

Τεχνητή Νοημοσύνη - Εργασία 3

Μηχανική Μάθηση – 2021

Σύνολο δεδομένων Titanic

Ονοματεπώνυμο: Μπαρακλilής Ιωάννης

AEM: 3685

email: imparakl@csd.auth.gr

Username στο kaggle: "Ioannis Baraklilis".

Άσκηση 1: Classification – Ταξινόμηση

Ερώτημα 1) Μεταβλητή με τον μεγαλύτερο αριθμό ελλιπών τιμών

Η μεταβλητή με τον μεγαλύτερο αριθμό ελλιπών τιμών είναι η μεταβλητή "cabin". Αυτό διαπιστώνεται στο Weka στις πληροφορίες σχετικά με την επιλεγμένη μεταβλητή (καρτέλα preprocess αφού έχουμε φορτώσει το αρχείο "titanic-train.arff" και επιλέξουμε την μεταβλητή "cabin") όπου δηλώνεται η τιμή της λείπει από 687 δείγματα (στα 891 συνολικά ή εναλλακτικά σε ποσοστό 77%).

| Selected attribute | | |
|--------------------|---------------|-------------------|
| Name: cabin | Type: Nominal | |
| Missing: 687 (77%) | Distinct: 147 | Unique: 101 (11%) |

Προβολή λεπτομερειών επιλεγμένης μεταβλητής στο Weka

Ερώτημα 2) Απαρίθμηση συνεχών και κατηγορικών μεταβλητών

Μπορούμε να δούμε τον τύπο μίας μεταβλητής στο τμήμα πληροφοριών της επιλεγμένης μεταβλητής στην καρτέλα preprocess, τμήμα "Selected attribute" αφού έχουμε φορτώσει το αρχείο "titanic-train.arff" και επιλέξουμε την επιθυμητή μεταβλητή. Το πεδίο "Type: ..." μας δίνει την ζητούμενη πληροφορία του τύπου. Ο τύπος είναι είτε Numeric (συνεχής) είτε Nominal (κατηγορικός).

Συνεχείς (Numeric) είναι οι μεταβλητές:

- age

| Selected attribute | | |
|--------------------|---------------|-----------------|
| Name: age | Type: Numeric | |
| Missing: 177 (20%) | Distinct: 88 | Unique: 16 (2%) |

- sibsp

| Selected attribute | | |
|--------------------|---------------|----------------|
| Name: sibsp | Type: Numeric | |
| Missing: 0 (0%) | Distinct: 7 | Unique: 0 (0%) |

- parch

| Selected attribute | | |
|--------------------|---------------|----------------|
| Name: parch | Type: Numeric | |
| Missing: 0 (0%) | Distinct: 7 | Unique: 1 (0%) |

- fare

| Selected attribute | | |
|--------------------|---------------|------------------|
| Name: fare | Type: Numeric | |
| Missing: 0 (0%) | Distinct: 248 | Unique: 97 (11%) |

Κατηγορικές (Nominal) είναι οι μεταβλητές:

- pclass

| Selected attribute | | |
|--------------------|-------------|----------------|
| Name: pclass | Distinct: 3 | Type: Nominal |
| Missing: 0 (0%) | | Unique: 0 (0%) |

- sex

| Selected attribute | | |
|--------------------|-------------|----------------|
| Name: sex | Distinct: 2 | Type: Nominal |
| Missing: 0 (0%) | | Unique: 0 (0%) |

- ticket

| Selected attribute | | |
|--------------------|---------------|-------------------|
| Name: ticket | Distinct: 681 | Type: Nominal |
| Missing: 0 (0%) | | Unique: 547 (61%) |

- cabin

| Selected attribute | | |
|--------------------|---------------|-------------------|
| Name: cabin | Distinct: 147 | Type: Nominal |
| Missing: 687 (77%) | | Unique: 101 (11%) |

- embarked

| Selected attribute | | |
|--------------------|-------------|----------------|
| Name: embarked | Distinct: 3 | Type: Nominal |
| Missing: 2 (0%) | | Unique: 0 (0%) |

- survived

| Selected attribute | | |
|--------------------|-------------|----------------|
| Name: survived | Distinct: 2 | Type: Nominal |
| Missing: 0 (0%) | | Unique: 0 (0%) |

Ερώτημα 3) Εφαρμογή αλγορίθμων ταξινόμησης για πρόβλεψη μεταβλητής “survived”

Για εκπαίδευση του μοντέλου (εφαρμογή αλγορίθμων ταξινόμησης) μεταβαίνουμε στην καρτέλα “Classify”, επιλέγουμε τον αλγόριθμο ταξινόμησης πατώντας το κουμπί “Choose” του τμήματος “Classifier”, επιλογή στο τμήμα “Test Options” το “Use training set”, επιλογή μεταβλητής πρόβλεψης την “survived” από την λίστα πάνω από κουμπί “Start” και μετά επιλογή του κουμπιού “Start”.

α. Αλγόριθμος Logistic Regression (Logistic)

Με εφαρμογή του αλγορίθμου αυτού έχουμε σαν (στατιστικό) αποτέλεσμα (τα υπόλοιπα αποτελέσματα δεν φαίνονται λόγω του πλήθους τους) από το Weka:

```
Time taken to build model: 9.13 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.15 seconds

=== Summary ===

Correctly Classified Instances      883          99.1021 %
Incorrectly Classified Instances      8           0.8979 %
Kappa statistic                    0.981
Mean absolute error                 0.0135
Root mean squared error             0.0814
Relative absolute error             2.8443 %
Root relative squared error         16.7381 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.991    0.009    0.995     0.991    0.993     0.981    1.000    1.000    0
          0.991    0.009    0.985     0.991    0.988     0.981    1.000    0.999    1
Weighted Avg.   0.991    0.009    0.991     0.991    0.991     0.981    1.000    1.000

=== Confusion Matrix ===

  a  b  <-- classified as
544  5 | a = 0
 3 339 | b = 1
```

b. Αλγόριθμος Dec. (Decision) Trees (J48)

Με εφαρμογή του αλγορίθμου αυτού έχουμε σαν (στατιστικά) αποτελέσματα (τα υπόλοιπα αποτελέσματα δεν φαίνονται λόγω του πλήθους τους) από το Weka:

```
=== Classifier model (full training set) ===

J48 pruned tree
-----

sex = male
| parch <= 0: 0 (484.0/80.0)
| parch > 0
| | age <= 3: 1 (19.69/6.27)
| | age > 3: 0 (73.31/15.58)
sex = female: 1 (314.0/81.0)

Number of Leaves :    4

Size of the tree :    7

Time taken to build model: 0.26 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances      709           79.5735 %
Incorrectly Classified Instances    182           20.4265 %
Kappa statistic                    0.5658
Mean absolute error                 0.3219
Root mean squared error            0.4008
Relative absolute error             68.0521 %
Root relative squared error        82.4144 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.843    0.281    0.828     0.843    0.836     0.566    0.786    0.802     0
               0.719    0.157    0.741     0.719    0.730     0.566    0.786    0.654     1
Weighted Avg.   0.796    0.233    0.795     0.796    0.795     0.566    0.786    0.745

=== Confusion Matrix ===

  a  b  <-- classified as
463 86 |  a = 0
 96 246 |  b = 1
```

c. Αλγόριθμος KNN (K Nearest Neighbors / IBk) με k = 10

Με εφαρμογή του αλγορίθμου αυτού μετά την αλλαγή της μεταβλητής του αριθμού των γειτόνων σε 10 (με δεξί κλικ στο πλαίσιο με το όνομα του αλγορίθμου > επιλογή "Show Properties" > αλλαγή πεδίου "KNN" σε 10) έχουμε σαν (στατιστικά) αποτελέσματα (τα υπόλοιπα αποτελέσματα δεν φαίνονται λόγω του πλήθους τους) από το Weka:

```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 10 nearest neighbour(s) for classification

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.51 seconds

=== Summary ===

Correctly Classified Instances      757           84.9607 %
Incorrectly Classified Instances    134           15.0393 %
Kappa statistic                    0.6697
Mean absolute error                 0.2406
Root mean squared error            0.3342
Relative absolute error            50.8692 %
Root relative squared error        68.7103 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.940    0.295    0.836     0.940    0.885     0.679    0.919    0.936     0
               0.705    0.060    0.880     0.705    0.782     0.679    0.919    0.876     1
Weighted Avg.   0.850    0.205    0.853     0.850    0.846     0.679    0.919    0.913

=== Confusion Matrix ===

  a  b  <-- classified as
516 33 |  a = 0
101 241 |  b = 1
```

Ερώτημα 4) Ποσοστό παραδειγμάτων σωστής ταξινόμησης για πρόβλεψη μεταβλητής “survived”

Η πληροφορία αυτή δίνεται από το πεδίο “Correctly Classified Instances” της κατηγορίας “Summary” των στατιστικών αποτελεσμάτων που δίνονται μετά την εκτέλεση ενός αλγορίθμου (εκπαίδευση μοντέλου). Τα αποτελέσματα αυτά φαίνονται στα στιγμιότυπα του προηγούμενου ερωτήματος.

Επομένως, έχουμε ότι από τα 891 συνολικά δείγματα ταξινομήθηκαν σωστά:

- a) Με τον αλγόριθμο Logistic Regression (Logistic), τα 883 με ποσοστό 99.1021 %.
- b) Με τον αλγόριθμο Dec. Trees (J48), τα 709 με ποσοστό 79.5735 %.
- c) Με τον αλγόριθμο KNN (IBk) με $k = 10$, τα 757 με ποσοστό 84.9607 %.

Ερώτημα 5) Αξιολόγηση επίδοσης μέσω μετρικής ακρίβειας accuracy (Correctly Classified Instances)

Η πληροφορία αυτή δίνεται από το πεδίο “Correctly Classified Instances” της κατηγορίας “Summary” των στατιστικών αποτελεσμάτων που δίνονται μετά την εκτέλεση ενός αλγορίθμου με την επιλογή Cross-validation ή Percentage split αντίστοιχα στο τμήμα “Test Options”. Τα αποτελέσματα αυτά φαίνονται στα παρακάτω στιγμιότυπα στα οποία φαίνονται τα (στατιστικά) αποτελέσματα των ελέγχων ταξινόμησης με Cross-validation (10 folds) ή Percentage split (66%) αντίστοιχα.

a. Με Percentage Split 66%

- a) Αλγόριθμος Logistic Regression (Logistic): Ακρίβεια (accuracy): 80.8581 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο “Correctly Classified Instances”):

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances      245          80.8581 %
Incorrectly Classified Instances    58          19.1419 %
Kappa statistic                    0.5893
Mean absolute error                 0.1915
Root mean squared error            0.4348
Relative absolute error             40.5432 %
Root relative squared error        89.7185 %
Total Number of Instances         303

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.857    0.272    0.839     0.857    0.848     0.590    0.808    0.840    0
      0.728    0.143    0.755     0.728    0.741     0.590    0.804    0.674    1
Weighted Avg.   0.809    0.223    0.807     0.809    0.808     0.590    0.806    0.778

=== Confusion Matrix ===

  a  b  <-- classified as
162 27 | a = 0
 31 83 | b = 1
```

- b) Αλγόριθμος Dec. Trees (J48): Ακρίβεια (accuracy): 78.5479 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο “Correctly Classified Instances”):

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances      238          78.5479 %
Incorrectly Classified Instances    65          21.4521 %
Kappa statistic                    0.5324
Mean absolute error                 0.3017
Root mean squared error            0.3906
Relative absolute error             63.8786 %
Root relative squared error        80.6043 %
Total Number of Instances         303

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.862    0.342    0.807     0.862    0.834     0.535    0.781    0.794    0
      0.658    0.138    0.743     0.658    0.698     0.535    0.781    0.722    1
Weighted Avg.   0.785    0.265    0.783     0.785    0.783     0.535    0.781    0.767

=== Confusion Matrix ===

  a  b  <-- classified as
163 26 | a = 0
 39 75 | b = 1
```

c) Αλγόριθμος KNN (IBk) με $k = 10$: Ακρίβεια (accuracy): 79.538 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.22 seconds

=== Summary ===

Correctly Classified Instances      241          79.538 %
Incorrectly Classified Instances    62           20.462 %
Kappa statistic                    0.5532
Mean absolute error                0.2777
Root mean squared error            0.3933
Relative absolute error            58.7927 %
Root relative squared error        81.1629 %
Total Number of Instances         303

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.873    0.333    0.813     0.873    0.842     0.556    0.825    0.833     0
          0.667    0.127    0.760     0.667    0.710     0.556    0.825    0.775     1
Weighted Avg.    0.795    0.256    0.793     0.795    0.792     0.556    0.825    0.811

=== Confusion Matrix ===

  a    b  <-- classified as
165  24 |  a = 0
 38  76 |  b = 1
```

b. Με 10-cross validation

a) Αλγόριθμος Logistic Regression (Logistic): Ακρίβεια (accuracy): 74.4108 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
Time taken to build model: 8.26 seconds

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

Correctly Classified Instances      663          74.4108 %
Incorrectly Classified Instances    228          25.5892 %
Kappa statistic                    0.4443
Mean absolute error                0.2574
Root mean squared error            0.491
Relative absolute error            54.4109 %
Root relative squared error        100.9558 %
Total Number of Instances         891

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.836    0.404    0.769     0.836    0.801     0.447    0.804    0.837     0
          0.596    0.164    0.694     0.596    0.642     0.447    0.804    0.747     1
Weighted Avg.    0.744    0.312    0.740     0.744    0.740     0.447    0.804    0.802

=== Confusion Matrix ===

  a    b  <-- classified as
459  90 |  a = 0
138 204 |  b = 1
```

b) Αλγόριθμος Dec. Trees (J48): Ακρίβεια (accuracy): 77.4411 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
Time taken to build model: 0.18 seconds

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

Correctly Classified Instances      690          77.4411 %
Incorrectly Classified Instances    201          22.5589 %
Kappa statistic                    0.5076
Mean absolute error                0.3234
Root mean squared error            0.407
Relative absolute error            68.3639 %
Root relative squared error        83.6962 %
Total Number of Instances         891

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.869    0.377    0.787     0.869    0.826     0.513    0.768    0.789     0
          0.623    0.131    0.747     0.623    0.679     0.513    0.768    0.687     1
Weighted Avg.    0.774    0.283    0.772     0.774    0.770     0.513    0.768    0.750

=== Confusion Matrix ===

  a    b  <-- classified as
477  72 |  a = 0
129 213 |  b = 1
```

c) Αλγόριθμος KNN (IBk) με $k = 10$: Ακρίβεια (accuracy): 80.3591 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances      716          80.3591 %
Incorrectly Classified Instances    175          19.6409 %
Kappa statistic                    0.5669
Mean absolute error                 0.2679
Root mean squared error             0.3723
Relative absolute error             56.6185 %
Root relative squared error         76.5589 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.909    0.365    0.800    0.909    0.851    0.577    0.856    0.866    0
      0.635    0.091    0.813    0.635    0.713    0.577    0.856    0.811    1
Weighted Avg.    0.804    0.260    0.805    0.804    0.798    0.577    0.856    0.845

=== Confusion Matrix ===
      a  b  <-- classified as
499  50 |  a = 0
125 217 |  b = 1
```

Ερώτημα 6) Επιλογή από τους παραπάνω για πρόβλεψη σε άγνωστα δεδομένα.

Από τους 3 (παραπάνω) αλγορίθμους, για πρόβλεψη σε άγνωστα δεδομένα θα διάλεγα τον αλγόριθμο των K κοντινότερων γειτόνων (KNN / K Nearest Neighbors / IBk). Αυτό, γιατί η επίδοση του στην περίπτωση της μέτρησης με Percentage split (66%) είναι μέση και όμοια με των άλλων αλγορίθμων, ενώ στην περίπτωση του Cross-validation (με 10 Folds) είναι σαφώς ανώτερη των άλλων δύο αλγορίθμων.

Συγκεκριμένα, όπως μπορεί κάποιος να δει:

- Στην περίπτωση μέτρησης ακρίβειας με Percentage split (66%) έχουμε:
Ο αλγόριθμος KNN έχει ακρίβεια 79.538% που είναι ανώτερο από την ακρίβεια του Dec. Trees με ακρίβεια 78.5479% (ανώτερο κατά 0.9901%) και κατώτερο από την ακρίβεια του Logistic Regression με ακρίβεια 80.8581% (κατώτερο κατά 1.3201%).
- Στην περίπτωση μέτρησης ακρίβειας με Cross-validation (με 10 Folds) έχουμε:
Ο αλγόριθμος KNN έχει ακρίβεια 80.3591% που είναι ανώτερο από την ακρίβεια του Dec. Trees με ακρίβεια 77.4411% (ανώτερο κατά 2.918%) και ανώτερο από την ακρίβεια του Logistic Regression με ακρίβεια 74.4108 % (ανώτερο κατά 5.9483%).

a. Πειραματισμός με επιπλέον αλγορίθμους.

- Αλγόριθμος ZeroR:

Ακρίβεια (accuracy): 61.6162 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

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

Correctly Classified Instances      549          61.6162 %
Incorrectly Classified Instances    342          38.3838 %
Kappa statistic                    0
Mean absolute error                 0.4731
Root mean squared error             0.4863
Relative absolute error             100 %
Root relative squared error         100 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1.000    1.000    0.616    1.000    0.763    ?      0.497    0.615    0
      0.000    0.000    ?      0.000    ?      ?      0.497    0.382    1
Weighted Avg.    0.616    0.616    ?      0.616    ?      ?      0.497    0.525

=== Confusion Matrix ===
      a  b  <-- classified as
549  0 |  a = 0
342  0 |  b = 1
```

- Αλγόριθμος SimpleLogistic:

Ακρίβεια (accuracy): 81.1448 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

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

Correctly Classified Instances      723          81.1448 %
Incorrectly Classified Instances    168          18.8552 %
Kappa statistic                    0.5938
Mean absolute error                 0.2765
Root mean squared error             0.3742
Relative absolute error             58.4489 %
Root relative squared error        76.9492 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| | 0.878 | 0.295 | 0.827 | 0.878 | 0.852 | 0.596 | 0.855 | 0.872 | 0 |
| | 0.705 | 0.122 | 0.782 | 0.705 | 0.742 | 0.596 | 0.855 | 0.823 | 1 |
| Weighted Avg. | 0.811 | 0.229 | 0.810 | 0.811 | 0.809 | 0.596 | 0.855 | 0.853 | |

```
=== Confusion Matrix ===

 a  b  <-- classified as
482 67 | a = 0
101 241 | b = 1
```

- Αλγόριθμος JRip:

Ακρίβεια (accuracy): 81.4815%.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

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

Correctly Classified Instances      726          81.4815 %
Incorrectly Classified Instances    165          18.5185 %
Kappa statistic                    0.5897
Mean absolute error                 0.2803
Root mean squared error             0.3835
Relative absolute error             59.2441 %
Root relative squared error        78.8526 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| | 0.925 | 0.363 | 0.804 | 0.925 | 0.860 | 0.603 | 0.787 | 0.794 | 0 |
| | 0.637 | 0.075 | 0.842 | 0.637 | 0.725 | 0.603 | 0.787 | 0.761 | 1 |
| Weighted Avg. | 0.815 | 0.252 | 0.818 | 0.815 | 0.809 | 0.603 | 0.787 | 0.781 | |

```
=== Confusion Matrix ===

 a  b  <-- classified as
508 41 | a = 0
124 218 | b = 1
```

- Αλγόριθμος SGD:

Ακρίβεια (accuracy): 83.5017 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

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

Correctly Classified Instances      744          83.5017 %
Incorrectly Classified Instances    147          16.4983 %
Kappa statistic                    0.6483
Mean absolute error                 0.165
Root mean squared error             0.4062
Relative absolute error             34.874 %
Root relative squared error        83.5209 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| | 0.880 | 0.237 | 0.856 | 0.880 | 0.868 | 0.649 | 0.821 | 0.828 | 0 |
| | 0.763 | 0.120 | 0.798 | 0.763 | 0.780 | 0.649 | 0.821 | 0.700 | 1 |
| Weighted Avg. | 0.835 | 0.192 | 0.834 | 0.835 | 0.834 | 0.649 | 0.821 | 0.779 | |

```
=== Confusion Matrix ===

 a  b  <-- classified as
483 66 | a = 0
81 261 | b = 1
```

- Αλγόριθμος Random Forest:
Ακρίβεια (accuracy): 76.7677 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο “Correctly Classified Instances”):

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 1.47 seconds

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

Correctly Classified Instances      684          76.7677 %
Incorrectly Classified Instances    207          23.2323 %
Kappa statistic                    0.4799
Mean absolute error                 0.3528
Root mean squared error             0.399
Relative absolute error             74.5648 %
Root relative squared error         82.0388 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| | 0.903 | 0.450 | 0.763 | 0.903 | 0.827 | 0.496 | 0.847 | 0.883 | 0 |
| | 0.550 | 0.097 | 0.780 | 0.550 | 0.645 | 0.496 | 0.847 | 0.793 | 1 |
| Weighted Avg. | 0.768 | 0.315 | 0.770 | 0.768 | 0.757 | 0.496 | 0.847 | 0.849 | |

```

=== Confusion Matrix ===
  a  b  <-- classified as
496  53 | a = 0
154 188 | b = 1

```

Όπως μπορούμε να δούμε, από τους επιπλέον αλγορίθμους, καλύτερη απόδοση επιτυγχάνει ο αλγόριθμος SGD με ακρίβεια 83.5017%.

Ερώτημα 7) Χρήση φίλτρου “ReplaceMissingValues”

Η εφαρμογή του φίλτρου “ReplaceMissingValues” γίνεται με επιλογή φίλτρου στο τμήμα “Filter” της καρτέλας “Preprocess” (το φίλτρο “ReplaceMissingValues” βρίσκεται στο μονοπάτι `weka > filters > unsupervised > attribute > ReplaceMissingValues`) και ενεργοποίηση του (επιλογή Apply) πριν την εφαρμογή των αλγορίθμων.

Αποτελέσματα:

- Με τον αλγόριθμο Logistic Regression (Logistic), ταξινομεί σωστά τα 883/891 με ποσοστό 99.1021 %.

Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο “Correctly Classified Instances”):

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances      883          99.1021 %
Incorrectly Classified Instances     6          0.8979 %
Kappa statistic                    0.981
Mean absolute error                 0.0135
Root mean squared error             0.0814
Relative absolute error             2.8443 %
Root relative squared error         16.7381 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| | 0.991 | 0.009 | 0.995 | 0.991 | 0.993 | 0.981 | 1.000 | 1.000 | 0 |
| | 0.991 | 0.009 | 0.985 | 0.991 | 0.988 | 0.981 | 1.000 | 0.999 | 1 |
| Weighted Avg. | 0.991 | 0.009 | 0.991 | 0.991 | 0.991 | 0.981 | 1.000 | 1.000 | |

```

=== Confusion Matrix ===
  a  b  <-- classified as
544   5 | a = 0
  3 339 | b = 1

```

Δεν παρατηρείται διαφορά στο ποσοστό των παραδειγμάτων που ταξινομείται σωστά (99.1021% με το φίλτρο έναντι 99.1021% χωρίς αυτό όπως είδαμε στο ερώτημα 4).

- b) Με τον αλγόριθμο Dec. Trees (J48), ταξινομεί σωστά τα 719/891 με ποσοστό 80.6958 %. Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances      719          80.6958 %
Incorrectly Classified Instances    172          19.3042 %
Kappa statistic                    0.5896
Mean absolute error                 0.3057
Root mean squared error             0.3909
Relative absolute error             64.6114 %
Root relative squared error         80.3864 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.852    0.266    0.837      0.852    0.845      0.590    0.799    0.809     0
                0.734    0.148    0.756      0.734    0.745      0.590    0.799    0.671     1
Weighted Avg.   0.807    0.221    0.806      0.807    0.806      0.590    0.799    0.756

=== Confusion Matrix ===

  a  b  <-- classified as
468 81 |  a = 0
 91 251 |  b = 1
```

Παρατηρούμε αύξηση στο ποσοστό των παραδειγμάτων που ταξινομείται σωστά (80.6958% με το φίλτρο έναντι 79.5735% χωρίς αυτό όπως είδαμε στο ερώτημα 4) κατά 1.1223%.

- c) Με τον αλγόριθμο KNN (Ibk) με $k = 10$, ταξινομεί σωστά τα 732/891 με ποσοστό 82.1549 %. Αυτό φαίνεται στα αποτελέσματα που δίνει το Weka (πεδίο "Correctly Classified Instances"):

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.3 seconds

=== Summary ===

Correctly Classified Instances      732          82.1549 %
Incorrectly Classified Instances    159          17.8451 %
Kappa statistic                    0.6023
Mean absolute error                 0.2304
Root mean squared error             0.3346
Relative absolute error             48.6993 %
Root relative squared error         68.793 %
Total Number of Instances          891

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.940    0.368    0.804      0.940    0.866      0.619    0.916    0.937     0
                0.632    0.060    0.867      0.632    0.731      0.619    0.916    0.874     1
Weighted Avg.   0.822    0.250    0.828      0.822    0.814      0.619    0.916    0.913

=== Confusion Matrix ===

  a  b  <-- classified as
516 33 |  a = 0
126 216 |  b = 1
```

Παρατηρούμε μείωση στο ποσοστό των παραδειγμάτων που ταξινομείται σωστά (82.1549% με το φίλτρο έναντι 84.9607% χωρίς αυτό όπως είδαμε στο ερώτημα 4) κατά 2.8058%.

Ερώτημα 8) Προβλέψεις στο Kaggle: Username: Ioannis Baraklilis.

Πρόβλεψη μεταβλητής survived στο test-set με το μοντέλο KNN με $K=10$.

Φορτώνουμε το αρχείο "titanic-test.arff" στην επιλογή "Supplied test set" του τμήματος "Test Options" (καρτέλα "Classifier") και επιλογή κάποιας μορφής εξόδου των προβλέψεων (επιλογή "More Options" > "Output Predictions" > "Choose" > κάποια επιλογή της κατηγορίας "prediction"). Προσωπική επιλογή το "PlainText".


Ένα τμήμα των προβλέψεων φαίνεται στην συνέχεια:

=== Predictions on test set ===

| inst# | actual | predicted | error | prediction |
|-------|--------|-----------|-------|------------|
| 1 | 1:? | 1:0 | 0.9 | |
| 2 | 1:? | 1:0 | 0.6 | |
| 3 | 1:? | 1:0 | 0.9 | |
| 4 | 1:? | 1:0 | 0.7 | |
| 5 | 1:? | 1:0 | 0.5 | |
| 6 | 1:? | 1:0 | 0.9 | |
| 7 | 1:? | 1:0 | 0.5 | |
| 8 | 1:? | 1:0 | 0.7 | |
| 9 | 1:? | 2:1 | 0.6 | |
| 10 | 1:? | 1:0 | 0.9 | |

στην τρίτη στήλη βρίσκονται οι προβλέψεις όπου η τιμή "1:0" αντιστοιχεί σε μη-επιβίωση ενώ η "2:1" σε επιβίωση. Μετά την κατάλληλη μορφοποίηση των τιμών αυτών σε κατάλληλη σύμφωνα με τις οδηγίες του Kaggle γίνεται υποβολή.

Βαθμολογία: 0.78229 / 1, δηλαδή προβλέφθηκε σωστά το 78.229% των περιπτώσεων:

| | | | | |
|------|--------------------|---------------------------------------------------------------------------------------|---------|---|
| 9010 | Ioannis Baraklilis |  | 0.78229 | 1 |
|------|--------------------|---------------------------------------------------------------------------------------|---------|---|

Αυτό είναι αναμενόμενο, εφόσον στην αξιολόγηση της επίδοσης του αλγορίθμου KNN (με $K = 10$) υπολογίστηκε η ακρίβεια του σε 79.538% και 80.3591% σε percentage split και cross-validation αντίστοιχα.

Άσκηση 2: Clustering – Ομαδοποίηση

α) Εκτέλεση του αλγορίθμου k-Means για $k = 3$

Αρχικά, αφαιρώ την μεταβλητή survived γιατί δεν χρειάζεται και πιθανώς να δημιουργήσει πρόβλημα καθώς μπορεί να οδηγήσει στην δημιουργία ομάδων με βάση την μεταβλητή survived.

Στην συνέχεια εκτελώ τον αλγόριθμο SimpleKMeans αφού ορίσω τον αριθμό των ομάδων σε 3 (με δεξί κλικ στο πλαίσιο με το όνομα του αλγορίθμου > επιλογή "Show Properties" > αλλαγή πεδίου "numClusters" σε 3).

Το αποτέλεσμα που δίνει το Weka φαίνεται στην συνέχεια:

Final cluster centroids:

| Attribute | Cluster# | | | |
|-----------|----------------------|--------------|--------------|--------------|
| | Full Data (891.0) | 0 (425.0) | 1 (256.0) | 2 (210.0) |
| ===== | | | | |
| pclass | 3 | 3 | 3 | 1 |
| sex | male | male | female | male |
| age | 29.6991 | 26.8524 | 25.1557 | 40.999 |
| sibsp | 0.523 | 0.4612 | 0.7305 | 0.3952 |
| parch | 0.3816 | 0.2094 | 0.7617 | 0.2667 |
| ticket | 347082.0 | 1601.0 | 347082.0 | 19950.0 |
| fare | 32.2042 | 13.8454 | 29.0102 | 73.2527 |
| cabin | G6 | G6 | G6 | G6 |
| embarked | S | S | S | S |

Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

```
0      425 ( 48%)
1      256 ( 29%)
2      210 ( 24%)
```

Βλέπουμε ότι δημιουργήθηκαν οι 3 ομάδες με:

- Ομάδα 0 με μέγεθος 425 στοιχεία (48% των δειγμάτων).
- Ομάδα 1 με μέγεθος 256 στοιχεία (29% των δειγμάτων).
- Ομάδα 2 με μέγεθος 210 στοιχεία (24% των δειγμάτων).

Άσκηση 3: Association Rules – Κανόνες Συσχέτισης

Μετά την αφαίρεση των αριθμητικών (Numeric) χαρακτηριστικών (age, sibsp, parch, fare), στην καρτέλα “Associate” επιλέγω τον αλγόριθμο Apriori και θέτω τις παραμέτρους lowerBoundMinSupport σε 0.1 (είναι η προκαθορισμένη τιμή) ώστε τα αποτελέσματα να ικανοποιούν τις προϋποθέσεις της εκφώνησης, minMetric (δεδομένης της επιλογής “metricType” σε “Confidence”) σε 0.8 ώστε να βρεθούν συνολικά τουλάχιστον πάνω από 5 κανόνες αλλά εκείνοι με το μεγαλύτερη δυνατή τιμή εμπιστοσύνης (confidence) και το numRules σε 20 ώστε να “αναγκαστεί” να βρει όσο το δυνατόν περισσότερους κανόνες και να μην “σταματήσει” έχοντας μικρότερο αριθμό κανόνων που όμως δεν έχουν μέγιστες τιμές εμπιστοσύνης.

Τα αποτελέσματα της εκτέλεσης στο Weka φαίνονται παρακάτω:

Apriori

=====

Minimum support: 0.1 (89 instances)

Minimum metric <confidence>: 0.8

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 21

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 1

Best rules found:

```
1. pclass=1 sex=female 94 ==> survived=1 91    <conf:(0.97)> lift:(2.52) lev:(0.06) [54] conv:(14.48)
2. pclass=2 survived=0 97 ==> sex=male 91    <conf:(0.94)> lift:(1.45) lev:(0.03) [28] conv:(4.88)
3. pclass=2 sex=male 108 ==> embarked=S 97    <conf:(0.9)> lift:(1.24) lev:(0.02) [18] conv:(2.49)
4. pclass=2 184 ==> embarked=S 164    <conf:(0.89)> lift:(1.23) lev:(0.03) [31] conv:(2.43)
5. pclass=3 sex=male embarked=S 265 ==> survived=0 231    <conf:(0.87)> lift:(1.41) lev:(0.08) [67] conv:(2.91)
6. pclass=3 sex=male 347 ==> survived=0 300    <conf:(0.86)> lift:(1.4) lev:(0.1) [86] conv:(2.77)
7. survived=0 549 ==> sex=male 468    <conf:(0.85)> lift:(1.32) lev:(0.13) [112] conv:(2.36)
8. embarked=S survived=0 427 ==> sex=male 364    <conf:(0.85)> lift:(1.32) lev:(0.1) [87] conv:(2.35)
9. pclass=2 sex=male 108 ==> survived=0 91    <conf:(0.84)> lift:(1.37) lev:(0.03) [24] conv:(2.3)
10. sex=male embarked=S 441 ==> survived=0 364    <conf:(0.83)> lift:(1.34) lev:(0.1) [92] conv:(2.17)
11. sex=male 577 ==> survived=0 468    <conf:(0.81)> lift:(1.32) lev:(0.13) [112] conv:(2.01)
12. pclass=3 embarked=S 353 ==> survived=0 286    <conf:(0.81)> lift:(1.31) lev:(0.08) [68] conv:(1.99)
13. pclass=3 embarked=S survived=0 286 ==> sex=male 231    <conf:(0.81)> lift:(1.25) lev:(0.05) [45] conv:(1.8)
14. pclass=3 survived=0 372 ==> sex=male 300    <conf:(0.81)> lift:(1.25) lev:(0.07) [59] conv:(1.8)
```

Βλέπουμε ότι έχει βρει 14 κανόνες που είναι περισσότεροι από 5 που χρειαζόμαστε αλλά λιγότεροι από 20 της επιλογής, γεγονός που δείχνει ότι έχει βρει όλους τους κανόνες με τιμή εμπιστοσύνης μεγαλύτερη ή ίση του κατωφλίου εμπιστοσύνης 0.8 άρα σε αυτούς τους κανόνες περιλαμβάνονται και οι 5 με την μεγαλύτερη εμπιστοσύνη.

Μάλιστα, οι κανόνες που βρέθηκαν από τον αλγόριθμο είναι ταξινομημένοι ως προς την τιμή της εμπιστοσύνης κατά φθίνουσα σειρά. Επομένως, οι 5 κανόνες με την μεγαλύτερη εμπιστοσύνη (confidence) και υποστήριξη (support) ≥ 0.1 (εξασφαλίζεται λόγω της παραμέτρου lowerBoundMinSupport) είναι οι 5 πρώτοι:

- 1) pclass=1, sex=female ==> survived=1, με εμπιστοσύνη 0.97.
- 2) pclass=2, survived=0 ==> sex=male, με εμπιστοσύνη 0.94.
- 3) pclass=2, sex=male ==> embarked=S, με εμπιστοσύνη 0.90.
- 4) pclass=2 ==> embarked=S, με εμπιστοσύνη 0.89.
- 5) pclass=3, sex=male, embarked=S ==> survived=0, με εμπιστοσύνη 0.87.