PROJECT 2 ΡΟΜΠΟΤΙΚΗ-ΛΕΥΤΕΡΗΣ ΑΜΙΤΣΗΣ

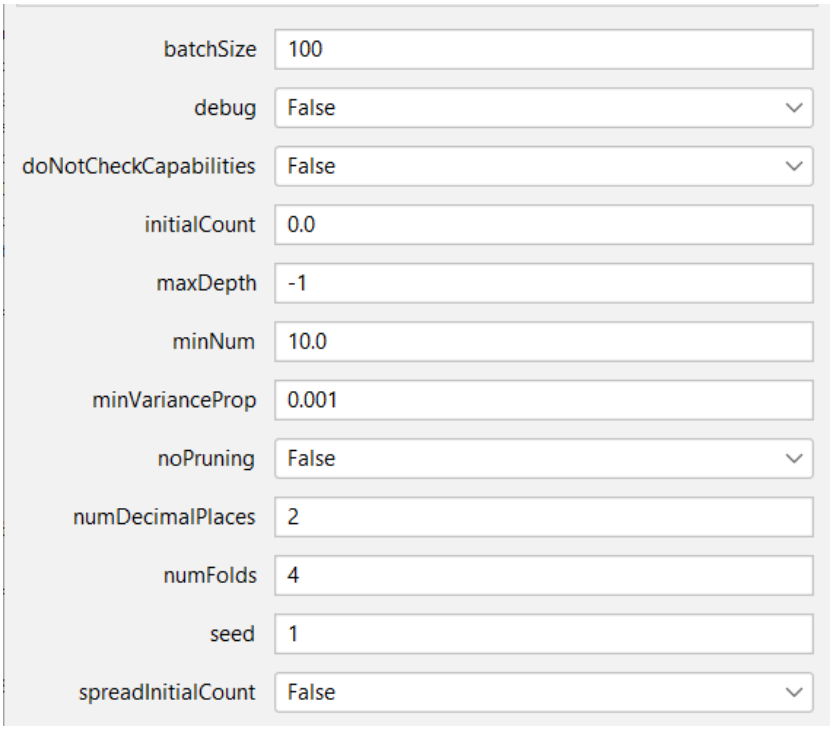
Εξαγωγή Κανόνων (WEKA)

- Προετοιμασία δεδομένων

Στα δεδομένα δεν πραγματοποιήθηκε καμία απολύτως αλλαγή. Δεν έλειπαν τιμές για να συμπληρωθούν και δεν υπήρχε πολύ μεγάλο εύρος τιμών ανά attribute οπότε η διακριτοποίηση έκανε περισσότερο κακό παρά καλό στο μοντέλο (μείωση σωστών προβλέψεων κατά 10%) . Η μόνη ίσως αλλαγή που θα μπορούσε να γίνει είναι να αφαιρεθεί η κλάση 3 που περιέχει πολύ μικρό αριθμό instances συγκριτικά με τις άλλες 3 κλάσεις (μόνο το 5% του dataset), παρόλα αυτά το accuracy ήταν ήδη πολύ υψηλό και η αύξηση του δεν ήταν αρκετά σημαντική ώστε να αφαιρεθεί.

WEKA REPTree

Παράμετροι:



BatchSize : Ο αριθμός των instances που επεξεργάζονται ταυτόχρονα σε κάθε iteration του αλγορίθμου.

debug : Εμφανίζει επιπλέον πληροφορίες σχετικά με την διαδικασία που εκτελέστηκε πράγμα που δεν ισχύει όμως στον REPTree.

doNotCheckCapabilities : Σε περίπτωση που το κάνουμε true δεν ελέγχεται το δέντρο προτού ολοκληρωθεί.

InitialCount - Spread Initial Count : Τα πεδία αυτά έχουν να κάνουν με τις γνώσεις πάνω στο dataset. Μέσω αυτών μπορούμε να επηρεάσουμε την δημιουργία του δέντρου. Δεν μπόρεσα να βρω παραπάνω πληροφορίες σχετικά με αυτά τα πεδία και τα μόνα πεδία που επηρέαζαν ήταν αυτά του υπολογισμού του λάθους.

maxDepth : Το μέγιστο βάθος του δέντρου.

minNum : Τα ελάχιστα instances που πρέπει να καλύπτει κάθε κανόνας.

minVarianceProp : Στο πεδίο αυτό ελέγχουμε το πόσο σύχνα θέλουμε να γίνονται splits στο δέντρο μας. Όταν το variance πέσει κάτω από την προβλεπόμενη τιμή δεν θα πραγματοποιηθούν περεταίρω διαχωρισμοί.

noPruning : Αν γίνει true δεν θα πραγματοποιηθεί το pruning.

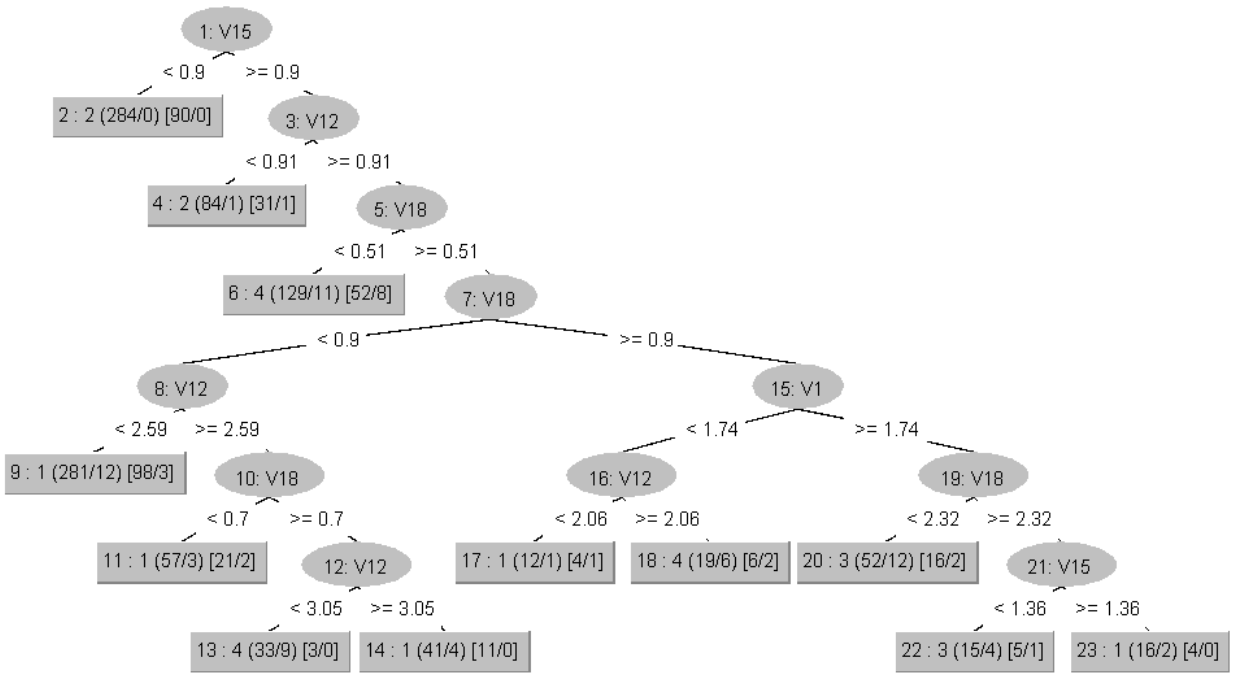
numDemicalPlaces : Πόσα δεκαδικά θα φαίνονται στο visualization του δέντρου.

numfolds : καθορίζει με πόσα δεδομένα θα πραγματοποιηθεί το pruning.

seed : Αυτό φροντίζει να παίρνουμε τα ίδια αποτελέσματα με τα συγκεκριμένα δεδομένα και configurations.

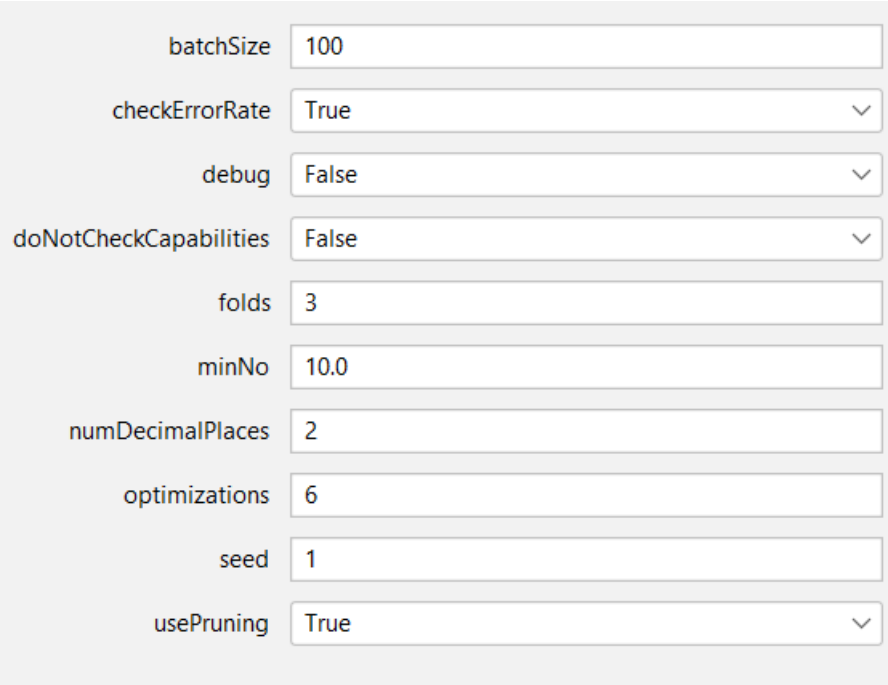
Τα αποτελέσματα είναι τα εξής :

```
=== Summary ===

Correctly Classified Instances        1243               91.129  %
Incorrectly Classified Instances       121                8.871  %
Kappa statistic                          0.868
Mean absolute error                      0.0829
Root mean squared error                  0.1869
Relative absolute error                 25.0302 %
Root relative squared error             45.9219 %
Total Number of Instances             1364

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.049    0.927      0.878    0.902      0.836    0.965     0.936     1
                 0.962    0.008    0.986      0.962    0.974      0.959    0.982     0.983     2
                 0.889    0.026    0.653      0.889    0.753      0.747    0.977     0.684     3
                 0.888    0.036    0.828      0.888    0.857      0.829    0.966     0.835     4
Weighted Avg.    0.911    0.030    0.918      0.911    0.913      0.876    0.972     0.923

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 495   6  34  29 |   a = 1
   8 486   0  11 |   b = 2
   6   1  64   1 |   c = 3
  25   0   0 198 |   d = 4
```



Συμπέρασμα 12 κανόνες με 91.129% των instances κατανεμημένες σωστά και accuracy 95.5%

WEKA JRip



Optimazations : Πόσου γύρους optimization θα πραγματοποιηθούν.(Εξακολουθεί να υπάρχει ο κίνδυνος του overfitting)

Τα αποτελέσματα είναι τα εξής :

```
(V18 >= 1.186) and (V18 <= 2.308) and (V1 >= 1.743) and (V15 >= 1.555) => Class=3 (34.0/1.0)
(V18 >= 0.904) and (V1 >= 2.1) and (V12 >= 1.425) => Class=3 (38.0/7.0)
(V18 <= 0.505) and (V15 >= 1.092) and (V15 <= 1.946) and (V12 >= 1.388) => Class=4 (133.0/0.0)
(V15 >= 1.327) and (V18 <= 0.49) and (V12 >= 1.698) => Class=4 (28.0/3.0)
(V12 >= 2.105) and (V18 >= 3.156) => Class=4 (19.0/4.0)
(V12 >= 2.591) and (V12 <= 2.994) and (V18 >= 0.7) and (V15 >= 0.933) => Class=4 (30.0/3.0)
(V15 <= 0.9) => Class=2 (372.0/0.0)
(V12 <= 0.909) => Class=2 (115.0/2.0)
 => Class=1 (595.0/46.0)

Number of Rules : 9
```

```
Correctly Classified Instances        1278               93.695  %
Incorrectly Classified Instances        86                6.305  %
Kappa statistic                          0.9043
Mean absolute error                      0.0507
Root mean squared error                  0.1713
Relative absolute error                 15.3106 %
Root relative squared error             42.1008 %
Total Number of Instances             1364

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.957    0.068    0.909      0.957    0.933      0.884    0.957     0.896     1
                 0.960    0.007    0.988      0.960    0.974      0.959    0.981     0.982     2
                 0.833    0.009    0.845      0.833    0.839      0.830    0.947     0.725     3
                 0.865    0.013    0.928      0.865    0.896      0.877    0.956     0.900     4
Weighted Avg.    0.937    0.033    0.938      0.937    0.937      0.908    0.965     0.920

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 540   5   8  11 |   a = 1
  14 485   2   4 |   b = 2
  11   1  60   0 |   c = 3
  29   0   1 193 |   d = 4
```

Ακολουθεί μια επεξήγηση του k-cross-validation:
Στο k-cross-validation που χρησιμοποιήθηκε κατά την υλοποίηση και των δύο αλγορίθμων διαχωρίζουμε το dataset σε k folds, τα k-1 folds χρησιμοποιούνται ως training set ενώ το άλλο ένα ως validation set. Αυτή η διαδικασία πραγματοποιείται k φορές και τα οι τελικές μετρικές θα βγουν από το σύνολο αυτών των διαδικασιών.

Σύγκριση JRip και REPTree για το συγκεκριμένο dataset:

Και στους δύο αλγορίθμους δημιουργήθηκαν κανόνες με ελάχιστο coverage 10 instances. Το κριτήριο για το ποιος αλγόριθμος θα αξιοποιηθεί για την δημιουργία έμπειρου συστήματος στο CLIPS ήταν το accuracy και το simplicity .

Ο JRip έχει 9 rules με 97% accuracy έναντι του REPTree που έχει 12 rules με 95.5% accuracy

Επομένως το σύστημα δημιουργήθηκε με τους κανόνες του JRip με το καλύτερο accuracy και simplicity.

ΕΞΥΠΝΟ ΣΥΣΤΗΜΑ CLIPS

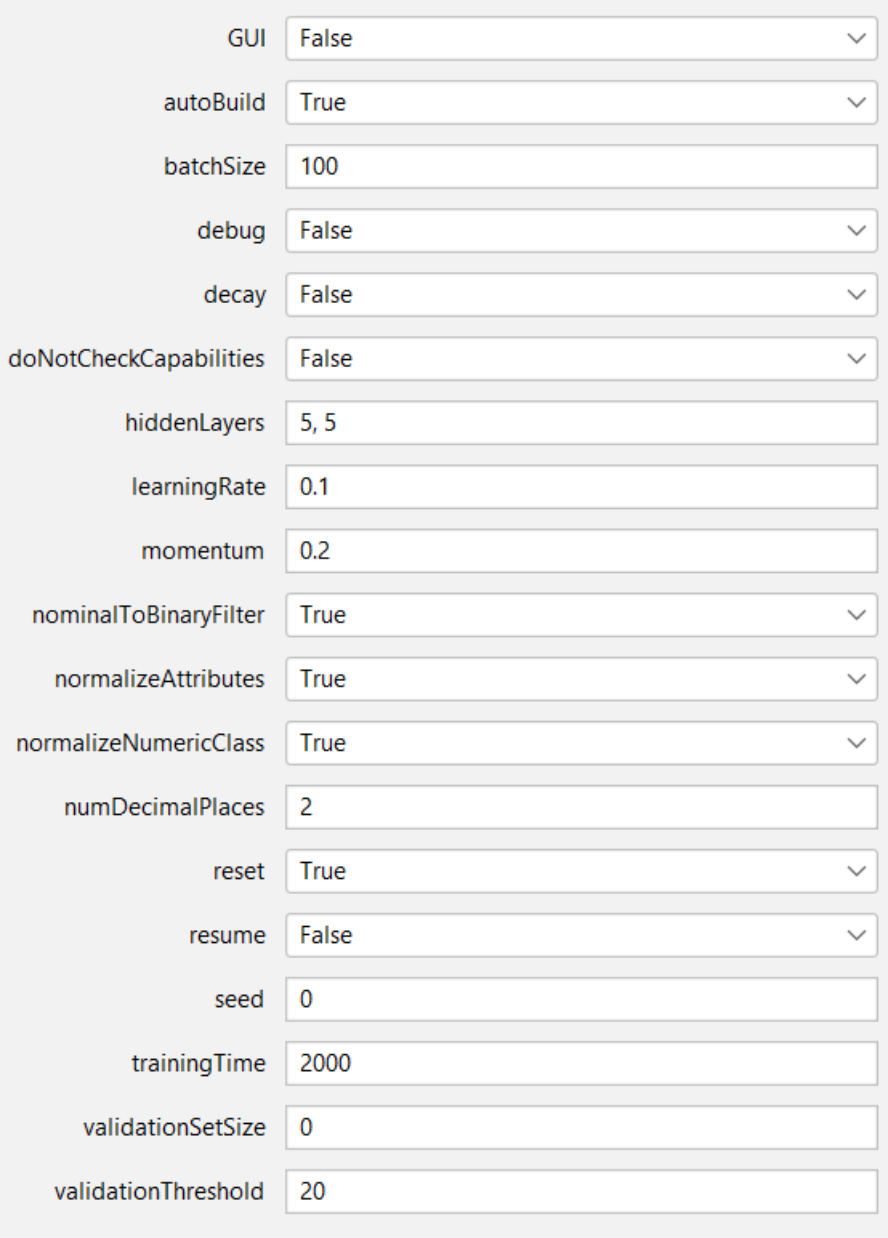
Τα αποτελέσματα που προέκυψαν από την δημιουργία του συστήματος με τους κανόνες του JRip είναι τα εξής:

```
CLIPS> (results)
Rule 1 : 33/1
Rule 2 : 31/7
Rule 3 : 133/0
Rule 4 : 25/3
Rule 5 : 15/4
Rule 6 : 27/3
Rule 7 : 372/0
Rule 8 : 113/2
Rule 9 : 549/46
Correcly Classified Instances : 1298/66
Correcly Classified Instances % : 95.1612903225807

CLIPS> (metrics)
The accuracy is 97.5806451612903%
The precision is 93.9963035179849%
The sensitivity is 92.9887542589771%
The specificity is 98.1303879733682%
CLIPS>
```

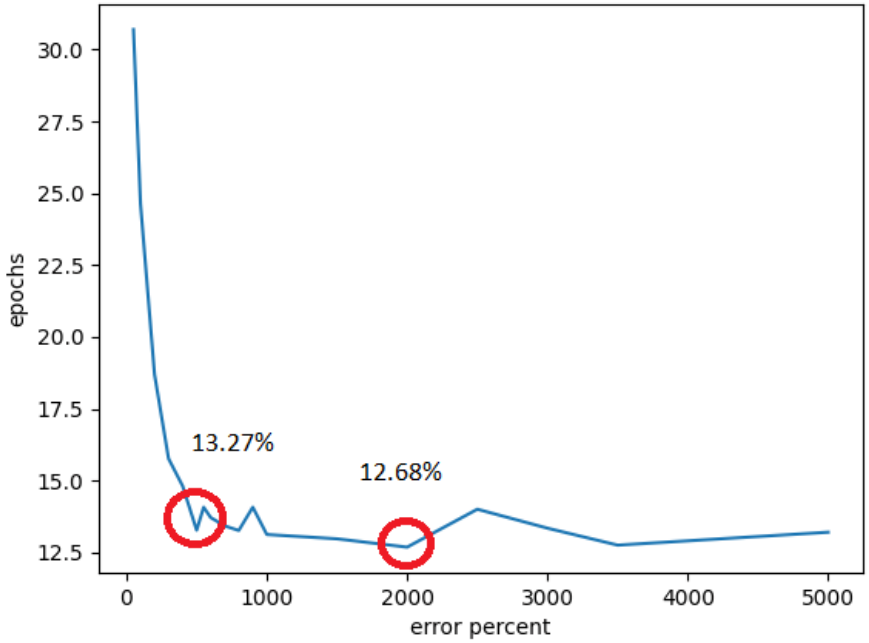
Μετά την ολοκλήρωση του συστήματος έγινε απόπειρα βελτιστοποίησής του με τροποποίηση ή και αφαίρεση κανόνων, η οποία όμως αποδείχθηκε μάταια. Αυτό οφείλεται στο ότι τα optimization runs του αλγόριθμου JRip εκτελούν ακριβώς αυτή την λειτουργία μέχρις ότου να μειωθεί το accuracy του rule set. Από το 7^{ο} optimization run ενώ βελτιώνεται ως ένα βαθμό το simplicity, μειώνεται σημαντικά το accuracy .

MLP

Μετά από πολλές δοκιμές τα βέλτιστο configuration για την εκπαίδευση του Multilayer Perceptron με αυτό το dataset είναι τα παρακάτω:



Ο αριθμός των εποχών που χρησιμοποιήθηκαν στο μοντέλο αυτό αποτελεί αποτέλεσμα πολυάριθμων δοκιμών που απεικονίζονται το παρακάτω γράφημα. Το γράφημα αυτό δείχνει πόσο μειώνεται το ποσοστό σφάλματος του μοντέλου ανάλογα με τον αριθμό των εποχών. Στις 500 εποχές παρατηρείται ένα τοπικό ελάχιστο με ποσοστό 13.27% ενώ στις 2000 εποχές βρίσκεται το ολικό ελάχιστο. Μετά τις 2000 εποχές ξεκινάνε να εμφανίζονται τα πρώτα δείγματα overfitting ενώ μετά τις 3500 εποχές ξεκινά η γραμμική αύξηση του error.



Παρακάτω είναι οι μετρικές του μοντέλου σύμφωνα με το weka:

```
Correctly Classified Instances        1191               87.3167 %
Incorrectly Classified Instances       173               12.6833 %
Kappa statistic                          0.8077
Mean absolute error                      0.0843
Root mean squared error                  0.2214
Relative absolute error                 25.4367 %
Root relative squared error             54.4018 %
Total Number of Instances             1364

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.876    0.115    0.843      0.876    0.859      0.757    0.939     0.906     1
                 0.947    0.026    0.956      0.947    0.951      0.923    0.976     0.953     2
                 0.736    0.019    0.679      0.736    0.707      0.690    0.972     0.682     3
                 0.744    0.030    0.830      0.744    0.785      0.747    0.954     0.854     4
Weighted Avg.    0.873    0.063    0.874      0.873    0.873      0.813    0.957     0.903

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 494  17  24  29 |   a = 1
  22 478   0   5 |   b = 2
  15   4  53   0 |   c = 3
  55   1   1 166 |   d = 4
```

Σύγκριση του MPL και του συστήματος που δημιουργήσαμε με το rule set του JRip:

Το μοντέλου του JRip έχει accuracy σχεδόν 97.5% που είναι αρκετά μεγαλύτερο έναντι του 93.65 % του MPL. Το ίδιο ισχύει και για τις άλλες μετρικές αφού στον MLP το Precision και το Sensitivity έχουν τιμή περίπου 87% έναντι του JRip που είναι στο 93%.

Συμπέρασμα

Υπάρχουν δύο βασικοί λόγοι για τους οποίους το έξυπνο σύστημα στο clips έχει καλύτερη απόδοση από αυτό του MLP, ο πρώτος και σημαντικότερος λόγος είναι το μέγεθος του dataset το οποίο είναι πάρα πολύ μικρό και δεν μπορεί να το νευρωνικό μας να παράγει τα

βέλτιστα αποτελέσματα. Ο δεύτερος λόγος είναι ότι τα instances της κάθε κλάσης έχουν διαφορετικό αριθμό με αποτέλεσμα το νευρωνικό να μην εκπαιδευτεί σωστά .

Αμίτσης Λευτέρης 1072464