# Τεχνητή Νοημοσύνη Μηχανική Μάθηση και εφαρμογές – Μέρος Β

**ΔΠΜΣ Πληροφορική Επιστημών Ζωής**

Πρέζα Παναγιώτα

Κουρούσια Μαρία

**Πάτρα Ιούνιος 2021**

## Μέρος Α - Ταξινόμηση

Να εκτελέσετε και να συγκρίνετε τους παρακάτω αλγορίθμους ταξινόμησης.

I k-nearest neighbors (KNN)

II Δένδρα Αποφάσεων (Decision Tree)

III Naive Bayes Classifiers

IV Support Vector Machines (SVM)

V Linear Discriminant Analysis (LDA)

VI Adaptive Boosting for Multiclass Classification

## Λύση

Τα dataset που έχουμε επιλέξει να εφαρμόσουμε τους παραπάνω αλγορίθμους είναι το GSE86469 και το GSE103334. Το πρώτο το οποίο περιέχει δεδομένα γονιδιώματος του ανθρώπου (Homo Sapiens) και εκφράζει την μεταγραφική ενός κυττάρου που ορίζει τις υπογραφές των ανθρώπινων νησιωτικών κυττάρων και αποκαλύπτει συγκεκριμένες κυτταρικές αλλαγές έκφρασης στον διαβήτη τύπου 2.

Τα δεδομένα είναι στη μορφή Matlab πίνακα, αριθμητικών τιμών με:

- 26616 Διαστάσεις, γνωρίσματα
- 638 Εγγραφές, δείγματα
- 8 Κλάσεις

Το δεύτερο dataset GSE103334 έχει δεδομένα γονιδιώματος ποντικού για τη χρονική παρακολούθηση της ενεργοποίησης των μικρογλοιών στον νευροεκφυλισμό σε ανάλυση ενός κυττάρου

Για την σύγκριση των αλγορίθμων χρησιμοποιήσαμε τις μετρικές

Ορθότητα (Accuracy) = TP+TN/(TP+FP+TN+FN) η οποία μας δείχνει τον λόγο των σωστά προβλεπόμενων παρατηρήσεων προς όλες τις παρατηρήσεις.

Εξειδίκευση (Specificity)=TN/(TN+FP) μετράει το πραγματικό ποσοστό των αρνητικών παρατηρήσεων που χαρακτηρίστηκαν ως negative .

F1-Score = 2 * (Specificity * Recall ) / Specificity + Recall (είναι ο αρμονικός μέσος των Specificity και Recall με τιμές ανάμεσα στο σε 0 (η χειρότερη ακρίβεια) και 1 (η τέλεια ακρίβεια).

Οι εκτελέσεις των αλγορίθμων έγιναν με 10-fold cross validation στην επιλογή των δειγμάτων χρησιμοποιώντας την ακόλουθη συνάρτηση και ορίσματα της Matlab: crossvalind('Kfold',Ytrain,10)[1], όπου Ytrain είναι το σύνολο των observations.

Για τον αλγόριθμο KNN (k-nearest neighbors) χρησιμοποιήσαμε την συνάρτηση fitcknn[2] τα εξής ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των δεδομένων απόκρισης του training set και τέλος, ορίζουμε το πεδίο 'NumNeighbors' που ισούται με k=5. Η επιλογή αυτή μας δείχνει τους 5 κοντινούς γείτονες σε μια εγγραφή, όπου k είναι ή προεπιλεγμένη τιμή και προσέχω ώστε να μην διαλέξω πολύ μικρή τιμή διότι τότε θα υπάρχει ευαισθησία στο σύστημα αλλά ούτε και μεγάλη τιμή διότι τότε η γειτονιά θα περιέχει σημεία από άλλες κλάσεις.

Για τον αλγόριθμο των Δένδρων Αποφάσεων (Decision Tree) χρησιμοποιήσαμε την συνάρτηση fitctree[3] με τα ακόλουθα ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των δεδομένων απόκρισης του training set.

Για τον αλγόριθμό του Naive Bayes χρησιμοποιήσαμε την συνάρτηση fitcnb[4] με τα ακόλουθα ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των δεδομένων απόκρισης του training set και τέλος, ορίζουμε το πεδίο 'DistributionNames' ως kernel, για καθορίσουμε την διανομή που θα μοντελοποιήσει τα δεδομένα.

Για τον αλγόριθμο SVM (Support Vector Machines) χρησιμοποιήσαμε την συνάρτηση fitcecoc[5] με τα ακόλουθα ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των δεδομένων απόκρισης του training set.

Για τον αλγόριθμο LDA (Linear Discriminant Analysis) χρησιμοποιήσαμε την συνάρτηση fitcdiscr[6] με τα ακόλουθα ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των

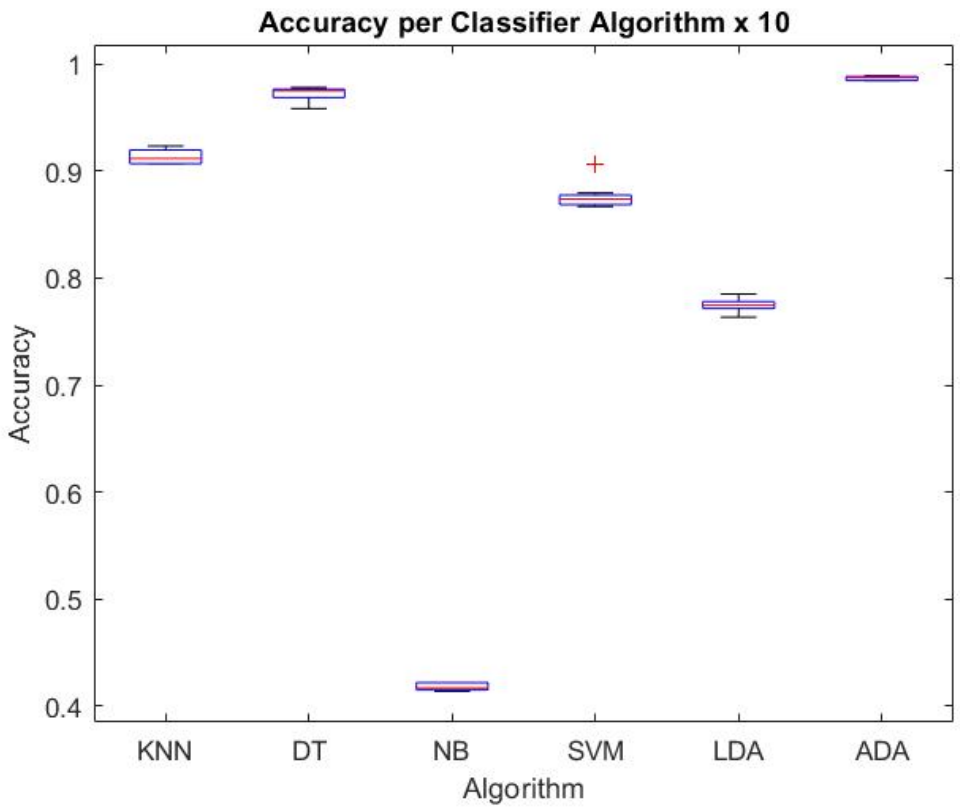
[1] https://www.mathworks.com/help/bioinfo/ref/crossvalind.html
[2] https://www.mathworks.com/help//stats/fitcknn.html
[3] https://www.mathworks.com/help/stats/fitctree.html
[4] https://www.mathworks.com/help/stats/fitcnb.html
[5] https://www.mathworks.com/help/stats/fitcecoc.html
[6] https://www.mathworks.com/help/stats/fitcdiscr.html

δεδομένων απόκρισης του training set και τέλος, ορίσαμε το πεδίο 'discrimType' ως diagLinear διότι το covariance μερικών δεδομένων του πίνακα ήταν μηδέν, με αποτέλεσμα να μην μπορεί να εκτελεστεί η συνάρτηση και άρα ο αλγόριθμός.

Για τον αλγόριθμο ADA (Adaptive Boosting for Multiclass Classification) χρησιμοποιήσαμε την συνάρτηση fitcensemble[7] με τα ακόλουθα ορίσματα: Xtrain(train,:) είναι το σύνολο των δεδομένων πρόσβλεψης του training set, Ytrain(train,:) είναι το σύνολο των δεδομένων απόκρισης του training set, ορίσαμε το πεδίο 'Method' ως AdaBoostM2 για multiclass classification και τέλος, ορίσαμε το πεδίο 'Learners' για τους tree learners.

## ACCURACY BOXPLOTS

## GSE86469




- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή της ακρίβειας είναι στο 0.94 ενώ

[7] https://www.mathworks.com/help/stats/fitcensemble.html

η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας η οποία είναι στο 0.92.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει πολύ καλή ακρίβεια 94% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Δέντρο Απόφασης (DT) συμπεραίνουμε ότι η μέγιστη τιμή της ακρίβειας είναι στο 0.97 , η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας η οποία είναι στο 0.96 και η ελάχιστη τιμή είναι 0.95.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει πολύ καλή ακρίβεια στο 97% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Naive Bayes Classifier (NBC) συμπεραίνουμε ότι η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.43 και η ελάχιστη τιμή ακρίβειας είναι 0.41.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει κακή ακρίβεια στο 0.41% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

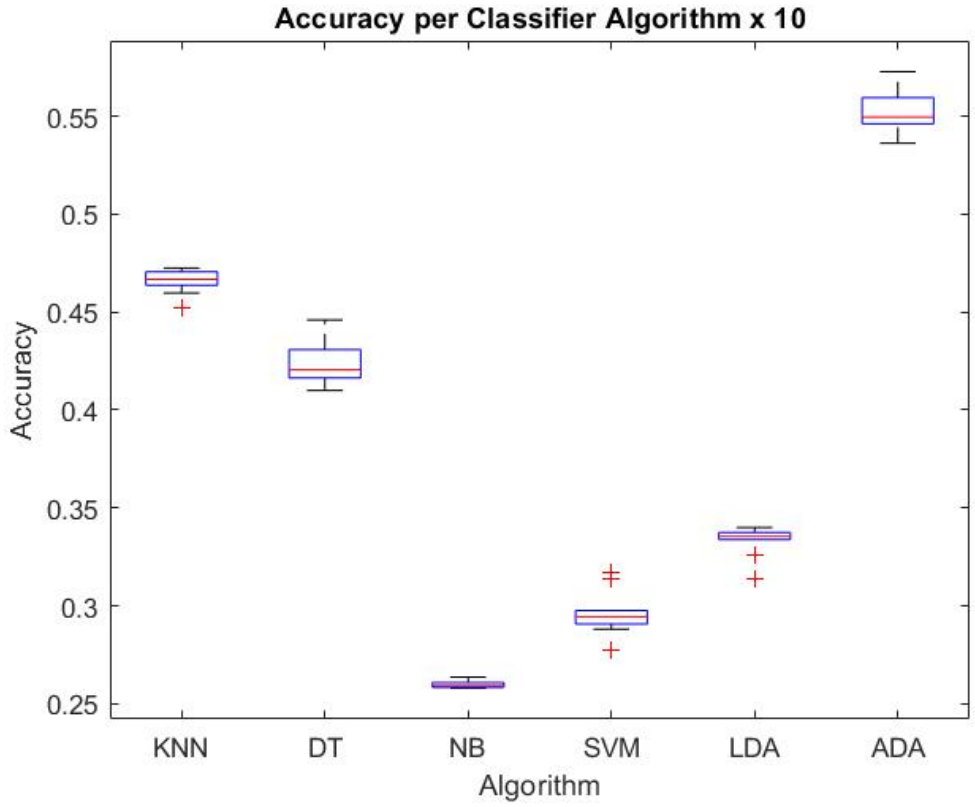
- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Support Vectors Machine (SVM) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.88 , η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.87 και η ελάχιστη τιμή ακρίβειας είναι 0.86 ,ακόμα το σύμβολο το σταυρού στην τιμή 0.90 συμβολίζει μια ακραία τιμή .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει καλή ακρίβεια στο 87% και άρα καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Linear Discriminant Analysis (LDA) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.78, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.77 και η ελάχιστη τιμή ακρίβειας είναι 0.76. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει μέτρια ακρίβεια στο 78% και άρα μέτρια πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469

- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Adaptive Boosting for Multiclass Classification (ADA) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.98, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.98 Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει πολύ καλή ακρίβεια στο 98% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

  Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική της ορθότητας συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς καλύτερη πρόβλεψη την δίνει ο αλγόριθμος Adaptive Boosting for Multiclass Classification (ADA) 98% έπειτα ο Decision Tree μια πολύ καλή πρόβλεψη στο 97%, ακολουθεί ο K-Nearest Neighbor με αρκετά καλή πρόβλεψη ακρίβειας 94%, ο Support Vectors Machine (SVM) με καλή πρόβλεψη ακρίβειας στο 87% , ο αλγόριθμος Linear Discriminant Analysis (LDA) με μέτρια πρόβλεψη ακρίβειας 78% και τέλος ο αλγόριθμος Naive Bayes Classifier (NBC) με κακή πρόβλεψη ακρίβειας 41% ο οποίος φαίνεται να μην ταιριάζει στο dataset GSE86469 .

**GSE103334**




- Από το boxplot accuracy για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή της ακρίβειας είναι στο 0.47 ενώ η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας η οποία είναι στο 0.46. Επιπλέον εμφανίζεται μια ακραία τιμή με το σύμβολο του σταυρού. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει κακή ακρίβεια 46% και αντίστοιχη απόδοση πρόβλεψης των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot accuracy για τον αλγόριθμο ταξινόμησης Δέντρο Απόφασης (DT) συμπεραίνουμε ότι η μέγιστη τιμή της ακρίβειας είναι στο 0.44 , η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας η οποία είναι στο 0.43 και η ελάχιστη τιμή είναι 0.41. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει ακρίβεια στο 44% εμφανίζει κακή ακρίβεια 46% και αντίστοιχη απόδοση πρόβλεψης των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot accuracy για τον αλγόριθμο ταξινόμησης Naive Bayes Classifier (NBC) συμπεραίνουμε ότι η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι

0.26 και η ελάχιστη τιμή ακρίβειας είναι 0.25. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει κακή ακρίβεια στο 25% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot accuracy για τον αλγόριθμο ταξινόμησης Support Vectors Machine (SVM) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.32 , η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.31 και η ελάχιστη τιμή ακρίβειας είναι 0.30. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει κακή ακρίβεια στο 31% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

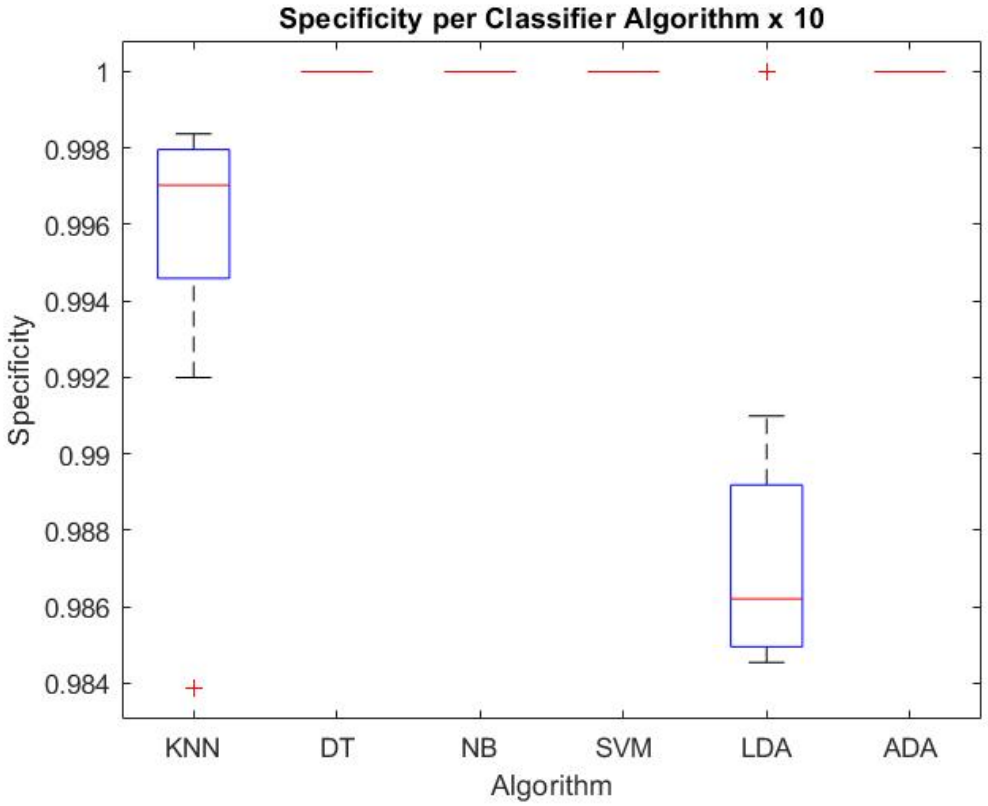
- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Linear Discriminant Analysis (LDA) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.34, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.33 και η ελάχιστη τιμή ακρίβειας είναι 0.31. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει κακή ακρίβεια στο 33% και άρα μέτρια πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot ή θηκόγραμμα accuracy για τον αλγόριθμο ταξινόμησης Adaptive Boosting for Multiclass Classification (ADA) συμπεραίνουμε ότι η μέγιστη τιμή ακρίβειας είναι 0.57, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της ακρίβειας και είναι 0.55. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης εμφανίζει μέτρια ακρίβεια στο 56% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική της ορθότητας συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς καλύτερη πρόβλεψη την δίνει ο αλγόριθμος Adaptive Boosting for Multiclass Classification (ADA) 56% έπειτα ο Decision Tree μια πολύ καλή πρόβλεψη στο 46%, ακολουθεί ο K-Nearest Neighbor με πρόβλεψη ακρίβειας 46%, ο Support Vectors Machine (SVM) με καλή πρόβλεψη ακρίβειας στο 33% , ο αλγόριθμος Linear Discriminant Analysis (LDA) με μέτρια πρόβλεψη ακρίβειας 33% και τέλος ο αλγόριθμος Naive Bayes Classifier (NBC) με κακή πρόβλεψη ακρίβειας 25%. Οι αλγόριθμοι όπως παραμετροποιήθηκαν για το πρώτο σύνολο φαίνεται να μην ταιριάζουν στο dataset GSE103334.

# SPECIFICITY BOX PLOT

## GSE86469




- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι στο 0.99 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.997,η ελάχιστη τιμή της εξειδίκευσης είναι 0.992 και εμφανίζεται και μια ακραία τιμή στο 0.98 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει πολύ καλή εξειδίκευση 99% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης Decision Tree συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης ,η κόκκινη γραμμή στο θηκόγραμμα που είναι η διάμεσος των τιμών της εξειδίκευσης καθώς και η ελάχιστη τιμή ταυτίζονται στην τιμή 0.1 γι'αυτό και δεν παρατηρούμε κάποιο εύρος τιμών αλλά μια κόκκινη γραμμή που δείχνει ότι όλες οι παρατηρήσεις συγκεντρώνονται γύρω από την τιμή 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης DT εμφανίζει

πολύ καλή εξειδίκευση 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης NBC συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης 0.991 ,η κόκκινη γραμμή στο θηκόγραμμα που είναι η διάμεσος των τιμών της εξειδίκευσης είναι καθώς και η ελάχιστη τιμή ταυτίζονται στην τιμή 0.1 γι'αυτό και δεν παρατηρούμε κάποιο εύρος τιμών αλλά μια κόκκινη γραμμή που δείχνει ότι όλες οι παρατηρήσεις συγκεντρώνονται γύρω από την τιμή 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης NBC εμφανίζει πολύ καλή εξειδίκευση 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης SVM συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης ,η κόκκινη γραμμή στο θηκόγραμμα που είναι η διάμεσος των τιμών της εξειδίκευσης καθώς και η ελάχιστη τιμή ταυτίζονται στην τιμή 0.1 γι'αυτό και δεν παρατηρούμε κάποιο εύρος τιμών αλλά μια κόκκινη γραμμή που δείχνει ότι όλες οι παρατηρήσεις συγκεντρώνονται γύρω από την τιμή 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης SVM εμφανίζει πολύ καλή εξειδίκευση 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

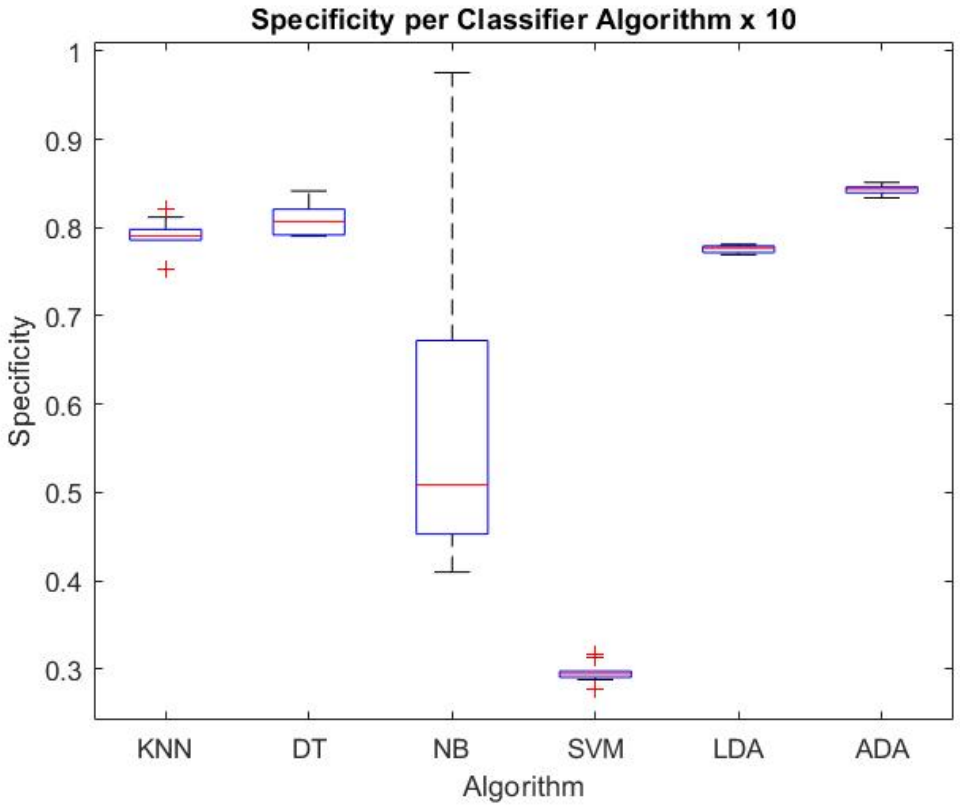
- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης LDA συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι 0.991 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.986,η ελάχιστη τιμή της εξειδίκευσης είναι 0.985 και εμφανίζεται και μια ακραία τιμή στο 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης LDA εμφανίζει πολύ καλή εξειδίκευση 99% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης ADA συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης ,η κόκκινη γραμμή στο θηκόγραμμα που είναι η διάμεσος των τιμών της εξειδίκευσης καθώς και η ελάχιστη τιμή ταυτίζονται στην τιμή 0.1 γι'αυτό και δεν παρατηρούμε κάποιο εύρος τιμών αλλά μια κόκκινη γραμμή που δείχνει ότι όλες οι παρατηρήσεις συγκεντρώνονται γύρω από την τιμή 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης ADA εμφανίζει

πολύ καλή εξειδίκευση 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική της εξειδίκευσης συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς την καλύτερη πρόβλεψη την δίνουν οι αλγόριθμοι DT, NVB, SVM, ADA με 100% και έπειτα ακολουθούν οι αλγόριθμοι KNN με ποσοστό 99% και ο LDA 99.1%.

**GSE103334**




- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι στο 0.82 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.79,η ελάχιστη τιμή της εξειδίκευσης είναι 0.75 και εμφανίζεται και μια ακραία τιμή στο 0.98 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει πολύ καλή εξειδίκευση 79% και άρα κάπως καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης Decision Tree συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι στο 0.84 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.82, η ελάχιστη τιμή της εξειδίκευσης είναι 0.79.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει καλή εξειδίκευση 82% και άρα καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης NBC συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι στο 0.67 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.55, η ελάχιστη τιμή της εξειδίκευσης είναι 0.47.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει καλή εξειδίκευση 55% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης SVM συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι στο 0.32 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.29, η ελάχιστη τιμή της εξειδίκευσης είναι 0.3.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει καλή εξειδίκευση 30% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης LDA συμπεραίνουμε ότι η μέγιστη τιμή της εξειδίκευσης είναι 0.34 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.32,η ελάχιστη τιμή της εξειδίκευσης είναι 0.3 και εμφανίζεται και μια ακραία τιμή στο 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης LDA εμφανίζει πολύ καλή εξειδίκευση 32% και άρα πολύ κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

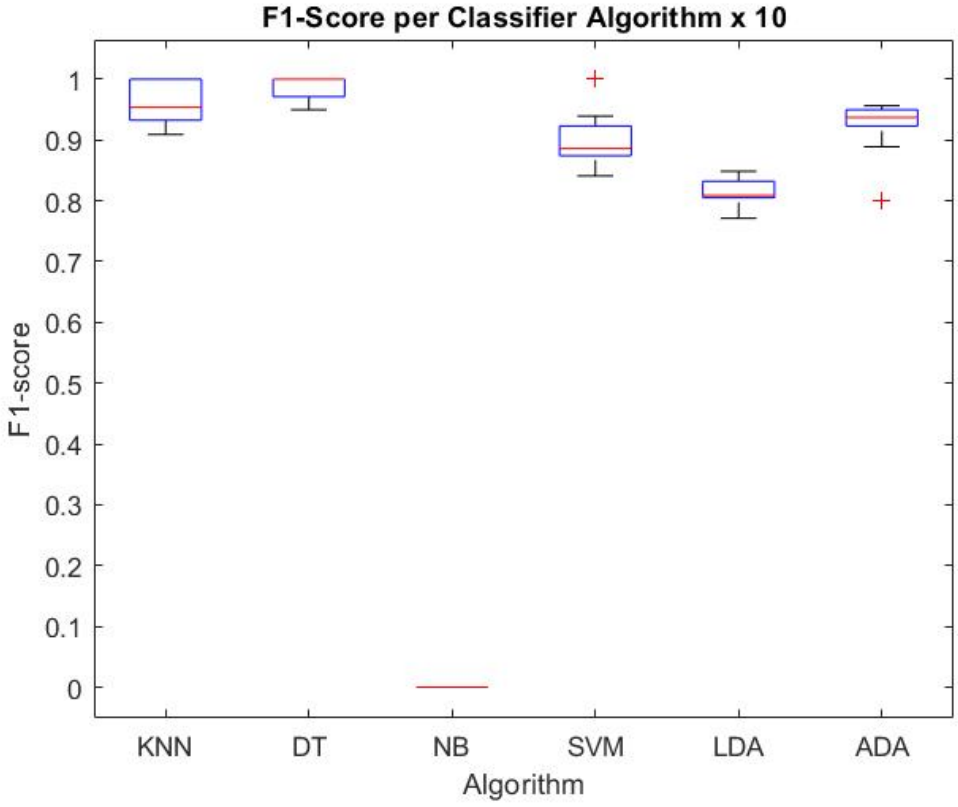
- Από το boxplot specificity για τον αλγόριθμο ταξινόμησης ADA συμπεραίνουμε ότι ότι η μέγιστη τιμή της εξειδίκευσης είναι 0.85 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.83,η ελάχιστη τιμή της εξειδίκευσης είναι 0.82 και εμφανίζεται και μια ακραία τιμή στο 0.1 .Επομένως

παρατηρούμε ότι ο αλγόριθμος ταξινόμησης LDA εμφανίζει πολύ καλή εξειδίκευση 83% και άρα καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική της εξειδίκευσης συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς την καλύτερη πρόβλεψη την δίνουν οι αλγόριθμοι DT, ADA με και έπειτα ακολουθούν οι αλγόριθμοι KNN με ποσοστό 82%.

## F1-SCORE BOX PLOT

**GSE86469**




- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.1, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.92,η ελάχιστη τιμή είναι 0.9 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει πολύ καλό

F1-Score 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης DT συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.1, η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών της εξειδίκευσης η οποία είναι 0.1και η ελάχιστη τιμή είναι 0.91 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης DT εμφανίζει πολύ καλό F1-Score 100% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης NBC συμπεραίνουμε ότι η μέγιστη τιμή , η διάμεσος και η ελάχιστη τιμή του F1-Score ταυτίζονται στο 0. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης NBC εμφανίζει κακό F1-Score 0% και άρα καμία πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης SVM συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.95 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών και είναι 0.81,και η ελάχιστη τιμή είναι 0.79 και παρατηρούμε μια ακραία τιμή στο 0.1 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης SVM εμφανίζει πολύ καλό F1-Score 91% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

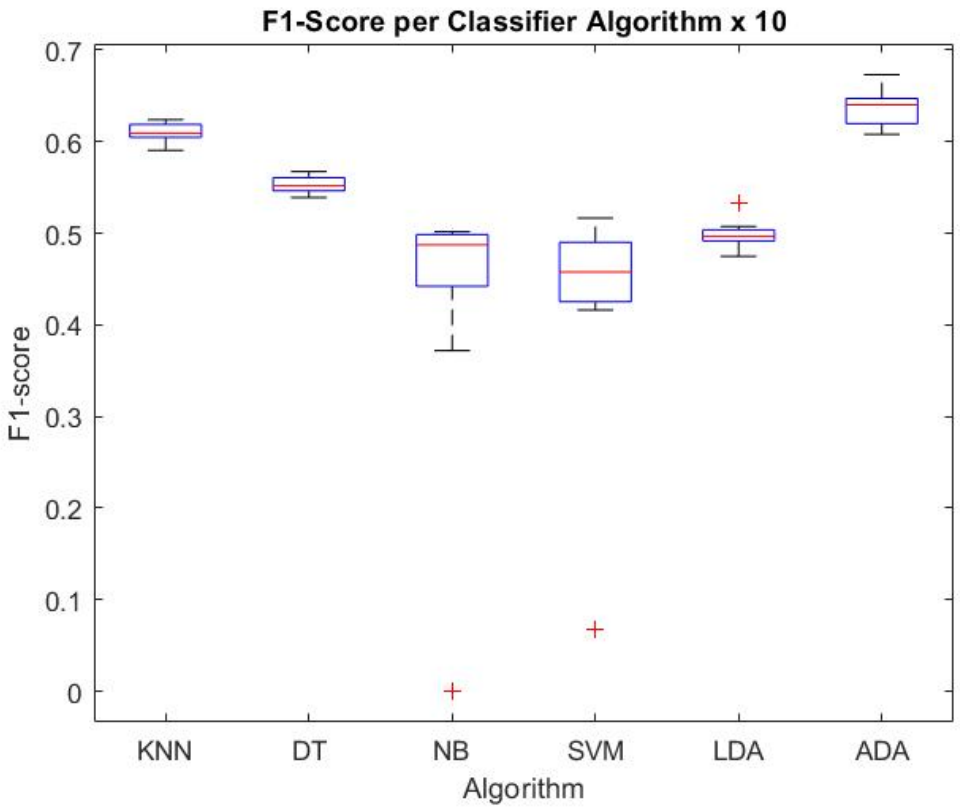
- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης LDA συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.84 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών και είναι 0.81 , η ελάχιστη τιμή είναι 0.77 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης LDA εμφανίζει καλό F1-Score 84% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης ADA συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.95 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών και είναι 0.93,και η ελάχιστη τιμή είναι 0.88 και παρατηρούμε μια ακραία τιμή στο 0.80 .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης

SVM εμφανίζει πολύ καλό F1-Score 91% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE86469.

Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική F1-Score συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς την καλύτερη πρόβλεψη την δίνουν οι αλγόριθμοι KNN, DT με ποσοστό 100% και επιτυγχάνουν την καλύτερη πρόβλεψη, έπειτα ακολουθούν οι αλγόριθμοι SVM, ADA με ποσοστό 91% και επιτυγχάνουν πολύ καλή πρόβλεψη, ο LDA 84% επιτυγχάνει καλή πρόβλεψη και τέλος ο NBC με ποσοστό 0% όπου φαίνεται ότι δεν επιτυγχάνει καμία πρόβλεψη στο συγκεκριμένο σετ δεδομένων .

**GSE103334**




- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης KNN συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score 0.620. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης KNN εμφανίζει μέτριο F1-Score 62% και άρα μέτρια πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης DT συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.55.Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης DT εμφανίζει πολύ καλό F1-Score 55% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης NBC συμπεραίνουμε ότι η διάμεσος του F1-Score ταυτίζονται στο 0.49. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης NBC εμφανίζει κακό F1-Score 49% και άρα μέτρια προς κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης SVM συμπεραίνουμε ότι η διάμεση τιμή του F1-Score είναι 0.45. Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης SVM εμφανίζει κακό F1-Score 45% και άρα κακή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης LDA συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.53 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών και είναι 0.5 , η ελάχιστη τιμή είναι 0.49. Επιπλέον βλέπουμε μια ακραία τιμή .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης LDA εμφανίζει μέτριο F1-Score 50% και άρα μέτρια πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

- Από το boxplot F1-Score για τον αλγόριθμο ταξινόμησης ADA συμπεραίνουμε ότι η μέγιστη τιμή του F1-Score είναι στο 0.67 ,η κόκκινη γραμμή στο θηκόγραμμα συμβολίζει την διάμεσο των τιμών και είναι 0.64, και η ελάχιστη τιμή είναι 0.6. .Επομένως παρατηρούμε ότι ο αλγόριθμος ταξινόμησης SVM εμφανίζει μέτριο F1-Score 64% και άρα πολύ καλή πρόβλεψη των παρατηρήσεων για το σετ δεδομένων GSE103334.

Επομένως από τις παραπάνω συγκρίσεις των αλγορίθμων ταξινόμησης με την μετρική F1-Score συμπεραίνουμε ότι τα καλύτερα αποτελέσματα και συνεπώς την καλύτερη πρόβλεψη την δίνουν ο αλγόριθμο ADA με ποσοστό 64% και επιτυγχάνουν την καλύτερη πρόβλεψη, έπειτα ακολουθούν οι αλγόριθμοι KNN

62%, και ο DT 50% και επιτυγχάνουν μέτρια πρόβλεψη, ο KNN και LDA με περίπου 50% και τέλος ο SVM με ποσοστό 45%.

# Μέρος Β – Ομαδοποίηση

## I. Kmeans

## II. Hierarchical Clustering – single linkage

Τα κριτήρια με τα οποία αξιολογήθηκαν οι αλγόριθμοι εξετάζουν τις περιπτώσεις των k=2,3,4,5 που αντιπροσωπεύει τον αριθμό των clusters που θα δημιουργηθούν από τον αλγόριθμο. Σε όλους τους συνδυασμούς Αλγορίθμου και κριτηρίου λαμβάνεται ο ίδιος βέλτιστος αριθμός συστάδων και είναι ίσος με 2. Επιπλέον, στα 10 ανεξάρτητα πειράματα που πραγματοποιήθηκαν δεν εμφανίζεται καθόλου διαφοροποίηση στις τιμές των αποτελεσμάτων για τα διάφορα κριτήρια.

## I. Kmeans

- **Davies Bouldin Index**

```
DaviesBouldinEvaluation with properties:

  NumObservations: 2208
      InspectedK: [2 3 4 5]
  CriterionValues: [0.0039 0.2046 1.1153 1.3114]
        OptimalK: 2
```

*Εικόνα 1. Kmeans, Davies-Bouldin, dataset 1*

Η τιμή OptimalK δείχνει ότι, με βάση το κριτήριο Davies-Bouldin, ο βέλτιστος αριθμός συστάδων είναι 2. Επομένως η αντίστοιχη τιμή που μας ενδιαφέρει για κ=2 είναι 0.0039 που δείχνει ότι έχει γίνει πολύ καλή συσταδοποίηση στο συγκεκριμένο dataset γι'αυτόν τον αριθμό συστάδων.

```
DaviesBouldinEvaluation with properties:

  NumObservations: 638
      InspectedK: [2 3 4 5]
  CriterionValues: [0.5702 0.6229 0.6295 0.9537]
        OptimalK: 2
```

*Εικόνα 2. Kmeans, Davies-Bouldin, dataset 2*

Στο δεύτερο dataset αντίθετα έχουμε μεγάλη τιμή για το κριτήριο 0.5702 που υπονοεί ότι ο kmeans έχει κάνει πολύ καλύτερη απόδοση συσταδοποίησης για τα δεδομένα του πρώτου dataset.

- **Silhouette**

Η ανάλυση του κριτηρίου Silhouette αναφέρεται θα βοηθήσει στην ερμηνεία και επικύρωση της συνέπειας εντός των συστάδων των δεδομένων. Η τιμή της είναι ένα μέτρο που δείχνει το πόσο παρόμοιο είναι ένα αντικείμενο με τη δική του ομάδα (συνοχή) και σε σύγκριση με άλλες ομάδες (διαχωρισμός) .

```
SilhouetteEvaluation with properties:

  NumObservations: 2208
      InspectedK: [2 3 4 5]
  CriterionValues: [0.9999 0.9936 0.2830 0.1034]
        OptimalK: 2
```

*Εικόνα 3. Kmeans, Silhouette, dataset 1*

Η τιμή OptimalK δείχνει ότι, με βάση το κριτήριο Silhouette, ο βέλτιστος αριθμός συστάδων είναι 2. Η αντίστοιχη τιμή για το κριτήριο είναι πολύ κοντά στο 1 (0.9999).

```
SilhouetteEvaluation with properties:

  NumObservations: 638
      InspectedK: [2 3 4 5]
  CriterionValues: [0.7999 0.7771 0.7255 0.4930]
        OptimalK: 2
```

*Εικόνα 4. Kmeans, Silhouette, dataset 2*

Για το δεύτερο dataset έχουμε λιγότερο καλά αποτελέσματα για τον ίδιο βέλτιστο αριθμό συστάδων (0.7999). Συνεπώς καταλαβαίνουμε ότι τα αντικείμενα ταιριάζουν πολύ καλά με τη δική τους σύστάδα και διαφοροποιούνται πολύ με τη γειτονική, ειδικά στο πρώτο σύνολο για το συγκεκριμένο κριτήριο.

**II. Hierarchical clustering – single linkage**

- **Davies Bouldin index**

```
DaviesBouldinEvaluation with properties:

  NumObservations: 2208
      InspectedK: [2 3 4 5]
  CriterionValues: [0.0039 0.2046 1.1153 1.3114]
        OptimalK: 2
```

*Εικόνα 5. **Hierarchical clustering – single linkage**, Davies-Bouldin, dataset 1*

Η τιμή OptimalK δείχνει ότι, με βάση το κριτήριο Davies-Bouldin, ο βέλτιστος αριθμός συστάδων είναι 2. Η τιμή του Davies-Bouldin Index για 2 συστάδες είναι 0,0039 που είναι ακριβώς η ίδια τιμή που πήραμε για το ίδιο κριτήριο και στον αλγόριθμο Kmeans.

```
DaviesBouldinEvaluation with properties:

  NumObservations: 638
      InspectedK: [2 3 4 5]
  CriterionValues: [0.5379 0.7254 0.6798 1.0037]
        OptimalK: 2
```

*Εικόνα 6. Hierarchical clustering – single linkage, Davies-Bouldin, dataset 2*

Η τιμή για το δεύτερο dataset είναι όπως και στον Kmeans και σε αυτή την περίπτωση αρκετά μεγάλη και συνεπώς καταλαβαίνουμε ότι δεν έχει γίνει καλή συσταδοποίηση σε αυτό το σύνολο δεδομένω για το συγκεκριμένο κριτήριο.

- **Silhouette**

```
SilhouetteEvaluation with properties:

  NumObservations: 2208
      InspectedK: [2 3 4 5]
  CriterionValues: [0.9999 0.9936 0.2830 0.1034]
        OptimalK: 2
```

*Εικόνα 7. Hierarchical clustering – single linkage, Silhouette, dataset 1*

Για το Silhouette κριτήριο παίρνουμε τιμή πολύ κοντά στο 1 για το πρώτο dataset και 0.8106 για το δεύτερο. Συνεπώς, καταλαβαίνουμε ότι τα αντικείμενα ταιριάζουν πολύ καλά με τη δική τους σύστάδα και διαφοροποιούνται πολύ με τη γειτονική.

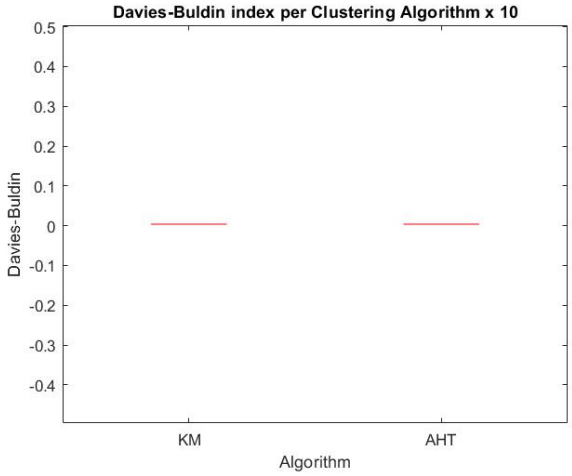
```
SilhouetteEvaluation with properties:

  NumObservations: 638
      InspectedK: [2 3 4 5]
  CriterionValues: [0.8106 0.6382 0.6854 0.4388]
        OptimalK: 2
```

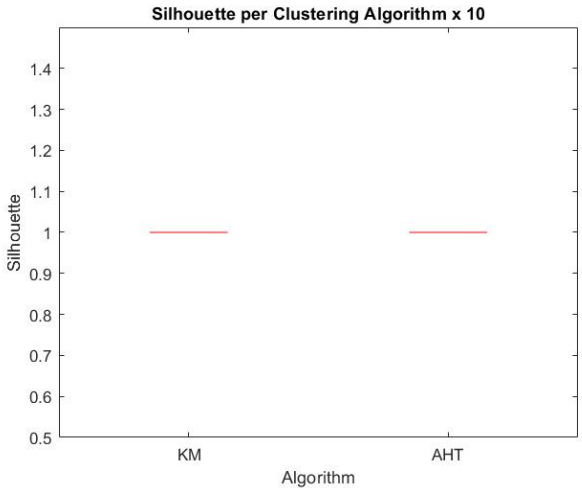
*Εικόνα 8. Hierarchical clustering – single linkage, Silhouette, dataset 2*

## Σύγκριση Αλγορίθμων

Παρακάτω έχουμε τη σύγκριση των δύο αλγορίθμων με βάση τις μετρικές που αναφέρθηκαν, για δύο ξεχωριστά datasets. Για το πρώτο dataset οι τιμές για τα κριτήρια Davies-Bouldin και Silhouette που προέκυψαν ήταν ταυτόσημες μεταξύ τους και για κάθε ένα από τα 10 πειράματα και συνεπώς βλέπουμε στα Boxplot αντίστοιχα τις δυο σταθερές γραμμές.

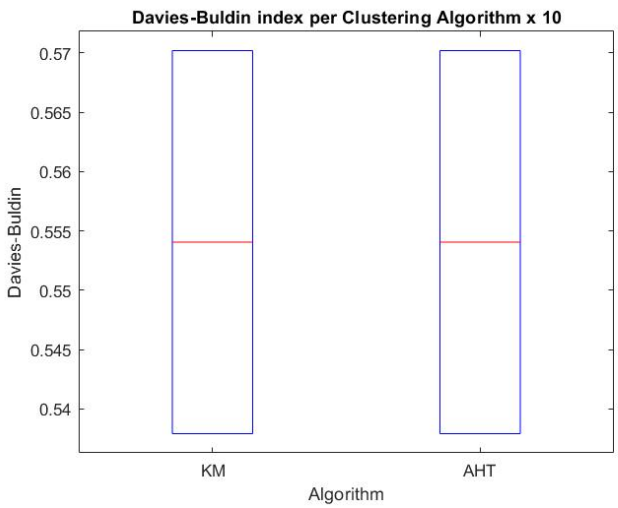



*Εικόνα 9 Σύγκριση αλγορίθμων με το κριτήριο Davies-Bouldin, Dataset 1*

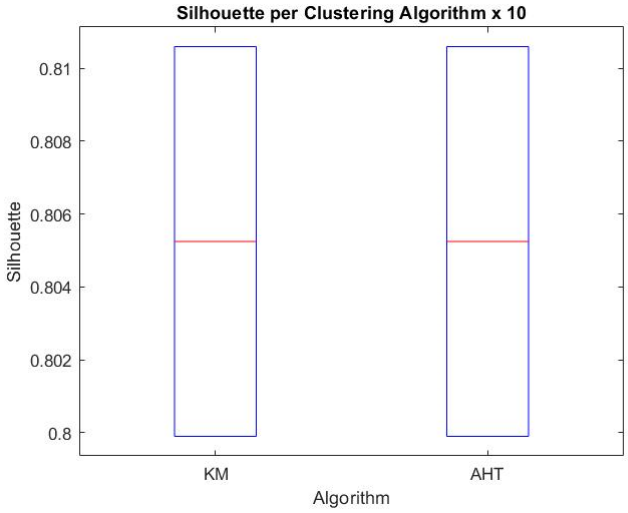



*Εικόνα 10 Σύγκριση αλγορίθμων με το κριτήριο Silhouette, Dataset 1*

Στο δεύτερο σύνολο δεδομένων έχουμε αρκετά κοντινές τιμές με τον Hierarchical clustering single linkage να εμφανίζει ελάχιστα καλύτερη απόδοση και για τα δύο κριτήρια.




*Εικόνα 11 Σύγκριση αλγορίθμων με το κριτήριο Davies-Bouldin, Dataset 2*




*Εικόνα 12 Σύγκριση αλγορίθμων με το κριτήριο Silhouette, Dataset 2*