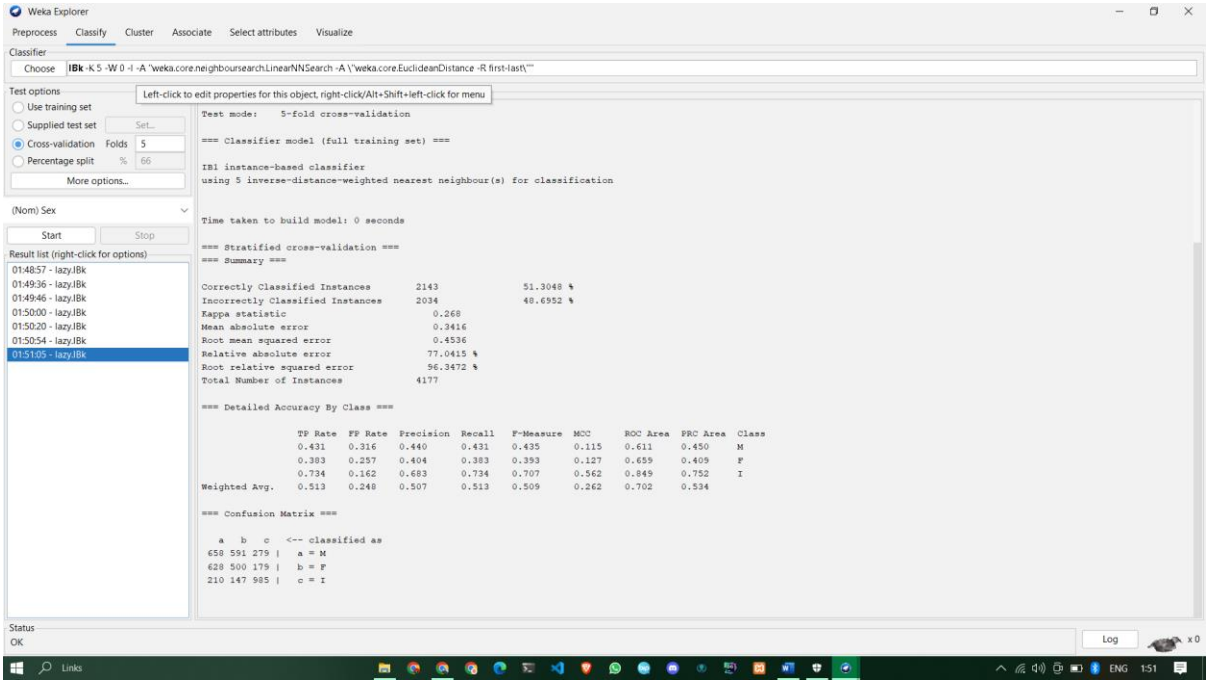
e. Supplied test set Abalone-testing.arff pada k = 11



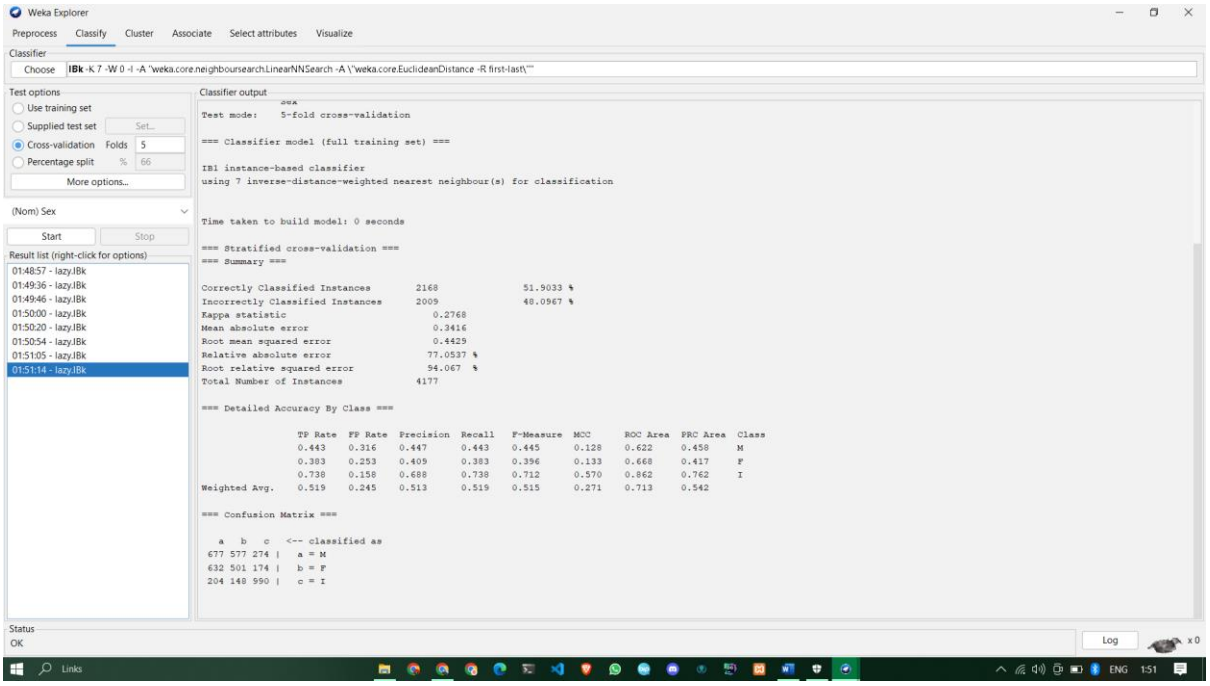
f. Cross-validation dengan file Abalone.arff pada k = 3



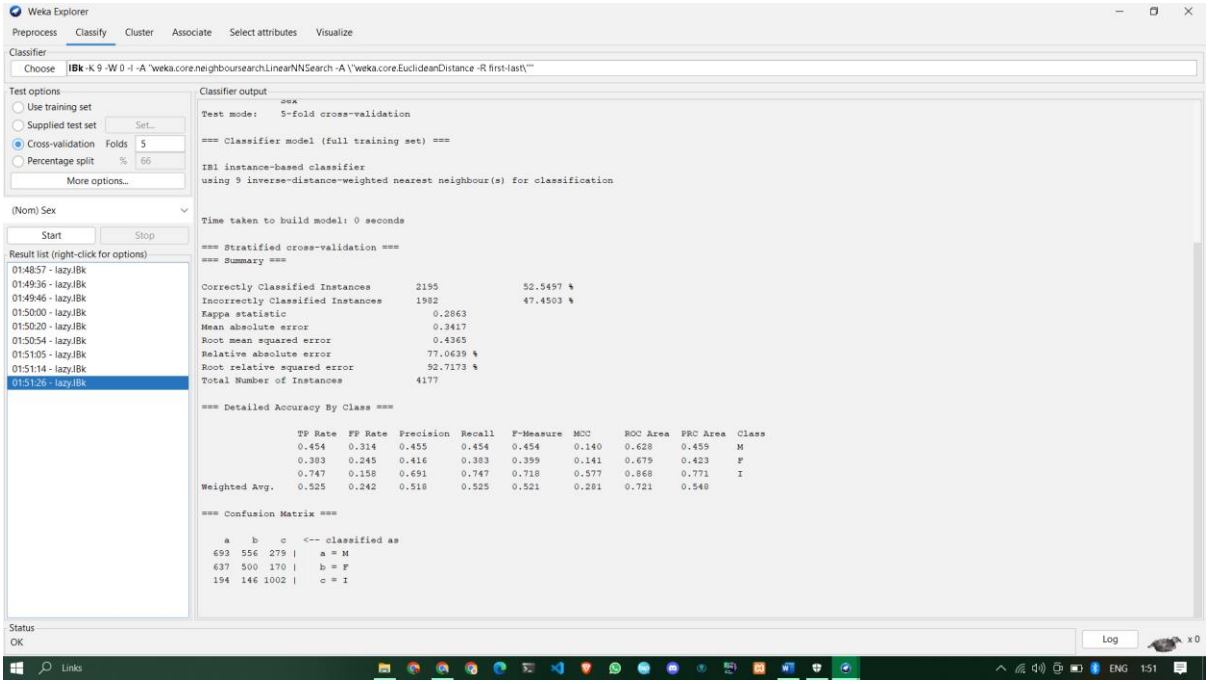
g. Cross-validation dengan file Abalone.arff pada k = 5



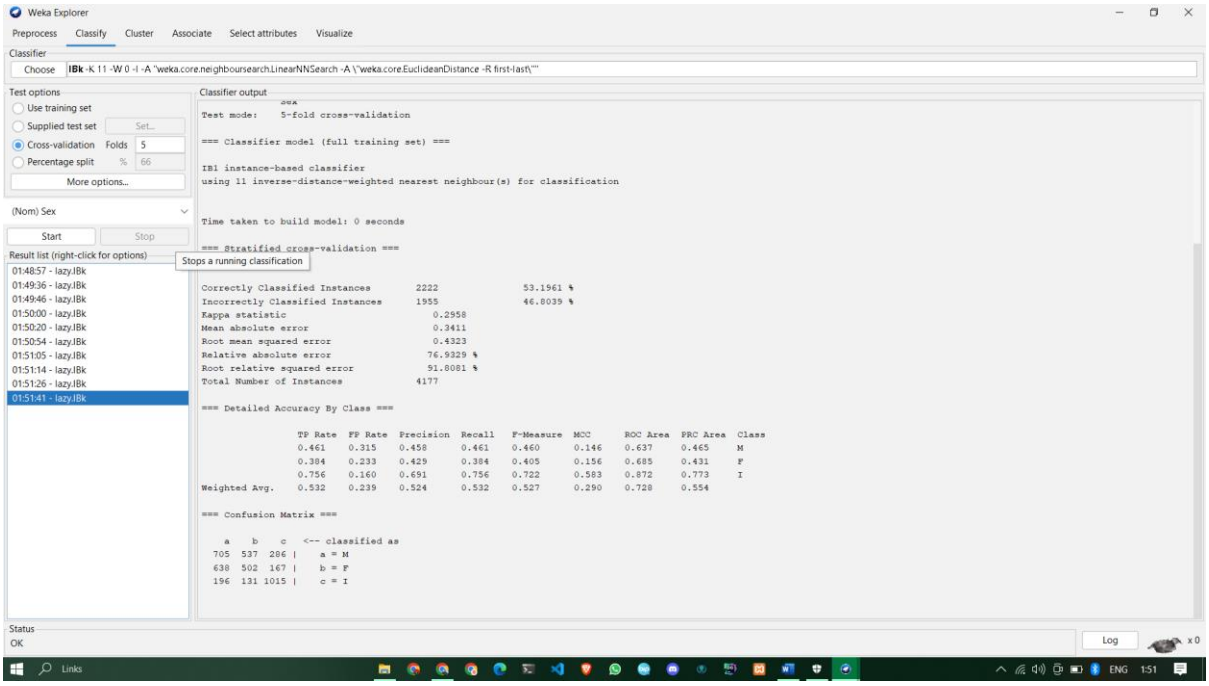
h. Cross-validation dengan file Abalone.arff pada k = 7



i. Cross-validation dengan file Abalone.arff pada k = 9



j. Cross-validation dengan file Abalone.arff pada k = 11



3. Kesimpulan.

Precision terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 7 berada pada angka 0.537. Recall terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 11 berada pada angka 0.545. F-Measure terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 7 dan k = 11 berada pada angka 0.540. Pada kasus dataset Abalone, menggunakan ke-5 k menunjukkan pengujian supplied test dapat menghasilkan akurasi yang lebih baik daripada cross-validation, selain itu dari ke-5 k menunjukkan penggunaan k = 7 dan k = 11 lebih baik digunakan untuk dataset ini.

C. Bank Data

1. Langkah-langkah untuk memisahkan data Bank.

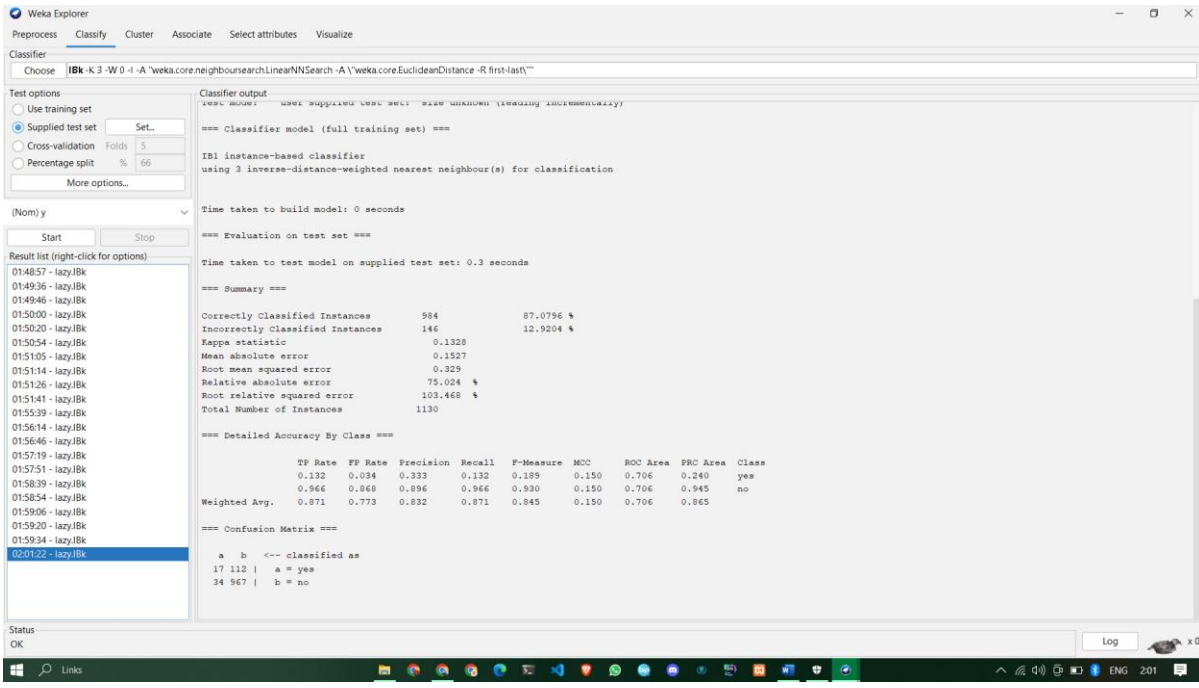
- Buka WEKA Explorer dan buka file Bank.arff.
- Kemudian pada bagian filter pilih Randomize pada package weka => filters => unsupervised=>instance => Randomize.
- Tanpa melakukan konfigurasi tambahan maka dataset tersebut telah teracak.
- Selanjutnya Kembali ke bagian filter pilih RemovePercentage pada package weka => filters =>unsupervised => instance => RemovePercentage.
- Pada bagian tersebut mengantikan presentase menjadi 75.0 untuk data testing maka data terhapus sebanyak 75 persen.
- Selanjutnya tekan apply dan simpan data tersebut dengan menekan save as di WEKA menjadi bank-testing.arrf.

- Kembali ke halaman utama weka dan klik undo, selanjutnya ganti nilai invertSelection menjadi True, dengan percentage masih sama maka data tersimpan sebanyak 75 persen.

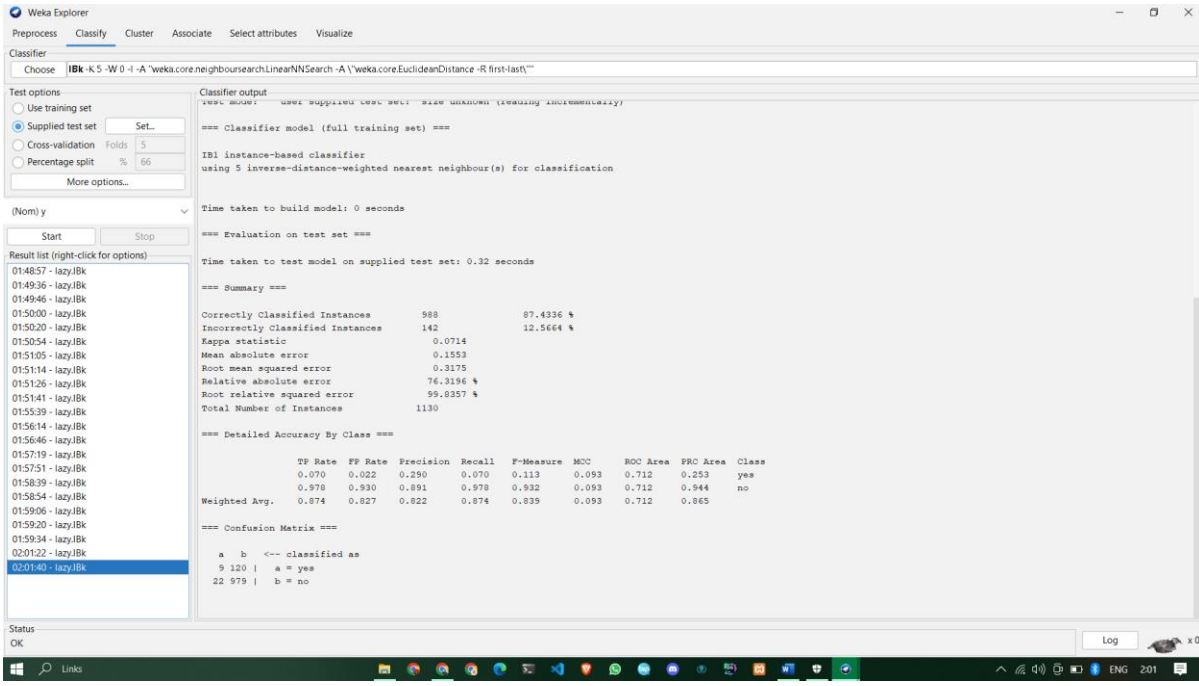
2. Uraikan hasil klasifikasi yang diperoleh.

Proses klasifikasi menggunakan Classifier IBk (Lazy/KNN) dengan k=3, k=5, k=7, k=9,dan k=11 serta pembobotan dalam voting penentuan class adalah distanceWeighting =Weight by 1/distance.

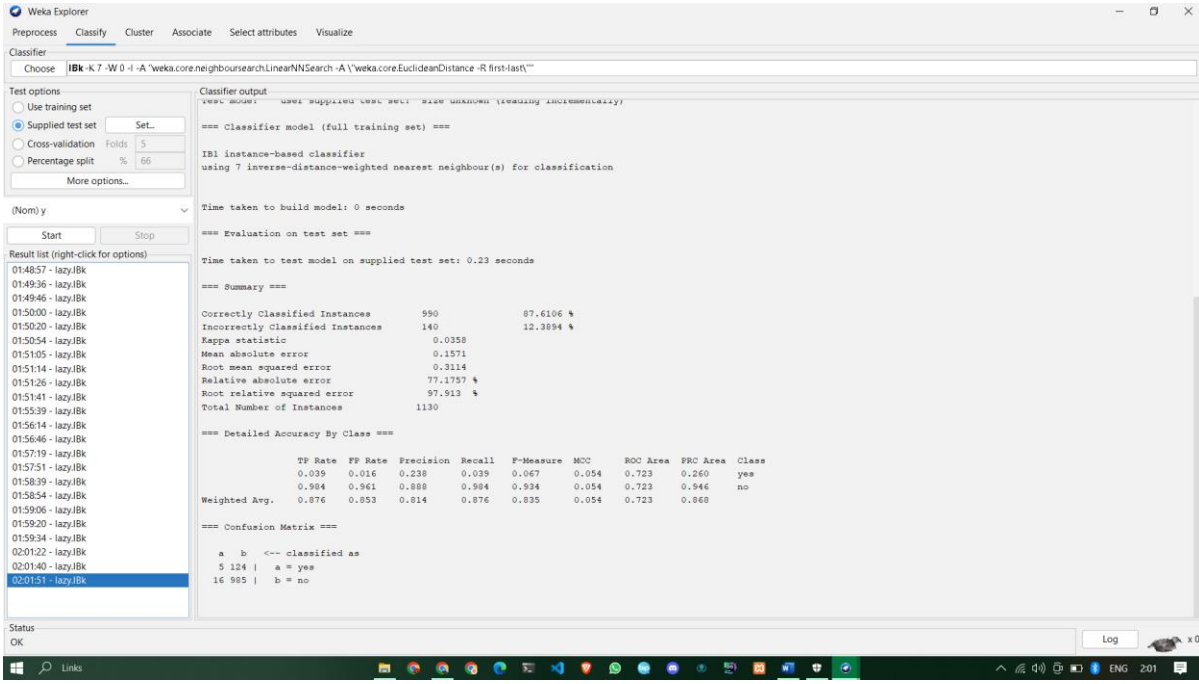
a. Supplied test set bank-testing.arff pada k = 3



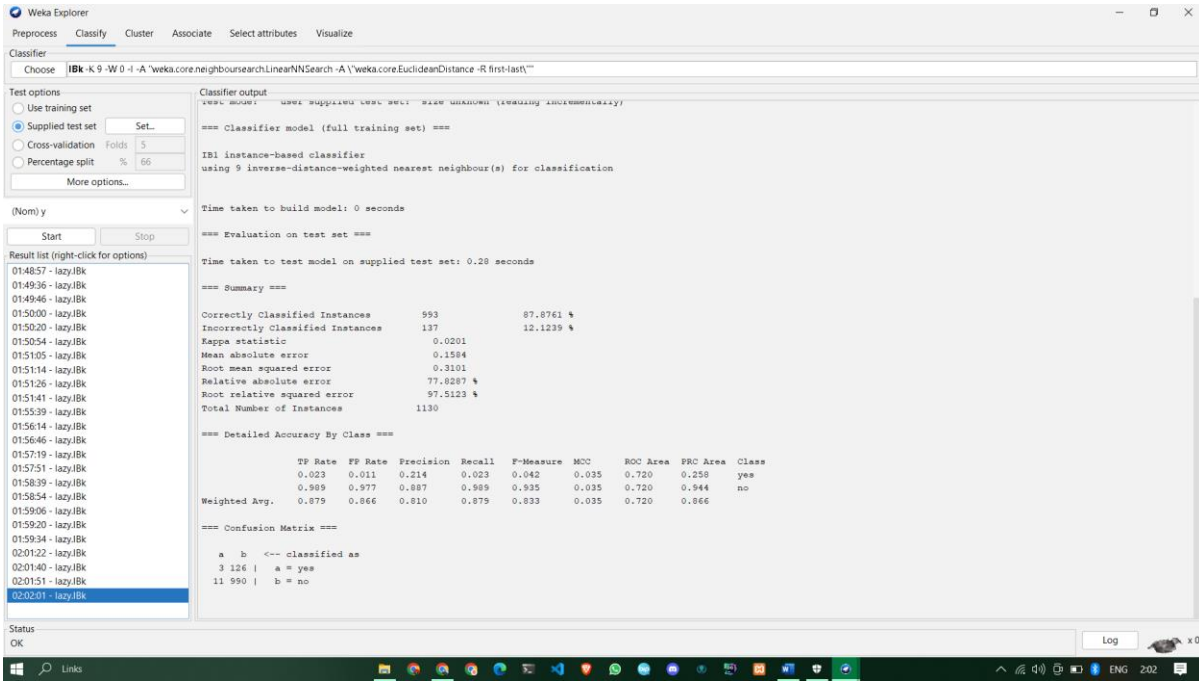
b. Supplied test set bank-testing.arff pada k = 5



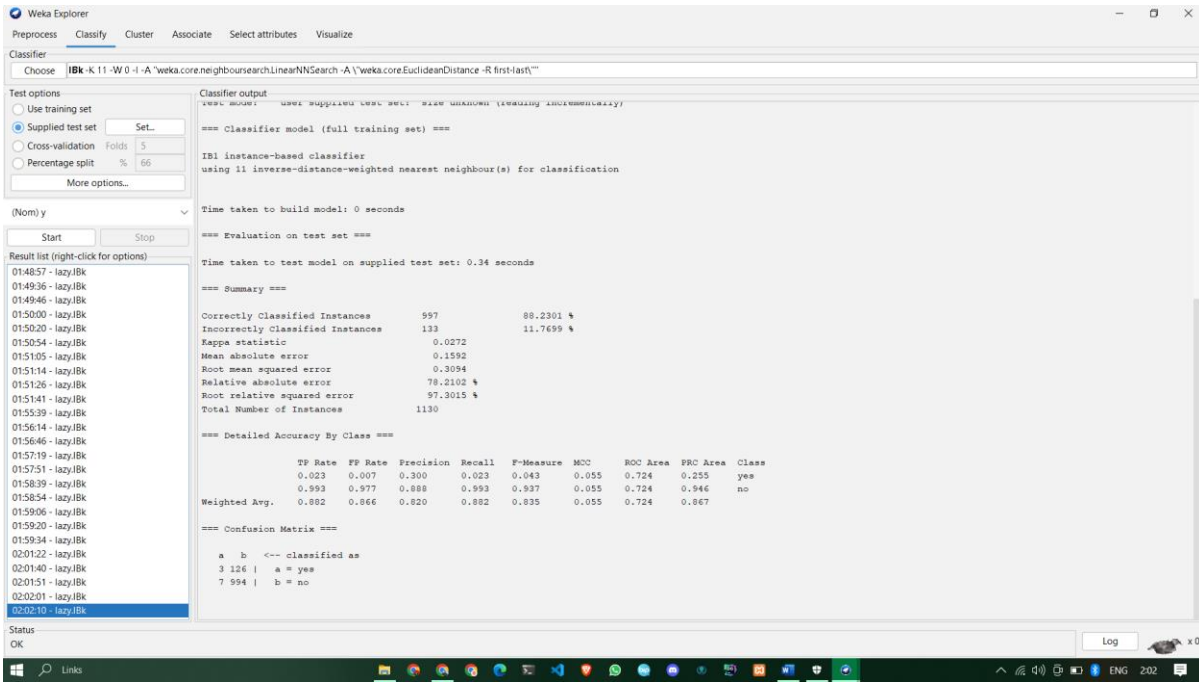
c. Supplied test set bank-testing.arff pada k = 7



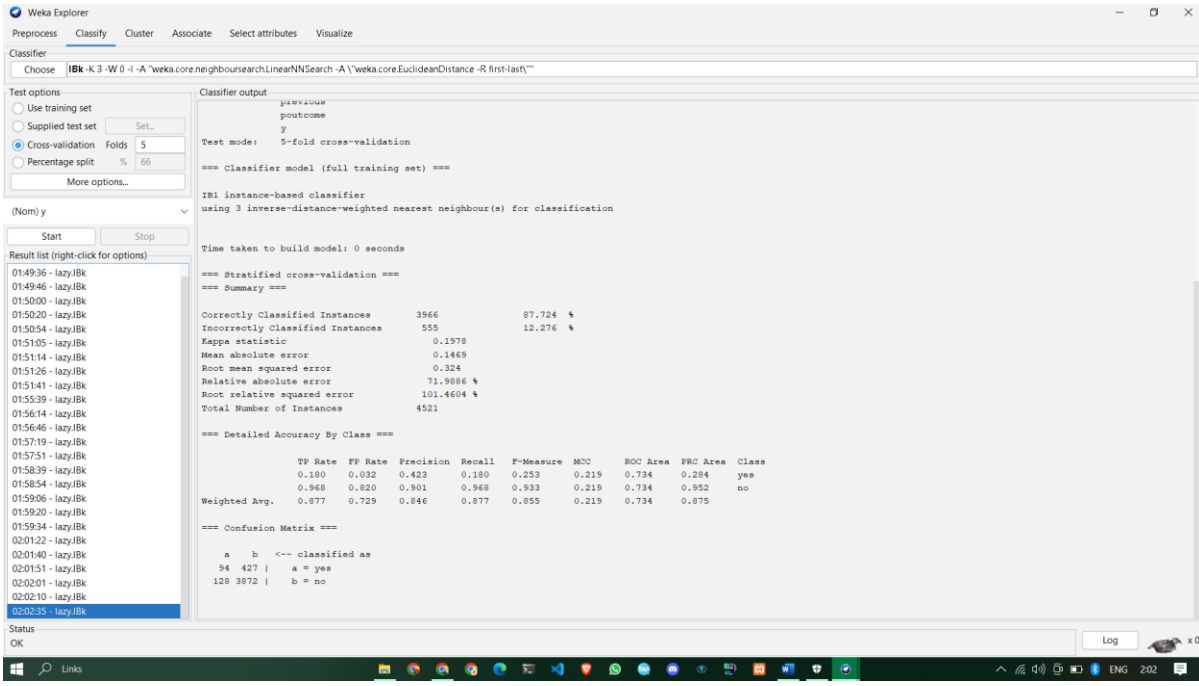
d. Supplied test set bank-testing.arff pada k = 9



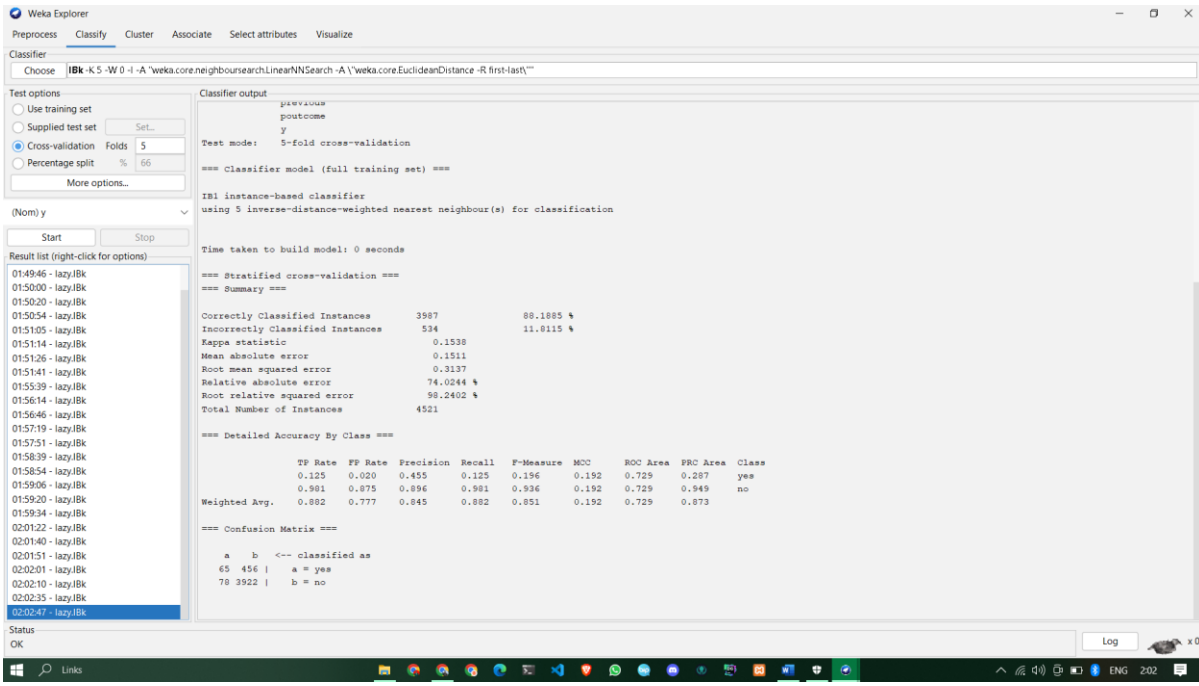
e. Supplied test set bank -testing.arff pada k = 11



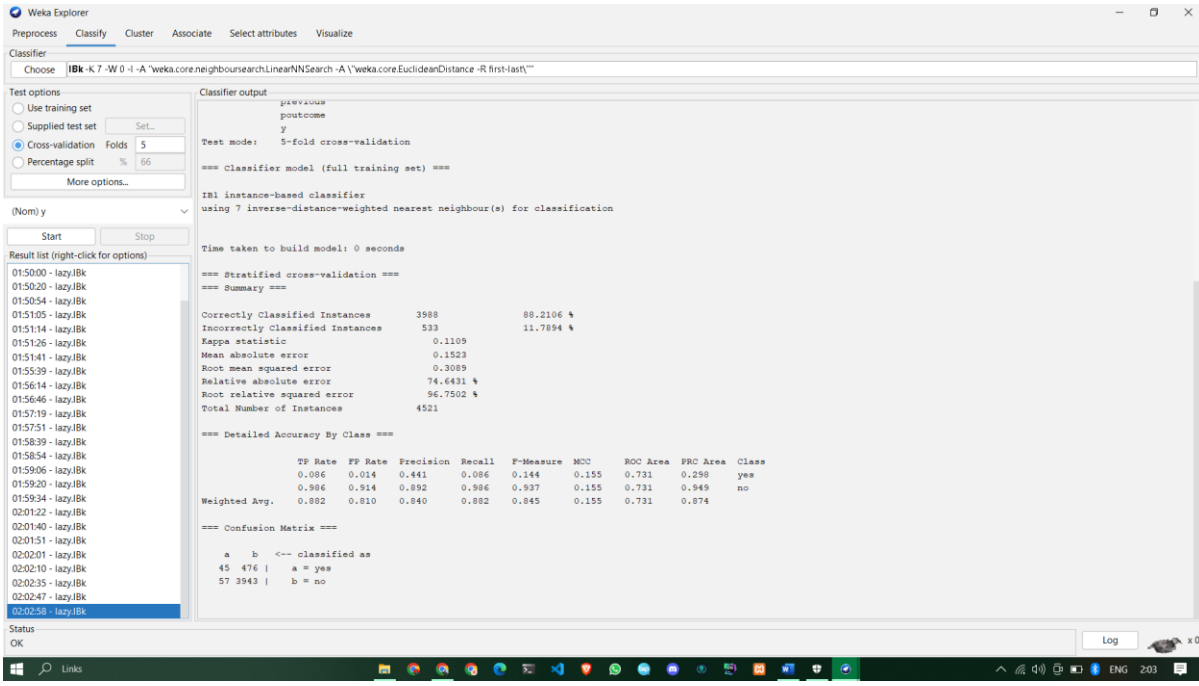
f. Cross-validation dengan file bank.arff pada k = 3



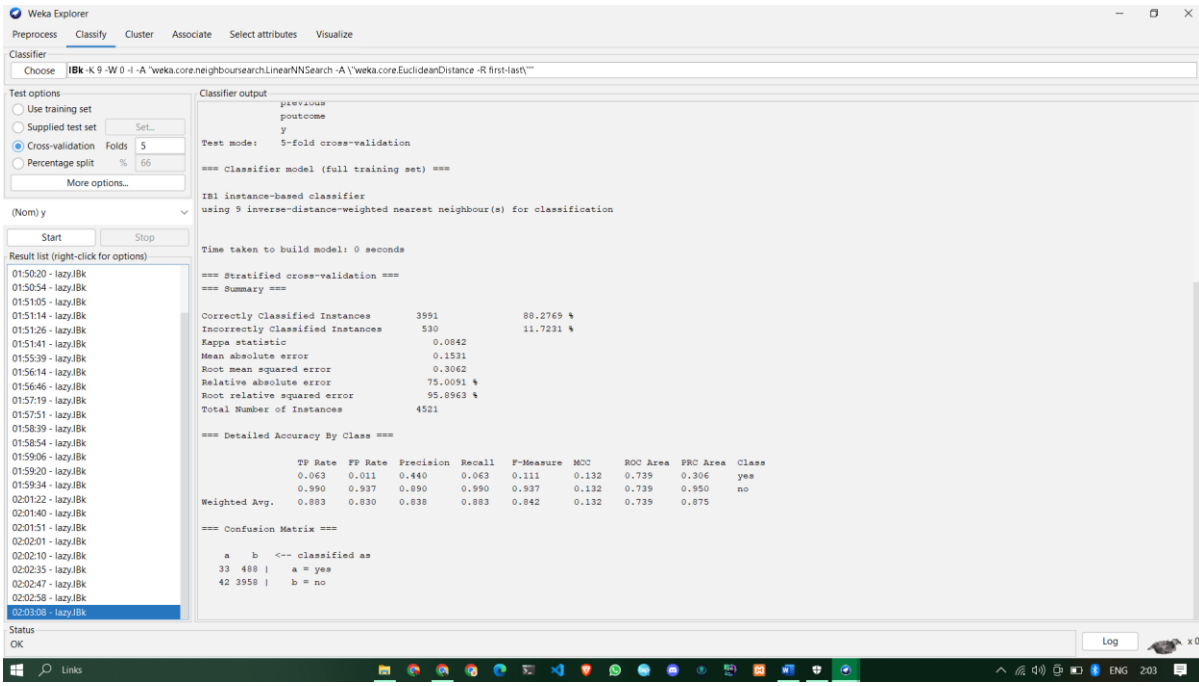
g. Cross-validation dengan file bank.arff pada k = 5



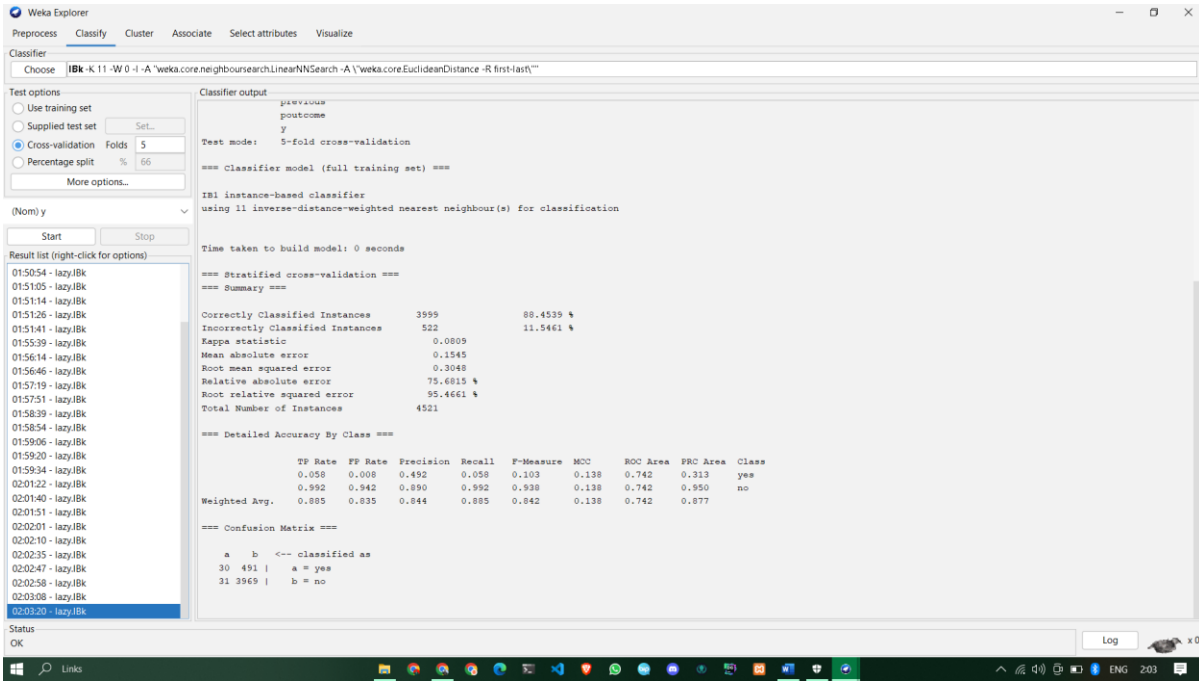
h. Cross-validation dengan file bank.arff pada k = 7



i. Cross-validation dengan file bank.arff pada k = 9



j. Cross-validation dengan file bank.arff pada k = 11



3. Kesimpulan.

Precision terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 3 berada pada angka 0.832. Recall terbesar yang di dapat dari uji coba adalah pada cross-validation dengan k = 11 berada pada angka 0.885. F-Measure terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 3 berada pada angka 0.845. Pada kasus dataset Bank Marketing, menggunakan ke-5 k menunjukkan hasil pengujian yang tidak memiliki selisih yang cukup besar baik pada pengujian supplied test maupun cross-validation. Nilai k yang digunakan pada supplied test lebih baik digunakan pada k = 3, lebih dari itu menghasilkan akurasi yang lebih rendah, namun pada cross-validation cenderung lebih baik menggunakan k yang lebih besar seperti 11.

D. Bank-full

1. Langkah-langkah untuk memisahkan data Bank-full.

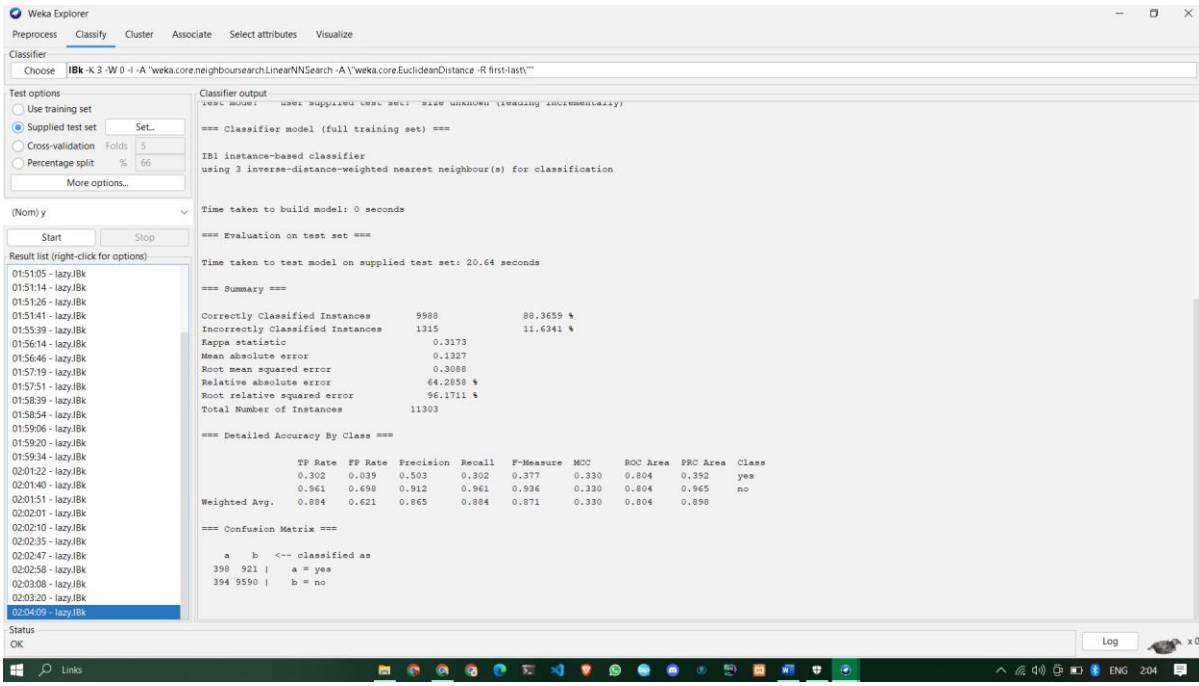
- Buka WEKA Explorer dan buka file Bank-full.arff.
- Kemudian pada bagian filter pilih Randomize pada package weka => filters => unsupervised=>instance => Randomize.
- Tanpa melakukan konfigurasi tambahan maka dataset tersebut telah teracak.
- Selanjutnya Kembali ke bagian filter pilih RemovePercentage pada package weka => filters =>unsupervised => instance => RemovePercentage.
- Pada bagian tersebut mengantikan presentase menjadi 75.0 untuk data testing maka data terhapus sebanyak 75 persen.

- Selanjutnya tekan apply dan simpan data tersebut dengan menekan save as di WEKA menjadi bank-full-testing.arrf.
- Kembali ke halaman utama weka dan klik undo, selanjutnya ganti nilai invertSelection menjadi True, dengan percentage masih sama maka data tersimpan sebanyak 75 persen.

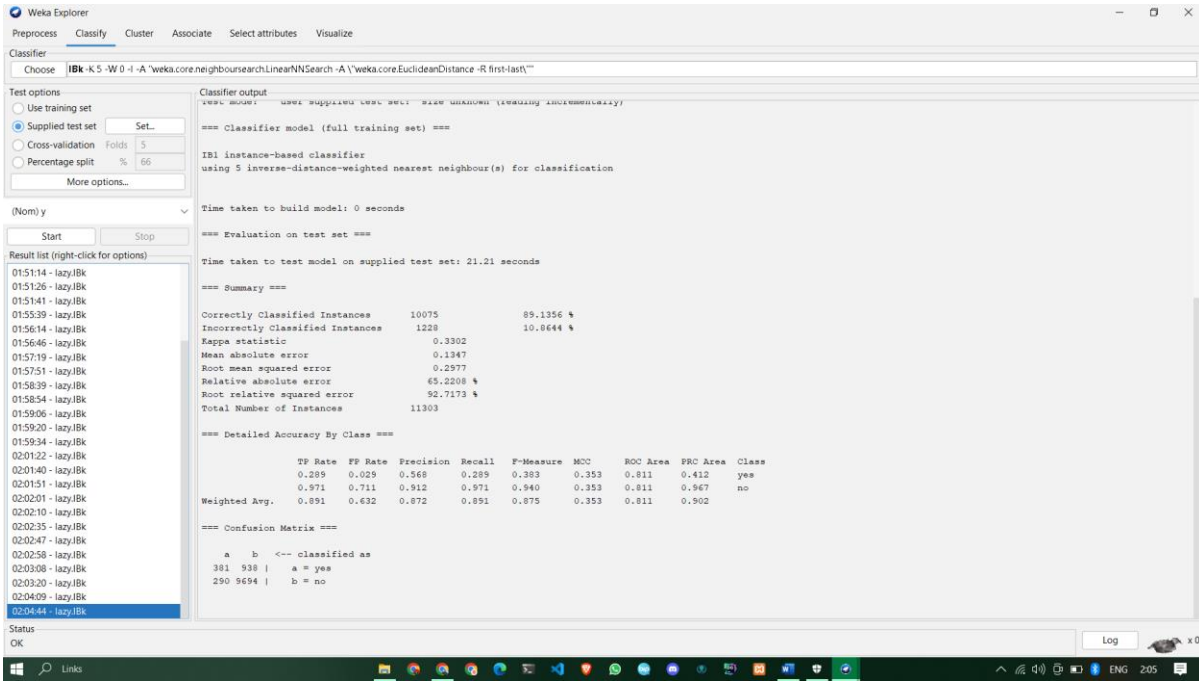
2. Uraikan hasil klasifikasi yang diperoleh.

Proses klasifikasi menggunakan Classifier IBk (Lazy/KNN) dengan k=3, k=5, k=7, k=9,dan k=11 serta pembobotan dalam voting penentuan class adalah distanceWeighting =Weight by 1/distance.

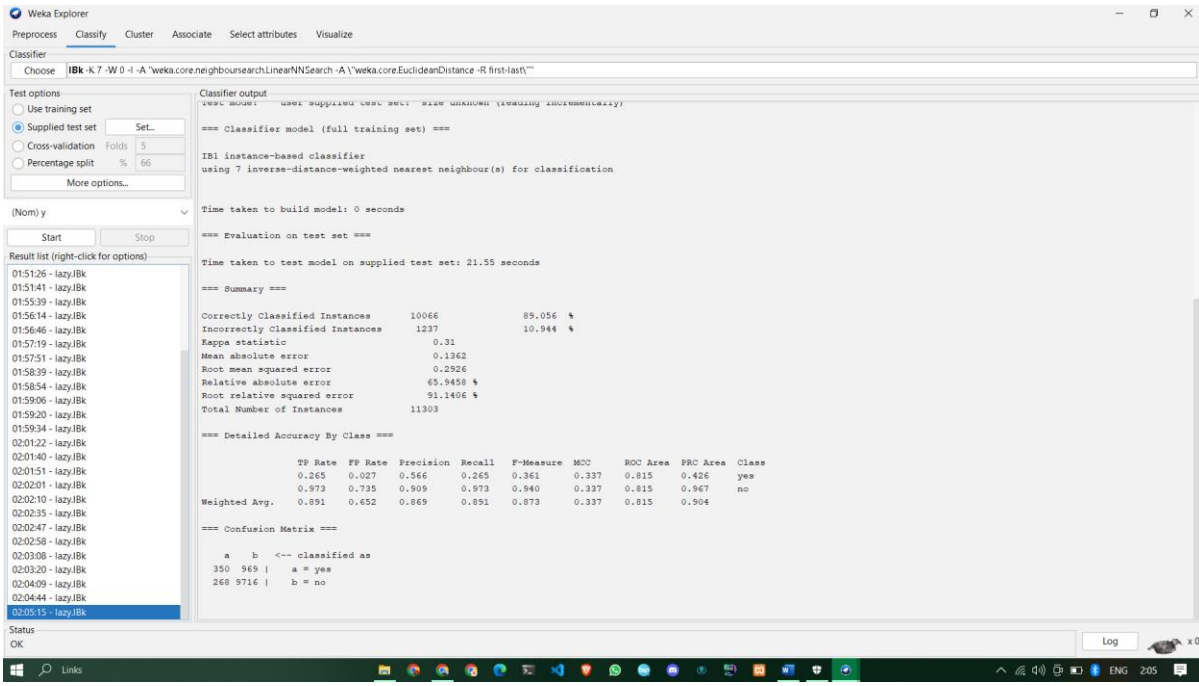
a. Supplied test set bank-full-testing.arff pada k = 3



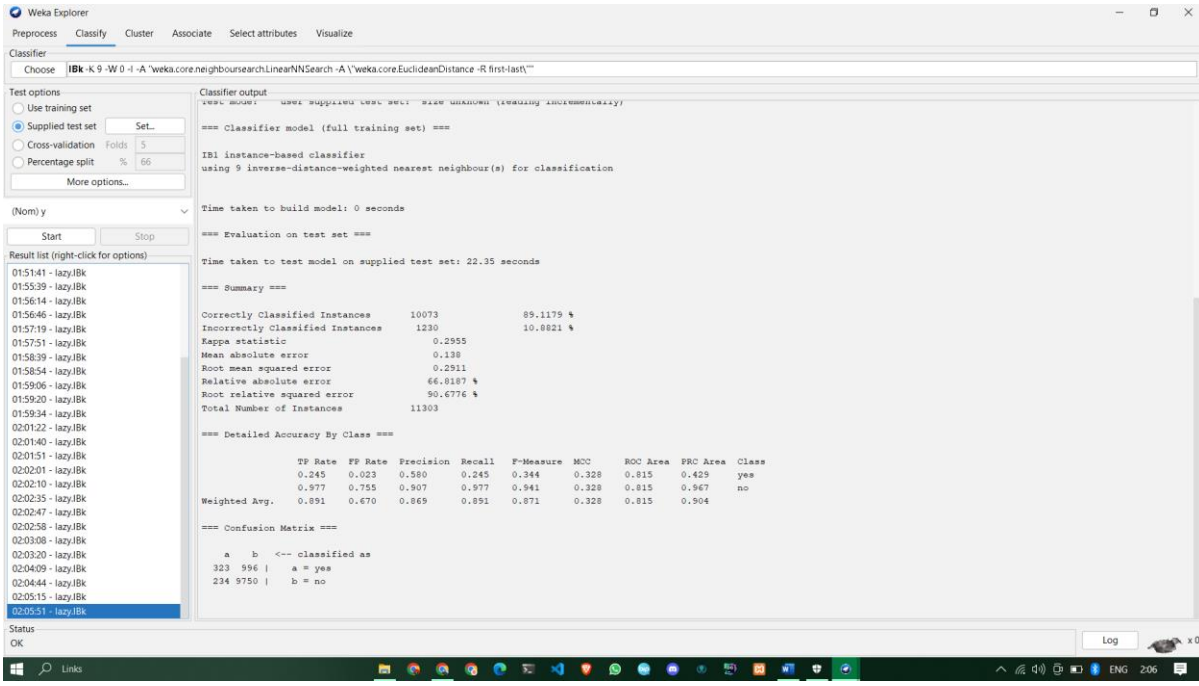
b. Supplied test set bank-full-testing.arff pada k = 5



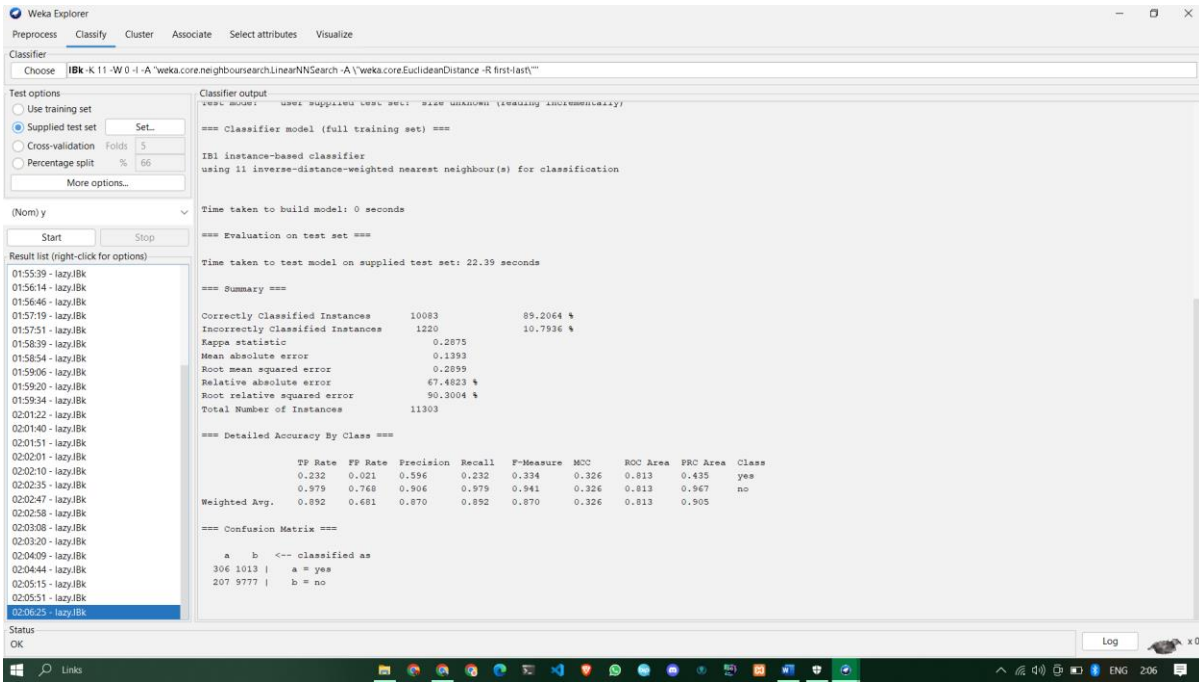
c. Supplied test set bank-full-testing.arff pada k = 7



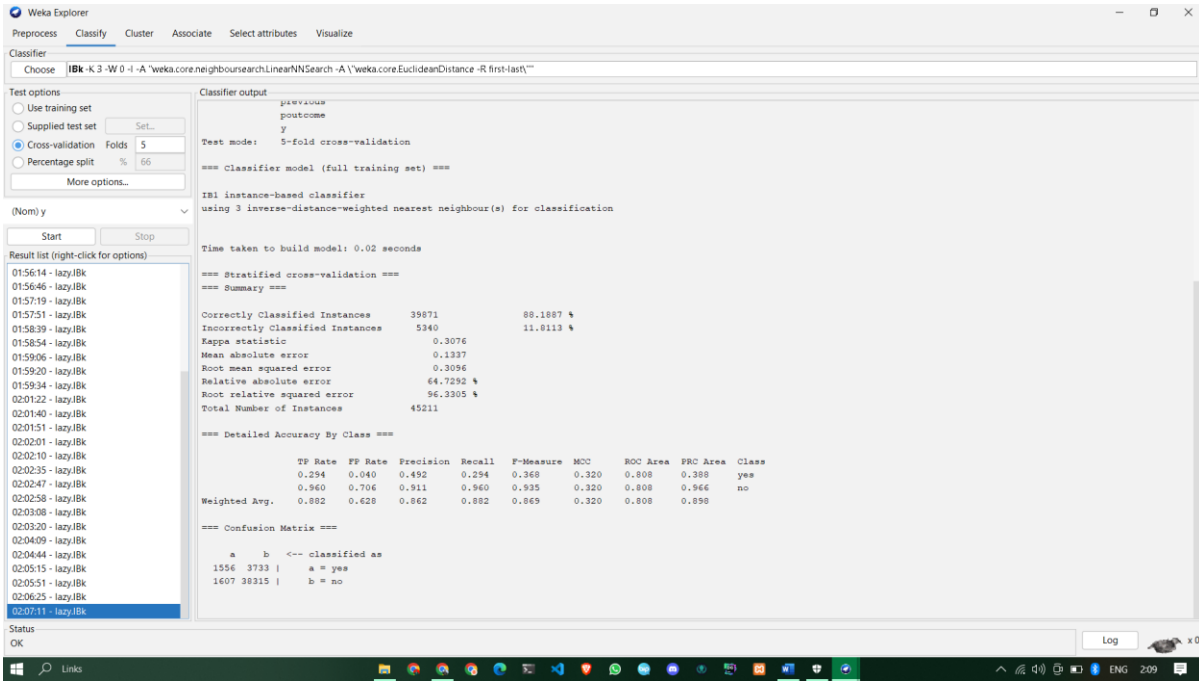
d. Supplied test set bank-full-testing.arff pada k = 9



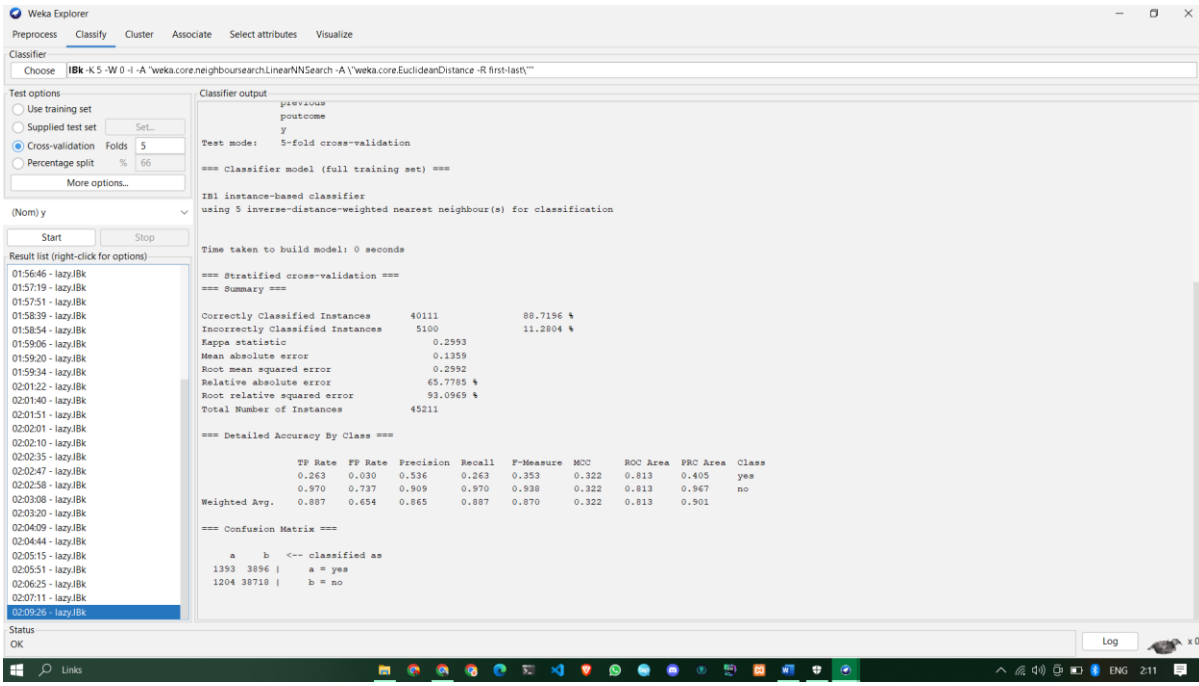
e. Supplied test set bank-full-testing.arff pada k = 11



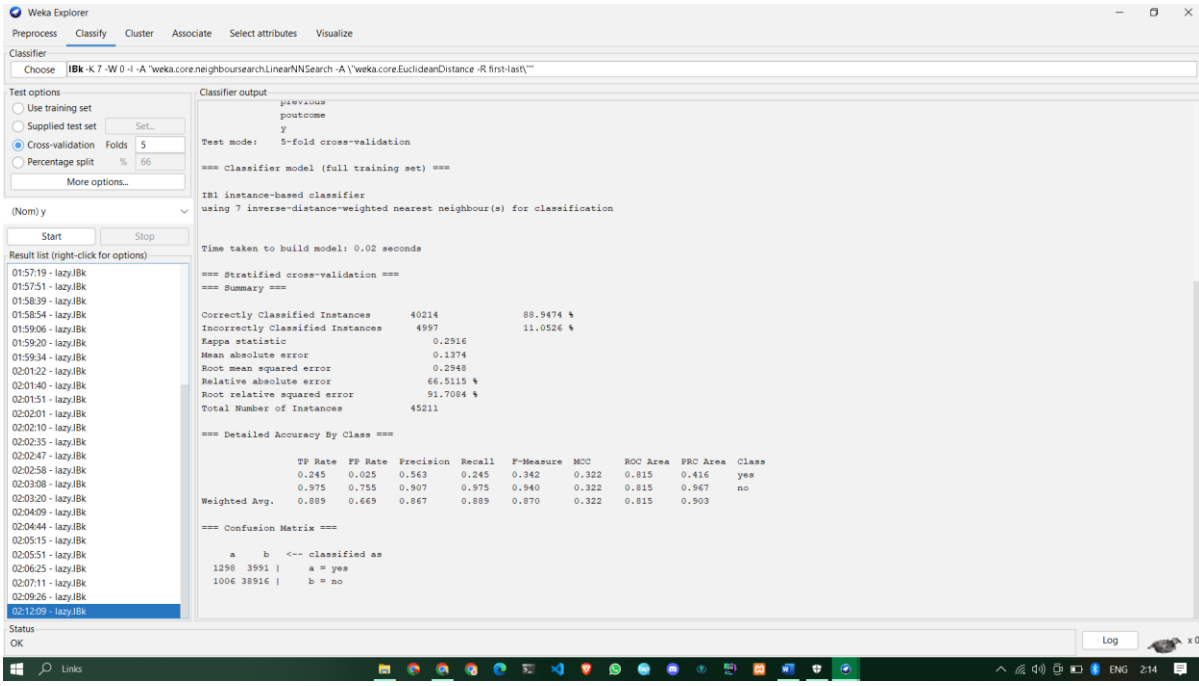
f. Cross-validation dengan file bank-full.arff pada k = 3



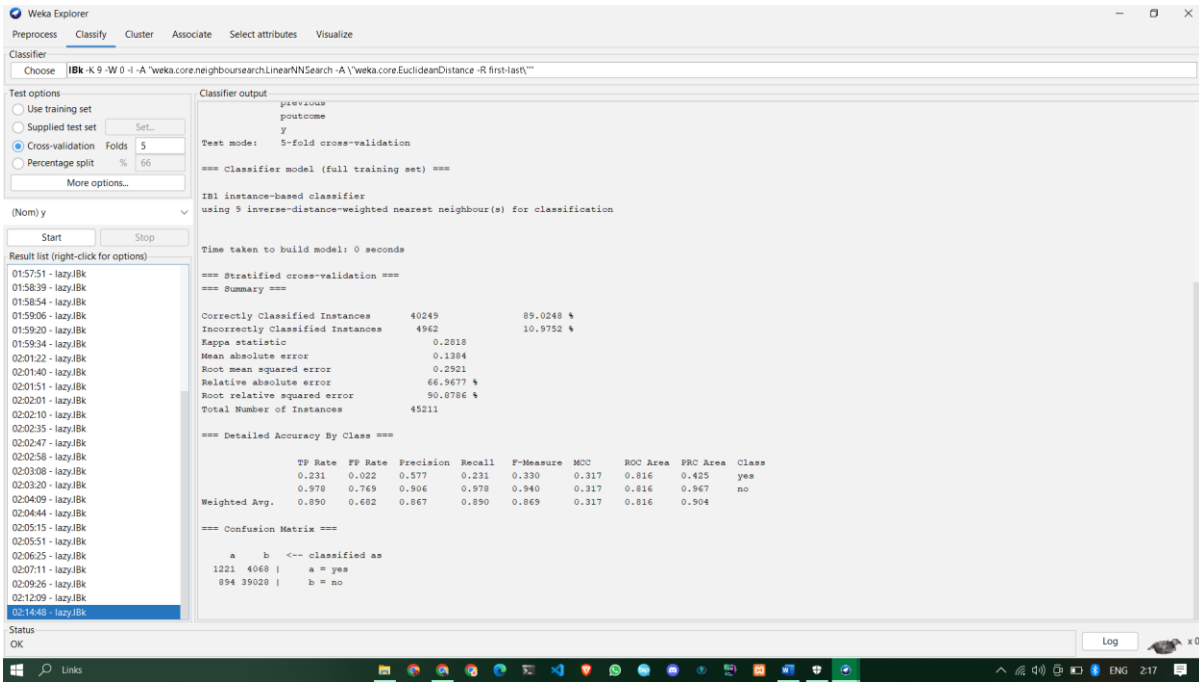
g. Cross-validation dengan file bank-full.arff pada k = 5



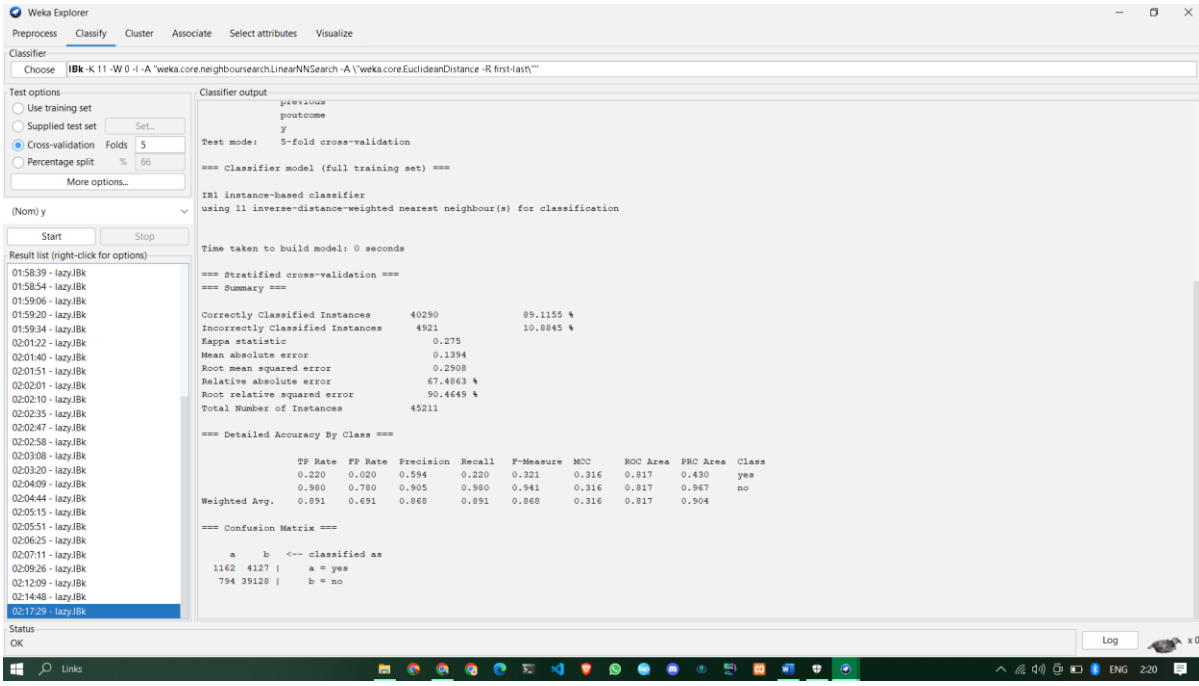
h. Cross-validation dengan file bank-full.arff pada k = 7



i. Cross-validation dengan file bank-full.arff pada k = 9



j. Cross-validation dengan file bank-full.arff pada k = 11



3. Kesimpulan.

Precision terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 5 berada pada angka 0.872. Recall terbesar yang di dapat dari uji coba adalah pada cross-validation dengan k = 11 berada pada angka 0.892. F-Measure terbesar yang di dapat dari uji coba adalah pada supplied test set dengan k = 5 berada pada angka 0.875. Pada kasus dataset Bank Marketing, menggunakan ke-5 k menunjukkan hasil pengujian yang tidak memiliki selisih yang cukup besar baik pada pengujian supplied test maupun cross-validation. Nilai k yang digunakan pada supplied test lebih baik digunakan pada k = 5, lebih dari itu menghasilkan akurasi yang lebih rendah, namun pada cross-validation cenderung lebih baik menggunakan k yang lebih besar seperti 9 atau 11.