ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΩΝ

ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

«Μελέτη ενός τοπολογικού μοντέλου διαβάθμισης πιστοληπτικής ικανότητας κάτω από το πρίσμα διαφορετικών μέτρων αξιολόγησης μοντέλων.»

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

Η παρούσα Εργασία εκπονήθηκε

ως μερική ικανοποίηση των απαιτήσεων για την απόκτηση του του αντιστοίχου τίτλου σπουδών στην

Στατιστική και Αναλογιστικά-Χρηματοοικονομικά Μαθηματικά

(Γιάγκου Σταματίνα)

Ημερομηνία

ΣΑΜΟΣ 2024

Γιάγκου Σταματίνα

«Μελέτη ενός τοπολογικού μοντέλου διαβάθμισης πιστοληπτικής ικανότητας κάτω από το πρίσμα διαφορετικών μέτρων αξιολόγησης μοντέλων»

Ημερομηνία

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Συγγραφέας: Γιάγκου Σταματίνα

Επιβλέπων: Ξανθόπουλος Στυλιανός

Μέλος Επιτροπής: Λάππας Παντελής

Μέλος Επιτροπής: Αραμπατζής Δημήτριος

ΣΑΜΟΣ 2024

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο Πανεπιστήμιο Αιγαίου, στο Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών κατά το έτος του 2023-2024.

Η ολοκλήρωση της μεταπτυχιακής εργασίας θα ήταν αδύνατη χωρίς την πολύτιμη υποστήριξη του καθηγητή μου, κύριου Ξανθόπουλου Στυλιανού. Του εκφράζω ένα βαθύ ευχαριστώ για όλη τη βοήθεια που μου προσέφερε.

Χρωστάω, επίσης, ένα μεγάλο ευχαριστώ στην Μυρτώ Χαρμπή (υποψήφια διδάκτορα του τμήματος) για την άριστη συνεργασία που είχαμε στα πλαίσια εκπόνησης αυτής της εργασίας.

Τέλος, θέλω να ευχαριστήσω πολύ την οικογένεια και τους φίλους μου, οι οποίοι υπήρξαν πάντα ένα ανεκτίμητο στήριγμα για μένα και στους οποίους οφείλω όλη την διαδρομή των σπουδών μου μέχρι σήμερα.

Περίληψη

Η "τοπολογική ανάλυση δεδομένων" είναι ένας νέος και αναπτυσσόμενος επιστημονικός κλάδος που συνδυάζει την ανάλυση δεδομένων με έννοιες και μεθόδους της αλγεβρικής τοπολογίας προκειμένου να διερευνήσει το ερώτημα «τι σχήμα έχουν τα δεδομένα». Τα εργαλεία που έχουν αναπτυχθεί προς αυτή την κατεύθυνση επιχειρούν να ανιχνεύσουν, να ποσοτικοποιήσουν και να οπτικοποιήσουν με χρήσιμο τρόπο την όποια τοπολογική πληροφορία μπορεί να κρύβεται στο χώρο από τον οποίο προέρχονται τα δεδομένα. Η παρούσα εργασία παρουσιάζει αρχικά τα θεωρητικά θεμέλια και τα βασικά εργαλεία της τοπολογικής ανάλυσης δεδομένων, και στη συνέχεια εξετάζει παραλλαγές και επεκτάσεις του τοπολογικού μοντέλου διαβάθμισης πιστοληπτικής ικανότητας που έχει προταθεί στο [1]. Το συγκεκριμένο μοντέλο ξεχωρίζει για την εννοιολογική του απλότητα, την επεξηγησιμότητα και την ερμηνευσιμότητά του και κυρίως τη δυνατότητα του για ικανοποιητική οπτικοποίηση δεδομένων υψηλής διάστασης. Αυτή η δυνατότητα οπτικοποίησης ενισχύει την κατανόηση της πιστοληπτικής ικανότητας και το καθιστά δυνητικά χρήσιμο συμπλήρωμα σε πιο κλασσικά μοντέλα διαβάθμισης πιστοληπτικής ικανότητας, όπως π.χ. η λογιστική παλινδρόμηση ή τα νευρωνικά δίκτυα, δεδομένου μάλιστα ότι η απόδοση του είναι ιδιαίτερα ικανοποιητική και επαρκώς συγκρίσιμη με αυτά παρά την εννοιολογική του απλότητα.

Λέξεις-κλειδιά: Τοπολογία, Ανάλυση Δεδομένων, Πιστοληπτική ικανότητα, Ball Mapper, Accuracy, Precision, Recall, F1 score, Hmeasure

Abstract

"Topological data analysis" is a new and growing field of science that combines data analysis with concepts and methods from algebraic topology in order to investigate the question "what shape is the data". Tools developed in this direction attempt to detect, quantify and visualise in a useful way any topological information that may be hidden in the space from which the data is originated. This paper firstly presents the theoretical foundations and basic tools necessary of topological data analysis, and then reviews variations and extensions of the topological model of credit rating proposed in [1]. This model stands out for its conceptual simplicity, its explicability and interpretability, and most importantly its ability to satisfactorily visualize high-dimensional data. This visualisation capability enhances our understanding of creditworthiness and makes it a potentially useful complement to more classical models of credit rating, such as logistic regression or neural networks, given that its performance is highly satisfactory and sufficiently comparable to them despite its conceptual simplicity.

Keywords: Topology, Data Analysis, Creditworthiness, Ball Mapper, Accuracy, Precision, Recall, F1 score, Hmeasure

Πίνακας περιεχομένων

Σύνοψη εικόνων & πινάκων

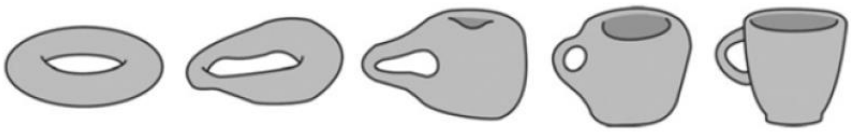
Κεφάλαιο1: Τοπολογικά Προαπαιτούμενα

1.1. Εισαγωγή

Η τοπολογία μελετά τις ιδιότητες των τοπολογικών χώρων, όπως η συνδεσιμότητα και η συμπαγής δομή τους, που διατηρούνται αναλλοίωτες κατά τις συνεχείς παραμορφώσεις. Η αλγεβρική τοπολογία μελετά τοπολογικούς χώρους χρησιμοποιώντας τεχνικές από την άλγεβρα προκειμένου να συσχετίζει τοπολογικά χαρακτηριστικά με αντίστοιχα αλγεβρικά αντικείμενα. Ένα από τα κύρια αλγεβρικά εργαλεία της είναι η θεωρία ομολογίας, που επιτρέπει την σύνδεση ακολουθιών αλγεβρικών αντικειμένων με την δομή ενός τοπολογικού χώρου [2]. Στο παρόν κεφάλαιο θα ορισθούν κάποιες βασικές έννοιες από την αλγεβρική τοπολογία που είναι χρήσιμες για την κατανόηση του τί είναι «σχήμα» των δεδομένων.

1.2. Τοπολογικές αναλλοίωτες και n-διάστατες οπές

Η τοπολογία αποτελεί έναν κλάδο των μαθηματικών που μελετά τις ιδιότητες των σχημάτων που παραμένουν αναλλοίωτες υπό συνεχείς παραμορφώσεις, όπως το τέντωμα, η περιστροφή ή το λύγισμα, αλλά όχι η αποκοπή και η επικόλληση. Για παράδειγμα, ένα ντόνατ με μια κούπα καφέ είναι τοπολογικά ισοδύναμα, αφού μπορεί κανείς να μετασχηματίσει το ένα στο άλλο συνεχώς [3].

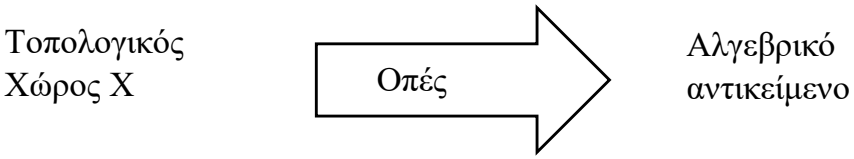


Εικόνα 1: Μια κούπα καφέ και ένας ντόνατ. Με την έννοια της αλγεβρικής τοπολογίας, αυτά τα δύο αντικείμενα είναι τα ίδια, επειδή μπορούν να μετασχηματιστούν το ένα στο άλλο διατηρώντας τα τοπολογικά τους χαρακτηριστικά [4].

Οι ιδιότητες αυτές των σχημάτων που δεν αλλάζουν μετά από συνεχείς μετασχηματισμούς ονομάζονται τοπολογικά αναλλοίωτες. Ο αριθμός των συνεκτικών συνιστωσών και το πλήθος οπών διαφόρων διαστάσεων αποτελούν παραδείγματα τοπολογικών αναλλοίωτων. Η αλγεβρική τοπολογία εξάγει τις αναλλοίωτες ενός αντικειμένου, τις καταμετρά και τις συσχετίζει με αλγεβρικά αντικείμενα, όπως διανυσματικούς χώρους [3].

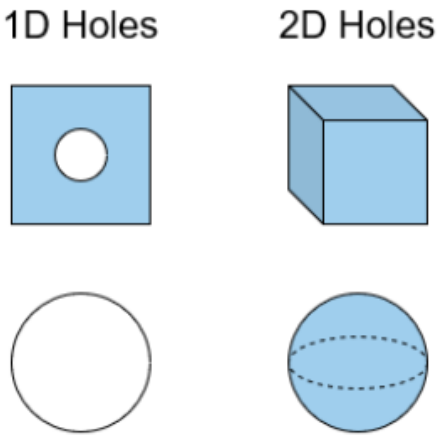
Η συσχέτιση ενός αλγεβρικού αντικειμένου, γίνεται με τρόπο που ενσωματώνει τοπολογικές πληροφορίες για τον τοπολογικό χώρο X στη δομή του αλγεβρικού αντικειμένου. Κατά συνέπεια, οι ερωτήσεις σχετικά με τον τοπολογικό χώρο X μπορούν να μεταφραστούν σε ερωτήσεις σχετικά με τη δομή του σχετιζόμενου αλγεβρικού αντικειμένου [5].

Όπως παρουσιάζεται στην εικόνα 2 η αντιστοίχιση από τον τοπολογικό χώρο X σε ένα αλγεβρικό αντικείμενο μπορεί να απεικονιστεί ως μια διαδικασία μεταφοράς πληροφοριών που δίνουν οι n-διάστατες οπές που υπάρχουν στον χώρο X.



Εικόνα 2: Χάρτης που μεταφέρει πληροφορίες για τις n-διάστατες τρύπες που υπάρχουν στον χώρο X [5].

Επιπρόσθετα, η διάσταση μιας οπής μπορεί να ορισθεί ως η διάσταση του συνόρου της. Για παράδειγμα στην εικόνα 3, ο χώρος που περικλείεται από τον μοναδιαίο κύκλο S^1 αποτελεί μια μονοδιάστατη οπή, καθώς το σύνορό της, ο ίδιος ο S^1, είναι ένα μονοδιάστατο αντικείμενο. Ομοίως, το κενό που περικλείεται από τη σφαίρα S^2 χαρακτηρίζεται ως 2-διάστατη οπή δεδομένου ότι το σύνορο της είναι μια 2-διάστατη επιφάνεια [5].



Εικόνα 3: Η διάσταση μιας οπής ορίζεται ως η διάσταση του συνόρου της [5].

Στην συνέχεια το κείμενο θα επικεντρωθεί αποκλειστικά σε μια συγκεκριμένη κατηγορία καλά συμπεριφερόμενων τοπολογικών χώρων, γνωστών ως simplicial complexes. Επιπρόσθετα θα περιγραφεί η μεθοδολογία για την κατασκευή των σχετικών αλγεβρικών αντικειμένων, γνωστών ως ομάδες ομολογίας. Αυτές οι ομάδες ομολογίας περικλείουν δομικές πληροφορίες που αφορούν n-διάστατες οπές.

1.3 Simplicial complexes

Τα simplicial complexes (πλεγματικά σύμπλοκα) εισάχθηκαν για πρώτη φορά το 1895 ως τριγωνοποίηση μιας πολλαπλότητας και χρησιμεύουν ως θεμελιώδη κατασκευάσματα στην εξερεύνηση και την δημιουργία τοπολογικών χώρων [3], [5]. Διαισθητικά, ένα simplicial complex γίνεται αντιληπτό ως ένα τοπολογικό αντικείμενο που κατασκευάζεται από την ένωση σημείων, ακμών, τριγώνων, τετραέδρων και ανάλογων υψηλότερης διάστασης πολυτόπων [3]. Δομείται από απλά κομμάτια που ονομάζονται simplices (πλέγματα) [6].

Ορισμός 1 (k-simplex): Για k ≥ 0, ένα k-simplex σ στον ευκλείδειο χώρο είναι η κυρτή θήκη (δηλαδή το μικρότερο κυρτό σύνολο) ενός συνόλου P={v_0, v_1, ..., v_k} που αποτελείται από k+1 ανεξάρτητα σημεία. Δηλαδή, το σ αποτελείται από όλα τα σημεία v τέτοια ώστε:

$$v=\sum_{i=0}^{k} \lambda_i v_i \text{ όπου } \sum_{i=0}^{k} \lambda_i = 1 \; και \;\; \lambda_i \geq 0 \text{ [5].}$$

Με τον όρο k+1 ανεξάρτητα σημεία εννοείται ότι τα v_0, v_1, ..., v_k δεν βρίσκονται σε κάποιο υπερεπίπεδο διάστασης μικρότερης από το k, κάτι που ισοδυναμεί με το ότι τα k διανύσματα v_1-v_0, v_2-v_0, ..., v_k - v_0 είναι γραμμικά ανεξάρτητα [5].

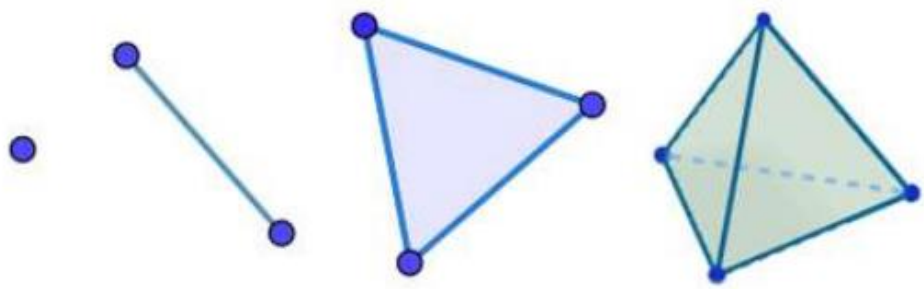
Αυτά τα σύμπλοκα αντιπροσωπεύουν τις απλούστερες μορφές n-διάστατων «τριγώνων» [5]. Για παράδειγμα:

-ένα 0-simplex είναι ένα σημείο,

-ένα 1- simplex είναι ένα ευθύγραμμο τμήμα,

-ένα 2- simplex είναι ένα τρίγωνο και

-ένα 3- simplex είναι ένα τετράεδρο



Εικόνα 4: . 0-simplex 1-simplex, 2-simplex, 3-simplex [5]

Ορισμός 2 (κορυφές και όψεις): Τα στοιχεία v_i του P, για i=0,1,…k, ονομάζονται κορυφές (vertices) του k-simplex σ και τα simplices που ορίζονται από τα υποσύνολα του P ονομάζονται όψεις (face) του σ.

Το ερώτημα που τίθεται τώρα είναι πως αυτά τα simplices μπορούν να κατασκευάσουν τα simplicial complexes. Αυτή η διαδικασία επιτυγχάνεται διαισθητικά μέσω ενός συγκεκριμένου τρόπου "συγκόλλησης" των simplices μεταξύ τους. Η κρίσιμη συνθήκη για αυτή τη συναρμολόγηση είναι ότι τα simplices πρέπει να είναι "κολλημένα" κατά μήκος των όψεων τους. Αξίζει να σημειωθεί ότι αν το P είναι ένα k-simplex, τότε το σύνολο όλων των όψεων του P, που συμβολίζεται face(P), δεν περιλαμβάνει μόνο τα (k-1)-simplex, αλλά και όλα τα simplex διαστάσεων από 0 μέχρι k. Αυτή η διαισθητική έννοια τυποποιείται επακριβώς στον ακόλουθο ορισμό [7].

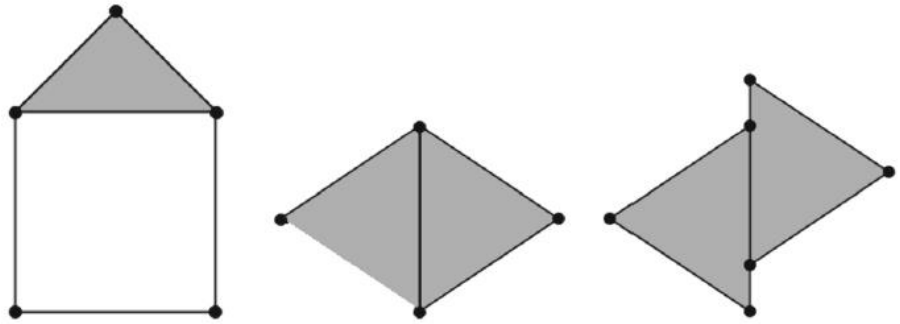
Ορισμός 3 (Γεωμετρικό Simplicial Complex): Έστω ένα simplicial complex K το οποίο είναι μια συλλογή από simplices, τέτοια ώστε:

- Κάθε όψη ενός simplex από το K να βρίσκεται επίσης στο K.

- Η μη κενή τομή οποιωνδήποτε δύο simplices σ_1, $\sigma_2 \in K$ είναι μια όψη τόσο του σ_1 όσο και του σ_2.

Η πρώτη συνθήκη λέει ότι αν ένα simplex, για παράδειγμα ένα τρίγωνο, είναι στο K, τότε οι όψεις του, όπως οι ακμές και οι κορυφές του, πρέπει επίσης να είναι στο K. Η δεύτερη συνθήκη λέει ότι μπορούν να ενωθούν simplex μόνο με τις κοινές τους όψεις. Για παράδειγμα, δύο τρίγωνα με μια κοινή όψη ή μια κοινή ακμή μπορούν να ενωθούν, αλλά μια κορυφή ενός τριγώνου δεν μπορεί να ενωθεί με μία από τις ακμές του άλλου τριγώνου [3].

Στην εικόνα 5 που ακολουθεί το πρώτο σχήμα είναι ένα παράδειγμα από simplicial complex, ενώ τα άλλα δυο δεν είναι αφού παραβιάζουν τη δεύτερη συνθήκη του ορισμού 3.



(α) (β) (γ)

Εικόνα 5:Το (α) είναι ένα simplicial complex, το (β) και το (γ) δεν είναι simplicial complex ,αφού στο (β) απουσιάζει μια ακμή και στο (γ) τα δύο τρίγωνα συναντώνται κατά μήκος μιας ακμής που δεν είναι ακμή κανενός τριγώνου [3]

Όσον αφορά την τοπολογία, το γεωμετρικό simplicial complex κωδικοποιεί επίσης πολλές πληροφορίες. Στην πραγματικότητα δεν είναι ουσιαστικό το πως ενσωματώνεται το αντικείμενο στο χώρο, αλλά κυρίως το πως είναι δομημένο. Με αυτό το δεδομένο εισάγεται η έννοια του αφηρημένου simplicial complex το οποίο απορρίπτει κάθε πληροφορία σχετικά με την θέση των σχημάτων στο χώρο και διατηρεί μόνο την έννοια της συνδεσιμότητας [7]. Το αφηρημένο simplicial complex είναι ένα καθαρά συνδυαστικό αντικείμενο όπως φαίνεται και από τον επόμενο ορισμό, κάτι που είναι ιδιαίτερα βολικό από υπολογιστικής σκοπιάς.

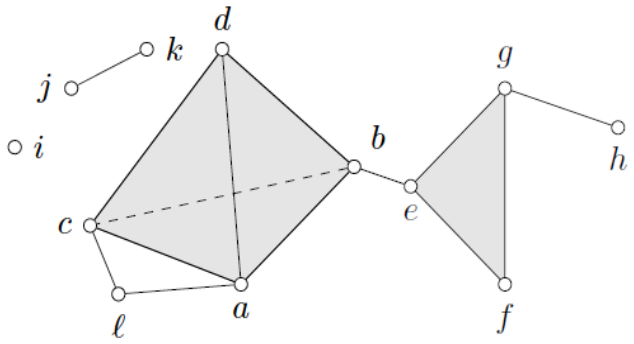
Ορισμός 4 (Αφηρημένο simplicial complex) : Ένα αφηρημένο simplicial complex K στο σύνολο των κορυφών $\{v_1,v_2,\dots v_n\}$, είναι μια συλλογή υποσυνόλων του $\{v_1,v_2,\dots v_n\}$,που είναι κλειστή ως προς τη σχέση εγκλεισμού, δηλαδή αν $\sigma \in K$ και τ μη κενό υποσύνολο του σ ($\tau \in \sigma$) , τότε $\tau \in K$ [7].

Τα στοιχεία του K είναι simplices. Αν $\sigma \in K$ η διάσταση του K ορίζεται ως εξής $\dim(\sigma) = |\sigma| - 1$. Δηλαδή η διάσταση ενός simplicial complex είναι απλά:

$$\dim(K) = max_{\sigma \in K}(\dim(\sigma))$$

Κάθε γεωμετρικό simplicial complex K έχει ένα αντίστοιχο αφηρημένο simplicial complex K' που δημιουργείται αφαιρώντας το κυρτό περίβλημα του K και διατηρώντας μόνο τις πληροφορίες των κορυφών. Αυτό το αντίστοιχο αφηρημένο simplicial complex K' ονομάζεται σχήμα κορυφών του K [7]. Από την άλλη κάθε αφηρημένο simplicial complex μπορεί να αναπαρασταθεί ως ένα γεωμετρικό simplicial complex, που ονομάζεται γεωμετρική του πραγματοποίηση, (αν και μπορεί να χρειάζεται μεγαλύτερη

διάσταση από αυτή του αφηρημένου simplicial complex) [8]. Στην εικόνα 6 παρουσιάζεται ένα 3-διάστατο simplicial complexes στο R^3.



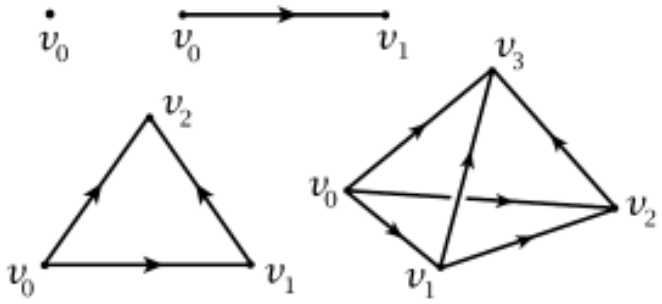
Εικόνα 6: Ένα γεωμετρικό simplicial complex στον R^3 είναι μια γεωμετρική πραγματοποιήση ενός αφηρημένου simplicial complex K = {[a, b, c, d], [e, f, g],…, [e,b],…, [a], [b],…,[l]} που περιλαμβάνει 32 simplices: ένα ενιαίο 3-simplex [a, b, c, d], 5 2-simplex όπως [a, c, d] και [e, f, g], δεκατέσσερα 1-simplices όπως τα [e,b] και [g, h], δώδεκα 0- simplicies [a],[b],…,[l]. Σημειώστε ότι στο γεωμετρικό simplicial complex, οι simplicies τέμνονται κατά μήκος των όψεων του [9].

Συνοψίζοντας, τα αφηρημένα simplicial complexes μπορούν να θεωρηθούν ως τοπολογική χώροι και τα γεωμετρικά simplicial complexes ως γεωμετρικές πραγματοποιήσεις της υποκείμενης συνδυαστικής δομής. Συνεπώς, μπορεί κάνεις να θεωρήσει τα simplicial complexes ταυτόχρονα ως συνδυαστικά αντικείμενα που είναι κατάλληλα για αποτελεσματικούς υπολογισμούς αλλά ταυτόχρονα και ως τοπολογικούς χώρους από τους οποίους μπορούν να εξαχθούν χρήσιμες τοπολογικές ιδιότητες. Αυτή η προσέγγιση αποτελεί κλειδί για το καθορισμό της "simplicial ομολογίας" μιας αλγεβρικής τοπολογικής μεθόδου που θα αναλυθεί λεπτομερώς στην επόμενη ενότητα.

1.4 Simplicial Homology

Η ομολογία είναι ένα αλγεβρικό εργαλείο για την ανίχνευση και την κωδικοποίηση n-διάστατων οπών σε ένα simplicial complex [7]. Στην ανάλυση ενός τέτοιου simplicial complex K, ο κεντρικός στόχος είναι ο υπολογισμός της ομολογίας του, δηλαδή μιας σειράς ομάδων που παρέχουν πληροφορίες για τις n-διάστατες οπές μέσα στο K. Αυτή η επιδίωξη καθιστά αναγκαία την εισαγωγή της έννοιας του προσανατολισμένου k-simplex [5].

Ορισμός 5 (προσανατολισμένο k-simplex): Έστω σ ένα simplex. Ορίζουμε δύο διαφορετικές διατάξεις των κορυφών του να είναι ισοδύναμες αν διαφέρουν κατά μία άρτια μετάθεση. Επομένως, στην περίπτωση $dim(\sigma)>0$ οι διατάξεις των κορυφών του k-simplex χωρίζονται σε δύο κλάσεις ισοδυναμίας. Κάθε μία από αυτές τις κλάσεις καλείται προσανατολισμός του σ. Αν το σ είναι 0-simplex υπάρχει μόνο ένας προσανατολισμός. Ένα προσανατολισμένο simplex σ είναι ένα simplex σ μαζί με κάποιον προσανατολισμό του σ [10].



Εικόνα 7 Παραδείγματα προσανατολισμένων simplicies [5]

Θεωρήθηκε ότι δύο προσανατολισμοί είναι ισοδύναμοι εάν και μόνο εάν οι διατάξεις τους διαφέρουν κατά μια άρτια μετάθεση. Αυτό σημαίνει ότι κάθε simplex διαθέτει ακριβώς δύο προσανατολισμούς. Μέσω αυτή της ισοδυναμίας επιτρέπεται να οριστεί το "αρνητικό" προσανατολισμένο simplex ως το simplex με τον αντίθετο προσανατολισμό. Παραδείγματος χάρη, θεωρώντας το 1-simplex $[v_i, v_j]$, το εκφράζουμε ως

$$[v_i, v_j] = -[v_j, v_i] \text{ [5].}$$

1.4.1 Οι ομάδες k-αλυσίδας C_k

Για την καλύτερη κατανόηση του κεφαλαίου αυτού εισάγονται αρχικά κάποιες βασικές αλγεβρικές έννοιες.

Ορισμός 6 (ομάδα): Ένα σύνολο G εφοδιασμένο με μία πράξη «+» καλείται ομάδα και συμβολίζεται $(G,+)$ εάν ικανοποιεί τις ακόλουθες ιδιότητες:

i. Για κάθε $a,b \in G$ το $a+b \in G$
ii. $(a+b)+c=a+(b+c)$
iii. Υπάρχει ουδέτερο στοιχείο $0 \in G$ τέτοιο ώστε $a+0=0+a=a$
iv. Για κάθε $a \in G$ υπάρχει αντίθετο στοιχείο $-a \in G$ τέτοιο ώστε $a+(-a)=(-a)+a=0$

Εάν ικανοποιείται η αντιμεταθετική ιδιότητα, δηλαδή $a+b=b+a \ \forall a,b \in G$ η ομάδα καλείται αβελιανή [10].

Ορισμός 7 (υποομάδα): Έστω $(G,+)$ μία ομάδα. Ένα υποσύνολο $H \subset G$ καλείται υποομάδα αν το $(H,+)$ είναι ομάδα.

Ορισμός 8 (Ελεύθερη αβελιανή ομάδα): Μια αβελιανή ομάδα $(G,+)$ ονομάζεται ελεύθερη εάν υπάρχει υποσύνολο $B \subseteq G$ ώστε κάθε στοιχείο του G να γράφεται με μοναδικό τρόπο ως πεπερασμένο άθροισμα στοιχείων του B και των αντιθέτων τους. Το σύνολο B καλείται βάση του G και το πλήθος των στοιχείων του ονομάζεται βαθμός. Στην περίπτωση που κάθε στοιχείο του G γράφεται ως πεπερασμένο άθροισμα στοιχείων του B αλλά όχι απαραίτητα μονοσήμαντα, το B λέγεται γεννήτορας του G [10].

Έστω ένα προσανατολισμένο simplicial complex K και έστω B_k το σύνολο όλων των k-διάστατων simplices του K. Η ομάδα C_k ορίζεται ως η ελεύθερη αβελιανή ομάδα με την B_k να χρησιμεύει ως βάση της.

Εννοιολογικά, μια ελεύθερη αβελιανή ομάδα κατασκευάζεται με την συναρμολόγηση των στοιχείων της βάσης της και σε αυτή την περίπτωση, κάθε στοιχείο της ομάδας C_k μπορεί να αναπαρασταθεί μοναδικά ως ακέραιος συνδυασμός ενός πεπερασμένου αριθμού στοιχείων βάσης. Κατά συνέπεια, τα στοιχεία εντός της C_k είναι πεπερασμένα τυπικά αθροίσματα γνωστά ως k-αλυσίδες.

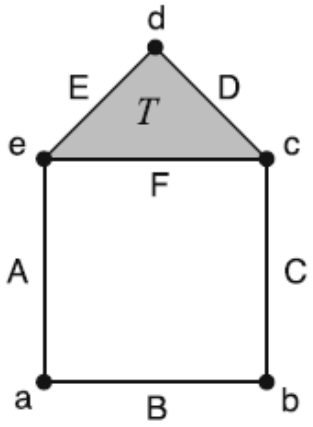
Ορισμός 9 (k-αλυσίδα): Μια k-αλυσίδα, που συμβολίζεται ως $c \in C_k$, έχει τη μορφή:

$$c = \sum_{i=0}^{N} a_i \sigma_i$$

όπου $\alpha_i \in Z$ (ακέραιοι αριθμοί), και το σ_i αντιπροσωπεύει ένα προσανατολισμένο simplex διάστασης k. Η k-αλυσίδα είναι ένα τυπικό άθροισμα των k-simplices ενός simplicial complex K με ακέραιους συντελεστές. Το σύνολο όλων των αλυσίδων του K είναι μια ομάδα και συμβολίζεται με C_k [5], [7].

Εάν $c' = \sum_{t=0}^{N} a'_i \sigma_i$ είναι μια άλλη k-αλυσίδα και $\lambda \in Z$, το άθροισμα c + c′ ορίζεται ως $c + c' = \sum_{t=0}^{N} (a_i + a'_i)\sigma_i$ και το γινόμενο λ*c ορίζεται ως $\lambda * c = \sum_{t=0}^{N} (\lambda a_i)\sigma_i$, καθιστώντας το C_k Z-module, δηλαδή μια δομή που μοιάζει με αυτή ενός διανυσματικού χώρου αλλά με συντελεστές στο Z. Αν από την άλλη το σύνολο των συντελεστών ήταν κάποιο $Z_p=\{0,1,\ldots,p-1\}$, δηλαδή οι ακέραιοι modulo p όπου p πρώτος αριθμός, τότε το Z_p θα ήταν σώμα και άρα το C_k θα μπορούσε να θεωρηθεί ως διανυσματικός χώρος επί του Z_p. Ιδιαίτερα στην περίπτωση όπου το σύνολο των συντελεστών είναι το Z_2 η έννοια του προσανατολισμού δεν έχει σημασία αφού 1+1=2=0(mod2) και άρα 1= -1(mod2).

Παράδειγμα : Έστω c_1 = a - 3b + 4d, c_2 = 2a + c - 2d + 3e είναι 0-αλυσίδες για το simplicial complexes της εικόνας 8. Μπορεί κανείς να προσθέσει δύο k-αλυσίδες προσθέτοντας απλά τον αντίστοιχο ακέραιο αριθμό συντελεστές, π.χ. $c_1 + c_2$ = 3a - 3b + c + 2d + 3e, και να πολλαπλασιάσει με κλιμάκια, π.χ. $2c_1$ = 2a-6b+8d [3].



Εικόνα 8: Ένα simplicial complexes με simplices [3].

1.4.2. Συνοριακός τελεστής ∂_k

Στην ενότητα αυτή εισάγεται η έννοια του συνοριακού τελεστή ενός simplex. Διαισθητικά ο συνοριακός τελεστής ενός k-simplex $\{v_0, v_1, ..., v_k\}$, μπορεί να θεωρηθεί ως το άθροισμα των (k-1)-διάστατων όψεών του. Αυτή η έννοια του τελεστή μπορεί να επεκταθεί και να οριστεί μια απεικόνιση μεταξύ ομάδων k-αλυσίδων [5].

Ορισμός 10 (k^{th} Boundary Map ∂_k): Για ένα προσανατολισμένο k-simplex $\sigma \epsilon C_k$ και έστω $\sigma=\{v_0, v_1, ..., v_k\}$, ο k-οστός συνοριακός τελεστής $\partial_k : C_k \rightarrow C_{k-1}$ δίνεται από την

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i (v_0, \dots, \hat{v}_i, \dots, v_k)$$

όπου ο συμβολισμός $(\hat{v}_i)$ δηλώνει ότι i-οστή κορυφή του σ παραλείπεται.

Ο ορισμός αυτός μπορεί να επεκταθεί σε k-αλυσίδες. Έστω μια k-αλυσίδα $c = c_i\sigma_i$ τότε ο συνοριακός τελεστής είναι $\vartheta_i(c) = \sum c_i \vartheta_i \sigma_i$. Στο παράδειγμα της εικόνας 8 ο συνοριακός τελεστής της ακμής A είναι $\vartheta_1(A) = e + a$ και της περιοχής T, $\vartheta_2(T) = E + D + F$.

Επίσης, πρέπει να σημειωθεί ότι το σύνορο ενός συνόρου είναι μηδέν, δηλαδή $\vartheta_i \vartheta_{i+1} = 0$. Αυτό αποδεικνύεται επίσης στο παράδειγμα της εικόνας 8 $\vartheta_1\vartheta_2(T) = \vartheta_1(E + D + F) = (d + e) + (e + c) + (c + d) = 2c + 2d + 2e = 0$, αφού $2=0$ αν υποθέσουμε ότι οι συντελεστές μας προέρχονται από το $Z_2 = Z/2Z = \{0,1\}$.

Μπορούν να διακριθούν τώρα δύο ειδικοί τύποι αλυσίδων χρησιμοποιώντας τον συνοριακό τελεστή.

Ορισμός 11 (k-κύκλος): Ένας k-κύκλος είναι μια k-αλυσίδα με μηδενικό σύνορο. Με άλλα λόγια μια k-αλυσίδα c είναι ένας k-κύκλος εάν και μόνο εάν το $\vartheta_i(c) = 0$, δηλαδή $c \in \ker(\vartheta_i)$.

Παραδείγματος χάριν, η 1-αλυσίδα A+B+C+F στην εικόνα 8 είναι ο 1-κύκλος αφού, $\vartheta_1(A + B + C + F) = \vartheta_1(A) + \vartheta_1(B) + \vartheta_1(C) + \vartheta_1(F) = (e + a) + (a + b) + (b + c) + (c + e) = 0$.

Το σύνολο όλων αυτών των i-κύκλων σχηματίζει έναν υποχώρο στο C_k, τον οποίο συμβολίζουμε Z_k.

Ορισμός 12 (k-όριο): Μια k-αλυσίδα c είναι k-σύνορο αν υπάρχει μια (k-1) αλυσίδα d της οποίας το σύνορο είναι η c, δηλαδή $c = \vartheta_{i+1}(d)$,ή με άλλα λόγια $c \in \mathrm{im}(\vartheta_{i+1})$.

Παραδείγματος χάριν, η 1-αλυσίδα E+D+F στην εικόνα 8 είναι 1-σύνορο αφού, $E + D + F = \vartheta_1(T)$.

Το σύνολο όλων αυτών των i-ορίων σχηματίζει έναν υποχώρο στο C_k, τον οποίο συμβολίζουμε B_k.

1.4.3 Το σύμπλεγμα αλυσίδων και οι ομάδες ομολογίας $H_k(K)$

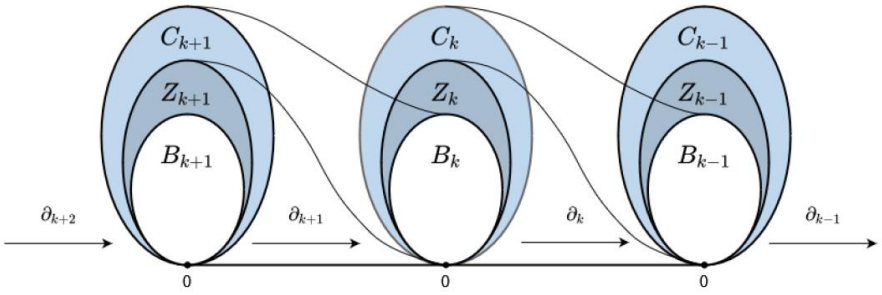
Με ένα απλό υπολογισμό, γίνεται φανερό ότι η σύνθεση δύο συνοριακών τελεστών $\partial_k \circ \partial_{k+1} : C_{k+1} \to C_{k-1}$ οδηγεί στη μηδενική απεικόνιση. Αυτή η ιδιότητα του των συνοριακών τελεστών, σε συνδυασμό με τις ομάδες k-αλυσίδων, κατασκευάζει ένα σύμπλεγμα αλυσίδων - μια ακολουθία αβελιανών ομάδων και ομομορφισμών που είναι διατεταγμένες ως εξής [4]:

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

Η ιδιότητα ότι η σύνθεση δύο διαδοχικών συνοριακών τελεστών είναι ο μηδενικός τελεστής, συνεπάγεται ότι η εικόνα του k+1 συνοριακού τελεστή πρέπει να περιλαμβάνεται στον πυρήνα του k-οστού συνοριακού τελεστή:

$$\partial_k \circ \partial_{k+1} = 0 \implies Im(\partial_{k+1}) \subset Ker(\partial_k)$$

Τα στοιχεία που ανήκουν στο $B_k := \mathrm{Im}(\partial_k+1)$ ονομάζονται σύνορα, ενώ εκείνα στο $Z_k := \mathrm{Ker}(\partial_k)$ αναφέρονται ως κύκλοι, όπως ορίσθηκαν και στο κεφάλαιο 1.4.2. Η σχέση αυτή, $B_k \subseteq Z_k$, απεικονίζεται στην εικόνα 9. Με αυτό το υπόβαθρο, μπορούμε να προχωρήσουμε στον ορισμό των ομάδων ομολογίας του simplicial complex K [5].



Εικόνα 9: Σύμπλεγμα αλυσίδων με οριακούς χάρτες [5].

Ορισμός 13 (k^{th} Homology Group H_k): Για ένα προσανατολισμένο simplicial complex K, η k-όστη ομάδα ομολογίας του K ορίζεται ως το πηλίκο δύο ομάδων

$$H_k = Z_k / B_k$$

Παρατηρείται ότι εφόσον η αρχική αλυσιδωτή ομάδα C_k ήταν αβελιανή, οι υποομάδες Z_k και B_k είναι αβελιανές. Είναι σημαντικό να αναγνωρίσουμε ότι η k-ομάδα ομολογίας κατέχει μη μηδενική τιμή ακριβώς όταν υπάρχουν k-κύκλοι στο K που δεν προέκυψαν ως σύνορα από k+1-simplices. Αρά όταν ορίζεται μια k-διάστατη οπή στο simplicial complex K, συνεπάγεται ότι ορίζεται μια k-οστή ομάδα ομολογίας που είναι μη τετριμμένη [5].

Η k-οστή ομάδα ομολογίας ανιχνεύει την παρουσία k-διάστατων οπών. Για παράδειγμα, η ομάδα ομολογίας $H_1(K)$ θα περιέχει πληροφορίες σχετικά με τις μονοδιάστατες οπές στο K, ενώ η $H_2(K)$ θα περιέχει πληροφορίες σχετικά με τις

δισδιάστατες οπές στο K. Στην περίπτωση που k = 0, η ομάδα ομολογίας $H_0(K)$ θα αποκαλύψει την παρουσία συνεκτικών συνιστωσών ή "0-διάστατων" οπών [5].

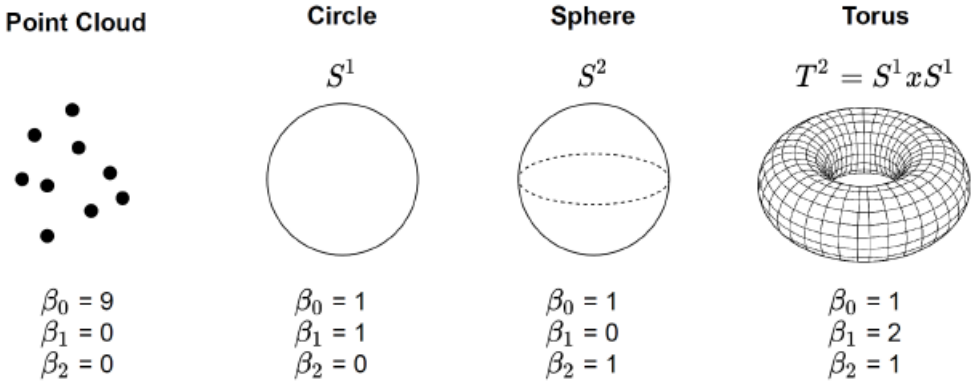
Επιπλέον, η k-ομάδα ομολογίας μεταφέρει πληροφορίες σχετικά με το πόσες k-διάστατες οπές υπάρχουν. Η ιδέα εδώ είναι ότι κάθε k-διάστατη οπή στο K θα αντιστοιχεί σε ένα γεννήτορα της ομάδας ομολογίας H_k. [5].

Ορισμός 14 (k-οστός αριθμός Betti β_k): Για ένα προσανατολισμένο simplicial complex K, ο k-οστός αριθμός Betti β_k ορίζεται ως ο βαθμός της ομάδας ομολογίας $H_k(K)$

$$\beta_k = \text{rank}(H_k) = \text{rank}(Z_k) - \text{rank}(B_k)$$

Ουσιαστικά, ο k-οστός αριθμός Betti β_k μετρά τον αριθμό των k-διάστατων οπών που υπάρχουν στο simplicial complex K. Οι πρώτοι αριθμοί Betti για μερικούς απλούς τοπολογικούς χώρους παρουσιάζονται στην εικόνα 10 [5], όπου

1. Το β_0 είναι ο αριθμός των συνδεδεμένων συνιστωσών.
2. Το β_1 είναι ο αριθμός των δισδιάστατων οπών.
3. Το β_2 είναι ο αριθμός των τρισδιάστατων οπών [11].



Εικόνα 10: Παρουσιάζονται οι πρώτοι αριθμοί Betti για ορισμένους κοινούς τοπολογικούς χώρους. Υπενθυμίζεται ότι το 0 μετράει τον αριθμό των συνδεδεμένων συνιστωσών, το 1 μετρά τον αριθμό των μονοδιάστατων οπών και το 2 μετρά τον αριθμό των αριθμό των δισδιάστατων οπών [5].

Κεφάλαιο 2: Τοπολογική Ανάλυση Δεδομένων

2.1. Εισαγωγή

Η τοπολογική ανάλυση δεδομένων χρησιμοποιεί έννοιες από την αλγεβρική τοπολογία για να αποκαλύψει την δομή και το σχήμα των δεδομένων, όπως συστάδες ή οπές, που ενδέχεται να μην είναι εμφανείς με την χρήση παραδοσιακών μεθόδων. [6] Αυτό ισχύει ακόμα και για υψηλής διάστασης και για θορυβώδη ή ελλιπή δεδομένα, όπου η συνδεσιμότητα και η παρουσία οπών αποτελούν πηγή χρήσιμων τοπολογικών πληροφοριών [5]. Στο παρόν κεφάλαιο παρουσιάζεται εν συντομία ένα θεμελιώδες εργαλείο της τοπολογικής ανάλυσης δεδομένων, η εμμένουσα ομολογία η οποία έχει σαν βασική ιδέα την χρήση της ομολογίας.

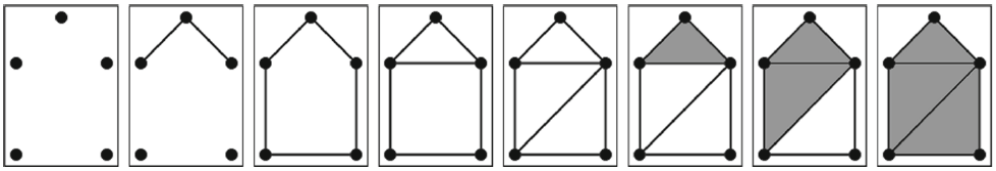
2.2 Εμμένουσα Ομολογία

Ένα ισχυρό εργαλείο της τοπολογικής ανάλυσης δεδομένων για την εξέταση της δομής των δεδομένων είναι η εμμένουσα ομολογία [12], μια αλγεβρική μέθοδος που μετράει τα τοπολογικά χαρακτηριστικά σχημάτων και συναρτήσεων [13]. Χρησιμεύει ως μεθοδολογία για τον υπολογισμό τοπολογικών χαρακτηριστικών ενός χώρου σε διάφορες κλίμακες, εστιάζοντας στον εντοπισμό της ύπαρξης n-διάστατων οπών σε ένα σύνολο δεδομένων. Κυρίως, παρακολουθεί τη χρονολογική εμφάνιση και εξαφάνιση αυτών των χαρακτηριστικών [7].

2.2.1 Διήθηση- Filtration

Σε προηγούμενο κεφάλαιο εισάχθηκε η έννοια της simplicial ομολογίας, ορίσθηκε το simplicial complex και κατασκευάστηκε μια ακολουθία ομάδων που περιείχε πληροφορίες σχετικά με το πλήθος των (ανεξάρτητων) n-διάστατων οπών του simplicial complex. Παρ' όλα αυτά, τα σύνολα δεδομένων εκδηλώνονται συνήθως ως νέφη σημείων δηλαδή ως ένα σύνολο διακριτών σημείων ενσωματωμένο σε κάποιο μετρικό χώρο. Για αυτά τα νέφη σημείων, η ομολογία δεν προσφέρει ενδιαφέρουσες πληροφορίες αφού μόνο το β_0 δίνει τον αριθμό των συνδεδεμένων συνιστωσών και οι άλλοι αριθμοί Betti είναι μηδενικοί, διότι δεν υπάρχουν n-διαστατές οπές στο σύνολο [3].

Ως εκ τούτου, αντί να εργάζεται κάποιος με το σύνολο των δεδομένων, μπορεί να δημιουργήσει μια οικογένεια από simplicial complexes, έστω K_X^δ για ένα εύρος τιμών δ ∈ R από το σύνολο X του νέφους των δεδομένων, έτσι ώστε το complex m να ενσωματώνεται στο complex n, για m ≤ n, δηλαδή $K_X^m \subseteq K_X^n$. Αυτή η φωλιασμένη (nested) οικογένεια simplicial complex ονομάζεται διήθηση [3].



Εικόνα 11: Παράδειγμα διήθησης για δ = 0, 1, 2, 3, 4, 5, 6, 7 [3].

Κατά τη διάρκεια αυτής της κατασκευής, ορισμένες οπές μπορεί να εμφανιστούν και στη συνέχεια να εξαφανιστούν, και η επιμονή αυτών των ομολογικών χαρακτηριστικών μπορεί να εκληφθεί ως ένδειξη για τα τοπολογικά χαρακτηριστικά του συνόλου δεδομένων. Σε μια διήθηση, μπορεί να καταγραφεί η γέννηση μιας οπής, δηλαδή η στιγμή που εμφανίζεται και ο θάνατος της, δηλαδή η στιγμή που εξαφανίζεται. Η ουσία της εμμένουσας ομολογίας είναι να εντοπίσει τη γέννηση και το θάνατο αυτών των ομολογικών χαρακτηριστικών στο K_X^{δ} για διαφορετικές τιμές του δ [3]. Η διάρκεια ζωής κάθε ομολογικού χαρακτηριστικού μπορεί να απεικονισθεί μέσα από γραφικές αναπαραστάσεις όπως τα Bar Codes και τα Persistence Diagrams.

2.2.2 Bar Codes and Persistence Diagrams

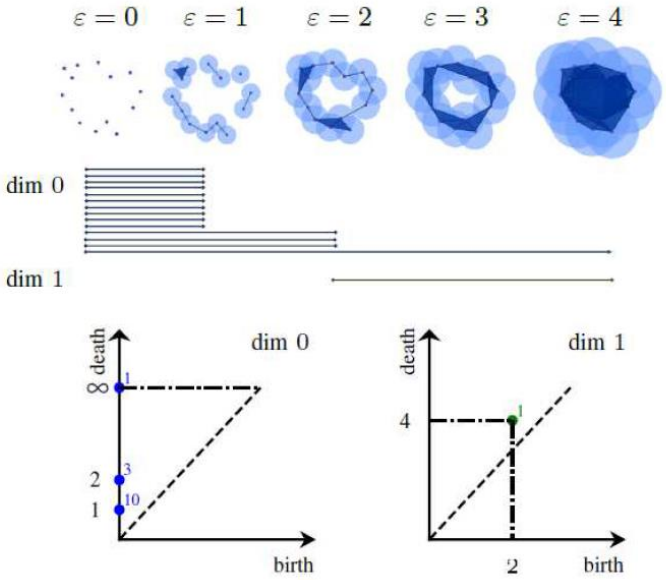
Όπως συζητήθηκε προηγουμένως, η εμμένουσα ομολογία ξεκινά την ανάλυσή της με ένα νέφος σημείων. Καθώς κινείται μέσα από μια διήθηση αυτού του συνόλου, μια ενσωματωμένη ακολουθία simplicial complex, καταγράφει το «πότε» μια ομάδα ομολογίας (οπή) εμφανίζεται και πότε εξαφανίζεται. Κατά συνέπεια για κάθε ομάδα ομολογίας που ανιχνεύεται προκύπτει ένα ζεύγος πραγματικών αριθμών που συμβολίζονται ως ϵ_1 και ϵ_2 και σηματοδοτούν τη γέννηση και το θάνατο του συγκεκριμένου χαρακτηριστικού εντός της διήθησης.

Η αναπαράσταση αυτών των ζευγών μπορεί να επιτευχθεί μέσω δύο κοινών μεθόδων.

1. **<u>Γραφική Παράσταση Ράβδων (Bar-Codes)</u>**: Μια διαδικασία που χρησιμοποιείται για την αναπαράσταση των αποτελεσμάτων της εμμένουσας ομολογίας είναι η χρήση ενός bar-code. Το bar-code είναι μια γραφική αναπαράσταση των κλάσεων ομολογίας (οπών) που εκδηλώνεται ως ένα σύνολο οριζόντιων γραμμών με τον άξονα x να αντιστοιχεί στην παράμετρο (ακτίνα) ε. Κάθε οριζόντια γραμμή αντιστοιχεί στη γέννηση και το θάνατο ενός συγκεκριμένου τοπολογικού χαρακτηριστικού. Η βασική ιδέα είναι ότι οι μακρύτερες ράβδοι, αυτές που παραμένουν σε ένα ευρύτερο φάσμα τιμών ε, συσχετίζονται με πιο ισχυρά τοπολογικά χαρακτηριστικά και είναι πιο αντιπροσωπευτικές της συνολικής δομής που υπάρχει στα δεδομένα

2. **<u>Διάγραμμα Εμμένουσας Ομολογίας (Persistence Diagrams):</u>** Μια εναλλακτική μέθοδος για την απεικόνιση αυτών των ζευγών είναι μέσω ενός διαγράμματος επιμονής. Το Persistence Diagrams (δισδιάστατο διάγραμμα) αναπαριστά τα σημεία γέννησης και θανάτου κάθε κλάσης ομολογίας (οπή), λαμβάνοντας υπόψη τις πολλαπλότητες. Σε αυτό το διάγραμμα είναι δυνατόν διαφορετικά χαρακτηριστικά να μοιράζονται τις ίδιες συντεταγμένες γέννησης και θανάτου. Ο άξονας x αντιστοιχεί στη γέννηση ενός

χαρακτηριστικού ενώ ο άξονας y αντιπροσωπεύει τον θάνατό του [6]. Αξίζει να σημειωθεί ότι, δεδομένου ότι το ϵ_1 πρέπει να είναι μικρότερο από το ϵ_2, ένα χαρακτηριστικό δεν μπορεί να πάψει να υπάρχει πριν από την εμφάνισή του. Κατά συνέπεια, όλα τα σημεία σε ένα Persistence Diagrams βρίσκονται πάνω από τη γραμμή y = x. Η βασική ιδέα είναι ότι τα σημεία που βρίσκονται πιο μακριά από τη γραμμή y = x, υποδεικνύοντας μεγαλύτερη "ζωή", αντιστοιχούν σε χαρακτηριστικά που συμβάλλουν πιο σημαντικά στη συνολική δομή των δεδομένων, επιμένοντας σε ένα ευρύτερο φάσμα κλιμάκων. Αντίθετα, τα σημεία που βρίσκονται κοντά στη γραμμή y = x υποδηλώνουν χαρακτηριστικά που είχαν σύντομη ύπαρξη και θεωρούνται λιγότερο κρίσιμα για τη δομή των δεδομένων.



Εικόνα 12: Bar Codes και Persistent Diagrams για διήθηση τεσσάρων βημάτων ενός νέφους σημείων [5].

2.2.3 Νεύρο, Cech και Vietoris-Rips Complex

Σε αυτό το σημείο τίθενται δύο ερωτήματα. Το πρώτο πώς μπορεί να κατασκευαστεί ένα simplicial complex πάνω σε ένα σύνολο δεδομένων; Υπάρχουν διάφοροι τρόποι για να προκύψει ένα simplicial complex από ένα σύνολο δεδομένων νέφους σημείων όπως το Cech complex & το Vietoris-Rips Complex. Το δεύτερο ερώτημα που τίθεται είναι πως αναγνωρίζεται αν η επιλογή του simplicial complex αποτελεί μια καλή προσέγγιση του χώρου από τον οποίο προέρχονται τα δεδομένα. Την απάντηση στο δεύτερο ερώτημα, δίνει το Θεώρημα του Νεύρου, το οποίο παρέχει τις συνθήκες υπό τις οποίες ένας χώρος και το νεύρο ενός καλύμματός του είναι ομοτοπικά ισοδύναμα.

Αρχικά αναφέρονται οι παρακάτω ορισμοί:

Ορισμός 15 (μετρικός χώρος) : Ένας μετρικός χώρος είναι ένα ζεύγος (X,d_x), όπου X είναι ένα σύνολο εφοδιασμένο με μία μετρική d_x: $X \times X \rightarrow$ R που ικανοποιεί τις ακόλουθες ιδιότητες για κάθε $x,y,z \in X$:

i. $d_x(x,y) \geq 0$ $και$ $d_x(x,y) = 0 \Leftrightarrow x = y$

ii. $d_x(x,y) = d_x(y,x)$

iii. $d_x(x,z) \leq d_x(x,y) + d_x(y,z)$ [10]

Ορισμός 16 (ανοιχτή μπάλα) : Έστω (X,d_x) μετρικός χώρος, $x \in X$ και $\varepsilon>0$. Θέτουμε $B_\varepsilon(x) = \{y \in X \mid d_x(x,y) < \varepsilon\}$. Το σύνολο $B_\varepsilon(x)$ ονομάζεται ανοιχτή μπάλα με κέντρο x και ακτίνα ε ή ε – περιοχή του x [10].

Ορισμός 17 (Νεύρο) : Δοθείσης μίας πεπερασμένης συλλογής συνόλων $\mathcal{U}=\{U_a\}, a \in A$, ορίζουμε το νεύρο του $\mathcal{U}$ να είναι το simplicial complex $N(\mathcal{U})$ το οποίο έχει ως σύνολο κορυφών το σύνολο δεικτών A και ένα υποσύνολο $\{a_0,a_1,\ldots,a_k\} \subseteq A$ ορίζει ένα k-simplex του $N(\mathcal{U})$ αν και μόνο αν $Ua_o \cap Ua_1 \cap \ldots \cap Ua_k \neq \emptyset$ [10].

Η ένωση των συνόλων ενός καλύμματος του υποκείμενου τοπολογικού χώρου αποτελεί την προσέγγισή για τον άγνωστο χώρο. Εάν το κάλυμμα είναι «καλό», το νεύρο του καλύμματος αιχμαλωτίζει την τοπολογία του άγνωστου χώρου από τον οποίο προέρχονται τα δεδομένα. Το παρακάτω θεώρημα συνδέει –υπό κάποιες προϋποθέσεις– την τοπολογία του νεύρου ενός καλύμματος με την τοπολογία της ένωσης των συνόλων του καλύμματος.

Το Θεώρημα του Νεύρου: Έστω $\mathcal{U}=\{U_a\}_{a \in A}$ ένα (πεπερασμένο) κάλυμμα ενός τοπολογικού χώρου X τέτοιο ώστε οποιαδήποτε τομή $\bigcap Ua_i k_i=0$ να είναι είτε κενή είτε συσταλτή. Τότε, ο X και το νεύρο $N(\mathcal{U})$ είναι ομοτοπικά ισοδύναμα [10].

Για τα παρακάτω θα θεωρήσουμε ότι ο άγνωστος χώρος από τον οποίο λάβαμε τα δεδομένα είναι ένας μετρικός χώρος. Δοθέντος ενός πεπερασμένου υποσυνόλου P ενός μετρικού χώρου (M,ρ) μπορούμε να κατασκευάσουμε ένα αφηρημένο simplicial complex με κορυφές στο P χρησιμοποιώντας την έννοια του νεύρου.

Ορισμός 18 (Čech complex) : Έστω (M,ρ) μετρικός χώρος, P ένα πεπερασμένο υποσύνολο του M και ένας πραγματικός αριθμός $r>0$. Το Čech complex, $\check{\mathbf{C}}\mathbf{ech_r(P)}$, είναι το νεύρο του συνόλου $\{B(p_i,r)\}$, όπου $B(p_i,r)=\{x \in \mathrm{M} \mid \rho(p_i,x) \leq r\}$ είναι η κλειστή μπάλα με κέντρο p_i και ακτίνα r.

Στην περίπτωση που το M είναι ο ευκλείδειος χώρος οι μπάλες που χρησιμοποιούνται για την κατασκευή του Čech complex είναι κυρτά σύνολα και συνεπώς οι τομές τους είναι συσταλτές. Επομένως, το Čech complex είναι ομοτοπικά ισοδύναμο με τον X. Το Čech complex δεν χρησιμοποιείται στην πράξη εξαιτίας της υψηλής υπολογιστικής πολυπλοκότητάς του. Αντίθετα, το Vietoris-Rips complex είναι ευρέως διαδεδομένο στην τοπολογική ανάλυση δεδομένων λόγω της ευκολίας κατασκευής του.

Ορισμός 19 (Vietoris-Rips Complex) : Έστω ο μετρικός χώρος (X, d_x). Για $\varepsilon>0$ ορίζεται ένα αφηρημένο simplicial complex $\boldsymbol{Rips_\varepsilon(X)}$ στο σύνολο των κορυφών του X από την ακόλουθη συνθήκη:

$$[x_0, x_1, \ldots, x_k] \in Rips_\varepsilon(X) \Leftrightarrow d_x(x_i, x_j) \leq \varepsilon, για\ κάθε\ i, j$$

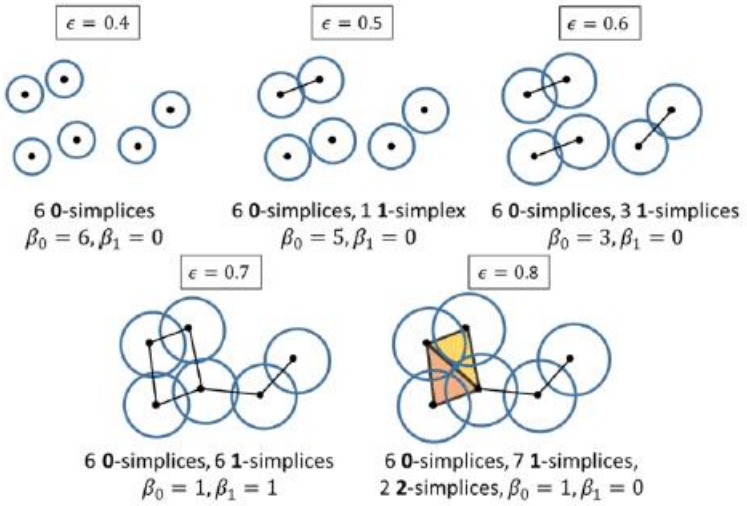
Υπάρχει ένας φυσικός εγκλεισμός $Rips_a(X) \subseteq Rips_b(X)$, για $a \leq b$. Έτσι, το simplicial complex $Rips_\varepsilon(X)$ σε συνδυασμό με τις σχέσεις εγκλεισμού ορίζουν ένα filtered complex στο X, το Vietoris – Rips complex.

Το Vietoris-Rips complex VR(S) αναπτύχθηκε αρχικά από τον Leopold Vietoris ως μέσο υπολογισμού της ομολογίας των μετρικών χώρων [70]. Για την κατασκευή του συμπλόκου Rips σε ένα πεπερασμένο υποσύνολο σημείων S, χρησιμοποιείται η ακόλουθη διαδικασία:

• Ορισμός μιας παραμέτρου ε

• Για όλα τα υποσύνολα s ⊆ S

• Αν η απόσταση s ≤ 2ε, συμπεριλαμβάνεται το πλέγμα με κορυφές στο s [14].



Εικόνα 13: Παράδειγμα συμπλόκων Vietoris-Rips και αντίστοιχοι αριθμοί Betti για διαφορετικές παραμέτρους κλίμακας [15].

Στο ερώτημα για το ποια είναι η κατάλληλη επιλογή του ε, το οποίο θα αποκαλύψει αποτελεσματικά τις πιο κρίσιμες πτυχές ενός δοθέντος συνόλου δεδομένων, θα απαντήσει η εμμένουσα ομολογία η οποία εξετάζει ένα εύρος ε, αξιοποιώντας την έννοια των διηθήσεων. Αρχικά όταν το ε=0, το simplicial complex VR(0) αποτελείται αποκλειστικά από τα σημεία του αρχικού νέφους σχηματίζοντας ένα σύνολο από 0-simplices. Καθώς το ε αυξάνεται στο σύμπλοκο κάποια από τα αρχικά simplices «ενώνονται» και σχηματίζουν καινούργια simplicial complexes [5].

Για μια αύξουσα ακολουθία τιμών ε δημιουργείται μια διήθηση η οποία αντιπροσωπεύει μια ακολουθία αναπτυσσόμενων simplicial complex:

$VR(\varepsilon_0) \subset V\,R(\varepsilon_1) \subset ...$

Η εμμένουσα ομολογία παρακολουθεί συστηματικά τις ομάδες ομολογίας καθώς εξελίσσεται αυτή η διήθηση. Καταγράφει τις τιμές του ε όπου εμφανίζονται τρύπες διάστασης n και σημειώνει τις τιμές όπου αυτές οι τρύπες παύουν να υπάρχουν. Στην ουσία καταγράφει τη διάρκεια ή την επιμονή αυτών των χαρακτηριστικών [5].

Κεφάλαιο 3: Ο αλγόριθμος MAPPER

3.1 . Εισαγωγή

Ο αλγόριθμος Mapper είναι ένα εργαλείο οπτικοποίησης, εξερεύνησης και σύνοψης πληροφοριών για σύνολα δεδομένων, το οποίο χρησιμοποιείται τόσο σε ακαδημαϊκό όσο και εμπορικό επίπεδο [4]. Πιο συγκεκριμένα, ο αρχικός αλγόριθμος mapper διαιρεί ένα σύνολο δεδομένων σε επικαλυπτόμενες "φέτες" και επιλέγει μια κλίμακα ανεξάρτητα για κάθε μια, προκειμένου να ομαδοποιήσει τα δεδομένα εντός κάθε "φέτας" [16]. Εκτελώντας την ομαδοποίηση στα επικαλυπτόμενα υποσύνολα των δεδομένων με πολλές διαστάσεις απλοποιούνται τα σύνθετα νέφη σημείων σε δύο ή τρείς μόνο διαστάσεις [17].

3.2 Η τοπολογική κατασκευή του Mapper

Η θεωρητική κατασκευή του αλγόριθμου Mapper έχει ως κίνητρο την ακόλουθη τοπολογική κατασκευή που συνδέει ένα simplicial complex με έναν τοπολογικό χώρο μέσω μιας συνεχούς συνάρτησης. Η κατασκευή του Mapper αναπτύσσεται σε έναν τοπολογικό χώρο X. Έστω, επίσης, Z ένας παραμετρικός χώρος με πεπερασμένη ανοικτή κάλυψη $U = \{U_\alpha\}_{\alpha \in A}$ και έστω $f : X \to Z$ συνεχής συνάρτηση. Το σύνολο $\{f^{-1}(U_\alpha)\}_{\alpha \in A}$, λόγω της συνέχειας, αποτελεί μια ανοικτή κάλυψη[1] του X. Στη συνέχεια, αναγνωρίζουμε τις συνδεδεμένες με μονοπάτια[2] συνιστώσες του $f^{-1}(U_\alpha)$ για όλα τα $\alpha \in A$. Κάθε συνεκτική συνιστώσα αντιστοιχεί σε μία κορυφή στο simplicial complex. Σε περίπτωση όπου οι συνεκτικές συνιστώσες από διαφορετικές προ-εικόνες $\{f^{-1}(U_\alpha)\}$ και $\{f^{-1}(U_\beta)\}$ αλληλεπικαλύπτονται, οι κορυφές που αντιστοιχούν στις συνδεόμενες με μονοπάτι συνιστώσες ενώνονται με μια ακμή [18].

3.3 Τοπολογική Ανάλυση Δεδομένων & Mapper

Με τον αλγόριθμο Mapper κατασκευάζεται ένα γράφημα δικτύου ή γενικότερα ένα simplicial complex από ένα σύνολο δεδομένων με τρόπο που να αποτυπώνει τα τοπολογικά χαρακτηριστικά των δεδομένων. Ανεπίσημα, ο αλγόριθμος Mapper λειτουργεί εκτελώντας μια τοπική ομαδοποίηση βασιζόμενη σε μια δοθείσα μέθοδο ομαδοποίησης που καθοδηγείται από μια συνάρτηση φίλτρου. Τα βήματα έχουν ως εξής [19]:

1. Κατά την επεξεργασία ενός πραγματικού συνόλου δεδομένων, το αρχικό βήμα περιλαμβάνει την επιλογή μιας μετρικής απόστασης (όπως η τρισδιάστατη ευκλείδεια απόσταση) για να περιγράφει η απόσταση μεταξύ κάθε ζεύγους σημείων. Δεν είναι απαραίτητο να χρησιμοποιηθεί η Ευκλείδεια μετρική, μπορεί επίσης να χρησιμοποιηθεί μια προκαθορισμένη μετρική.

[1] Μια ανοιχτή κάλυψη του τοπολογικού χώρου σημαίνει ότι μπορούμε να βρούμε μια συλλογή ανοιχτών υποσυνόλων του X, των οποίων η ένωση είναι ολόκληρος ο X.

[2] Ένα υποσύνολο $A \subseteq X$ ενός τοπολογικού χώρου X ονομάζεται συνδεδεμένο με μονοπάτια, αν για κάθε $x, y \in A$ υπάρχει ένα συνεχές μονοπάτι $\gamma : [0, 1] \to A$ τέτοιο ώστε $\gamma(0) = x$ και $\gamma(1) = y$.

Αυτή η επιλεγμένη μετρική χρησιμοποιείται στα επόμενα βήματα από μια συνάρτηση φίλτρου με σκοπό την προβολή των δεδομένων ή τη μείωση των διαστάσεων [19].

2. Στη συνέχεια, ο Mapper προχωρά στην προβολή του αρχικού συνόλου δεδομένων σε έναν χώρο χαμηλότερης διάστασης χρησιμοποιώντας μια συνάρτηση φίλτρου. Αυτή η συνάρτηση φίλτρου θα μπορούσε να είναι η PCA, η MDS, η t-SNE ή οποιαδήποτε άλλη μέθοδος μείωσης διαστάσεων. Είναι επίσης δυνατό να συνδυαστούν διαφορετικά φίλτρα για τη δημιουργία ενός φακού (lens) ή ενός χώρου χαμηλής διάστασης [19].

3. Μετά το φιλτράρισμα, το σύνολο δεδομένων διαχωρίζεται σε ομάδες (υπερκύβους) σημείων δεδομένων μέσω ενός καλύμματος, το οποίο είναι ουσιαστικά μια συλλογή διαστημάτων κατά μήκος των φίλτρων. Ο προσδιορισμός ενός καλύμματος περιλαμβάνει δύο παραμέτρους: τον αριθμό των διαστημάτων (ανάλυση) και το ποσοστό επικάλυψης μεταξύ γειτονικών διαστημάτων (επικάλυψη) [19].

4. Το τελικό βήμα του αλγορίθμου Mapper περιλαμβάνει την ομαδοποίηση και την κατασκευή δικτύου. Η ομαδοποίηση εφαρμόζεται σε κάθε υπερκύβο σημείων δεδομένων, προσδιορίζοντας ομάδες στον αρχικό χώρο δεδομένων και όχι στον προβαλλόμενο χώρο δεδομένων. Στη συνέχεια κατασκευάζεται το δίκτυο με βάση αυτά τα αποτελέσματα της ομαδοποίησης, όπου κάθε ομάδα αποτελεί ένα κόμβο στο δίκτυο. Οι ακμές συνδέουν δύο κόμβους εάν μοιράζονται κοινά δείγματα που αντιστοιχούν στις αντίστοιχες ομάδες τους [19].

Από τα παραπάνω προκύπτει ότι ένα μεγάλο πλεονέκτημα αλλά ταυτόχρονα και μειονέκτημα του αλγόριθμου Mapper, είναι η εξάρτησή του από διάφορες παραμέτρους: (1) τα δεδομένα εισόδου , (2) την συνάρτηση φίλτρου f: X→R, (3) το κάλυμμα του R το οποίο περιλαμβάνει παραμέτρους όπως το ποσοστό τομής των διαστημάτων και (4) τον αλγόριθμο ομαδοποίησης. Η κατάλληλη διαμόρφωση αυτών των παραμέτρων επιτρέπει μια ουσιαστική εικόνα των δεδομένων, καθιστώντας τον αλγόριθμο ένα ευέλικτο εργαλείο. Ωστόσο, η δεξιότητα έγκειται στην σωστή επιλογή αυτών των παραμέτρων και θεωρείται η κύρια τεχνογνωσία στη χρήση του Mapper.

3.4 Επιλογή συνάρτησης φίλτρου

Η παραγωγή του Mapper εξαρτάται σε μεγάλο βαθμό από την επιλογή της συνάρτησης φίλτρου. Ανάλογα με τις γεωμετρικές ιδιότητες για τις οποίες ενδιαφέρεται κανείς, ορισμένες συναρτήσεις φίλτρου είναι περισσότερο ή λιγότερο πιθανό να εμφανίσουν ενδιαφέροντα χαρακτηριστικά στα δεδομένα. Η ευελιξία του αλγόριθμου Mapper με τη συνάρτηση φίλτρου επιτρέπει να αναλύονται σύνολα δεδομένων από διάφορους τομείς και να τίθενται διάφορα ερευνητικά ερωτήματα [17]. Στη συνέχεια παρουσιάζονται ορισμένες από τις πιο γνωστές συναρτήσεις φίλτρου.

-Ο Gaussian πυρήνας: Είναι μια συνάρτηση φίλτρου που χρησιμοποιείται για την εκτίμηση της πυκνότητας ενός δείγματος δεδομένων. Για ένα σύνολο δεδομένων X, σημεία x,y ∈ X και παράμετρο ϵ > 0, ένα τέτοιο φίλτρο δίνεται από τη σχέση [17]:

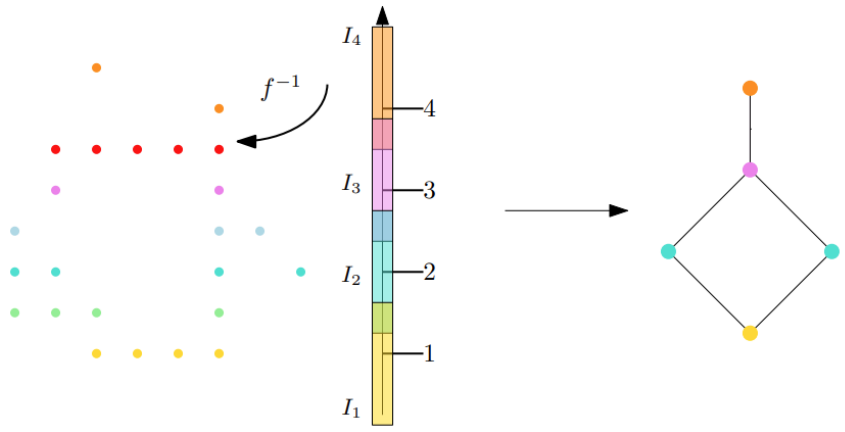
$$f_e(\mathrm{x}) = C_e * \sum_y \exp\left(\frac{-d(x,y)^2}{e}\right)$$

- Eccentricity function -Συνάρτηση εκκεντρότητας: είναι ένα μαθηματικό εργαλείο που χρησιμοποιείται για να ποσοτικοποιήσει το πόσο κοντά βρίσκονται τα σημεία δεδομένων στο κέντρο ενός συνόλου δεδομένων. Η έννοια αυτή συνδέεται συχνά με αυτό που είναι γνωστό ως «βάθος δεδομένων». Αυτές οι συναρτήσεις εκκεντρότητας ορίζονται ως προς μια παράμετρο "p", όπου "p" είναι μια τιμή εντός του εύρους [1, ∞). Η συνάρτηση αναπαρίσταται συνήθως [8].

$$E_p(\mathrm{x}) = \left(\frac{\sum_{yeX} d(x,y)^p}{|X|}\right)^{\frac{1}{p}}$$

- Projection-Προβολή: Αυτή η συνάρτηση φίλτρου επιστρέφει την προβολή σε επιλεγμένες στήλες των δεδομένων. Όταν χρειάζεται η εξαγωγή μιας συγκεκριμένης παράμετρού από το σύνολο δεδομένων, η συνάρτηση φίλτρου προβολής μπορεί να είναι χρήσιμη [8].

- Entropy-Εντροπία: Ένας τρόπος για να μετρήσουμε το πόσο «χαοτικό» είναι ένα σύνολο δεδομένων που αναπαρίστανται ως νέφος σημείων στον χώρο είναι η εντροπία [8].



Εικόνα 14: Παράδειγμα με συνάρτηση φίλτρου μια συνάρτηση ύψους [17].

Στην εικόνα 14 απεικονίζεται η μέθοδος που χρησιμοποιεί μια συνάρτηση ύψους ως συνάρτηση φίλτρου. Η εικόνα im(f) χωρίζεται σε τέσσερα επικαλυπτόμενα διαστήματα $I_1,\ldots,I_4$(κίτρινο, μπλε, βιολετή, πορτοκαλί). Τα δεδομένα $x_j \in X$ του νέφους σημείων χρωματίζονται σύμφωνα με την εικόνα της συνάρτησης φίλτρου $f(x_j) \in I_i$. Εάν για ένα σημείο x ∈ X η εικόνα βρίσκεται σε δύο διαστήματα, χρωματίζεται σύμφωνα με

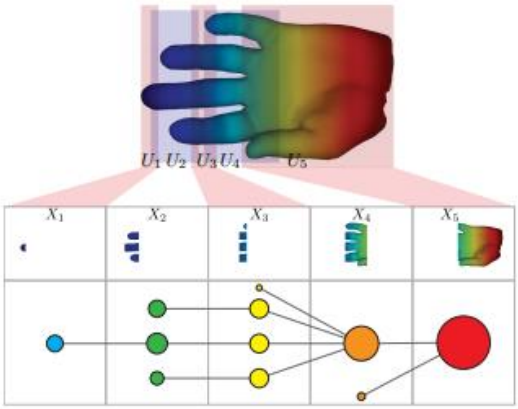
το επικαλυπτόμενο διάστημα (πράσινο, σκούρο μπλε, κόκκινο). Στη συνέχεια, εφαρμόζεται ένας αλγόριθμος ομαδοποίησης στην κάλυψη $\{f^{-1}(I_1), \ldots, f^{-1}(I_4)\}$ του X. Κάθε συστάδα αντιστοιχεί σε μία κορυφή του συμπλέγματος στα δεξιά. Οι ακμές μεταξύ των κορυφών σχεδιάζονται σύμφωνα με την επικάλυψη των διαστημάτων I_j [17].

3.5 Επιλογή Καλύμματος

Στην επιλογή του καλύμματος U, όταν η συνάρτηση f είναι πραγματική, συνιστάται το U να αποτελείται από πολλά διαστήματα με ίσο μήκος $r > 0$ που καλύπτουν το f(X). Το r είναι η ανάλυση του καλύμματος, και το ποσοστό επικάλυψης g μεταξύ δύο διαδοχικών διαστημάτων ονομάζεται κέρδος. Επιλέγοντας g κάτω από 50%, κάθε σημείο καλύπτεται, το πολύ, από 2 ανοιχτά σύνολα του U, δημιουργώντας ένα γράφημα. Το Mapper είναι ευαίσθητο στην επιλογή του U, με μικρές αλλαγές σε παραμέτρους να επηρεάζουν δραστικά το αποτέλεσμα. Η κλασική στρατηγική περιλαμβάνει τη δοκιμή διαφόρων παραμέτρων για την επιλογή των πιο ενημερωτικών αποτελεσμάτων [8].

3.6 Επιλογή αλγόριθμου ομαδοποίησης

Ο αλγόριθμος Mapper περιλαμβάνει την συσταδοποιήση των σημείων, του συνόλου δεδομένων, που προκύπτουν ως αντίστροφη εικόνα των διαστημάτων που χρησιμοποιήθηκαν για την κάλυψη του πεδίου τιμών. Οι διάφορες μέθοδοι συσταδοποίησης μπορούν να χωριστούν στις εξής κατηγορίες: representative based clustering, ιεραρχική συσταδοποίηση και συσταδοποίηση βάσει πυκνότητας [10].



Εικόνα 15 Οι τιμές της συνάρτησης σε ένα μοντέλο χεριού χωρίζονται σε διαστήματα όπως υποδεικνύεται από διαφορετικά χρώματα. Κατασκευή του Mapper για ένα τρισδιάστατο χέρι που αναπαρίσταται ως ένα νέφος σημείων. Η συνάρτηση φίλτρου f είναι η τετμημένη κάθε σημείου. Το πεδίο τιμών της χωρίζεται σε αλληλεπικαλυπτόμενα διαστήματα και τα σημεία χρωματίζονται βάσει της τιμής f στα διαστήματα αυτά. Τα σημεία χωρίζονται χρησιμοποιώντας την αντίστροφη εικόνα της f και χωρίζονται σε συστάδες με χρήση κάποιου αλγορίθμου συσταδοποίησης. Κάθε συστάδα αναπαρίσταται ως κορυφή (χρωματισμένη ανάλογα με την τιμή της συνάρτησης φίλτρου) και προστίθενται ακμές ανάμεσα στις συστάδες που έχουν κοινά στοιχεία. [20]

Κεφάλαιο 4 : Ο αλγόριθμος Ball Mapper

4.1Εισαγωγή

Στην παρούσα ενότητα, παρουσιάζεται ένας αλγόριθμος εμπνευσμένος από τον Mapper, ο οποίος επιτυγχάνει παρόμοια αποτελέσματα, αλλά απαιτεί μόνο μία παράμετρο, απλοποιώντας τη διαδικασία. Ο αλγόριθμος αυτός που αναπτύχθηκε από τον Dlotko το 2019, ονομάζεται Ball Mapper και χαρακτηρίζεται από την ευκολία υπολογισμού . Αυτή η επερχόμενη ενότητα βασίζεται κυρίως στις πηγές [21], [22] & [23].

4.2 Η κατασκευή του Ball Mapper

Ο αλγόριθμος Ball Mapper κατασκευάζει ένα μοντέλο ενός χώρου X βασιζόμενο σε ένα γράφημα, το οποίο συνήθως συνοδεύεται από μια συνάρτηση $f : X \rightarrow R$. Η διαδικασία περιλαμβάνει τη δημιουργία ενός καλύμματος του X χρησιμοποιώντας μπάλες με σταθερή ακτίνα ε και μια μετρική (συνήθως επιλέγεται η Ευκλείδεια). Επιλέγοντας την ακτίνα ε, δημιουργείται ένα κάλυμμα B(X) όπου $X \subset B(X)$. Αυτό το κάλυμμα κατασκευάζεται μέσω ενός επαναλαμβανόμενου αλγόριθμου, ξεκινώντας με ένα κενό κάλυμμα $B(X) = Ø$ και επιλέγοντας ένα σημείο $p \in X$,που δεν καλύπτεται από καμία μπάλα B(X). Μια μπάλα, B(p,ε) κατασκευάζεται και τα σημεία $x \in B(p,\varepsilon)$ θεωρούνται τώρα ότι καλύπτονται. Οι επαναλήψεις τελειώνουν όταν δεν υπάρχουν πλέον ακάλυπτα σημεία.

Παρατηρήσεις:

- Η επανάληψη και η τυχαία επιλογή δηλώνει ότι υπάρχουν πολλές διαφορετικές καλύψεις που μπορεί να εμφανιστούν για τα ίδια τα δεδομένα, για αυτό καλό είναι ο αλγόριθμος να εκτελείται αρκετές φορές εκ νέου για τροποποιημένα δεδομένα, έτσι ώστε να επαληθεύονται οι πληροφορίες που αποκτόνται.
- Η επιλογή της ακτίνας της μπάλας είναι ένας κρίσιμος παράγοντας για το τελικό αποτέλεσμα. Η επιλογή μικρών ακτινών παρέχει λεπτομερείς γνώσεις, αλλά ενέχει τον κίνδυνο να επικεντρωθεί υπερβολικά σε τοπικά φαινόμενα, εμποδίζοντας ενδεχομένως την παρουσίαση του ευρύτερου πλαισίου. Αντίθετα, η επιλογή μιας πολύ μεγάλης σφαίρας μπορεί να παραβλέψει κρίσιμες λεπτομέρειες και να υπονομεύσει την εξαγωγή ουσιαστικών συμπερασμάτων.

4.3 Εκδοχές αλγόριθμου Ball Mapper

Θα συμβολίζεται με C το σύνολο των κέντρων και με $B(C) = \cup_{p \in C} B(p, \varepsilon)$ το κάλυμμα που προκύπτει για ένα συγκεκριμένο σύνολο κέντρων και μια συγκεκριμένη ακτίνα. Επιπλέον, με B(X,ε) ορίζεται το σύνολο σημείων του X μαζί με μία λίστα κέντρων μπαλών έτσι ώστε για κάθε σημείο p του X, οι μπάλες ακτίνας ε που έχουν το κέντρο τους στο B(X,ε) [x] να καλύπτουν το x, ενώ οι υπόλοιπες μπάλες όχι. Δηλαδή, το

σύνολο B(X,ε) δηλώνει σε ποιες μπάλες ανήκει κάθε σημείο του X. Υπάρχουν διάφοροι τρόποι με τους οποίους η συλλογή των κέντρων των σφαιρών C μπορεί να επιλεγεί μεταξύ των X:

1. Χτίζοντας ένα ε-δίκτυο στο X (ε είναι μια παράμετρος)

2. Με τη χρήση του αλγόριθμου ομαδοποίησης k-means (όπου η παράμετρος της μεθόδου είναι ο αριθμός συστάδων) και στη συνέχεια τον εντοπισμό της απόστασης «ενός πιο απομακρυσμένου σημείου δεδομένων από τα κέντρα των επιλεγμένων συστάδων».

Αλγόριθμος 1: Greedy ε-δίκτυο

Είσοδος: Σύνολο δεδομένων X, ε > 0

Δημιουργία ενός αρχικά κενού διανύσματος B(X ,ε)

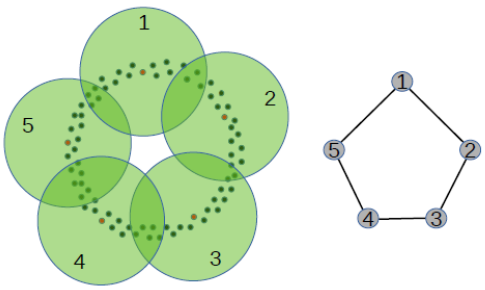
Επανάλαβε: Διάλεξε ένα σημείο p ∈ X, το οποίο δεν ανήκει σε καμία μπάλα

Για κάθε x ∈ B(p ,ε) ∩ X, πρόσθεσε το p στο B(X ,ε) [x]

Μέχρι να καλυφθούν όλα τα σημεία του X

Έξοδος: B(X, ε)

Σε αυτή την κατασκευή η ακτίνα της σφαίρας είναι η μόνη εξωγενής είσοδος. Η διαδοχική διαδικασία του αλγορίθμου 1 μπορεί να παράγει ελαφρώς διαφορετικά αποτελέσματα με βάση την τυχαία επιλογή κάθε επόμενου ακάλυπτου σημείου p, αλλά επειδή όλα τα πιθανά ε-δίκτυα είναι κοντά το ένα στο άλλο, ο αντίκτυπος είναι αμελητέος. Λόγω του τρόπου με τον οποίο σχηματίζονται οι μπάλες, η μέγιστη απόσταση των σημείων από το κέντρο της σφαίρας περιορίζεται στην ακτίνα ε. Επίσης, δεν είναι δυνατόν να αποφασιστεί μια "βέλτιστη" τιμή της παραμέτρου ε, και γι' αυτό αφήνεται στον ερευνητή να την καθορίσει ανάλογα με τις ανάγκες της ανάλυσης.



Εικόνα 16: Παράδειγμα κατασκευής Ball Mapper με τον αλγόριθμο Greedy-ε. [24]

Αλγόριθμος 2: Max-min ε- δίκτυο

Είσοδος: Σύνολο δεδομένων X, $\varepsilon > 0$

Διάλεξε ένα αυθαίρετο σημείο $p \in X$ και θέσε $C = \{p\}$

Επανάλαβε: Βρες το σημείο $p^* \in X \backslash C$ που βρίσκεται πιο μακριά από το C

$d = dist(p^*, C)$

Θέσε $C = C \cup \{p^*\}$

Μέχρι $d \leq \varepsilon$

Δημιουργία ενός αρχικά κενού διανύσματος B(X, ε)

Για κάθε $p \in C$ βρες όλα τα σημεία $x \in B(p, \varepsilon) \cap X$ και πρόσθεσε το p στο B(X, ε) $[x]$

Έξοδος: B(X,ε)

Η μετατροπή της εξόδου από τον Αλγόριθμο 1 σε γράφημα TDA Ball Mapper απαιτεί ένα περαιτέρω στάδιο κατασκευής γραφήματος. Ο αλγόριθμος 2 παρέχει ένα τέτοιο στάδιο όταν κατασκευάζεται ένα αφηρημένο γράφημα που συνοψίζει το σχήμα του X. Όπως ορίζεται από τον αλγόριθμο, σχεδιάζεται μια ακμή μεταξύ των κέντρων κάθε δύο σφαιρών που έχουν σημεία νεφών (data points) στην τομή τους- οι ακμές αυτές βοηθούν να προσδιοριστεί σε ποιο σημείο του νέφους βρίσκεται κάθε σφαίρα σε σχέση με τις άλλες. Λόγω του τρόπου με τον οποίο κατασκευάζεται το γράφημα, αναμένεται ότι θα εμφανιστούν περισσότερες κορυφές στο Ball Mapper απ' ό,τι κατά τη χρήση του συμβατικού αλγόριθμου Mapper. Κατά συνέπεια, μπορεί να υπάρχουν πρόσθετες πληροφορίες οι οποίες απεικονίζονται στο γράφημα Ball Mapper. Ενώ και οι δύο αλγόριθμοι παράγουν ένα ε-δίκτυο, ο αλγόριθμος 2 μπορεί να τροποποιηθεί κατάλληλα έτσι ώστε να σταματάει μετά από έναν καθορισμένο αριθμό επαναλήψεων. Από την άλλη ο αλγόριθμος 1 είναι πολύ εύκολος στην εφαρμογή.

Εάν το ε ,που παρέχεται από τον αλγόριθμο 1 ή 2 , είναι κατάλληλο μπορεί κανείς να κατασκευάσει ένα καλό κάλυμμα του X, $B(C) = \bigcup_{x \in C} B(x, \varepsilon)$, υπό την προϋπόθεση ότι οι τομές των συνόλων του καλύμματος είναι είτε κενές είτε συσταλτές. Συνεπώς, μπορούμε να κατασκευάσουμε το νεύρο του καλύμματος θεωρώντας κάθε κέντρο c_1, c_2, … , c_n ως κορυφή και προσθέτοντας ένα k-simplex $[c_{i0}, c_{i1}, \dots, c_{ik}]$ όταν $B(c_{i0},\varepsilon) \cap \dots \cap B(c_{ik}, \varepsilon) = \emptyset$. Σύμφωνα με το Θεώρημα του Νεύρου, το X και το νεύρο N είναι ομοτοπικά ισοδύναμα υπό την προϋπόθεση ότι το κάλυμμα είναι "καλό". Το αποτέλεσμα αυτής της κατασκευής είναι το simplicial complex/γράφημα Ball-Mapper, δηλαδή ένα γράφημα με κορυφές τις μπάλες που προέκυψαν και ακμές μεταξύ των μπαλών που έχουν μη κενή τομή. Το μέγεθος κάθε μπάλας/κορυφής αντικατοπτρίζει τον αριθμό των παρατηρήσεων που περιέχει. Τέλος, οι μπάλες μπορούν να χρωματιστούν

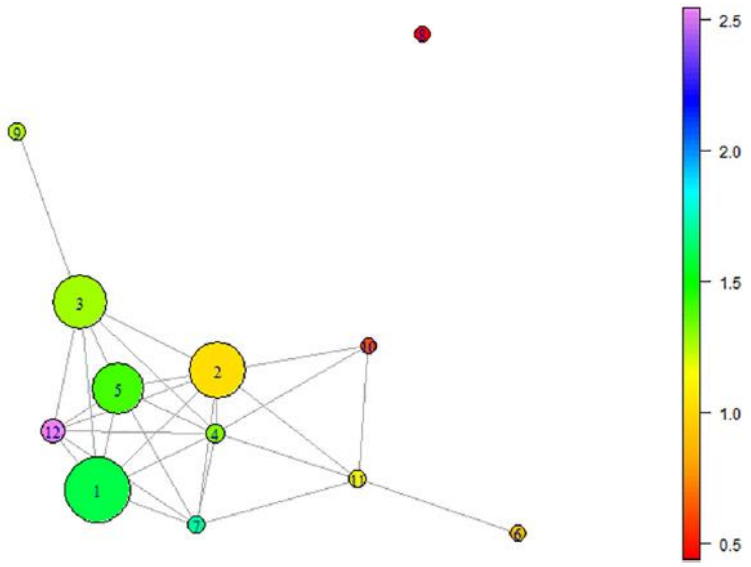
χρησιμοποιώντας μια συνάρτηση, όπως οι συντεταγμένες. Καθώς καταλήγουμε στον 1-σκελετό του simplicial complex, αναφερόμαστε σε αυτό ως γράφημα ως Ball Mapper [21] [22] [23].

4.4 Σύνδεση Mapper και Ball Mapper

Σε αυτή την ενότητα, εξετάζεται η σχέση μεταξύ των αλγορίθμων Mapper και Ball Mapper. Κεντρικό στοιχείο είναι η συνάρτηση φίλτρου f που χρησιμοποιείται στον αλγόριθμο Mapper και η οποία πρέπει να είναι συνεχής γιατί αλλιώς οι γραφικές παραστάσεις των δύο αλγορίθμων θα είναι προφανώς διαφορετικές. Έστω μια συνεχής συνάρτηση φίλτρου f: X → R και ένας μετρικός χώρος (X,ρ), που επιλέχθηκε για την κατασκευή του Mapper, που είναι ομοιόμορφα συνεχής δηλαδή για κάθε x, y και για κάθε $\varepsilon > 0$ υπάρχει $\delta > 0$ τέτοιο ώστε $\rho(x, y) < \delta \Rightarrow |f(x) - f(y)| < \varepsilon$. Έστω επίσης ότι χρησιμοποιείται ο αλγόριθμος ομαδοποίησης single linkage παραμέτρου ε και ότι το κάλυμμα του R αποτελείται από διαστήματα μήκους 6δ που αλληλοκαλύπτονται σε υποδιαστήματα μήκους δ. Τότε, τα σημεία που βρίσκονται κοντά στο Ball Mapper αναμένεται να βρίσκονται κοντά και στο Mapper, δεδομένου ότι η συνάρτηση φίλτρου που έχει επιλεχθεί είναι συνεχής. Ο Ball Mapper παρέχει ακριβέστερες πληροφορίες σχετικά με τα δεδομένα, αν και δυνητικά είναι πιο δύσκολος στην ερμηνεία, καθιστώντας τον ένα πολύτιμο εργαλείο για την διερευνητική ανάλυση δεδομένων [10].

4.5 Ερμηνεία των γραφημάτων Ball Mapper

Τα γραφήματα Ball Mapper έχουν χαρακτηριστικά που διευκολύνουν την κατανόηση των απεικονιζόμενων δεδομένων. Αν και αφηρημένα, είναι τοπολογικά πιστά στις υποκείμενες διαστάσεις. Προερχόμενα από την ιδέα ότι τα σημεία δεδομένων είναι τυχαίες επιλογές από μια υποκείμενη πολλαπλότητα, άρα η ύπαρξη αρκετών σημείων επιτρέπει την ανάκτηση της ομολογίας με υψηλή εμπιστοσύνη. Στην συνέχεια επεξηγείτε ένα γράφημα Ball Mapper και αποδίδονται τα κύρια χαρακτηριστικά του.



Εικόνα 17: Ερμηνεία των γραφημάτων του Ball Mapper [23] .

Ο χρωματισμός του γραφήματος βοηθά στην κατανόηση του πως κατανέμεται ένα συγκεκριμένο αποτέλεσμα στον χώρο. Αυτό μπορεί να περιλαμβάνει τον υπολογισμό της μέσης τιμής για όλα τα σημεία δεδομένων σε μια συγκεκριμένη περιοχή (όπως μια μπάλα). Ωστόσο, είναι επίσης δυνατό να χρησιμοποιηθούν και άλλα μέτρα, όπως αριθμοί, τυπικές αποκλίσεις, ελάχιστες και μέγιστες τιμές κ.α. Οι τιμές παρουσιάζονται σε μια κλίμακα στα δεξιά του διαγράμματος, η οποία δείχνει τις βαθμολογίες Z από το μοντέλο Altman (1968). Αυτό επιτρέπει να παρατηρηθεί ότι στην εικόνα 17 οι χαμηλότερες βαθμολογίες τοποθετούνται στα δεξιά του γραφήματος και η μόνη μπάλα με μέσο όρο πάνω από 2 είναι η μπάλα 12 [23].

Το μέγεθος των σφαιρών στο γράφημα παρέχει πληροφορίες σχετικά με την ποσότητα των σημείων δεδομένων σε μια συγκεκριμένη περιοχή του γραφήματος. Οι μεγαλύτερες μπάλες υποδηλώνουν μεγαλύτερο αριθμό σημείων και πιο συγκεντρωμένη κατανομή δεδομένων εντός της ακτίνας του κεντρικού σημείου της μπάλας. Στην Εικόνα 17, η σφαίρα νούμερο 1 περιέχει τα περισσότερα σημεία, ακολουθούμενη στενά από τις σφαίρες 2, 3 και 5. Υπάρχουν μικρότερες, λιγότερο πυκνοκατοικημένες μπάλες που εκτείνονται προς τη δεξιά πλευρά του σχήματος, και ορισμένες, όπως οι μπάλες 4 και 12, βρίσκονται σε κοντινή απόσταση από τις μεγαλύτερες μπάλες [23].

Η τρίτη πολύτιμη πτυχή ενός γραφήματος Ball Mapper είναι η συνδεσιμότητα μεταξύ των μπαλών. Στην εικόνα 17, πολλές ακμές εκτείνονται από τις περισσότερες μπάλες, υποδεικνύοντας ένα σημαντικό επίπεδο επικάλυψης. Αυτό υποδηλώνει ότι ολόκληρο το γράφημα καλύπτει ένα νέφος παρόμοιων σημείων δεδομένων. Ωστόσο, υπάρχουν μικρότερες προεκτάσεις ή βραχίονες που συνδέονται με τις μπάλες 6 και 9, αντιπροσωπεύοντας τμήματα του νέφους δεδομένων που εκτείνονται πέρα από την κύρια συλλογή σημείων. Αξίζει να σημειωθεί ότι, όταν μια μπάλα δεν είναι συνδεδεμένη, σηματοδοτεί μια πιθανή ακραία τιμή. Για παράδειγμα, η σφαίρα 8 στην επάνω δεξιά γωνία της εικόνας 17 είναι ένα τέτοιο ακραίο σημείο. Είναι ουσιώδες να σημειωθεί ότι οι θέσεις των μη συνδεδεμένων μπαλών δεν αποκαλύπτουν τη σχέση τους με τα κύρια συνδεδεμένα στοιχεία, δεδομένου ότι το διάγραμμα είναι αφηρημένο [23].

Τέταρτον, τα διαγράμματα Ball Mapper αποκαλύπτουν συσχετίσεις μεταξύ των μεταβλητών. Για παράδειγμα στην δισδιάστατη περίπτωση τα συσχετιζόμενα σημεία σχηματίζουν μια στενή ζώνη και το μοτίβο αυτό επεκτείνεται σε πολλαπλές διαστάσεις. Κατά συνέπεια, το γράφημα Ball Mapper τείνει να παίρνει σχήμα μακρύ και λεπτό. Να σημειωθεί ότι η αφηρημένη φύση του γραφήματος Ball Mapper υποδηλώνει ότι η σχεδιαζόμενη γραμμή είναι απίθανο να είναι απόλυτα ευθεία. Αντίθετα μπορεί να λυγίσει για να χωρέσει μέσα στο γράφημα. Όταν οι μεταβλητές συσχετίζονται λιγότερο, η κάλυψη θα πρέπει να απλωθεί, με αποτέλεσμα μια πιο δικτυωτή εμφάνιση, όπως παρατηρείται στο κάτω αριστερό μέρος την εικόνα 17 [23].

Ακόμη και αν δεν εμφανίζεται άμεσα στο διάγραμμα, η ποσότητα των μπαλών παρέχει στον αναλυτή μια αίσθηση του επιπέδου λεπτομέρειας των δεδομένων. Όταν χρησιμοποιούνται μικρότερης ακτίνας σφαίρας, απαιτούνται περισσότερες σφαίρες για να περιλάβουν το σύνολο δεδομένων. Ο ακριβής προσδιορισμός της κατάλληλης ε για μια συγκεκριμένη εφαρμογή εξαρτάται από την κρίση του αναλυτή. Στην εικόνα 17 θα μπορούσε να συναχθεί από τα δεδομένα ότι η επιλεγμένη ακτίνα ε είναι πολύ μεγάλη, καθώς δεν καταγράφονται πολλές λεπτομέρειες στο κέντρο του διαγράμματος [23].

Ο Ball Mapper (BM) διαθέτει μια σειρά από διάφορα πολύτιμα χαρακτηριστικά που μπορούν να βοηθήσουν στην ερμηνεία της σχέση μεταξύ των χαρακτηριστικών των δεδομένων. Όπως συμβαίνει με όλες τις μεθοδολογίες, η τελική επιλογή των (input) εισόδων θα είναι ο καθοριστικός παράγοντας για την αποτελεσματικότητα της ανάλυσης. Στην ουσία, η επιτυχία της χρήσης του Ball Mapper για την κατανόηση της σχέσης μεταξύ των χαρακτηριστικών των δεδομένων και της αποτυχίας εξαρτάται σημαντικά από την προσεκτική επιλογή των παραμέτρων εισόδου [23].

Κεφάλαιο 5:Μέτρα αξιολόγησης μοντέλων

5.1 Εισαγωγή

Ένα από τα συχνότερα ερωτήματα που τίθεται μόλις κατασκευαστεί ένα μοντέλο μηχανικής μάθησης είναι το πόσο αξιόπιστό και ακριβές είναι το αποτέλεσμα που έδωσε. Την απάντηση σε αυτό το ερώτημα την προσφέρουν οι μετρικές αξιολόγησης. Υπάρχουν πολλές μετρικές αξιολόγησης για διαφορετικά σύνολα αλγορίθμων μηχανικής μάθησης. Παραδείγματος χάρη για την αξιολόγηση μοντέλων ταξινόμησης χρησιμοποιούνται μετρικές ταξινόμησης και αντίστοιχα για την αξιολόγηση μοντέλων παλινδρόμησης χρησιμοποιούνται μετρικές παλινδρόμησης [25].

Στις ενότητες που ακολουθούν, θα περιγράφουν μερικά από τα πιο βασικά μέτρα αξιολόγησης που αφορούν την ταξινόμηση όπως Accuracy, Precision, Recall, F1-score, ROC Area και Hmeasure. Πολλές μετρικές βασίζονται στον πίνακα σύγχυσης, δεδομένου ότι περικλείει όλες τις σχετικές πληροφορίες για τον αλγόριθμο και την απόδοση του κανόνα ταξινόμησης [25].

5.2 Confusion matrix- Πίνακας σύγχυσης/ταξινόμησης

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Εικόνα 18: Πίνακας σύγχυσης για πρόβλημα δυαδικής ταξινόμησης [27]

Ο πίνακας ταξινόμησης είναι ένας διαγώνιος πίνακας που καταγράφει τον αριθμό των περιστατικών μεταξύ δύο βαθμολογητών, την αληθινή/πραγματική ταξινόμηση και την προβλεπόμενη ταξινόμηση [25]. Οι τιμές True (1) και False (0) χρησιμοποιούνται για την αναπαράσταση των πραγματικών τιμών, ενώ οι τιμές Positive (1) και Negative (0) χρησιμοποιούνται για την αναπαράσταση των προβλεπόμενων τιμών. Οι πιθανότητες που σχετίζονται με τα μοντέλα ταξινόμησης προέρχονται από τις μετρικές που βρίσκονται στον πίνακα σύγχυσης, συγκεκριμένα υπάρχουν τέσσερις σημαντικοί όροι TP (True Positives), TN (True Negatives), FP (False Positives) και FN (False Negatives) [26] οι οποίοι αναλύονται ως εξής:

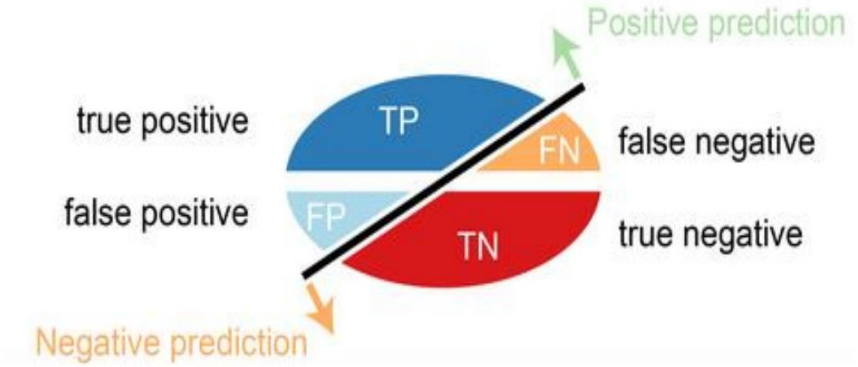
- TP (True Positive) - Το σημείο δεδομένων στον πίνακα σύγχυσης είναι True Positive (TP) όταν προβλέπεται θετικό αποτέλεσμα και αυτό που συνέβη είναι το ίδιο. [27]

- FP (False Positive) - Το σημείο δεδομένων στον πίνακα σύγχυσης είναι ψευδώς θετικό, όταν προβλέπεται θετικό αποτέλεσμα και αυτό που συνέβη είναι ένα αρνητικό αποτέλεσμα. Αυτό το σενάριο είναι γνωστό ως Σφάλμα τύπου 1 [27].

- FN (False Negative) - Το σημείο δεδομένων στον πίνακα σύγχυσης είναι ψευδώς αρνητικό όταν προβλέπεται ένα αρνητικό αποτέλεσμα και αυτό που συνέβη είναι ένα θετικό αποτέλεσμα. Αυτό το σενάριο είναι γνωστό ως σφάλμα τύπου 2 [27].

- TN (True Negative) - Το σημείο δεδομένων στον πίνακα σύγχυσης είναι True Negative (TN) όταν προβλέπεται αρνητικό αποτέλεσμα και αυτό που συμβαίνει είναι το ίδιο. Τα αποτελέσματα της ταξινόμησης αυτής παρουσιάζονται στην εικόνα 19 [27].

Τέσσερα αποτελέσματα της ταξινόμησης



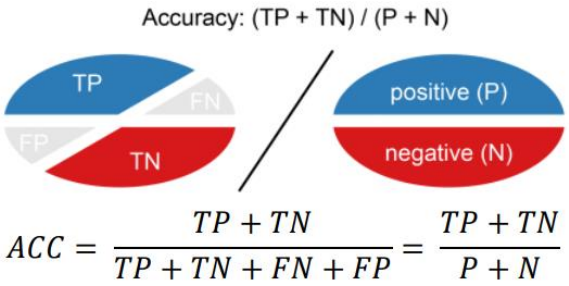
Εικόνα 19 Ελλειπτική αναπαράσταση τεσσάρων δυαδικών αποτελεσμάτων της ταξινόμησης του συνόλου δοκιμών [27]

Συνοψίζοντας, ένας πίνακα σύγχυσης έχει πολλά πλεονεκτήματα. Πρώτον, προσφέρει μια ολοκληρωμένη ανάλυση των αποτελεσμάτων ταξινόμησης, παρέχοντας πολύτιμες πληροφορίες σχετικά με την απόδοση ενός μοντέλου [28]. Αυτές οι λεπτομερείς πληροφορίες είναι καθοριστικές για την αξιολόγηση της ακρίβειας του μοντέλου και τον εντοπισμό των τομέων στους οποίους μπορεί να χρειάζεται βελτίωση. Δεύτερον, τα παράγωγα του πίνακα σύγχυσης, όπως η ακρίβεια, η ανάκληση και το F1-score, χρησιμοποιούνται ευρέως στη μηχανική μάθηση και τη στατιστική για τη μέτρηση της απόδοσης του μοντέλου από διάφορες οπτικές γωνίες, καθιστώντας το ένα ευέλικτο εργαλείο αξιολόγησης [28]. Τέλος, η οπτικοποίηση του πίνακα σύγχυσης μέσω εργαλείων όπως οι θερμικοί χάρτες ενισχύει την ερμηνευσιμότητα των αποτελεσμάτων, διευκολύνοντας τους αναλυτές και τους ενδιαφερόμενους να αντιληφθούν τα μοτίβα και τις πιθανές προβληματικές περιοχές εντός των δεδομένων, βοηθώντας τελικά στην καλύτερη λήψη αποφάσεων και την τελειοποίηση του μοντέλου [28].

5.3 Accuracy-ακρίβεια

Η ακρίβεια ταξινόμησης είναι μια ευρέως χρησιμοποιούμενη μετρική απόδοσης στη μηχανική μάθηση και τη στατιστική, η οποία αντιπροσωπεύει το ποσοστό των σωστών προβλέψεων επί του συνόλου των περιπτώσεων που αξιολογήθηκαν. [29]. Υπολογίζεται προσθέτοντας τον αριθμό των αληθώς θετικών (TP) και αληθώς αρνητικών

(TN) προβλέψεων και διαιρώντας το αποτέλεσμα με τον συνολικό αριθμό των σημείων δεδομένων (P + N), όπου το P αντιπροσωπεύει θετικές περιπτώσεις και το N αντιπροσωπεύει αρνητικές περιπτώσεις. [27]. Αυτή η απλή μετρική κυμαίνεται μεταξύ 0 και 1, με το 1 να αντιπροσωπεύει τέλεια ακρίβεια και το 0 να υποδηλώνει πλήρως λανθασμένη ταξινόμηση [30].



Εικόνα 20 : Δύο ελλείψεις δείχνουν πώς υπολογίζεται η ακρίβεια [27]

Παράδειγμα: Η ακρίβεια όπως υπολογίζεται στα μοντέλα ταξινόμησης προκύπτει από ένα πίνακα σύγχυσης ο οποίος συνοψίζει τις προβλέψεις του μοντέλου και τα πραγματικά αποτελέσματα. Σε ένα πίνακα σύγχυσης τα αληθώς θετικά και τα αληθώς αρνητικά είναι τα στοιχεία που ταξινομούνται σωστά από το μοντέλο και βρίσκονται στην κύρια διαγώνιο του (εικόνα 21) [25].

Αυτά τα στοιχεία θα τοποθετούνται στον αριθμητή, ενώ στον παρονομαστή θα τοποθετηθούν όλες οι καταχωρήσεις, τόσο οι σωστές προβλέψεις όσο και οι περιπτώσεις που το μοντέλο ταξινόμησε εσφαλμένα. Συμπεριλαμβάνοντας όλα αυτά τα στοιχεία στον παρονομαστή, η ακρίβεια προσφέρει ένα ολοκληρωμένο μέτρο της συνολικής ορθότητας του μοντέλου και της ικανότητάς του να κάνει ακριβείς προβλέψεις σε όλες τις κλάσεις [25].

		PREDICTED classification				
	Classes	a	b	c	d	Total
ACTUAL classification	a	6	0	1	2	9
	b	3	9	1	1	14
	c	1	0	10	2	13
	d	1	2	1	12	16
	Total	11	11	13	17	**52**

Εικόνα 21: Παράδειγμα confusion matrix [25]

Υπολογισμός ακρίβειας μοντέλου:

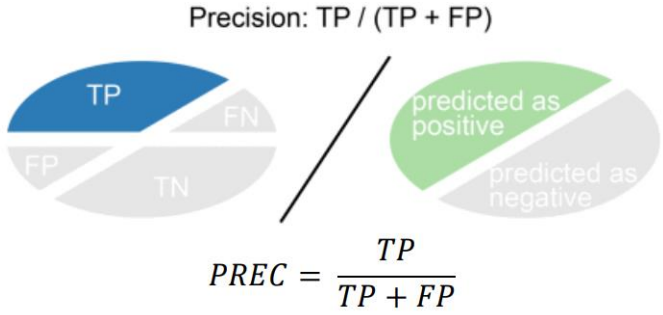
$$Accuracy = \frac{6 + 9 + 10 + 12}{52} = \frac{37}{52} \approx 71\%$$

Ένα από τα βασικά πλεονεκτήματα της ακρίβειας ως μέτρο αξιολόγησης σε μοντέλα ταξινόμησης είναι η απλότητα και η ευκολία στην κατανόηση της. Επιπλέον, η ακρίβεια εφαρμόζεται όχι μόνο σε προβλήματα δυαδικής ταξινόμησης αλλά και σε σενάρια πολλαπλών κατηγοριών και πολλαπλών ετικετών, προσφέροντας ευελιξία στην αξιολόγηση διαφόρων τύπων ταξινομητών. [29]

Παρόλο που είναι ένα βολικό και διαισθητικό μέτρο, έχει ορισμένους περιορισμούς που μπορεί να οδηγήσουν σε παραπλανητικά συμπεράσματα, ιδίως σε σενάρια που περιλαμβάνουν μη ισορροπημένα σύνολα δεδομένων ή δεδομένα λανθασμένης ταξινόμησης. [29].

5.4 Positive Predictive Value or Precision

Η Precision επίσης γνωστή και ως Positive Predictive Value είναι μια σημαντική μετρική για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης [29]. Χρησιμοποιείται για να υπολογιστεί ο αριθμός των σωστών θετικών προβλέψεων (TP) διαιρούμενο με το συνολικό αριθμό των θετικών προβλέψεων (TP + FP) [27].



Εικόνα 22: Δύο ελλείψεις δείχνουν πώς υπολογίζεται η precision [27]

Ειδικότερα, τα True Positive είναι τα στοιχεία που έχουν επισημανθεί ως θετικά από το μοντέλο και είναι πράγματι θετικά, ενώ τα False Positive είναι τα στοιχεία που έχουν επισημανθεί ως θετικά από το μοντέλο, αλλά στην πραγματικότητα είναι αρνητικά [25].

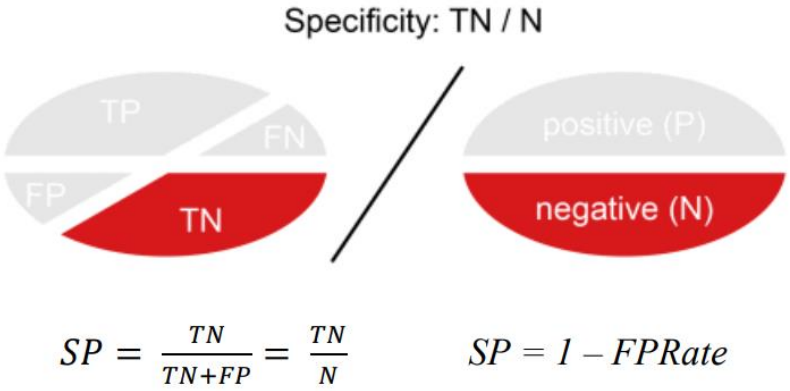
Το Precision είναι ιδιαίτερα πολύτιμο όταν πρόκειται για λοξά και μη ισορροπημένα σύνολα δεδομένων όπου η μια κλάση υπερτερεί σημαντικά έναντι της άλλης. Σε τέτοιες περιπτώσεις, όπου τα ψευδώς θετικά αποτελέσματα μπορεί να είναι προβληματικά το precision γίνεται ένα καίριο εργαλείο διότι μπορεί να αποφύγει να κάνει εσφαλμένες θετικές προβλέψεις [31].

Το κύριο μειονέκτημα της είναι ότι δεν λαμβάνει υπόψη όλα τα στοιχεία του πίνακα σύγχυσης και έτσι δεν μπορεί να χρησιμοποιηθεί ξεχωριστά για την αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης. Καλό θα ήταν να συνδυαστεί με άλλες μετρικές όπως True Negative Rate ή Specificity, με την False Positive Rate και με την ευαισθησία γνωστή ως Recall ή True Positive Rate (TPR) [32].

5.4.1 True Negative Rate - Specificity

Το True Negative Rate, γνωστό και ως Specificity, προσδιορίζεται με τον υπολογισμό του αριθμού των σωστών αρνητικών προβλέψεων (TN) και τη διαίρεσή του με τον συνολικό αριθμό των αρνητικών (N). Μια τέλεια βαθμολογία εξειδίκευσης είναι

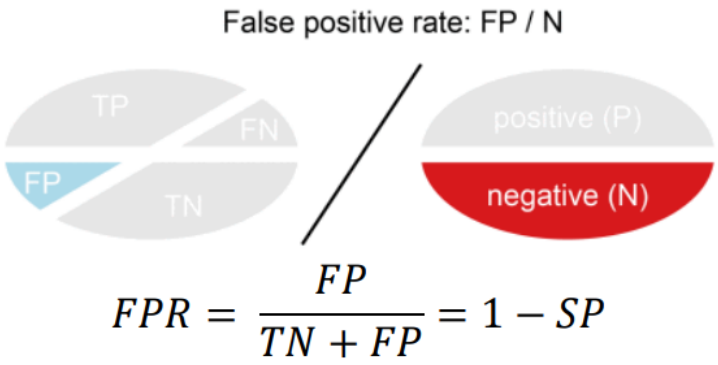
1, υποδεικνύοντας ότι όλες οι αρνητικές προβλέψεις είναι σωστές, ενώ η χαμηλότερη δυνατή βαθμολογία είναι 0, υποδεικνύοντας ότι καμία από τις αρνητικές προβλέψεις δεν είναι ακριβής [30].



Εικόνα 23 Δύο ελλείψεις δείχνουν τον τρόπο υπολογισμού του Specificity [27]

5.4.2 False Positive Rate

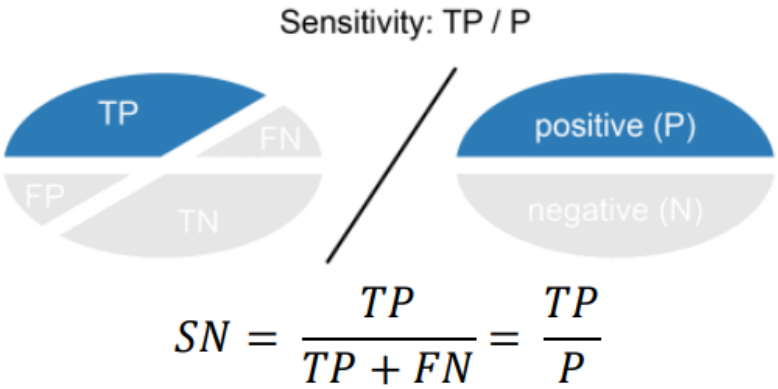
Το False Positive Rate, υπολογίζεται με τη διαίρεσή του αριθμού των ψευδώς θετικών προβλέψεων (FP) με το συνολικό αριθμό των αρνητικών προβλέψεων (N). Ένα τέλειο ποσοστό ψευδώς θετικών ισούται με 0, υποδεικνύοντας ότι καμία αρνητική πρόβλεψη δεν είναι εσφαλμένη, ενώ το χειρότερο δυνατό αποτέλεσμα είναι 1, υποδεικνύοντας ότι όλες οι αρνητικές προβλέψεις είναι ανακριβείς. Επιπλέον, το ποσοστό ψευδώς θετικών προβλέψεων μπορεί να υπολογιστεί ως 1 μείον την ειδικότητα (1-specificity) [30].



Εικόνα 24 Δύο ελλείψεις δείχνουν πώς υπολογίζεται το False Positive Rate [27]

5.5 Sensitivity or Recall

Η ανάκληση που αναφέρεται επίσης ως True Positive Rate ή ευαισθησία (sensitivity) είναι μια σημαντική μετρική για την αξιολόγηση της ακρίβειας ενός μοντέλου ταξινόμησης στον ορθό εντοπισμό θετικών προτύπων. [29]. Η μετρική αυτή υπολογίζεται διαιρώντας τον αριθμό των σωστά προβλεπόμενων θετικών περιπτώσεων (True Positives, TP) με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων (P). [27].



$$SN = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Εικόνα 25: Δύο ελλείψεις δείχνουν πώς υπολογίζεται η ευαισθησία [27]

Μια τέλεια βαθμολογία ανάκλησης συμβολίζεται με 1, υποδεικνύοντας ότι όλες οι θετικές περιπτώσεις αναγνωρίζονται σωστά, ενώ η χαμηλότερη δυνατή βαθμολογία είναι 0, υποδεικνύοντας ότι καμία από τις θετικές περιπτώσεις δεν ταξινομείται με ακρίβεια. [27]

Η Ανάκληση επικεντρώνεται στην καταγραφή των Αληθώς Θετικών και στην αποφυγή των Ψευδώς Αρνητικών αποτελεσμάτων, δηλαδή των περιπτώσεων που λανθασμένα χαρακτηρίζονται ως αρνητικές από το μοντέλο, αλλά στην πραγματικότητα είναι θετικές. [25]. Αυτή η μετρική, μετρά την ικανότητα του μοντέλου να εντοπίζει και να ταξινομεί σωστά όλες τις θετικές μονάδες στο σύνολο δεδομένων, καθιστώντας την ιδιαίτερα πολύτιμη σε σενάρια όπου η έλλειψη θετικών περιπτώσεων μπορεί να έχει σημαντικές συνέπειες όπως σε ιατρικές διαγνώσεις ή στην ανίχνευση απάτης. [25]

5.6 F1-score

Τόσο το Precision (ακρίβεια) όσο και το Recall (ανάκληση) είναι σημαντικές μετρικές για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης, αλλά η καθεμία προσφέρει μοναδικές γνώσεις. Η ακρίβεια μας λέει πόσο χρήσιμα είναι τα αποτελέσματα μετρώντας την ακρίβεια των θετικών προβλέψεων, ενώ η ανάκληση μας ενημερώνει για το πόσο πλήρη είναι τα αποτελέσματα αξιολογώντας την ικανότητα του μοντέλου να καταγράφει όλες τις θετικές περιπτώσεις. [28]

Για να επιτευχθεί ισορροπία μεταξύ ακρίβειας και ανάκλησης, χρησιμοποιείται η βαθμολογία F1, ένας αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, σχεδιασμένος για να παρέχει μια ενιαία μετρική που λαμβάνει υπόψη και τις δύο πτυχές της απόδοσης ενός μοντέλου. [28].

Ο τύπος για το F1-Score έχει ως εξής:

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Εικόνα 26: Εξίσωση για το f1-score [33]

Η συμβολή της ακρίβειας και της ανάκλησης είναι ίση στην F1-score και ο αρμονικός μέσος όρος είναι χρήσιμος για την εύρεση του καλύτερου συμβιβασμού μεταξύ των δύο μεγεθών [34]. Για να εξετάσουμε την συμπεριφορά του F1-score, ελέγχουμε την επίδραση του αρμονικού μέσου όρου στην τελική βαθμολογία με ένα παράδειγμα.

Παράδειγμα :Έστω το μοντέλο Α με precision ίσο με το recall και ίσο με 80%, και το μοντέλο Β του οποίου το precision είναι 60% και το recall 100%. . Αριθμητικά, ο μέσος όρος της ακρίβειας και της ανάκλησης είναι ο ίδιος και για τα δύο μοντέλα, αλλά χρησιμοποιώντας τον αρμονικό μέσο, δηλαδή υπολογίζοντας το F1-Score, το μοντέλο Α λαμβάνει βαθμολογία 80%, ενώ το μοντέλο Β έχει βαθμολογία μόνο 75% [35].

Επιπλέον, είναι σημαντικό να σημειωθεί ότι η Ακρίβεια και η Ανάκληση εμπίπτουν και οι δύο στο εύρος [0, 1], και αν κάποια από αυτές πλησιάσει μια τιμή κοντά στο 0, το τελικό F1-Score παρουσιάζει σημαντική μείωση [25]. Το φαινόμενο αυτό συμβαίνει επειδή ο αρμονικός μέσος όρος τείνει να αποδίδει μεγαλύτερη βαρύτητα σε χαμηλές τιμές, γεγονός που μπορεί να επηρεάσει σημαντικά τη συνολική βαθμολογία. [25].

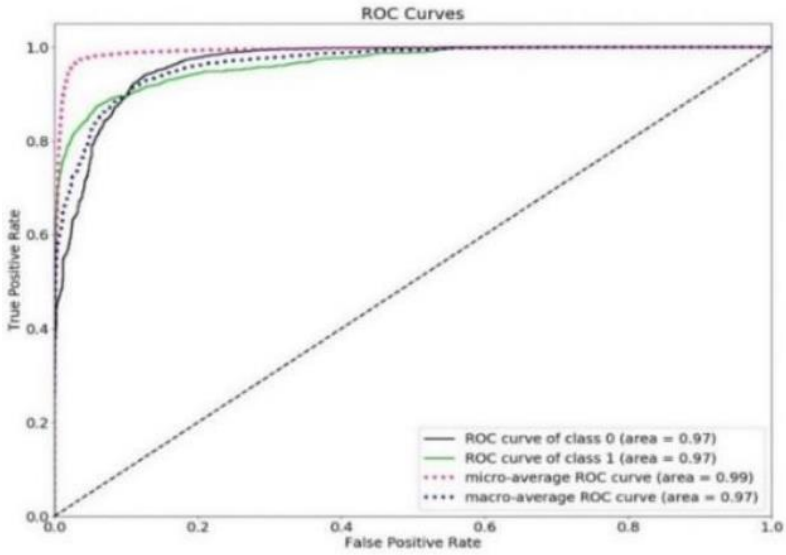
5.7 Περιοχή κάτω από την καμπύλη ROC (AUC)

Η καμπύλη ROC είναι μια τυπική γραφική μέθοδος για αξιολόγησης της απόδοσης ή της ακρίβειας ενός ταξινομητή, που αντιστοιχεί στο συνολικό ποσοστό των σωστά ταξινομημένων παρατηρήσεων [36]. Ουσιαστικά είναι ένα δισδιάστατο γράφημα που απεικονίζει το συμβιβασμό μεταξύ True Positive Rate (άξονας y) και False Positive Rate(άξονας x) για κάθε δυνατό κατώφλι (threshold). Όσο υψηλότερο είναι το True Positive Rate τόσο χαμηλότερο είναι το False Positive Rate δηλαδή το ποσοστό των ψευδώς θετικών αποτελεσμάτων αρά και τόσο καλύτερο το αποτέλεσμα [27].

Για προβλήματα διπλών κλάσεων [37], η τιμή της AUC μπορεί να υπολογιστεί ως εξής:

$$AUC = \frac{S_p - n_p\,(n_n + 1)/2}{n_p n_n}$$

Όπου S_p είναι το άθροισμα όλων των θετικών παραδειγμάτων που κατατάσσονται, ενώ n_p & n_n υποδηλώνουν τον αριθμό των θετικών και αρνητικών παραδειγμάτων αντίστοιχα [29]. Η AUC αποδείχθηκε θεωρητικά και εμπειρικά καλύτερη από την μετρική της ακρίβειας (Accuracy) [38].



Εικόνα 27 Καμπύλη ROC [27]

Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC ονομάζεται AUC (Area under the Curve) και είναι ένας αριθμός που καθορίζει πόσο καλή είναι η καμπύλη ROC [39]. Αυτή η καμπύλη παράγει δύο βασικές μετρικές: ευαισθησία (sensitivity) και ειδικότητα(specificity).

Η καμπύλη ROC για έναν επιτυχημένο μοντέλο θα πρέπει να αυξάνεται απότομα, υποδηλώνοντας ότι όταν το κάτω όριο πιθανότητας πέφτει, το αληθές θετικό ποσοστό (άξονας y) ανεβαίνει ταχύτερα από το ψευδώς θετικό ποσοστό (άξονας x) [36]. Ως αποτέλεσμα, το "ιδανικό σημείο" βρίσκεται στην επάνω αριστερή γωνία του γραφήματος, με ποσοστό ψευδώς θετικών μηδέν και ποσοστό αληθώς θετικού αποτελέσματος ίσο με ένα. Παρόλο που αυτό δεν είναι ιδιαίτερα ρεαλιστικό, συνεπάγεται ότι όσο μεγαλύτερη είναι η AUC, τόσο καλύτερος είναι ο ταξινομητής. Η μετρική AUC κυμαίνεται από 0,50 έως 1,00. Ένας καλός ταξινομητής έχει τιμή μεγαλύτερη από 0,80 [32].

5.8 H-measure

Το H-measure είναι ένα μέτρο αξιολόγησης που προτείνεται για την αντιμετώπιση του προβλήματος της ανισορροπίας των κλάσεων σε σύνολα δεδομένων. Η ανισορροπία των κλάσεων εμφανίζεται όταν η κατανομή των κλάσεων διαφέρει σημαντικά από την ομοιόμορφη κατανομή. Εμφανίζεται σε πολλές εφαρμογές μηχανικής μάθησης, όπως η ανίχνευση απάτης και άλλες. Οι περισσότεροι ταξινομητές είναι σχεδιασμένοι για να μεγιστοποιούν την ακρίβεια των μοντέλων τους. Ωστόσο, όταν τα δεδομένα είναι ανισόρροπά, συνήθως οδηγούνται στο να δίνουν υπερβολική έμφαση στα δεδομένα της πλειοψηφίας και να αγνοούν τα δεδομένα της μειοψηφίας. Το πρόβλημα αυτό επηρεάζει αρνητικά την απόδοση των ταξινομητών. Η προτεινόμενη λύση είναι το "H-measure", το οποίο χρησιμοποιεί μια συμμετρική κατανομή Beta για να αντικαταστήσει την ασυμμετρία που εισάγει η AUC [40].

Το παρακάτω παράδειγμα θα κάνει πιο κατανοητό την έννοια «δεδομένα πλειοψηφίας» και «δεδομένα μειοψηφίας»

Παράδειγμα: Στην ανίχνευση απάτης σε πιστωτικές κάρτες, υποθέτουμε ότι το σύνολο δεδομένων έχει 999 νόμιμες συναλλαγές (δεδομένα πλειοψηφίας) και μόνο 1 δόλια (δεδομένα μειοψηφίας, αυτά που θέλουμε να ανιχνεύσουμε). Για να μεγιστοποιηθεί η ακρίβεια σε αυτήν την περίπτωση, οι ταξινομητές που έχουν βελτιστοποιηθεί για την ακρίβεια θα ταξινομήσουν όλες τις συναλλαγές ως ανήκουσες στην πλειοψηφική τάξη για να λάβουν ακρίβεια 99,9%. Ωστόσο, αυτό το αποτέλεσμα δεν έχει νόημα επειδή η δόλια συναλλαγή ταξινομείται εσφαλμένα [40].

5.8.1 Το μέτρο H - Αντικατάσταση της AUC

Το "H-measure" προτείνεται ως αντικατάσταση του AUC (Area Under the ROC Curve) για να αντιμετωπίσει την ασυνέπεια που παρουσιάζει. Το μέτρο αυτό καθορίζεται από την παρακάτω εξίσωση:

$$H = 1 - \frac{\int Q(c); b, c) u_{a,b}(c) dc}{\pi_0 \int_0^{\pi_1} c u_{a,b}(c) dc + \pi_1 \int_{\pi_1}^{0} (1-c) u_{a,b}(c) dc}$$

όπου π_0 και π_1 είναι οι έκτων προτέρων πιθανότητες, οι c_0 και c_1 είναι το κόστος λανθασμένης ταξινόμησης για την κλάση 0 (πλειοψηφία) και την κλάση 1 (μειονότητα), b = c_0 + c_1 και c = c_1/ (c_0 + c_1) , f_0(s) και f_1(s) είναι οι συναρτήσεις πυκνότητας πιθανότητας- και F_0(s) και F_1(s) είναι οι αθροιστικές συναρτήσεις κατανομής για την τάξη 0 και την τάξη 1, αντίστοιχα.

$$Q(t; b, c) \triangleq \{c\pi_1 \left(1 - F_1(t)\right) + (1-c)\pi_0 \left(\pi_0 F_0(t)\right)\} b$$

είναι η απώλεια για μια τυχαία επιλογή του κατωφλιού (threshold) t και

$$u_{a,b} = beta(c; a, b) = \frac{c^{a-1}(1-c)^{b-1}}{B(1; a, b)}$$

είναι η συμμετρική κατανομή Beta.

Συνοψίζοντας, το H-measure είναι μια προσέγγιση που προσπαθεί να λύσει τα προβλήματα που εμφανίζονται με την AUC, εισάγοντας κόστη για διάφορα είδη λανθασμένης ταξινόμησης και λαμβάνοντας υπόψη την αβεβαιότητα ως προς τα κόστη κατά την αξιολόγηση των ταξινομητών [41].

Κεφάλαιο 6 Εφαρμογή

6.1 Credit scoring – Βαθμολόγηση της πιστοληπτικής ικανότητας

Η αξιολόγηση του πιστωτικού κινδύνου αποτελεί ένα κρίσιμο παράγοντα στην διαχείριση του χρηματοοικονομικού κινδύνου και πρόσφατα έχει γίνει ένας κύριος στόχος του τραπεζικού και χρηματοπιστωτικού τομέα [42]. Στις μέρες μας, μια τέτοια ανάλυση θεωρείται πολύ σημαντική καθώς είναι σε θέση να παρέχει αξιόπιστη αξιολόγηση των αιτούντων τραπεζικών πιστώσεων [43].

Συγκεκριμένα, η βαθμολόγηση της πιστοληπτικής ικανότητας (credit scoring) είναι μια μέθοδος αξιολόγησης της αξιοπιστίας ενός προσώπου ή μιας εταιρείας που υποβάλλει αίτηση για ένα τραπεζικό δάνειο. Η βαθμολογία παρουσιάζεται συνήθως σε μορφή πόντων, όπου υψηλότερη βαθμολογία υποδηλώνει μεγαλύτερη αξιοπιστία του δανειολήπτη. Παράγοντες όπως το επάγγελμα, η εκπαίδευση, η κατάσταση κατοικίας, η περιοχή κατοικίας, το μηνιαίο εισόδημα, η ηλικία, η οικογενειακή κατάσταση, ο αριθμός εξαρτώμενων μελών, η ασφάλιση ζωής, η περίοδος απασχόλησης στην τρέχουσα θέση λαμβάνονται υπόψη κατά την αξιολόγηση. Η τράπεζα με βάση τα προαναφερθέντα χαρακτηριστικά, ορίζει ένα προκαθορισμένο όριο, γνωστό ως cut-off και μέσα από αυτό αποφασίζει να χορηγήσει ή να αρνηθεί την αίτηση του δανείου [43].

Δύο βασικοί τύποι εργασιών βαθμολόγησης της πιστοληπτικής ικανότητας είναι: α) η βαθμολόγηση αίτησής και β) η βαθμολόγηση συμπεριφοράς. Η πρώτη ταξινομεί τους αιτούντες ως "καλούς" ή "κακούς" πιστωτές βάσει δημογραφικών και οικονομικών πληροφοριών, ενώ η δεύτερη λαμβάνει υπόψη το ιστορικό πληρωμών του πελάτη [44].

Κατά την αξιολόγηση της πιστοληπτικής ικανότητας ενός δανειολήπτη δίνεται ιδιαίτερη προσοχή στα ακόλουθα:

1. Εξαγωγή και επιλογή χαρακτηριστικών [44], [45]
2. Επιλογή κατάλληλων ταξινομητών ή συνδυασμών τους [46]
3. Βελτιστοποίηση των παραμέτρων των ταξινομητών [47]
4. Χρήση αξιόπιστων τεχνικών cross validation κατά την εκπαίδευση των ταξινομητών για την επίτευξη αξιόπιστων αποτελεσμάτων [48].

Παρά την πολυπλοκότητα της πιστοληπτικής βαθμολόγησης λόγω της αβεβαιότητας στα οικονομικά δεδομένα, οι παραδοσιακές στατιστικές μέθοδοι , όπως η διακριτική ανάλυση και η γραμμική παλινδρόμηση, παρουσιάζουν περιορισμούς. Τα τελευταία χρόνια, οι αλγόριθμοι μηχανικής μάθησης, όπως τα Νευρωνικά δίκτυα, support vector machine, k-nearest neighbors αλγόριθμοι, δέντρα αποφάσεων και άλλα, έχουν κερδίσει δημοτικότητα λόγω της αποτελεσματικότητάς τους στη χρήση πραγματικών ιστορικών δεδομένων [43]. Παρά την υψηλή απόδοσή τους σε πολλά προβλήματα βαθμολόγησης σε σύγκριση με τις στατιστικές μεθόδους, οι αλγόριθμοι μηχανικής μάθησης έχουν μειονεκτήματα, όπως για παράδειγμα ο μεγάλος αριθμός παραμέτρων προς βελτιστοποίηση, η ευαισθησία σε τοπικά ελάχιστα, η τάση για υπερπροσαρμογή και η υψηλή υπολογιστική πολυπλοκότητα που απαιτείται για τα συστήματα μάθησης [47].

6.2 Μεθοδολογία

6.2.1 Στόχος

Στόχος της παρούσας ερευνάς είναι να μελετηθεί το μοντέλο το οποίο έχει εισαχθεί στην εργασία [1]. Συγκεκριμένα αναλύεται ένα σύνολο δεδομένων με 690 παρατηρήσεις, όπου κάθε παρατήρηση αντιστοιχεί σε έναν δανειολήπτη. Το σύνολο δεδομένων περιέχει πληροφορίες για κάθε δανειολήπτη μέσα από 14 χαρακτηριστικά/γνωρίσματα τα οποία αποτελούν τις επεξηγηματικές μεταβλητές. Όταν αυτά τα χαρακτηριστικά είναι συναφή (δηλαδή διαθέτουν επαρκή επεξηγηματική ικανότητα όσον αφορά την αξιοπιστία του δανειολήπτη), είναι λογικό να αναμένεται το εξής : Παρόμοιοι δανειολήπτες, δηλαδή αυτοί που βρίσκονται κοντά ο ένας στον άλλον, στο χώρο των 14-διάστατων χαρακτηριστικών, αναμένεται να παρουσιάζουν συγκρίσιμα πιστωτικά προφίλ. Κατά συνέπεια, μια συλλογή σημείων δεδομένων των δανειολήπτων μπορεί να θεωρηθεί ως ένα νέφος σημείων που προέρχεται από μια υποκείμενη πολλαπλότητα[3].

Όταν δημιουργείται μια "κατάλληλη" κάλυψη για αυτή την υποκείμενη πολλαπλότητα, δημιουργείται και η δυνατότητα να ανακτηθούν πληροφορίες για την υποκείμενη τοπολογία κατασκευάζοντας το νεύρο της κάλυψης. Σύμφωνα με το θεώρημα του νεύρου, αυτό το νεύρο θα είναι ομοτοπικό με την υποκείμενη τοπολογία. Για να προσεγγιστεί μια τέτοια κάλυψη, λαμβάνονται υπόψη σφαίρες/μπάλες ακτίνας ε, με κέντρα διάφορα σημεία δεδομένων, οι οποίες καλύπτουν όλο το σύνολο δεδομένων.

Η επιλογή της ε είναι σημαντική και αποτελεί πρόκληση, διότι η κάλυψη που θα κατασκευαστεί μέσω αυτής θα πρέπει να αντιπροσωπεύει επαρκώς την υποκείμενη πολλαπλότητα. Μόλις κατασκευαστεί μια κάλυψη, η αξιολόγηση παρέχεται με απλό τρόπο: Η θέση ενός νέου δανειολήπτη ψ στον χώρο των χαρακτηριστικών καλύπτεται από κάποιον μέγιστο αριθμό k(ψ) των καλυπτόμενων σφαιρών. Τότε η βαθμολογία του δανειολήπτη ψ είναι το ποσοστό των δανειοληπτών που αθέτησαν των δάνειο σε σχέση με όλους τους δανειολήπτες που ανήκουν στην ένωση αυτών των k(ψ) σφαιρών. Τα μέτρα απόδοσης που χρησιμοποιήθηκαν είναι το Accuracy, Hmeasure, Precision, Recall, F1score, AUC. Είναι σαφές ότι η απόδοση αυτής της προσέγγισης εξαρτάται από την ε. Αντιμετωπίζουμε το πρόβλημα της επιλογής του κατάλληλου ε βελτιστοποιώντας το κάθε μέτρο απόδοσης εντός του συνόλου εκπαίδευσης.

Το νεύρο αυτής της κάλυψης έχει τη μορφή ενός simplicial complex (στην πραγματικότητα ενός γραφήματος). Οι κορυφές του γραφήματος αντιπροσωπεύουν τις κατασκευασμένες μπάλες κάλυψης, ενώ οι ακμές ορίζονται μεταξύ αυτών των κορυφών όποτε οι αντίστοιχες μπάλες κάλυψης έχουν μη κενές τομές. Όταν παρουσιάζεται το γράφημα, το μέγεθος κάθε κορυφής μπορεί να αντικατοπτρίζει τον αριθμό των σημείων δεδομένων που ανήκουν στην αντίστοιχη σφαίρα. Επιπλέον, οι κορυφές μπορούν να χρωματιστούν σε σχέση με τις τιμές κάποιας συνάρτησης που ορίζεται στο σύνολο των

3 Με τον όρο υποκείμενη πολλαπλότητα εννοείται ότι τα δεδομένα που αναλύνονται (δηλαδή, η θέση κάθε δανειολήπτη στον χώρο των δεδομένων) δεν είναι απλά τυχαία σκορπισμένα, αλλά ακολουθούν κάποια κρυμμένη δομή/μοτίβο που συνδέει όλα αυτά τα γνωρίσματα εντός του 14-διάστατου χώρου.

σημείων δεδομένων [1]. Στην περίπτωσή μας η συνάρτηση χρωματισμού θα είναι δύο τιμών, αντιστοιχίζοντας τους defaulters (αθετούντες) στο 1 και τους non-defaulters (μη αθετούντες) στο 0 και ο χρωματισμός των κορυφών θα αντιστοιχεί στο ποσοστό των defaulters εντός της αντίστοιχης μπάλας κάλυψης.

Το γράφημα Ball Mapper προσφέρει μια καλή απεικόνιση της πιστωτικής αξιολόγησης των διαφόρων «γειτονιών» του χώρου των χαρακτηριστικών και των διασυνδέσεων μεταξύ αυτών των γειτονιών. Με αυτό τον τρόπο, επιτρέπει τον εντοπισμό και την περαιτέρω εξέταση περιοχών στις οποίες θα πρέπει να δοθεί μεγαλύτερη προσοχή από τον δανειστή. Επιπλέον, αυτό το γράφημα έχει τη δυνατότητα να χρησιμεύσει ως πλατφόρμα για την παροχή οδηγιών και συμβουλών στους δανειολήπτες, όσον αφορά τα χαρακτηριστικά που θα μπορούσαν να εξετάσουν για βελτίωση, προκειμένου να μεταβούν σε πιο ευνοϊκές πιστωτικές γειτονιές του χώρου των χαρακτηριστικών.

6.2.2 Τοπολογικό μοντέλο

Κατασκευή του Ball Mapper και πιθανότητες αθέτησης: Όπως έχει ήδη αναφερθεί, η κατασκευή του Ball Mapper απαιτεί τον καθορισμό μίας μετρικής και μίας παραμέτρου, της ακτίνας ε [21]. Για την εφαρμογή επιλέχθηκε η Ευκλείδεια μετρική. Θεωρείται ότι το ε που θα επιλεγεί είναι κατάλληλο για τη δημιουργία ενός «καλού» καλύμματος του υποκείμενου τοπολογικού χώρου από τον οποίο έγινε η δειγματοληψία. Η επιλογή του ε γίνεται από τον χρήστη χωρίς ωστόσο να υπάρχει κάποια ιδανική τιμή [22]. Η ιδέα είναι η εξής: αφού γίνει η επιλογή του ε, χρησιμοποιούνται οι επεξηγηματικές μεταβλητές για να κατασκευαστεί το γράφημα Ball Mapper και η μεταβλητή απόκρισης (δηλαδή, αθέτηση=1/όχι αθέτηση=0) για τον χρωματισμό του. Το χρώμα κάθε μπάλας αντιστοιχεί στο ποσοστό «κακών» που βρίσκονται σε αυτήν.

Στη συνέχεια, η πιθανότητα αθέτησης εκτιμάται για κάθε νέο δανειολήπτη ως εξής:

- Εξετάζοντας σε ποια μπάλα ή μπάλες ανήκει ο νέος δανειολήπτης και μετά προσδιορίζονται οι δανειολήπτες που ανήκουν ήδη στην ένωση αυτών των μπαλών.

-Έπειτα, ανατίθεται η μέση τιμή της μεταβλητής απόκρισης αυτών των δανειοληπτών/παρατηρήσεων ως η πιθανότητα αθέτησης του νέου δανειολήπτη.

Στην περίπτωση που ένας νέος δανειολήπτης δεν ανήκει σε καμία μπάλα, μπορούμε να ακολουθήσουμε δύο διαφορετικές προσεγγίσεις:

- Μέθοδος Α (πλησιέστερος γείτονας):

Εντοπίζεται η πλησιέστερη παρατήρηση και βρίσκονται οι μπάλες κάλυψης που την περιέχουν. Στη συνέχεια, στον νέο δανειολήπτη αποδίδεται η μέση τιμή της μεταβλητής απόκρισης για τις παρατηρήσεις που περιέχονται στις μπάλες αυτές.

- Μέθοδος Β (γενικός πληθυσμός):

Ανατίθεται η μέση τιμή της μεταβλητής απόκρισης για όλες τις παρατηρήσεις που περιέχονται στο σύνολο δεδομένων εκπαίδευσης.

Οι παραπάνω διαδικασίες επαναλαμβάνονται για όλους τους νέους δανειολήπτες και το αποτέλεσμα είναι μια εκτίμηση της πιθανότητας αθέτησης για καθέναν από αυτούς.

6.2.3 Προσδιορισμός των παραμέτρων

Όπως προαναφέρθηκε, η μόνη παράμετρος που απαιτείται για την κατασκευή του μοντέλου Ball Mapper είναι η ακτίνα ε. Καθώς δεν υπάρχει συγκεκριμένη μεθοδολογία για τον προσδιορισμό της βέλτιστης τιμής, ακολουθήθηκε η εξής διαδικασία: από το αρχικό Training σύνολο, δημιουργήθηκαν πολλαπλά βοηθητικά σύνολα Training-Testing με επαναληπτική διαδικασία.

Συγκεκριμένα, το αρχικό Training σύνολο χωρίστηκε 50 φορές σε σύνολο Training_2 και Testing_2 με την αναλογία 70/30 αυτών που αθέτησαν προς αυτών που δεν αθέτησαν να παραμένει η ίδια όπως αυτή του αρχικού πληθυσμού.

Για κάθε t = 1, ..., 50 και για διάφορες τιμές του ε (από 0.40-1.79 με βήμα 0.01) κατασκευάστηκε ο αλγόριθμος Ball Mapper(t, ε) του αντίστοιχου $\text{Training}_2(t)$ με τις προκύπτουσες πιθανότητες αθέτησης να αποδίδονται σε κάθε παρατήρηση του $\text{Testing}_2(t)$. Στην συνέχεια υπολογίστηκαν τα αντίστοιχα μέτρα Accuracy(t, ε), Hmeasure (t, ε), Precision(t, ε), Recall(t, ε), F1score(t,ε), AUC(t, ε). Έπειτα, για κάθε ε, υπολογίστηκε ο μέσος όρος (ως προς t) μαζί με την ελάχιστη και μέγιστη τιμή για κάθε μέτρο αξιολόγησης σε όλες τις t επαναλήψεις. Επιλέγουμε να χρησιμοποιήσουμε την τιμή του ε που μεγιστοποιεί τη μέση τιμή για κάθε μέτρο αξιολόγησης ξεχωριστά, ως κριτήριο για την επιλογή του ε.

6.3 Λεπτομερής περιγραφή της εφαρμογής

6.3.1 Περιγραφή συνόλου δεδομένων

Για τους σκοπούς της εργασίας χρησιμοποιήθηκε ένα ευρέως γνωστό σύνολο δεδομένων, το οποίο προέρχεται από το UCI Machine Learning Repository. Αυτό το σύνολο δεδομένων περιλαμβάνει πραγματικές πληροφορίες για δανειολήπτες και είναι γνωστό ως Statlog Australian Credit Approval. Συγκεκριμένα περιέχει πληροφορίες για 690 δανειολήπτες, οι οποίοι κατηγοριοποιούνται σε δύο κατηγορίες: αποδεκτές/καλοί (307 περιπτώσεις) ή απορριφθείσες/κακοί (383 περιπτώσεις) αιτούμενη δάνειου/πίστωσης. Οι περιπτώσεις του συνόλου δεδομένων αποτελούνται από 14 χαρακτηριστικά, με 8 κατηγορικές και 6 αριθμητικές, και 1 χαρακτηριστικό κλάσης που δείχνει αν η αίτηση έγινε δεκτή ή απορρίφθηκε.

6.3.2 Προεπεξεργασία δεδομένων

Αρχικά, για να διασφαλιστεί ότι οι τιμές όλων των χαρακτηριστικών του συνόλου δεδομένων θα κυμαίνονται μεταξύ το [0,1], τα δεδομένα κανονικοποιήθηκαν σύμφωνα με τον τύπο:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)}$$

, όπου x^* αντιπροσωπεύει την κανονικοποιημένη τιμή του χαρακτηριστικού x. Εκτός από αυτή την κανονικοποίηση, δεν υπήρξε καμία άλλη προεπεξεργασία του συνόλου δεδομένων αφού και τα 14 χαρακτηριστικά χρησιμοποιήθηκαν ως επεξηγηματικές μεταβλητές.

6.3.3 Μέτρα αξιολόγησης

Για να αξιολογήσουμε την απόδοση του μοντέλου και να το συγκρίνουμε με άλλα μοντέλα χρησιμοποιούμε τις εξής έξι μετρικές Accuracy, Hmeasure, Precision, Recall, F1score, AUC οι οποίες περιγράφονται αναλυτικά στο κεφάλαιο 5.

6.3.4 Η επιλογή της παραμέτρου ε.

Για τον προσδιορισμό μιας κατάλληλης τιμής για την παράμετρο ε και για την κατασκευή του σχετικού BallMapper, εφαρμόστηκε η μεθοδολογία που περιγράφεται στην ενότητα 6.2.2. Συγκεκριμένα πραγματοποιήθηκε 50 φορές η διαδικασία, όπου κάθε επανάληψη αποτελείτο από έναν διαχωρισμό 70/30 σε Training_2 και Testing_2, διατηρώντας παράλληλα την αναλογία αθετούντων και μη αθετούντων όπως στο αρχικό σύνολο δεδομένων. Χρησιμοποιήθηκε η εντολή set.seed για να διασφαλιστεί η αναπαραγωγιμότητα των αποτελεσμάτων. Τα γραφήματα Ball Mapper κατασκευάστηκαν με τη χρήση της R και το πακέτο BallMapper και οι μετρικές υπολογίστηκαν με τα πακέτα Metrics και Hmeasure. Επιπλέον, υπολογίστηκαν οι μέσες, ελάχιστες και μέγιστες τιμές του accuracy, hmeasure, precision, recall, f1score, AUC για το κάθε ένα ξεχωριστά για κάθε τιμή ε. Οι προκύπτουσες accuracy, hmeasure, precision, recall, f1score, AUC τιμές χρησιμοποιήθηκαν για την επιλογή των βέλτιστων $\varepsilon*$ για κάθε μετρική αξιολόγησης.

6.3.5 Cross validation

Χρησιμοποιήθηκε η διασταυρωμένη επικύρωση (cross validation) για να αξιολογηθεί η απόδοση του μοντέλου στο σύνολο δεδομένων. Συγκεκριμένα, το αρχικό σύνολο δεδομένων χωρίστηκε σε 40 σύνολα εκπαίδευσης και δοκιμών, διατηρώντας παράλληλα το ποσοστό του αρχικού συνόλου δεδομένων των δανειοληπτών που αθέτησαν και των δανειοληπτών που δεν αθέτησαν. Δηλαδή κάθε διαχωρισμός ήταν 70/30. Έπρεπε να επιλεγεί μια κατάλληλη τιμή για την παράμετρο ε σε κάθε κύκλο. Για να επιτευχθεί αυτό, χωρίστηκε κάθε ένα από τα 40 σύνολα εκπαίδευσης σε σύνολα Training_2 και Testing_2 δέκα φορές το καθένα, διατηρώντας παράλληλα το ίδιο ποσοστό defaulters και non-defaulters. Χρησιμοποιήθηκαν accuracy, hmeasure, precision, recall, f1score, AUC για να καθοριστούν οι κατάλληλες τιμές του ε.

Για τα ε που προέκυψαν εφαρμόστηκε ο αλγόριθμος του Ball Mapper για το training set, αποδίδοντας πιθανότητες αθέτησης στις νέες παρατηρήσεις και με τις δύο μεθόδους και υπολογίστηκαν τα accuracy, hmeasure, precision, recall, f1score, AUC.

Επιπρόσθετα, αποθηκεύτηκαν σε ένα πίνακα οι ακτίνες που έδιναν την βέλτιστη μέση τιμή για κάθε μέτρο σε κάθε επανάληψη.

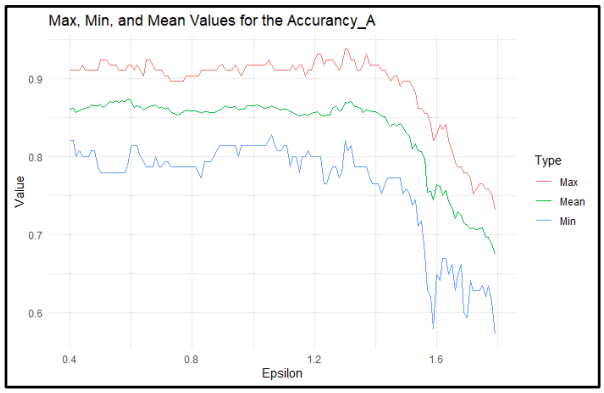
6.3.6 Σύγκριση με άλλα μοντέλα

Χρησιμοποιώντας τα ίδια σύνολα δεδομένων εκπαίδευσης και δοκιμής και τα ίδια χαρακτηριστικά και σε άλλα μοντέλα αναφοράς όπως τα Νευρωνικά δίκτυα, την λογιστική παλινδρόμηση και τα support vector machines. Στην συνέχεια συγκρίθηκαν οι επιδόσεις τους με το μοντέλο Ball Mapper.

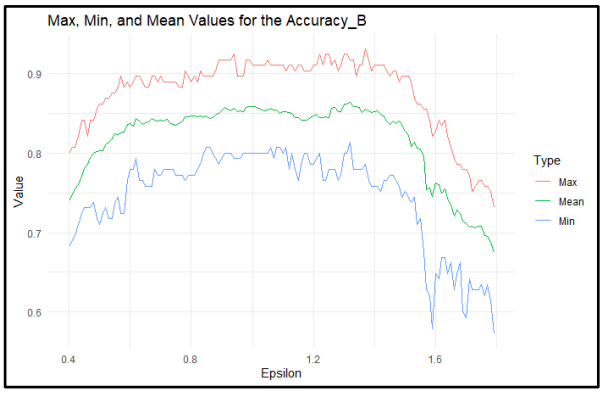
Κεφάλαιο 7 : Αποτελέσματα

Η διαδικασία που εφαρμόστηκε είχε ως στόχο τον καθορισμό του ε και κατ' επέκταση την κατασκευή του Ball Mapper χωρίζοντας το σύνολο εκπαίδευσης σε training$_2$ και testing$_2$ πενήντα φορές. Για την επίτευξη του ίδιου αποτελέσματος κάθε φορά έγινε χρήση της εντολής set.seed. Στη συνέχεια, κατασκευάστηκε το BM για 140 τιμές του ε, που κυμαίνονταν από 0,40-1,79 με βήμα 0,01. Μετά από κάθε κατασκευή, αποδόθηκαν οι πιθανότητες αθέτησης καθώς και τα accuracy, hmeasure, precision, recall, f1score, AUC και υπολογίστηκαν: η μέση, η ελάχιστη και η μέγιστη τιμή για κάθε ένα από αυτά. Το πακέτο "BallMapper" στην R χρησιμοποιήθηκε για την κατασκευή του Ball Mapper, το πακέτο "Metrics" χρησιμοποιήθηκε για τον υπολογισμό του Accuracy και το πακέτο "hmeasure" για όλα τα υπόλοιπα μέτρα.

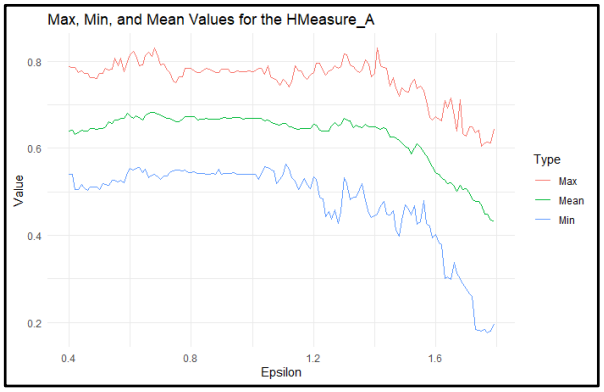
Τα ακόλουθα διαγράμματα[4] απεικονίζουν τα ευρήματα:



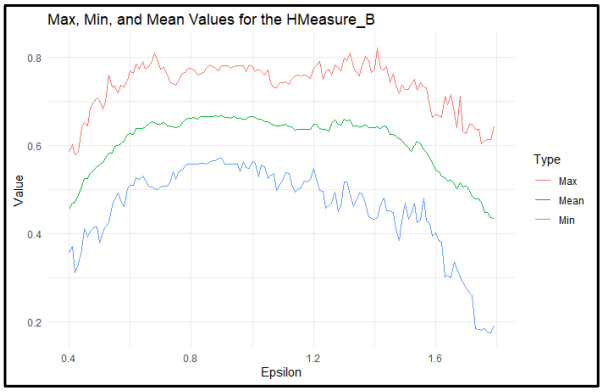
max=0.94,mean=0.87,min=0.57



max=0.93,mean=0.86,min=0.57

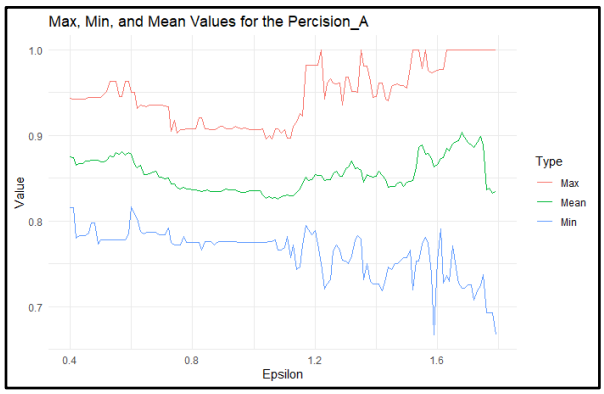


max=0.83, mean=0.68, min=0.17

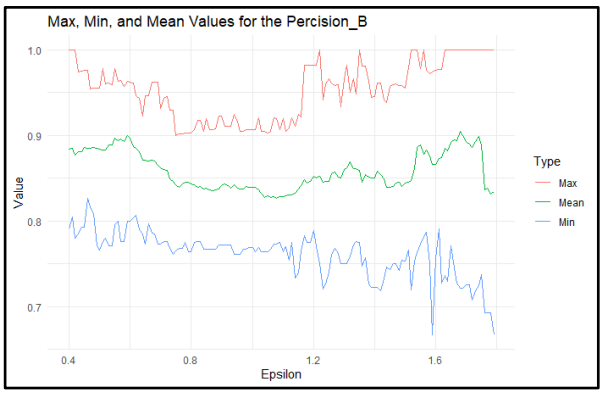


max=0.82,mean=0.67,min=0.17

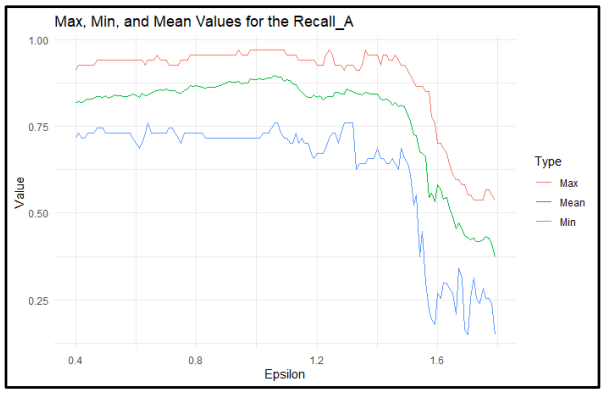
[4] Στο παράτημα οι πίνακες 3 μέχρι 8 απεικονίζουν τα αποτελέσματα αναλυτικά



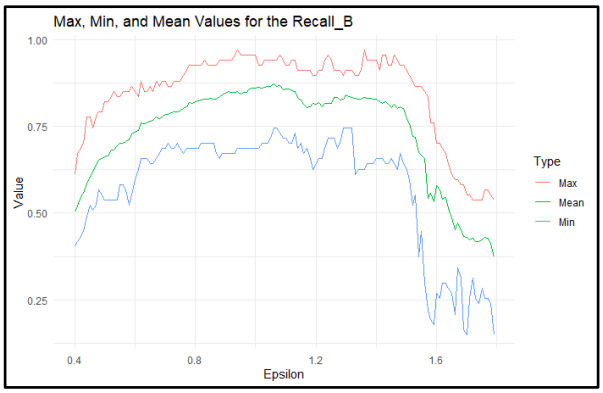
max=1.00,mean=0.90, min=0.67



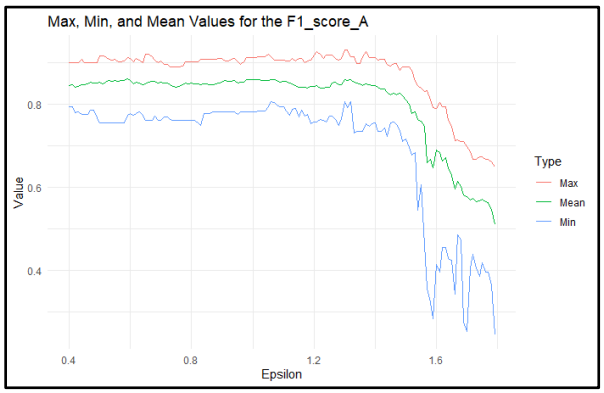
max=1.00, mean=0.90, min=0.67



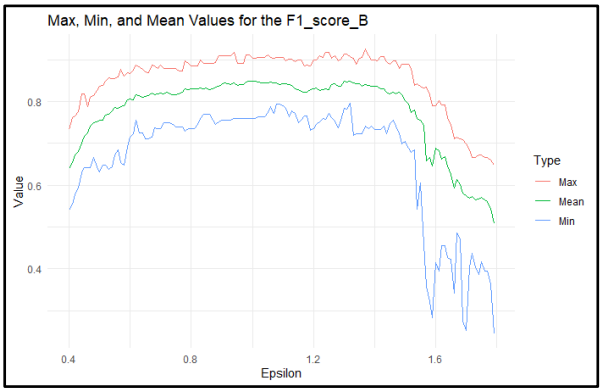
max=0.98,mean=0.90, min= 0.15



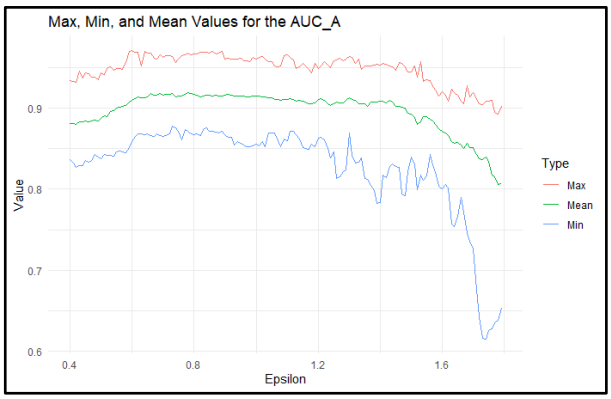
max=0.99, mean=0.87, min=0.14

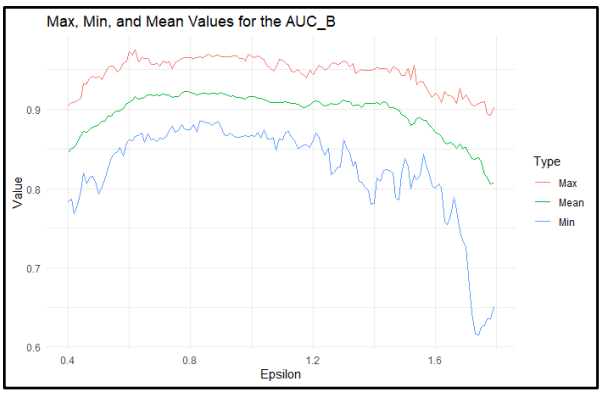


max=0.93,mean=0.86 , min=0.24



max=0.93, mean=0.85, min=0.24





max=0.97,mean=0.92, min=0.62

max=0.98, mean=0.92, min=0.62

Σε όλα τα μέτρα απόδοσης με μια πρώτη ματιά φαίνεται ότι οι τιμές του μέγιστου, μέσου μέγιστου και ελάχιστου δεν έχουν μεγάλες διαφορές μεταξύ των δύο μεθόδων. Η ακρίβεια κάθε μοντέλου για τις μέγιστες τιμές και για τις δύο μεθόδους βρίσκεται μεταξύ 0,82 και 1,00 και αντιστοιχεί σε διάφορες ακτίνες από 0,40 μέχρι και 1,44.

Το μοναδικό μέτρο που φτάνει την ακρίβεια του μέχρι το 1 είναι το Precision και η ακτίνα, η οποία έδωσε αυτήν την ακρίβεια, είναι 1.22. Να σημειωθεί ότι η τιμή 1 στο Precision συμβολίζει τους δανειολήπτες, οι οποίοι είχαν προβλεφθεί ότι δε θα αθετήσουν το δάνειο, και, τελικά, δεν το αθέτησαν.

Επιπρόσθετα, παρατηρείται μεγάλη πτώση του ελάχιστου σε όλα τα μέτρα για τις τιμές του ε από 1.60-1.79.

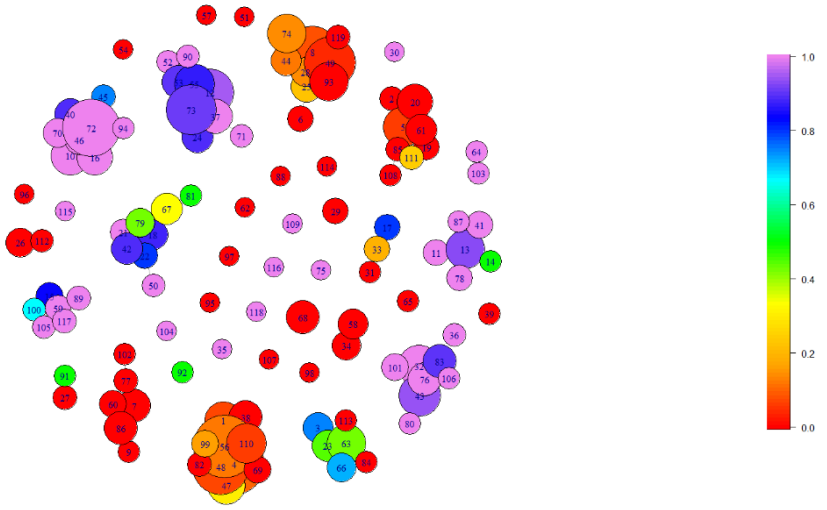
Όσον αφορά τη μέση μέγιστη τιμή, ο ακόλουθος πίνακα σύνοψης των τιμών κάθε μέτρου βοηθά στην κατανόηση των δεδομένων που παρουσιάστηκαν προηγουμένως. Ακόμα, στον ίδιο πίνακα φαίνεται και η ακτίνα που απέδωσε την εκάστοτε τιμή.

Πίνακας 1: Παρουσιάζει τη μέγιστη μέση τιμή κάθε μέτρου αξιολόγησης κατά την διάρκεια της εκπαίδευσης καθώς και την ακτίνα που έδωσε την τιμή αυτή.

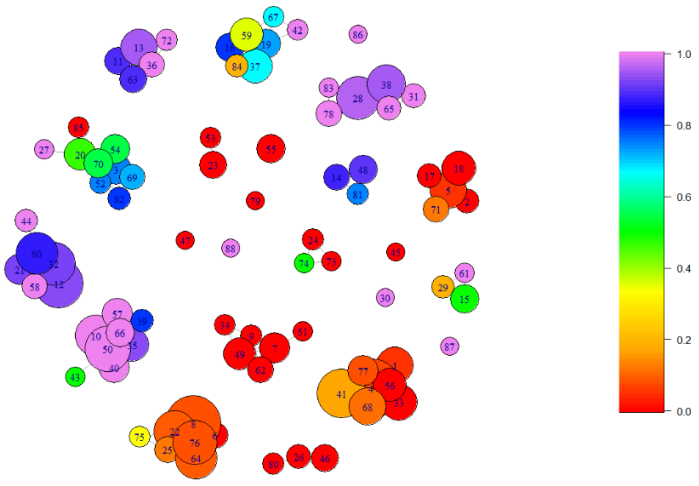
ΜΕΘΟΔΟΣ	Accuracy	Hmeasure	Precision	Recall	F1 score	AUC
A	0.87 ε=0.59	0.68 ε=0.67	0.90 ε=1.68	0.90 ε=1.06	0.86 ε=0.59	0.92 ε=0.78
B	0.86 ε=1.32	0.67 ε=0.90	0.90 ε=1.68	0.87 ε=1.06	0.85 ε=1.32	0.92 ε=0.78

Οι τιμές του Accuracy, Hmeasure, Recall και F1score με την μέθοδο A είναι ελαφρώς μεγαλύτερες από ότι στη μέθοδο B. Οι τιμές του Precision και AUC είναι παρόμοιες μεταξύ των δύο μεθόδων.

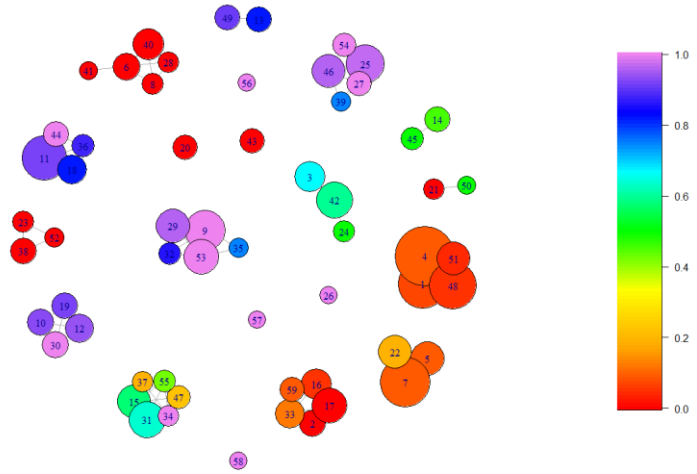
Στην συνέχεια, για καθεμιά από αυτές τις ακτίνες κατασκευάστηκε το γράφημα Ball Mapper για το αρχικό training & testing και έγινε επαναληπτικός υπολογισμός των 6 μέτρων απόδοσης που ελέγχθηκαν.



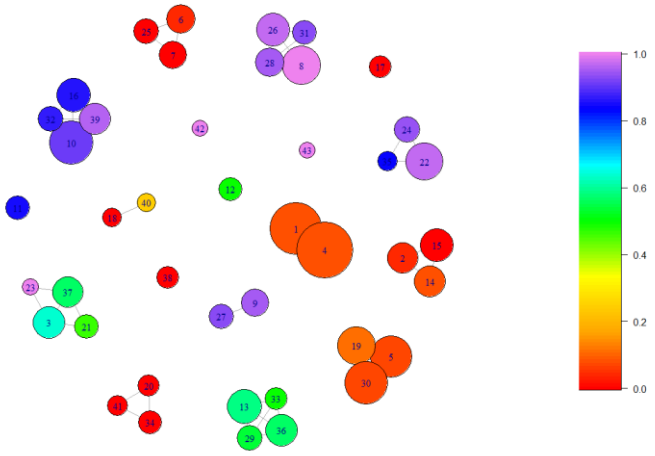
Εικόνα 28: Ακτίνα 0.59



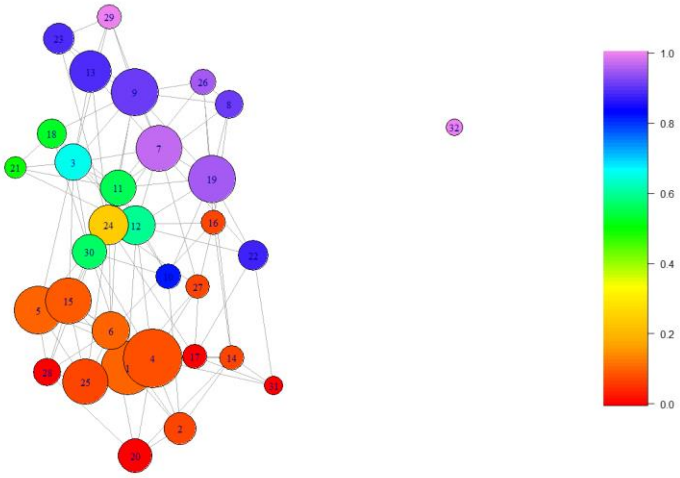
Εικόνα 29: Ακτίνα 0.67



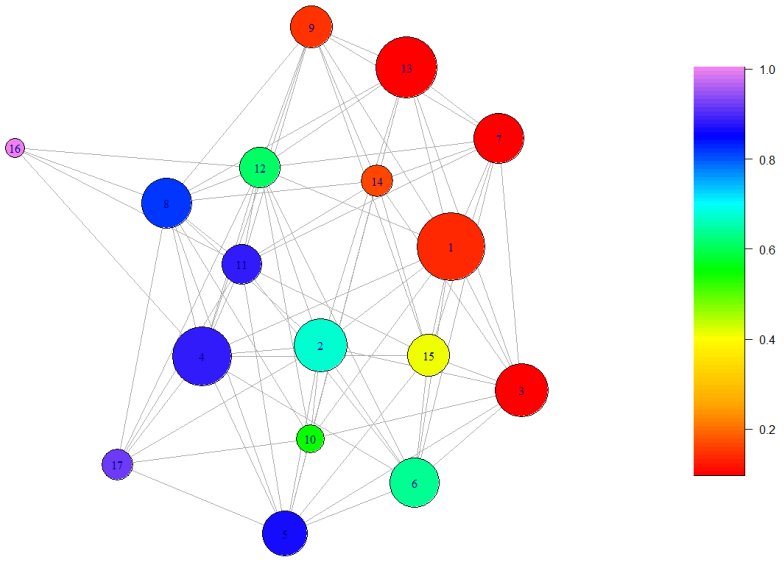
Εικόνα 30: Ακτίνα 0.78



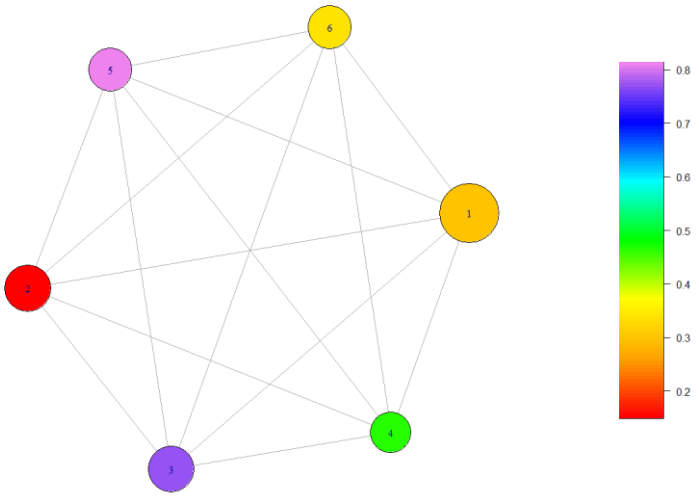
Εικόνα 31: Ακτίνα 0.90



Εικόνα 32: Ακτίνα 1.06



Εικόνα 33: Ακτίνα 1.32



Εικόνα 34: Ακτίνα 1,68

Πίνακας 2: Κατασκεύη Ball Mapper στο αρχικό Training & Testing και η απόδοση τους στα 6 μέτρα αξιολόγησης.

Σκόρ εκπαίδευσης	**Μέθοδος Προσδιορισμού ε***	**Ακτίνα ε***	**Accuracy**	**Hmeasure**	**Precision**	**Recall**	**F1 score**	**AUC**
0.87 **0.86**	Accuracy_A/ F1_score_A	**0.59**	0.80	0.50	0.75	0.78	0.76	0.85
0.68	Hmeasure_A	**0.67**	0.82	0.55	0.74	0.85	0.79	0.88
0.92 **0.92**	AUC_A/ AUC_B	**0.78**	0.82	0.53	0.73	0.87	0.80	0.87
0.67	Hmeasure_B	**0.90**	0.81	0.54	0.72	0.87	0.79	0.87
0.90 **0.87**	Recall_A/ Recall_B	**1.06**	0.81	0.53	0.70	0.92	0.80	0.87
0.85 **0.86**	F1_Score_B/ Accuracy_B	**1.32**	0.80	0.48	0.74	0.80	0.77	0.84
0.90 **0.90**	Precision_A Precision_B	**1.68**	0.73	0.41	0.85	0.41	0.56	0.83

Για το Accuracy οι ακτίνες με τη μεγαλύτερη απόδοση ήταν 0,67 και 0,78. Παρατηρείται ότι για το hmeasure οι αποδόσεις είναι μικρότερες σε σχέση με τα άλλα μέτρα. Στο μέτρο αυτό τη μεγαλύτερη απόδοση την έδωσε η ακτίνα 0,67 δείχνοντας ισχυρό διαχωρισμό μεταξύ των κλάσεων ανισορροπίας. Όσον αφορά το Percision, η μεγαλύτερη απόδοση βρίσκεται στην ακτίνα 1.68 και στο Recall στην ακτίνα 1.06. Επιπρόσθετα, για το f1score οι ακτίνες με τη μεγαλύτερη απόδοση ήταν 0,78 και 1,06. Τέλος, το AUC φέρει τη μεγαλύτερη απόδοση στην ακτίνα 0,67.

Παρατηρείται, ακόμα, ότι τη μεγαλύτερη απόδοση από όλα τα μέτρα τη δίνει η μετρική Recall με 92% και αντιστοιχεί στην ακτίνα 1.06. Συνολικά, όμως, η ακτίνα που έχει τις μεγαλύτερες αποδόσεις σε όλα τα μέτρα είναι 0.67.

7.1 Cross Validation

Προκειμένου να γίνει η επικύρωση του μοντέλου (Ball Mapper) με τους τρόπους Α και Β, πραγματοποιήσαμε 40 επαναλήψεις, στις οποίες το αρχικό σύνολο δεδομένων χωρίστηκε σε training_1 και testing_1, σε αναλογία 70/30, με αναλογία κακών αυτή του αρχικού πληθυσμού (δηλαδή 0.46). Σε κάθε επανάληψη χρειάστηκε ο καθορισμός της παραμέτρου του μοντέλου, δηλαδή της ακτίνας ε. Για το σκοπό αυτό, το κάθε training_1 χωρίστηκε 10 φορές σε training_2 και testing_2 τηρώντας τις παραπάνω αναλογίες. Σε καθεμία από αυτές τις επαναλήψεις, εφαρμόστηκε ο αλγόριθμος Ball Mapper για ε, από 0.55-1.80 με βήμα 0.05, υπολογίστηκαν οι πιθανότητες αθέτησης για το αντίστοιχο testing_2 και βρέθηκαν τα: accuracy, hmeasure, precision, recall, f1score, AUC . Μετά από τις 10 επαναλήψεις έγινε ο υπολογισμός για τη μεγαλύτερη μέση τιμή accuracy, hmeasure, precision, recall, f1score, AUC και προσδιορίστηκε για κάθε μέτρο το ε, στο οποίο αντιστοιχεί το μεγαλύτερο μέσο accuracy, hmeasure, precision, recall, f1score, AUC, αντίστοιχα.

Οι τιμές των accuracy, hmeasure, precision, recall, f1score, AUC καθώς και τα ε που επιλέχθηκαν και με τις δύο μεθόδους παρουσιάζονται στους Πίνακες του παραρτήματος.

Το μέσο Accuracy βρέθηκε 0.85 με τον τρόπο Α και 0.84 με τον τρόπο Β.

Το μέσο Hmeasure βρέθηκε 0.64 με τον τρόπο Α και 0.63 με τον τρόπο Β.

Το μέσο Precision βρέθηκε 0.87 με τον τρόπο Α και 0.87 με τον τρόπο Β.

Το μέσο Recall βρέθηκε 0.88 με τον τρόπο Α και 0.85 με τον τρόπο Β.

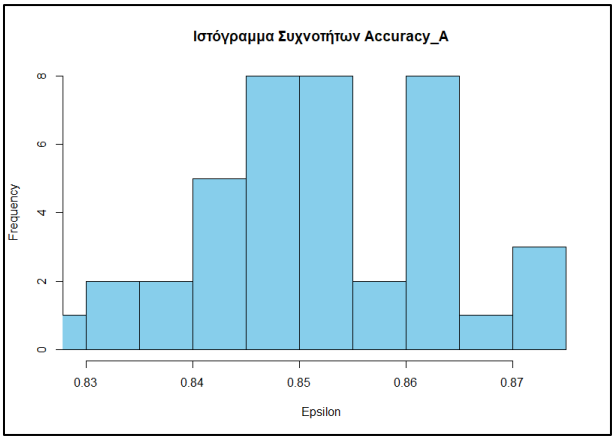
Το μέσο F1score βρέθηκε 0.84 με τον τρόπο Α και 0.83 με τον τρόπο Β.

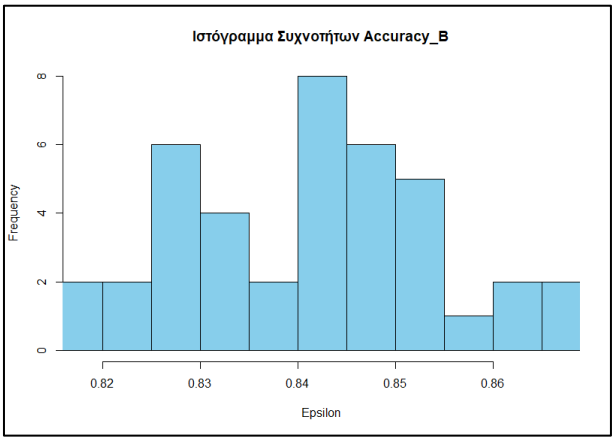
Το μέσο AUC βρέθηκε 0.90 με τον τρόπο Α και 0.91 με τον τρόπο Β.

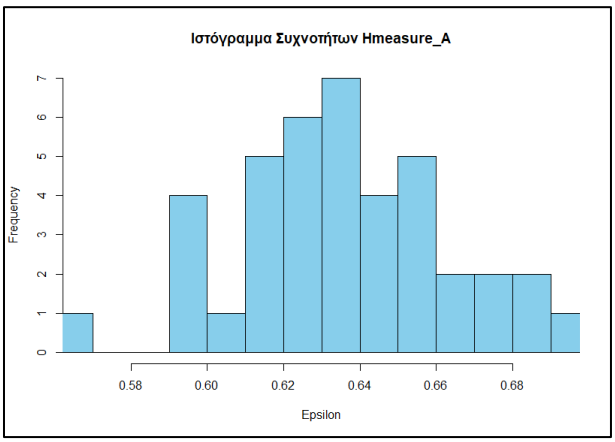
Συγκρίνοντας τα αποτελέσματα του cross validation με τον πίνακα 1 γίνεται ξεκάθαρο ότι η μέση μέγιστη τιμή κάθε μέτρου απόδοσης δε διαφέρει πολύ από των υπολοίπων.

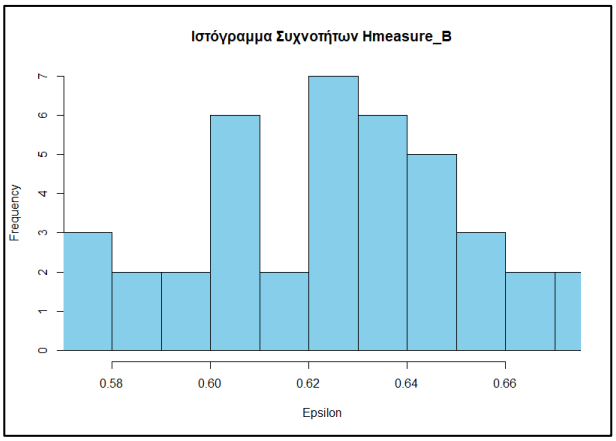
ΜΕΘΟΔΟΣ	A	B	A	B
	1 ΕΠΑΝΑΛΗΨΗ		CROSS VALIDATION	
Accuracy	0.87 ε=0.59	0.86 ε=1.32	0.85 ε=0.88	0.84 ε=0.73
Hmeasure	0.68 ε=0.67	0.67 ε=0.90	0.64 ε=0.89	0.63 ε=0.60
Precision	0.90 ε=1.68	0.90 ε=1.68	0.87 ε=1.37	0.87 ε=1.33
Recall	0.90 ε=1.06	0.87 ε=1.06	0.88 ε=0.95	0.85 ε=0.79
F1 score	0.86 ε=0.59	0.85 ε=1.32	0.84 ε=0.93	0.83 ε=0.79
AUC	0.92 ε=0.78	0.92 ε=0.78	0.90 ε=0.85	0.91 ε=0.81

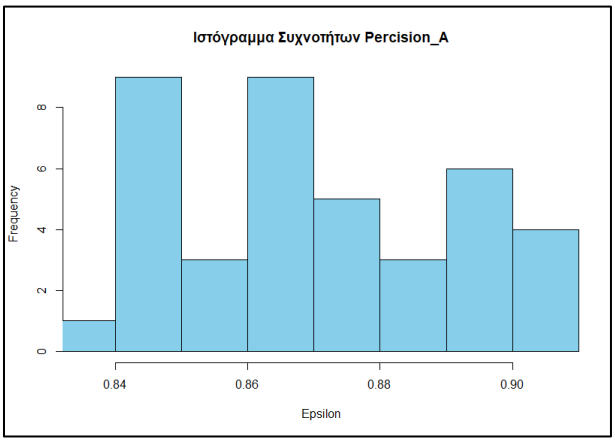
Τα αποτελέσματα απεικονίζονται και στο παρακάτω ιστόγραμμα συχνοτήτων:

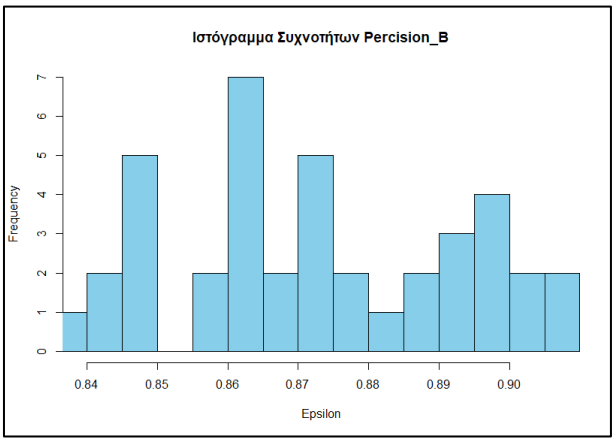


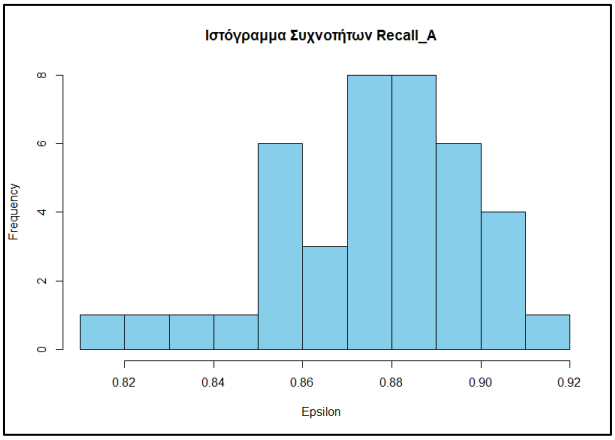


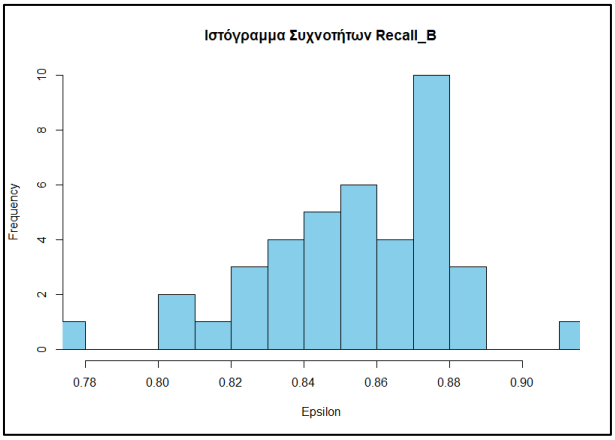


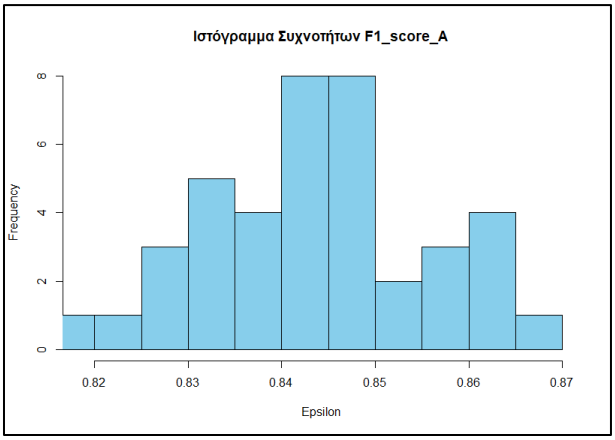


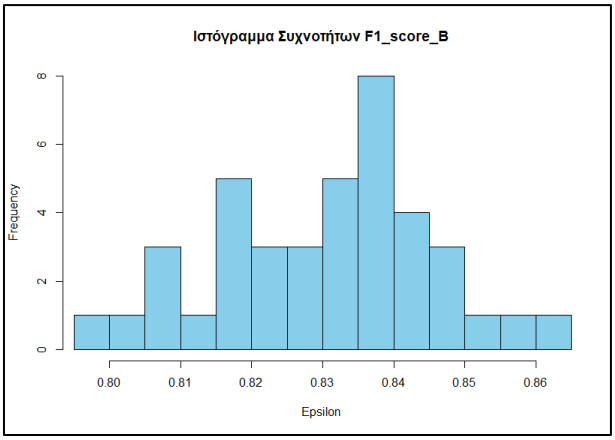

Ιστόγραμμα Συχνοτήτων Percision_A
Frequency
0
2
4
6
8
0.84
0.86
0.88
0.90
Epsilon

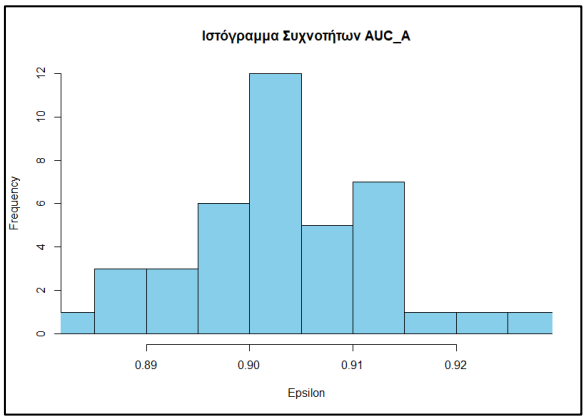

Ιστόγραμμα Συχνοτήτων Percision_B
Frequency
0
1
2
3
4
5
6
7
0.84
0.85
0.86
0.87
0.88
0.89
0.90
Epsilon

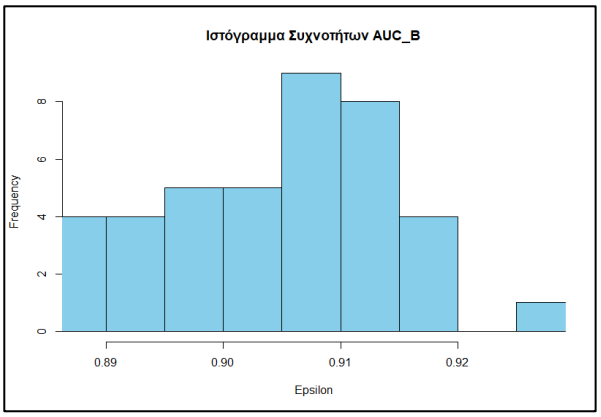

Ιστόγραμμα Συχνοτήτων Recall_A
Frequency
0
2
4
6
8
0.82
0.84
0.86
0.88
0.90
0.92
Epsilon


Ιστόγραμμα Συχνοτήτων Recall_B
Frequency
0
2
4
6
8
10
0.78
0.80
0.82
0.84
0.86
0.88
0.90
Epsilon


Ιστόγραμμα Συχνοτήτων F1_score_A
Frequency
0
2
4
6
8
0.82
0.83
0.84
0.85
0.86
0.87
Epsilon


Ιστόγραμμα Συχνοτήτων F1_score_B
Frequency
0
2
4
6
8
0.80
0.81
0.82
0.83
0.84
0.85
0.86
Epsilon





Στην συνέχεια, δημιουργήθηκαν οι πίνακες συχνοτήτων για τις ακτίνες που παρουσίασαν τη μέση μέγιστη τιμή στις 40 επαναλήψεις. Με τον τρόπο αυτό διαπιστώθηκε ποιες είναι οι επικρατούσες ακτίνες για κάθε μέτρο. Επιπρόσθετος, υπολογίστηκε και η μέση ακτίνα για κάθε μέτρο.

Μέθοδος Α													Μέσο ε
ACCURACY	**ΑΚΤΙΝΑ**	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,2		0,88
	ΣΥΧΝΟΤΗΤΑ	1	2	8	2	8	1	12	1	3	2		
HMEASURE	**ΑΚΤΙΝΑ**	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,15	1,25		0,89
	ΣΥΧΝΟΤΗΤΑ	2	8	1	9	2	13	2	1	1	1		
PERSISION	**ΑΚΤΙΝΑ**	0,55	0,6	0,65	1,2	1,4	1,45	1,5	1,55	1,6	1,65	1,7	1,37
	ΣΥΧΝΟΤΗΤΑ	5	2	1	1	1	1	5	7	9	4	4	
RECALL	**ΑΚΤΙΝΑ**	0,85	0,9	0,95	1	1,05							0,95
	ΣΥΧΝΟΤΗΤΑ	4	3	27	2	4							
F1_SCORE	**ΑΚΤΙΝΑ**	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,2	1,3	0,93
	ΣΥΧΝΟΤΗΤΑ	1	1	2	1	7	2	19	1	3	2	1	
AUC	**ΑΚΤΙΝΑ**	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,25		0,85
	ΣΥΧΝΟΤΗΤΑ	2	2	13	1	7	1	7	4	2	1		

Μέθοδος Β												Μέσο ε
ACCURACY	**ΑΚΤΙΝΑ**	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1,2	0,73
	ΣΥΧΝΟΤΗΤΑ	8	4	5	8	4	1	2	3	3	2	
HMEASURE	**ΑΚΤΙΝΑ**	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,95			0,60
	ΣΥΧΝΟΤΗΤΑ	2	2	6	12	9	3	3	3			
PERSISION	**ΑΚΤΙΝΑ**	0,55	0,65	1,2	1,4	1,45	1,5	1,55	1,6	1,65	1,7	1,33
	ΣΥΧΝΟΤΗΤΑ	2	1	1	1	2	7	7	8	5	6	
RECALL	**ΑΚΤΙΝΑ**	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05		0,79
	ΣΥΧΝΟΤΗΤΑ	2	2	2	2	5	10	13	1	3		
F1_SCORE	**ΑΚΤΙΝΑ**	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95		0,79
	ΣΥΧΝΟΤΗΤΑ	5	3	4	7	8	2	2	4	5		
AUC	**ΑΚΤΙΝΑ**	0,6	0,65	0,7	0,75	0,8	0,9	0,95	1	1,05		0,81
	ΣΥΧΝΟΤΗΤΑ	1	7	3	11	3	3	2	7	3		

Τέλος, με το αρχικό training και testing κατασκευάστηκαν κι άλλα μοντέλα όπως: Νευρωνικά δίκτυα, λογιστική παλινδρόμηση και support vector machine και εξετάστηκε η απόδοση των 6 μέτρων σε σχέση με τον ball mapper τις διάφορες ακτίνες.

Μεθοδος	Accuracy	Hmeasure	Precision	Recall	F1 score	AUC
Ball Mapper 0.59	80%	50%	75%	78%	76%	85%
Ball Mapper 0.67	82%	55%	74%	85%	79%	88%
Ball Mapper 0.78	82%	53%	73%	87%	80%	87%
Ball Mapper 0.90	81%	54%	72%	87%	79%	87%
Ball Mapper 1.06	81%	53%	70%	92%	80%	87%
Ball Mapper 1.32	80%	48%	74%	80%	77%	84%
Ball Mapper 1.68	73%	41%	85%	41%	56%	83%
Logistic Regression	83%	58%	74%	86%	80%	91%
Neural Network	83%	61%	80%	78%	79%	91%
Support Vector Machine	81%	48%	71%	92%	80%	83%

Το συμπέρασμα είναι ότι η Λογιστική Παλινδρόμηση και τα Νευρωνικά δίκτυα έχουν υψηλότερα ποσοστά απόδοσης σε σχέση με τα άλλα μοντέλα. Παρόλα αυτά, η διαφορά τους δεν είναι τόσο μεγάλη.

Κεφάλαιο 8 Επίλογος-Συμπεράσματα

Στο πλαίσιο της παρούσας διπλωματικής εργασίας παρουσιάστηκε μια ολοκληρωμένη επισκόπηση της Τοπολογική Ανάλυσης Δεδομένων (TDA), ένα σχετικά νέο πεδίο που προσφέρει μια μοναδική προοπτική για την κατανόηση περίπλοκων και πολυδιάστατων δεδομένων. Η βασική ιδέα της είναι ότι οι τοπολογικές ιδιότητες ενός συνόλου δεδομένων, όπως η συνδεσιμότητα και η παρουσία οπών, περιέχουν χρήσιμες πληροφορίες. Η TDA χρησιμοποιεί αυτή την ιδέα για να αποκαλύψει τη δομή σε σύνολα δεδομένων υψηλής διάστασης.

Επιπλέον μελετήθηκαν ορισμένες έννοιες από την αλγεβρική τοπολογία που παρείχαν τα θεμέλια για την κατανόηση των εργαλείων και των μεθόδων της TDA. Συγκεκριμένα, αναπτύχθηκε η θεμελιώδης έννοια των n-διάστατων "οπών" και ο τρόπος με τον οποίο η simplicial ομολογία ανιχνεύει την παρουσία τους(Κεφάλαιο 1). Ακόμα, αναφέρθηκε πώς η εμμένουσα ομολογία επεκτείνει αυτήν την έννοια για να αντιμετωπίζει δεδομένα νέφους σημείων και να εξάγει χαρακτηριστικά από πολλαπλές κλίμακες(Κεφάλαιο 2).Ακόμα, παρουσιάστηκαν δύο σημαντικοί αλγόριθμοι ο Mapper(Κεφάλαιο 3) και ο Ball Mapper (Κεφάλαιο 4) ως εργαλεία οπτικοποίησης των δεδομένων.

Στην συνέχεια με την βοήθεια της γλώσσας προγραμματισμού R, έγινε η μελέτη ενός τοπολογικού μοντέλου διαβάθμισης πιστοληπτικής ικανότητας κάτω από το πρίσμα διαφορετικών μέτρων αξιολόγησης μοντέλων(Κεφάλαιο 6). Στα αποτελέσματα της εφαρμογής προέκυψε ότι οι δύο μέθοδοι, Μέθοδος Α και Μέθοδος Β, που χρησιμοποιήθηκαν για την κατασκευή του Ball Mapper, παρουσιάζουν παρόμοιες επιδόσεις σε όλα τα μετρικά αξιολόγησης, όπως Accuracy, Hmeasure, Precision, Recall, F1 score, και AUC.

Η σύγκριση των μέγιστων, ελάχιστων, και μέσων τιμών κάθε μετρικής δεν αποκαλύψε σημαντικές διαφορές μεταξύ των δύο μεθόδων. Το Ball Mapper με ακτίνα 0.67 παρουσίασε συνολικά ικανοποιητική απόδοση, με το Accuracy να φτάνει το 82%, το Hmeasure 55%, το Precision το 74%, το Recall το 85%, το F1_score το 79% και το AUC το 88%. Επιπλέον, η σύγκριση με άλλα μοντέλα όπως Νευρωνικά Δίκτυα, Λογιστική Παλινδρόμηση και Support Vector Machine αποκάλυψε παρόμοιες επιδόσεις, με τη Λογιστική Παλινδρόμηση και τα Νευρωνικά Δίκτυα να έχουν ελαφρώς υψηλότερα ποσοστά ακρίβειας.

Επιπρόσθετα, το γεγονός ότι τα αποτελέσματα της μεθόδου που βασίσθηκε στον αλγόριθμο Ball Mapper είναι επαρκώς συγκρίσιμα με τα αποτελέσματα των υπόλοιπων μοντέλων, οδηγεί στο συμπέρασμα ότι η μέθοδος αυτή θα μπορούσε να λειτουργήσει συμπληρωματικές στους υπόλοιπους 3 αλγόριθμούς μηχανικής μάθησης, αφού το συγκεκριμένο μοντέλο ξεχωρίζει για την εννοιολογική του απλότητα, την επεξηγησιμότητα και την ερμηνευσιμότητά του.

Συνολικά, η εργασία αναδεικνύει την αποτελεσματικότητα του Ball Mapper στην ανάλυση και την αξιολόγηση δεδομένων, προσφέροντας προοπτικές για εφαρμογές σε διάφορους τομείς, όπως στην περίπτωση μας στο credit scoring (Κεφάλαιο 7).

Βιβλιογραφία

[1] M. Charmpi, P. Giannouli και S. Xanthopoulos, «A Topological Approach to Credit Scoring,» Available at SSRN 4629581, 2023.

[2] V. Dimitrievska Ristovska και S. Petar , «Application of presistent Homology on Bio-Medical data -- A case study,» Mathematical Modeling, τόμ. 3, αρ. 4, pp. 109--112, 2019.

[3] M. E. Aktas, E. Akbas and A. E. Fatmaoui, "Persistence homology of networks: methods and applications," Applied Network Science, vol. 4, no. 1, pp. 1-28, 2019.

[4] R. Kraft , Illustrations of data analysis using the mapper algorithm and persistent homology, Master of Science - Applied and Computational Mathematics, 2016.

[5] J. S. Garcia, Applications of topological data analysis to natural language processing and computer vision., Colorado State University, 2022.

[6] C. Shultz, «Applications of Topological Data Analysis in Economics,» Available at SSRN 4378151, 2023.

[7] E. A. Rooke, On the Efficient Implementation and Parameter Selection for TDA Mapper, The University of Iowa, 2023.

[8] F. Chazal και M. Bertrand , «An introduction to topological data analysis: fundamental and practical aspects for data scientists,» Frontiers in artificial intelligence, αρ. 4, p. 108, 2021.

[9] G. Naitzat, A. Zhitnikov και L.-H. Lim, «Topology of deep neural networks,» The Journal of Machine Learning Research, τόμ. 21, αρ. 1, pp. 7503--7542, 2020.

[10] Μ. Χαρμπή, Τοπολογική ανάλυση δεδομένων και εφαρμογή στην βαθμονόμηση πιστοληπτικής ικανότητας, Πανεπιστήμιο Αιγαίου, 2023.

[11] M. Khoury, «Introduction to Simplicial Homology,» [Ηλεκτρονικό]. Available: https://web.cse.ohio-state.edu/~wang.1016/courses/788/Lecs/lec6-marc.pdf.

[12] E. Munch, «A user’s guide to topological data analysis,» Journal of Learning Analytics, τόμ. 4, αρ. 2, pp. 47--61, 2017.

[13] . H. Edelsbrunner και J. Harer, «Persistent homology - A survey,» Journal of symbolic computation, τόμ. 453, αρ. 26, pp. 257--282, 2008.

[14] Ε. Κ. Καραπέτσας, «Τοπολογική Ανάλυση Δεδομένων και Χρηματοοικονομικές Χρονοσειρές,» Πανεπιστήμιο Αιγαίου, 2023.

[15] K. Mittal και S. Gupta, «Topological characterization and early detection of bifurcations and chaos in complex systems using persistent homology,» Chaos: An Interdisciplinary Journal of Nonlinear Science, τόμ. 27, αρ. 5, 2017.

[16] G. Carlsson, R. Jardine, D. Feichtner-Kozlov, D. Morozov, F. Chazal, V. de Silva, B. Fasy, J. Johnson, M. Kahle και G. Lerman, «Topological data analysis and machine learning theory,» σε Proc. BIRS Workshop, 2012, pp. 1--11.

[17] M. Mohnhaupt , The Nerve Theorem and its Applications in Topological Data Analysis, Zurich: Bachelor of Science ETH in Mathematics Swiss Federal Institute of Technology (ETH), 2023.

[18] G. Singh, M. Facundo και E. C. Gunnar, «Topological methods for the analysis of high dimensional data sets and 3d object recognition,» PBG@ Eurographics, αρ. 2, pp. 091--100, 2007.

[19] M. A. Joshi και D. L. Joshi, «Topological Data Analysis Based Feature Selection for Predicting Fatigue Strength of Steel Using Machine Learning.,» σε IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2020, p. 012083.

[20] T. K. Dey και W. Yusu , Computational topology for data analysis, Cambridge University Press, 2022.

[21] P. Dłotko, «Ball mapper: A shape summary for topological data analysis,» arXiv preprint arXiv:1901.07410, 2019.

[22] P. Dlotko, Q. Wanling και S. Rudkin, «Topological Data Analysis Ball Mapper for Finance,» arXiv preprint arXiv:2206.03622, 2022.

[23] W. Qiu, R. Simon και P. Dłotko, «Refining understanding of corporate failure through a topological data analysis mapping of Altman's Z-score model,» Expert Systems with Applications, αρ. 156, p. 113475, 2020.

[24] P. Dlotko, D. Gurnari και R. Sazdanovic, «Mapper-type algorithms for complex data and relations,» arXiv e-prints, pp. arXiv--2109, 2021.

[25] G. Margherita, B. Enrico και V. Giorgio , «Metrics for multi-class classification: an overview,» arXiv preprint arXiv:2008.05756, 2020.

[26] S. Narkhede, «Understanding Confusion Matrix,» Towards Data Science, τόμ. 180, αρ. 1, pp. 1--12, 2018.

[27] V. Zeljko, «Classification Model Evaluation Metrics,» International Journal of Advanced Computer Science and Applications, τόμ. 12, αρ. 6, pp. 599--606, 2021.

[28] R. Vidiyala, Performance metrics for classification machine learning problems, 2020.

[29] M. Hossin και S. Md Nasir , «A Review on Evaluation Metrics for Data Classification Evaluations,» International journal of data mining & knowledge management process, τόμ. 5, αρ. 2, p. 1, 2015.

[30] A. Tharwat, «Classification assessment methods,» Applied computing and informatics, τόμ. 17, αρ. 1, pp. 168--192, 2020.

[31] . S. Goyal, «Evaluation Metrics for Classification Models.,» Published in Analytics Vidhva, 2021.

[32] I. M. De Diego, R. Ana R. , F. Rubén R. , . N. Jorge και M. Javier M. , «General Performance Score for classification problems,» Applied Intelligence, τόμ. 52, αρ. 10, pp. 12049-12063, 2022.

[33] T. Srivastava, «Important model evaluation metrics for machine learning everyone should know,» Commonly Used Machine Learning Algorithms: Data Science, 2019.

[34] Y. Sasaki, «The truth of the F-measure,» Teach tutor mater, τόμ. 1, αρ. 5, pp. 1--5, 2007.

[35] B. Shmueli, «Multi-class metrics made simple,,» σε The F1-score, 2019.

[36] O. M. Adesanya, S. O. Abam και I. O. Muraina, Data analytics evaluation metrics essentials: Measuring model performance in cassification and regression, 2022.

[37] D. J. Hand και R. J. Till , «A simple generalization of the area under the ROC curve to multiple class classification problems,» Machine learning, αρ. 45, pp. 171--186, 2001.

[38] J. Huang και C. X. Ling, «Using AUC and accuracy in evaluating learning algorithms,» IEEE Transactions on knowledge and Data Engineering, τόμ. 17, αρ. 3, pp. 299--310, 2005.

[39] A. Mishra, «Metrics to evaluate your machine learning algorithm,» Towards data science, pp. 1--8, 2018.

[40] N. Thai-Nghe, Z. Gantner και L. Schmidt-Thieme, «A new evaluation measure for learning from imbalanced data,» σε The 2011 International Joint Conference on Neural Networks, IEEE, 2011, pp. 537--542.

[41] D. J. Hand και C. Anagnostopoulos, «Notes on the H-measure of classifier performance,» Advances in Data Analysis and Classification, τόμ. 17, αρ. 1, pp. 109--124, 2023.

[42] K. K. Lai, L. Yu, S. Wang και L. Zhou, «Credit risk analysis using a reliability-based neural network ensemble model.,» σε International Conference on Artificial Neural Networks, Springer, 2006, pp. 682--690.

[43] P. Pławiak, M. Abdar και A. U. Rajendra , «Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring,» Applied Soft Computing, τόμ. 84, p. 105740, 2019.

[44] A. Khashman, «Credit risk evaluation using neural networks: Emotional versus conventional models,» Applied Soft Computing, 2011.

[45] S. Oreski, D. Oreski και G. Oreski, «Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment,» Expert systems with applications, τόμ. 39, αρ. 16, 2012.

[46] S. Finlay, «Multiple classifier architectures and their application to credit risk assessment,» European Journal of Operational Research, 2011.

[47] C.-L. Huang, M.-C. Chen και . C. Wang, «Credit scoring with a data mining approach based on support vector machines,» Expert systems with applications, 2007.

[48] G. Zhang, M. Y. Hu, B. E. Patuwo και D. C. Indro, «Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis,» European journal of operational research, 1999.

[49] T. Saito και R. Marc , «Basic evaluation measures from the confusion matrix,» 2017. [Ηλεκτρονικό]. Available: https://classeval. wordpress. com/introduction/basic-evaluation-measures.

[50] D. Leykam και D. G. Angelakis, «Topological data analysis and machine learning,» 2023.

[51] G. Carlsson, «Topology and data,» Bulletin of the American Mathematical Society, 2009.

[52] C. F. Loughrey, P. Fitzpatrick, N. Orr και A. Jurek-Loughrey, «The topology of data: Opportunities for cancer research,» Bioinformatics, 2021.

[53] E. R. Love, «Machine Learning with Topological Data Analysis,» 2021.

[54] D. J. Hand, «Measuring classifier performance: a coherent alternative to the area under the ROC curve,» Machine learning, τόμ. 77, αρ. 1, pp. 103--123, 2009.

ΠΑΡΑΡΤΗΜΑ

Πίνακες

Πίνακας 3 : Μαχ τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Α

ΑΚΤΙΝΑ	Accuracy	H	Precision	Recall	F	AUC
0.4	0.9103448	0.7883860	0.9433962	0.9402985	0.9064748	0.9334099
0.41	0.9103448	0.7866221	0.9423077	0.9402985	0.9064748	0.9332185
0.42	0.9103448	0.7866221	0.9423077	0.9402985	0.9064748	0.9313050
0.43	0.9103448	0.7747626	0.9423077	0.9402985	0.9064748	0.9453693
0.44	0.9172414	0.7770296	0.9423077	0.9402985	0.9076923	0.9372369
0.45	0.9103448	0.7712209	0.9423077	0.9402985	0.9064748	0.9429774
0.46	0.9103448	0.7712209	0.9444444	0.9402985	0.9064748	0.9421163
0.47	0.9103448	0.7632144	0.9444444	0.9402985	0.9064748	0.9385763
0.48	0.9103448	0.7632144	0.9444444	0.9402985	0.9064748	0.9381936
0.49	0.9103448	0.7611066	0.9444444	0.9402985	0.9022556	0.9352277
0.5	0.9241379	0.7715021	0.9444444	0.9402985	0.9160305	0.9435515
0.51	0.9241379	0.7723046	0.9482759	0.9402985	0.9160305	0.9416380
0.52	0.9241379	0.7833068	0.9516129	0.9402985	0.9147287	0.9482396
0.53	0.9172414	0.7798123	0.9636364	0.9402985	0.9076923	0.9504401
0.54	0.9172414	0.7822590	0.9636364	0.9402985	0.9062500	0.9467088
0.55	0.9172414	0.8076371	0.9629630	0.9402985	0.9090909	0.9487179
0.56	0.9172414	0.7888848	0.9454545	0.9402985	0.9090909	0.9487179
0.57	0.9172414	0.8075748	0.9454545	0.9402985	0.9090909	0.9478569
0.58	0.9103448	0.7751823	0.9636364	0.9402985	0.9037037	0.9562763
0.59	0.9172414	0.7993601	0.9629630	0.9402985	0.9117647	0.9703406
0.6	0.9172414	0.8152863	0.9500000	0.9402985	0.9104478	0.9710103
0.61	0.9103448	0.8222391	0.9500000	0.9402985	0.9022556	0.9690011
0.62	0.9172414	0.8133919	0.9310345	0.9402985	0.9090909	0.9684271
0.63	0.9103448	0.7901906	0.9354839	0.9253731	0.9051095	0.9518752
0.64	0.9034483	0.7916236	0.9344262	0.9402985	0.8985507	0.9701493
0.65	0.9241379	0.8127756	0.9333333	0.9402985	0.9197080	0.9680444
0.66	0.9241379	0.8200540	0.9354839	0.9402985	0.9197080	0.9672790
0.67	0.9172414	0.8113171	0.9354839	0.9552239	0.9142857	0.9612514
0.68	0.9103448	0.8298458	0.9354839	0.9552239	0.9064748	0.9602947
0.69	0.9103448	0.8102514	0.9354839	0.9552239	0.8992248	0.9659395
0.7	0.9103448	0.7918514	0.9354839	0.9552239	0.9037037	0.9625909
0.71	0.9034483	0.7940895	0.9344262	0.9552239	0.8955224	0.9648871
0.72	0.9034483	0.7798399	0.9333333	0.9253731	0.8955224	0.9646958
0.73	0.8965517	0.7733582	0.9166667	0.9253731	0.8888889	0.9634520
0.74	0.8965517	0.7545991	0.9180328	0.9402985	0.8888889	0.9564677
0.75	0.8965517	0.7506000	0.9032258	0.9552239	0.8888889	0.9607731
0.76	0.8965517	0.7633420	0.9062500	0.9701493	0.8965517	0.9648871

0.77	0.9034483	0.7642915	0.9062500	0.9402985	0.8955224	0.9655568
0.78	0.9103448	0.7832641	0.9076923	0.9552239	0.9037037	0.9673747
0.79	0.9034483	0.7834037	0.9076923	0.9552239	0.9014085	0.9660352
0.8	0.9034483	0.7828446	0.9076923	0.9701493	0.9014085	0.9664179
0.81	0.9034483	0.7776554	0.9076923	0.9552239	0.9014085	0.9667049
0.82	0.9034483	0.7737484	0.9206349	0.9552239	0.9014085	0.9689055
0.83	0.9103448	0.7747360	0.9206349	0.9552239	0.9078014	0.9686184
0.84	0.9103448	0.7793771	0.9076923	0.9552239	0.9078014	0.9686184
0.85	0.9103448	0.7835778	0.9076923	0.9552239	0.9078014	0.9681401
0.86	0.9103448	0.7824775	0.9062500	0.9552239	0.9078014	0.9700536
0.87	0.9103448	0.7778831	0.9062500	0.9552239	0.9078014	0.9671833
0.88	0.9103448	0.7733103	0.9076923	0.9552239	0.9078014	0.9678530
0.89	0.9172414	0.7739550	0.9104478	0.9552239	0.9104478	0.9694795
0.9	0.9172414	0.7810715	0.9104478	0.9552239	0.9104478	0.9603904
0.91	0.9103448	0.7808533	0.9076923	0.9552239	0.9037037	0.9609644
0.92	0.9103448	0.7808387	0.9076923	0.9552239	0.9037037	0.9601990
0.93	0.9103448	0.7730676	0.9076923	0.9552239	0.9078014	0.9599120
0.94	0.9172414	0.7734029	0.9104478	0.9701493	0.9104478	0.9607731
0.95	0.9103448	0.7776296	0.9090909	0.9552239	0.9037037	0.9610601
0.96	0.9103448	0.7767215	0.9076923	0.9552239	0.9037037	0.9577114
0.97	0.9103448	0.7767017	0.9090909	0.9552239	0.9037037	0.9581898
0.98	0.9172414	0.7767017	0.9076923	0.9701493	0.9117647	0.9572331
0.99	0.9172414	0.7771366	0.9062500	0.9701493	0.9117647	0.9624952
1	0.9172414	0.7767887	0.9062500	0.9701493	0.9117647	0.9607731
1.01	0.9172414	0.7769632	0.9062500	0.9701493	0.9117647	0.9623039
1.02	0.9172414	0.7836406	0.9062500	0.9701493	0.9130435	0.9631649
1.03	0.9172414	0.7831431	0.9076923	0.9850746	0.9130435	0.9604860
1.04	0.9172414	0.7701831	0.8955224	0.9701493	0.9130435	0.9570417
1.05	0.9241379	0.7898206	0.9000000	0.9701493	0.9197080	0.9573287
1.06	0.9172414	0.7642565	0.8955224	0.9850746	0.9130435	0.9514925
1.07	0.9172414	0.7602027	0.9076923	0.9701493	0.9154930	0.9509185
1.08	0.9103448	0.7565053	0.9076923	0.9701493	0.9090909	0.9523536
1.09	0.9103448	0.7537492	0.9090909	0.9701493	0.9090909	0.9643131
1.1	0.9172414	0.7580417	0.9090909	0.9701493	0.9154930	0.9653655
1.11	0.9103448	0.7530615	0.9090909	0.9850746	0.9103448	0.9625909
1.12	0.9103448	0.7466571	0.9090909	0.9552239	0.9037037	0.9593379
1.13	0.9172414	0.7530163	0.9104478	0.9552239	0.9104478	0.9484309
1.14	0.9103448	0.7900839	0.9152542	0.9402985	0.9037037	0.9509185
1.15	0.9172414	0.7786590	0.9259259	0.9402985	0.9090909	0.9552239
1.16	0.9172414	0.7773606	0.9230769	0.9701493	0.9090909	0.9523536
1.17	0.9034483	0.7647160	0.9818182	0.9402985	0.9000000	0.9475698
1.18	0.9103448	0.7575702	0.9818182	0.9402985	0.9037037	0.9437428
1.19	0.9103448	0.7682873	0.9821429	0.9402985	0.9064748	0.9550325
1.2	0.9241379	0.7746449	0.9821429	0.9253731	0.9172932	0.9485266

1.21	0.9310345	0.7950770	0.9821429	0.9253731	0.9253731	0.9514925
1.22	0.9310345	0.7949728	10.000.000	0.9253731	0.9193548	0.9570417
1.23	0.9172414	0.7774683	0.9423077	0.9552239	0.9090909	0.9528320
1.24	0.9241379	0.7681059	0.9607843	0.9701493	0.9185185	0.9500574
1.25	0.9241379	0.7771530	0.9666667	0.9552239	0.9185185	0.9581898
1.26	0.9241379	0.7806052	0.9615385	0.9253731	0.9172932	0.9595293
1.27	0.9172414	0.7894637	0.9600000	0.9253731	0.9104478	0.9610601
1.28	0.9103448	0.7838906	0.9607843	0.9402985	0.9051095	0.9586682
1.29	0.9172414	0.7867188	0.9354839	0.9253731	0.9104478	0.9601033
1.3	0.9379310	0.8174952	0.9677419	0.9552239	0.9302326	0.9638347
1.31	0.9379310	0.8153564	0.9677419	0.9552239	0.9302326	0.9623995
1.32	0.9241379	0.7911750	0.9516129	0.9552239	0.9147287	0.9546498
1.33	0.9241379	0.7872049	0.9516129	0.9552239	0.9147287	0.9602947
1.34	0.9103448	0.7799412	0.9500000	0.9552239	0.9064748	0.9480482
1.35	0.9172414	0.7734664	10.000.000	0.9402985	0.9130435	0.9518752
1.36	0.9172414	0.7800034	0.9814815	0.9701493	0.9142857	0.9517796
1.37	0.9310345	0.8025702	0.9811321	0.9552239	0.9264706	0.9516839
1.38	0.9172414	0.7929549	0.9636364	0.9701493	0.9117647	0.9529277
1.39	0.9172414	0.7644955	0.9444444	0.9850746	0.9117647	0.9535974
1.4	0.9172414	0.7722006	0.9454545	0.9850746	0.9166667	0.9532147
1.41	0.9172414	0.8306251	0.9615385	0.9701493	0.9062500	0.9546498
1.42	0.9103448	0.7893438	0.9615385	0.9552239	0.9078014	0.9543628
1.43	0.9103448	0.7861514	0.9423077	0.9701493	0.9078014	0.9514925
1.44	0.9034483	0.7842068	0.9400000	0.9701493	0.9027778	0.9513012
1.45	0.8965517	0.7430646	0.9574468	0.9402985	0.8905109	0.9468044
1.46	0.9034483	0.7632619	0.9591837	0.9701493	0.8970588	0.9557023
1.47	0.9034483	0.7392852	0.9600000	0.9701493	0.8970588	0.9549369
1.48	0.8896552	0.7198390	0.9583333	0.9552239	0.8888889	0.9513012
1.49	0.8965517	0.7388128	0.9583333	0.9402985	0.8888889	0.9442212
1.5	0.8965517	0.7307853	0.9555556	0.9402985	0.8888889	0.9440299
1.51	0.8965517	0.7296117	0.9800000	0.9552239	0.8951049	0.9523536
1.52	0.8896552	0.7483078	10.000.000	0.8805970	0.8805970	0.9382893
1.53	0.8827586	0.7580223	10.000.000	0.8656716	0.8547009	0.9567547
1.54	0.8620690	0.7365056	10.000.000	0.8656716	0.8437500	0.9329315
1.55	0.8620690	0.7428486	0.9782609	0.8656716	0.8372093	0.9350364
1.56	0.8551724	0.7325493	10.000.000	0.8507463	0.8307692	0.9341753
1.57	0.8551724	0.7102227	0.9761905	0.8507463	0.8321168	0.9259472
1.58	0.8413793	0.6796956	0.9729730	0.7761194	0.8099174	0.9208764
1.59	0.8206897	0.6710690	0.9743590	0.7611940	0.7906977	0.9146575
1.6	0.8551724	0.6830389	0.9761905	0.7462687	0.8235294	0.9200153
1.61	0.8482759	0.6937212	0.9772727	0.7164179	0.8135593	0.9157099
1.62	0.8344828	0.6921488	0.9767442	0.6865672	0.7931034	0.9085343
1.63	0.8413793	0.7327565	10.000.000	0.6716418	0.7927928	0.9228856
1.64	0.8206897	0.7113981	10.000.000	0.6417910	0.7636364	0.9185802

1.65	0.8068966	0.7148005	10.000.000	0.6119403	0.7454545	0.9159969
1.66	0.7931034	0.6750002	10.000.000	0.5970149	0.7115385	0.9084386
1.67	0.7862069	0.6892349	10.000.000	0.5970149	0.7142857	0.9064294
1.68	0.7862069	0.7292436	10.000.000	0.5820896	0.7102804	0.9272866
1.69	0.7793103	0.7289576	10.000.000	0.5820896	0.7090909	0.9178148
1.7	0.7793103	0.7512470	10.000.000	0.5522388	0.6981132	0.9213548
1.71	0.7724138	0.7323512	10.000.000	0.5522388	0.6857143	0.9133180
1.72	0.7586207	0.7295243	10.000.000	0.5373134	0.6666667	0.9111175
1.73	0.7586207	0.6798068	10.000.000	0.5373134	0.6666667	0.9075775
1.74	0.7655172	0.6935704	10.000.000	0.5522388	0.6730769	0.9083429
1.75	0.7655172	0.6378742	10.000.000	0.5522388	0.6730769	0.9083429
1.76	0.7586207	0.6117487	10.000.000	0.5671642	0.6666667	0.9095867
1.77	0.7586207	0.6138457	10.000.000	0.5671642	0.6666667	0.8952354
1.78	0.7517241	0.6123389	10.000.000	0.5522388	0.6607143	0.8926521
1.79	0.7379310	0.6426956	10.000.000	0.5522388	0.6607143	0.9019326

Πίνακας 4: Μαχ τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Β

AKTINA	Accuracy	H	Precision	Recall	F	AUC
0.4	0.8068966	0.5394244	0.8153846	0.6865672	0.7666667	0.8360122
0.41	0.8000000	0.5394244	0.8153846	0.7164179	0.7680000	0.8322809
0.42	0.8000000	0.5051021	0.7794118	0.7164179	0.7786260	0.8272101
0.43	0.8068966	0.5058073	0.7826087	0.7164179	0.7812500	0.8287409
0.44	0.8000000	0.5172012	0.7826087	0.7313433	0.7751938	0.8289323
0.45	0.8000000	0.5090721	0.7826087	0.7313433	0.7751938	0.8349598
0.46	0.8000000	0.5030847	0.7857143	0.7313433	0.7751938	0.8332377
0.47	0.8068966	0.5114671	0.7972973	0.7462687	0.7846154	0.8354382
0.48	0.8068966	0.5114671	0.7972973	0.7462687	0.7846154	0.8426139
0.49	0.7862069	0.5118845	0.7727273	0.7462687	0.7669173	0.8400306
0.5	0.7793103	0.5045002	0.7777778	0.7313433	0.7538462	0.8379258
0.51	0.7793103	0.5187431	0.7777778	0.7313433	0.7538462	0.8423268
0.52	0.7793103	0.5174608	0.7777778	0.7313433	0.7538462	0.8413701
0.53	0.7793103	0.5151550	0.7777778	0.7313433	0.7538462	0.8420398
0.54	0.7793103	0.5270056	0.7777778	0.7313433	0.7538462	0.8407003
0.55	0.7793103	0.5266770	0.7777778	0.7313433	0.7538462	0.8455798
0.56	0.7793103	0.5219468	0.7777778	0.7014925	0.7538462	0.8472063
0.57	0.7793103	0.5264099	0.7777778	0.7014925	0.7538462	0.8460582
0.58	0.7793103	0.5215142	0.7777778	0.7014925	0.7538462	0.8450057
0.59	0.7931034	0.5433961	0.7846154	0.7164179	0.7727273	0.8508419
0.6	0.8137931	0.5546274	0.8153846	0.7014925	0.7768595	0.8589744
0.61	0.8137931	0.5502421	0.8026316	0.6865672	0.7731092	0.8653846
0.62	0.8137931	0.5532118	0.8026316	0.7014925	0.7768595	0.8680635
0.63	0.8000000	0.5551618	0.7878788	0.7313433	0.7819549	0.8686376
0.64	0.7931034	0.5450806	0.7846154	0.7611940	0.7727273	0.8672981
0.65	0.7862069	0.5527807	0.7866667	0.7313433	0.7596899	0.8687333

0.66	0.7862069	0.5331273	0.7866667	0.7313433	0.7596899	0.8661500
0.67	0.7862069	0.5387640	0.7866667	0.7313433	0.7596899	0.8652889
0.68	0.8000000	0.5403776	0.7866667	0.7313433	0.7716535	0.8681592
0.69	0.7862069	0.5341021	0.7837838	0.7313433	0.7596899	0.8657673
0.7	0.7862069	0.5292266	0.7837838	0.7313433	0.7596899	0.8652889
0.71	0.7931034	0.5372563	0.7837838	0.7462687	0.7692308	0.8662457
0.72	0.7931034	0.5370919	0.7922078	0.7462687	0.7692308	0.8680635
0.73	0.7862069	0.5437843	0.7750000	0.7313433	0.7633588	0.8774397
0.74	0.7862069	0.5481472	0.7721519	0.7164179	0.7596899	0.8754305
0.75	0.7862069	0.5496378	0.7721519	0.7014925	0.7596899	0.8718905
0.76	0.7862069	0.5504765	0.7721519	0.7313433	0.7596899	0.8611749
0.77	0.7862069	0.5481950	0.7820513	0.7313433	0.7596899	0.8739954
0.78	0.7862069	0.5500094	0.7750000	0.7313433	0.7596899	0.8713165
0.79	0.7862069	0.5444365	0.7750000	0.7313433	0.7596899	0.8677765
0.8	0.7862069	0.5439705	0.7750000	0.7313433	0.7596899	0.8667241
0.81	0.7862069	0.5455364	0.7750000	0.7313433	0.7596899	0.8682549
0.82	0.7793103	0.5429894	0.7750000	0.7313433	0.7538462	0.8661500
0.83	0.7724138	0.5423328	0.7656250	0.7164179	0.7480916	0.8736127
0.84	0.7931034	0.5427159	0.7761194	0.7164179	0.7751938	0.8758132
0.85	0.7931034	0.5411468	0.7761194	0.7164179	0.7751938	0.8710295
0.86	0.7931034	0.5411468	0.7761194	0.7164179	0.7751938	0.8714122
0.87	0.8000000	0.5414842	0.7714286	0.7164179	0.7751938	0.8707424
0.88	0.8068966	0.5411468	0.7746479	0.7164179	0.7804878	0.8697857
0.89	0.8137931	0.5524140	0.7763158	0.7164179	0.7804878	0.8715078
0.9	0.8137931	0.5403511	0.7763158	0.7164179	0.7804878	0.8668197
0.91	0.8137931	0.5426181	0.7763158	0.7164179	0.7804878	0.8643322
0.92	0.8137931	0.5424899	0.7763158	0.7164179	0.7804878	0.8636625
0.93	0.8068966	0.5424766	0.7763158	0.7164179	0.7804878	0.8541906
0.94	0.8068966	0.5432222	0.7763158	0.7164179	0.7804878	0.8589744
0.95	0.8000000	0.5399704	0.7750000	0.7164179	0.7751938	0.8564868
0.96	0.8068966	0.5399704	0.7750000	0.7164179	0.7804878	0.8549560
0.97	0.8068966	0.5399704	0.7750000	0.7164179	0.7804878	0.8521814
0.98	0.8137931	0.5399704	0.7750000	0.7164179	0.7804878	0.8526598
0.99	0.8137931	0.5404282	0.7750000	0.7164179	0.7804878	0.8535209
1	0.8137931	0.5404282	0.7750000	0.7164179	0.7804878	0.8551473
1.01	0.8137931	0.5398324	0.7750000	0.7164179	0.7804878	0.8535209
1.02	0.8137931	0.5292349	0.7750000	0.7313433	0.7840000	0.8588787
1.03	0.8137931	0.5403075	0.7750000	0.7313433	0.7840000	0.8523728
1.04	0.8137931	0.5570100	0.7682927	0.7313433	0.7840000	0.8692116
1.05	0.8206897	0.5552727	0.7682927	0.7462687	0.7936508	0.8688289
1.06	0.8275862	0.5514092	0.7750000	0.7611940	0.8062016	0.8691160
1.07	0.8137931	0.5479843	0.7692308	0.7611940	0.8029197	0.8592614
1.08	0.8068966	0.5190866	0.7662338	0.7313433	0.7971014	0.8528511
1.09	0.8068966	0.5279704	0.7662338	0.7164179	0.7933884	0.8621316

1.1	0.8137931	0.5405415	0.7500000	0.7164179	0.7933884	0.8602181
1.11	0.8137931	0.5627189	0.7625000	0.7014925	0.7833333	0.8717949
1.12	0.7862069	0.5515438	0.7571429	0.7014925	0.7737226	0.8715078
1.13	0.8000000	0.5325494	0.7654321	0.7313433	0.7878788	0.8668197
1.14	0.8000000	0.5252892	0.7435897	0.7014925	0.7899160	0.8621316
1.15	0.7793103	0.5059510	0.7464789	0.7164179	0.7681159	0.8514160
1.16	0.8000000	0.5146849	0.7714286	0.7014925	0.7846154	0.8499809
1.17	0.8000000	0.5307213	0.7922078	0.7014925	0.7716535	0.8490241
1.18	0.8000000	0.5175090	0.7746479	0.6716418	0.7642276	0.8556257
1.19	0.8000000	0.5080411	0.7837838	0.6567164	0.7521368	0.8522771
1.2	0.8000000	0.5338243	0.7887324	0.6716418	0.7563025	0.8628014
1.21	0.8000000	0.5285125	0.7733333	0.6716418	0.7563025	0.8635668
1.22	0.8000000	0.4881730	0.7500000	0.6716418	0.7627119	0.8607922
1.23	0.7655172	0.4844095	0.7200000	0.6865672	0.7605634	0.8501722
1.24	0.7655172	0.4436636	0.7260274	0.7164179	0.7571429	0.8388825
1.25	0.7793103	0.4542913	0.7307692	0.7313433	0.7714286	0.8465365
1.26	0.7862069	0.4370124	0.7647059	0.7313433	0.7703704	0.8137199
1.27	0.7862069	0.4588633	0.7714286	0.7014925	0.7633588	0.8148680
1.28	0.7724138	0.4266136	0.7656250	0.7313433	0.7480916	0.8219480
1.29	0.7793103	0.4545685	0.7536232	0.7164179	0.7647059	0.8242442
1.3	0.8206897	0.5331793	0.7530864	0.7164179	0.7868852	0.8695943
1.31	0.8068966	0.5183959	0.7500000	0.7164179	0.7840000	0.8413701
1.32	0.8137931	0.4811677	0.7594937	0.7611940	0.8057554	0.8324723
1.33	0.7862069	0.4870018	0.7777778	0.6268657	0.7304348	0.8330463
1.34	0.7862069	0.4886479	0.7826087	0.6417910	0.7350427	0.8383085
1.35	0.7862069	0.5051752	0.7792208	0.6417910	0.7350427	0.8134328
1.36	0.7862069	0.5181111	0.7307692	0.6417910	0.7350427	0.8122847
1.37	0.7862069	0.4803211	0.7500000	0.6567164	0.7521368	0.8052047
1.38	0.7724138	0.4513648	0.7297297	0.6567164	0.7457627	0.7985075
1.39	0.7655172	0.4414152	0.7260274	0.6567164	0.7521368	0.7824340
1.4	0.7655172	0.4452839	0.7260274	0.6865672	0.7540984	0.7835821
1.41	0.7655172	0.4485629	0.7260274	0.6567164	0.7333333	0.8176426
1.42	0.7517241	0.4679926	0.7183099	0.6567164	0.7333333	0.8147723
1.43	0.7655172	0.4787844	0.7323944	0.6417910	0.7413793	0.8280712
1.44	0.7724138	0.4483337	0.7464789	0.6417910	0.7226891	0.8307501
1.45	0.7724138	0.4463519	0.7428571	0.6567164	0.7540984	0.8285496
1.46	0.7724138	0.4559573	0.7500000	0.6417910	0.7555556	0.8264447
1.47	0.7724138	0.4115262	0.7500000	0.6268657	0.7500000	0.7940107
1.48	0.7724138	0.3985811	0.7536232	0.6865672	0.7360000	0.7921929
1.49	0.7517241	0.4275013	0.7571429	0.6567164	0.7096774	0.8217566
1.5	0.7586207	0.4143862	0.7567568	0.6417910	0.7154472	0.8173555
1.51	0.7517241	0.4089640	0.7500000	0.6119403	0.6949153	0.8221393
1.52	0.7379310	0.3553170	0.7195122	0.5223881	0.6779661	0.7792767
1.53	0.7448276	0.4188420	0.7534247	0.5522388	0.6837607	0.8084577

1.54	0.7103448	0.4244309	0.7532468	0.3731343	0.5434783	0.8110409
1.55	0.7172414	0.4307865	0.7733333	0.4477612	0.6060606	0.8163031
1.56	0.6758621	0.4616440	0.7808219	0.2985075	0.4597701	0.8285496
1.57	0.6275862	0.4278836	0.7741935	0.2238806	0.3571429	0.8297933
1.58	0.6206897	0.4220898	0.7384615	0.1940299	0.3209877	0.8173555
1.59	0.5793103	0.3940592	0.6666667	0.1791045	0.2823529	0.8040566
1.6	0.6482759	0.3472665	0.7543860	0.2686567	0.4137931	0.7904707
1.61	0.6413793	0.3560850	0.7906977	0.2537313	0.3953488	0.7995599
1.62	0.6689655	0.3630546	0.7272727	0.2985075	0.4545455	0.8016648
1.63	0.6689655	0.3013486	0.7358491	0.2985075	0.4545455	0.7572713
1.64	0.6482759	0.3044988	0.7291667	0.2835821	0.4269663	0.7490432
1.65	0.6620690	0.2993571	0.7708333	0.2686567	0.4235294	0.7663605
1.66	0.6137931	0.3381125	0.7446809	0.2089552	0.3333333	0.7640643
1.67	0.6482759	0.3116041	0.7272727	0.3432836	0.4842105	0.7630119
1.68	0.6620690	0.3013813	0.7209302	0.3134328	0.4719101	0.7455989
1.69	0.6000000	0.2891808	0.7209302	0.1641791	0.2750000	0.7355530
1.7	0.5931034	0.2771013	0.7250000	0.1492537	0.2531646	0.7266552
1.71	0.6413793	0.2651489	0.7250000	0.2537313	0.4047619	0.6769996
1.72	0.6275862	0.2590095	0.7073171	0.3134328	0.4375000	0.6426521
1.73	0.6275862	0.1846643	0.7179487	0.2537313	0.4047619	0.6168197
1.74	0.6275862	0.1825594	0.7241379	0.2388060	0.3855422	0.6150976
1.75	0.6344828	0.1800780	0.7368421	0.2835821	0.4175824	0.6254305
1.76	0.6206897	0.1836807	0.6923077	0.2537313	0.3953488	0.6275354
1.77	0.6344828	0.1757416	0.6923077	0.2537313	0.3953488	0.6358592
1.78	0.6137931	0.1797997	0.6923077	0.2388060	0.3636364	0.6386338
1.79	0.5724138	0.1950666	0.6666667	0.1492537	0.2439024	0.6529851

Πίνακας 5: Μεαn τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Α

ΑΚΤΙΝΑ	Accuracy	H	Precision	Recall	F	AUC
0.4	0.8611034	0.6391575	0.8749344	0.8176119	0.8441401	0.8811596
0.41	0.8619310	0.6411279	0.8735933	0.8214925	0.8456116	0.8812017
0.42	0.8568276	0.6313650	0.8655593	0.8191045	0.8404923	0.8803310
0.43	0.8588966	0.6358611	0.8666189	0.8229851	0.8430437	0.8826139
0.44	0.8611034	0.6419194	0.8671214	0.8280597	0.8459630	0.8835343
0.45	0.8620690	0.6402214	0.8698598	0.8271642	0.8467341	0.8838366
0.46	0.8633103	0.6393866	0.8697723	0.8304478	0.8484383	0.8830406
0.47	0.8664828	0.6450395	0.8714827	0.8364179	0.8523255	0.8841638
0.48	0.8655172	0.6448383	0.8709346	0.8346269	0.8511707	0.8851454
0.49	0.8646897	0.6440200	0.8706662	0.8325373	0.8501351	0.8836338
0.5	0.8664828	0.6449039	0.8695327	0.8385075	0.8525561	0.8876196
0.51	0.8630345	0.6453531	0.8692380	0.8301493	0.8480279	0.8906238
0.52	0.8666207	0.6501465	0.8710131	0.8367164	0.8523536	0.8896958
0.53	0.8702069	0.6600730	0.8759404	0.8394030	0.8562062	0.8953100
0.54	0.8688276	0.6578084	0.8747019	0.8379104	0.8546551	0.8967413

0.55	0.8714483	0.6658667	0.8803681	0.8370149	0.8570582	0.8996192
0.56	0.8693793	0.6649048	0.8779001	0.8352239	0.8547152	0.9010180
0.57	0.8714483	0.6688448	0.8807524	0.8367164	0.8568692	0.9019786
0.58	0.8704828	0.6696454	0.8773454	0.8388060	0.8563214	0.9036280
0.59	0.8740690	0.6801844	0.8803420	0.8441791	0.8605352	0.9074665
0.6	0.8711724	0.6752365	0.8780954	0.8397015	0.8571247	0.9094604
0.61	0.8626207	0.6686238	0.8663434	0.8334328	0.8481534	0.9115557
0.62	0.8652414	0.6741422	0.8619135	0.8459701	0.8526146	0.9142346
0.63	0.8640000	0.6707576	0.8647019	0.8391045	0.8505169	0.9123861
0.64	0.8590345	0.6654981	0.8545239	0.8402985	0.8460650	0.9126119
0.65	0.8612414	0.6741112	0.8544714	0.8462687	0.8489870	0.9142097
0.66	0.8634483	0.6804429	0.8553708	0.8504478	0.8516208	0.9174895
0.67	0.8653793	0.6820966	0.8567680	0.8534328	0.8539077	0.9163012
0.68	0.8663448	0.6818246	0.8580892	0.8540299	0.8548515	0.9161060
0.69	0.8619310	0.6785160	0.8512144	0.8525373	0.8506359	0.9179812
0.7	0.8638621	0.6763079	0.8510781	0.8576119	0.8531867	0.9158285
0.71	0.8608276	0.6729220	0.8491459	0.8528358	0.8495453	0.9165385
0.72	0.8615172	0.6689959	0.8502395	0.8528358	0.8502381	0.9165365
0.73	0.8575172	0.6691837	0.8436144	0.8519403	0.8463351	0.9175775
0.74	0.8550345	0.6656287	0.8434399	0.8459701	0.8430963	0.9139399
0.75	0.8528276	0.6606367	0.8390457	0.8462687	0.8409350	0.9146977
0.76	0.8543448	0.6619306	0.8369639	0.8531343	0.8432983	0.9152583
0.77	0.8571034	0.6675147	0.8390340	0.8573134	0.8465736	0.9168102
0.78	0.8595862	0.6728287	0.8373920	0.8665672	0.8502986	0.9186586
0.79	0.8587586	0.6729833	0.8373676	0.8644776	0.8492097	0.9178626
0.8	0.8587586	0.6723559	0.8361470	0.8662687	0.8495245	0.9169384
0.81	0.8586207	0.6717818	0.8360833	0.8659701	0.8492416	0.9162438
0.82	0.8569655	0.6655641	0.8349179	0.8632836	0.8474065	0.9133104
0.83	0.8555862	0.6666523	0.8341955	0.8608955	0.8456794	0.9143743
0.84	0.8572414	0.6677406	0.8354149	0.8635821	0.8475647	0.9159414
0.85	0.8575172	0.6695294	0.8363145	0.8629851	0.8477244	0.9158152
0.86	0.8561379	0.6674435	0.8345879	0.8620896	0.8463887	0.9149158
0.87	0.8565517	0.6673935	0.8340996	0.8638806	0.8470282	0.9154133
0.88	0.8579310	0.6670485	0.8340457	0.8677612	0.8488393	0.9147742
0.89	0.8598621	0.6675073	0.8346077	0.8722388	0.8512438	0.9150402
0.9	0.8612414	0.6690501	0.8354589	0.8746269	0.8528105	0.9154401
0.91	0.8645517	0.6702254	0.8372379	0.8805970	0.8568376	0.9165710
0.92	0.8630345	0.6696473	0.8368308	0.8770149	0.8548145	0.9161041
0.93	0.8628966	0.6686431	0.8364584	0.8773134	0.8547105	0.9149464
0.94	0.8637241	0.6713482	0.8359348	0.8802985	0.8558971	0.9151225
0.95	0.8600000	0.6705852	0.8341253	0.8734328	0.8514778	0.9146116
0.96	0.8602759	0.6706209	0.8336887	0.8749254	0.8520565	0.9145235
0.97	0.8602759	0.6677133	0.8330343	0.8758209	0.8521479	0.9134347
0.98	0.8649655	0.6697524	0.8344730	0.8859701	0.8579576	0.9138270

0.99	0.8652414	0.6693288	0.8353696	0.8853731	0.8581194	0.9144508
1	0.8652414	0.6691736	0.8353696	0.8853731	0.8581194	0.9143877
1.01	0.8660690	0.6690522	0.8354792	0.8871642	0.8591750	0.9146403
1.02	0.8648276	0.6684219	0.8351879	0.8844776	0.8576763	0.9138959
1.03	0.8630345	0.6681755	0.8302531	0.8874627	0.8564569	0.9138002
1.04	0.8616552	0.6620049	0.8264935	0.8898507	0.8557083	0.9124531
1.05	0.8628966	0.6655199	0.8286051	0.8895522	0.8567620	0.9127937
1.06	0.8646897	0.6611287	0.8266836	0.8973134	0.8595267	0.9103636
1.07	0.8633103	0.6567900	0.8272906	0.8925373	0.8576486	0.9104956
1.08	0.8615172	0.6554421	0.8251476	0.8913433	0.8559324	0.9098354
1.09	0.8590345	0.6533977	0.8279214	0.8797015	0.8519281	0.9103483
1.1	0.8611034	0.6562442	0.8289231	0.8838806	0.8545033	0.9104038
1.11	0.8611034	0.6574315	0.8300076	0.8817910	0.8540087	0.9116341
1.12	0.8575172	0.6497549	0.8292116	0.8734328	0.8496968	0.9101493
1.13	0.8565517	0.6494938	0.8298136	0.8701493	0.8483268	0.9084654
1.14	0.8537931	0.6462736	0.8331566	0.8585075	0.8441320	0.9094068
1.15	0.8521379	0.6439186	0.8372548	0.8474627	0.8408332	0.9079851
1.16	0.8529655	0.6463922	0.8444945	0.8402985	0.8401980	0.9069939
1.17	0.8544828	0.6457066	0.8515056	0.8334328	0.8404139	0.9052947
1.18	0.8522759	0.6461577	0.8474316	0.8340299	0.8384345	0.9057061
1.19	0.8551724	0.6456514	0.8484318	0.8394030	0.8419470	0.9067968
1.2	0.8561379	0.6548412	0.8544750	0.8328358	0.8418984	0.9105071
1.21	0.8572414	0.6530495	0.8528295	0.8382090	0.8438556	0.9114830
1.22	0.8526897	0.6423307	0.8536591	0.8265672	0.8378621	0.9098374
1.23	0.8517241	0.6399871	0.8472200	0.8325373	0.8379736	0.9055166
1.24	0.8529655	0.6387490	0.8477921	0.8352239	0.8397534	0.9035247
1.25	0.8533793	0.6389195	0.8478910	0.8355224	0.8403429	0.9056410
1.26	0.8620690	0.6516375	0.8555428	0.8474627	0.8500606	0.9070513
1.27	0.8642759	0.6597206	0.8583104	0.8489552	0.8520797	0.9067623
1.28	0.8584828	0.6511051	0.8522668	0.8420896	0.8457949	0.9067451
1.29	0.8593103	0.6545803	0.8522844	0.8441791	0.8471680	0.9096269
1.3	0.8692414	0.6681087	0.8611465	0.8588060	0.8584878	0.9127803
1.31	0.8681379	0.6644592	0.8627205	0.8534328	0.8564713	0.9109070
1.32	0.8702069	0.6623322	0.8697564	0.8492537	0.8580690	0.9098909
1.33	0.8644138	0.6481619	0.8610734	0.8453731	0.8516499	0.9048699
1.34	0.8638621	0.6526369	0.8621293	0.8420896	0.8505657	0.9052181
1.35	0.8612414	0.6484262	0.8586570	0.8405970	0.8480108	0.9056736
1.36	0.8560000	0.6472063	0.8448059	0.8471642	0.8438717	0.9017030
1.37	0.8606897	0.6555418	0.8546746	0.8447761	0.8475998	0.9077631
1.38	0.8586207	0.6492298	0.8519668	0.8441791	0.8456175	0.9076024
1.39	0.8571034	0.6490212	0.8509152	0.8417910	0.8436669	0.9077459
1.4	0.8579310	0.6484450	0.8519786	0.8423881	0.8446734	0.9082070
1.41	0.8555862	0.6464550	0.8583543	0.8292537	0.8404357	0.9084979
1.42	0.8514483	0.6426602	0.8531929	0.8262687	0.8361294	0.9062227

1.43	0.8503448	0.6479341	0.8483774	0.8295522	0.8354277	0.9091581
1.44	0.8430345	0.6434603	0.8390569	0.8223881	0.8273724	0.9077842
1.45	0.8391724	0.6254335	0.8398884	0.8116418	0.8223943	0.9020417
1.46	0.8423448	0.6267055	0.8405693	0.8197015	0.8263342	0.9020455
1.47	0.8394483	0.6231423	0.8442639	0.8062687	0.8212145	0.9008266
1.48	0.8423448	0.6184401	0.8457046	0.8104478	0.8250059	0.9000631
1.49	0.8376552	0.6116137	0.8403864	0.8053731	0.8198229	0.8940815
1.5	0.8307586	0.6050040	0.8452379	0.7808955	0.8086608	0.8924856
1.51	0.8257931	0.6000886	0.8464285	0.7656716	0.8003144	0.8888787
1.52	0.8100690	0.5872221	0.8475990	0.7256716	0.7774676	0.8801607
1.53	0.8158621	0.6006231	0.8618309	0.7226866	0.7820521	0.8825871
1.54	0.8077241	0.6105148	0.8858731	0.6770149	0.7618583	0.8889476
1.55	0.8063448	0.6024208	0.8886966	0.6713433	0.7586727	0.8892499
1.56	0.7977931	0.5905007	0.8777339	0.6611940	0.7459955	0.8867202
1.57	0.7536552	0.5825937	0.8793405	0.5453731	0.6585047	0.8846537
1.58	0.7555862	0.5694195	0.8734574	0.5576119	0.6662464	0.8788653
1.59	0.7442759	0.5582654	0.8630566	0.5340299	0.6453272	0.8748737
1.6	0.7634483	0.5440617	0.8658631	0.5808955	0.6894560	0.8710505
1.61	0.7612414	0.5401496	0.8731593	0.5674627	0.6827574	0.8694948
1.62	0.7506207	0.5318996	0.8742333	0.5408955	0.6626377	0.8661730
1.63	0.7558621	0.5274817	0.8849439	0.5453731	0.6701230	0.8590624
1.64	0.7431724	0.5186055	0.8818718	0.5155224	0.6462053	0.8566915
1.65	0.7347586	0.5203798	0.8900288	0.4889552	0.6272019	0.8581439
1.66	0.7209655	0.5121292	0.8927580	0.4540299	0.5945404	0.8544413
1.67	0.7289655	0.5006602	0.8941671	0.4713433	0.6141450	0.8507023
1.68	0.7246897	0.5148829	0.9035769	0.4537313	0.6013977	0.8561998
1.69	0.7150345	0.5044856	0.8982407	0.4343284	0.5812804	0.8512093
1.7	0.7114483	0.5073057	0.8914587	0.4307463	0.5763091	0.8512878
1.71	0.7071724	0.4965073	0.8884140	0.4229851	0.5690204	0.8427937
1.72	0.7081379	0.4837294	0.8855954	0.4271642	0.5730748	0.8379258
1.73	0.7060690	0.4770714	0.8921920	0.4188060	0.5656138	0.8368159
1.74	0.7078621	0.4777091	0.8986578	0.4194030	0.5669236	0.8393016
1.75	0.7084138	0.4699238	0.8888164	0.4241791	0.5716789	0.8347493
1.76	0.6968276	0.4481615	0.8361453	0.4319403	0.5649017	0.8176789
1.77	0.6960000	0.4484216	0.8383240	0.4286567	0.5624714	0.8148967
1.78	0.6881379	0.4359659	0.8320488	0.4125373	0.5459110	0.8058419
1.79	0.6753103	0.4328216	0.8343900	0.3740299	0.5085620	0.8069843

Πίνακας 6: Mean τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Β

ΑΚΤΙΝΑ	Accuracy	H	Precision	Recall	F	AUC
0.4	0.8000000	0.5839892	10.000.000	0.6119403	0.7339450	0.9042289
0.41	0.8068966	0.6015144	10.000.000	0.6716418	0.7627119	0.9090126
0.42	0.8068966	0.5787482	10.000.000	0.6865672	0.7666667	0.9085343
0.43	0.8206897	0.5860295	0.9743590	0.7014925	0.7768595	0.9123613

0.44	0.8413793	0.6406276	0.9756098	0.7761194	0.8188976	0.9140834
0.45	0.8413793	0.6519088	0.9761905	0.7761194	0.8188976	0.9326445
0.46	0.8206897	0.6436196	0.9761905	0.7462687	0.7874016	0.9322618
0.47	0.8413793	0.6838678	0.9545455	0.7761194	0.8099174	0.9382893
0.48	0.8413793	0.6985180	0.9555556	0.7910448	0.8153846	0.9415423
0.49	0.8551724	0.7073106	0.9555556	0.7910448	0.8264463	0.9395331
0.5	0.8620690	0.7025513	0.9555556	0.8208955	0.8360656	0.9421163
0.51	0.8620690	0.6831621	0.9777778	0.8208955	0.8396947	0.9371412
0.52	0.8689655	0.6998421	0.9600000	0.8358209	0.8484848	0.9452736
0.53	0.8689655	0.7587917	0.9607843	0.8507463	0.8571429	0.9531190
0.54	0.8758621	0.7319387	0.9591837	0.8358209	0.8548387	0.9549369
0.55	0.8758621	0.7335686	0.9782609	0.8358209	0.8548387	0.9532147
0.56	0.8827586	0.7186371	0.9629630	0.8507463	0.8595041	0.9468044
0.57	0.8965517	0.7364230	0.9642857	0.8507463	0.8780488	0.9508228
0.58	0.8827586	0.7334368	0.9574468	0.8507463	0.8617886	0.9579985
0.59	0.8896552	0.7477835	0.9636364	0.8656716	0.8688525	0.9615385
0.6	0.8827586	0.7692734	0.9622642	0.8507463	0.8682171	0.9729238
0.61	0.8896552	0.7641331	0.9615385	0.8358209	0.8750000	0.9690968
0.62	0.8965517	0.7852051	0.9464286	0.8805970	0.8872180	0.9752201
0.63	0.8965517	0.7734250	0.9433962	0.8507463	0.8837209	0.9602947
0.64	0.8896552	0.7903961	0.9411765	0.8507463	0.8769231	0.9659395
0.65	0.8827586	0.7733371	0.9464286	0.8656716	0.8721805	0.9646001
0.66	0.8827586	0.7794415	0.9464286	0.8656716	0.8721805	0.9641217
0.67	0.8965517	0.7902471	0.9622642	0.8805970	0.8872180	0.9573287
0.68	0.8965517	0.8099469	0.9622642	0.8656716	0.8854962	0.9571374
0.69	0.8896552	0.7910911	0.9622642	0.8805970	0.8787879	0.9580941
0.7	0.8965517	0.7727093	0.9310345	0.8805970	0.8872180	0.9552239
0.71	0.8896552	0.7761658	0.9433962	0.8656716	0.8787879	0.9610601
0.72	0.8896552	0.7621204	0.9454545	0.8656716	0.8787879	0.9590509
0.73	0.8896552	0.7443878	0.9298246	0.8805970	0.8805970	0.9601990
0.74	0.8896552	0.7402554	0.9298246	0.8955224	0.8805970	0.9511098
0.75	0.8896552	0.7373301	0.9000000	0.9104478	0.8805970	0.9592423
0.76	0.8827586	0.7476687	0.9016393	0.9253731	0.8794326	0.9616341
0.77	0.8827586	0.7632038	0.9016393	0.9104478	0.8740741	0.9646958
0.78	0.9034483	0.7645062	0.9032258	0.9253731	0.8985507	0.9654612
0.79	0.8965517	0.7745577	0.9032258	0.9253731	0.8920863	0.9651741
0.8	0.8896552	0.7739986	0.9032258	0.9253731	0.8857143	0.9655568
0.81	0.8965517	0.7708791	0.9062500	0.9253731	0.8857143	0.9646001
0.82	0.8896552	0.7604354	0.9180328	0.9253731	0.8857143	0.9650785
0.83	0.9034483	0.7609096	0.9180328	0.9402985	0.9000000	0.9663222
0.84	0.8965517	0.7656330	0.9047619	0.9253731	0.8920863	0.9655568
0.85	0.8965517	0.7753651	0.9193548	0.9253731	0.8920863	0.9676617
0.86	0.8965517	0.7817372	0.9062500	0.9253731	0.8920863	0.9695752
0.87	0.8965517	0.7789508	0.9062500	0.9253731	0.8920863	0.9667049

0.88	0.9034483	0.7748081	0.9076923	0.9402985	0.8939394	0.9673747
0.89	0.9172414	0.7697928	0.9230769	0.9402985	0.9090909	0.9688098
0.9	0.9172414	0.7813339	0.9230769	0.9402985	0.9090909	0.9663222
0.91	0.9172414	0.7810998	0.9104478	0.9402985	0.9104478	0.9662266
0.92	0.9172414	0.7810755	0.9104478	0.9402985	0.9104478	0.9662266
0.93	0.9172414	0.7750883	0.9104478	0.9552239	0.9104478	0.9662266
0.94	0.9241379	0.7779877	0.9242424	0.9701493	0.9172932	0.9672790
0.95	0.9034483	0.7803214	0.9166667	0.9552239	0.8955224	0.9675660
0.96	0.9034483	0.7800538	0.9056604	0.9552239	0.8955224	0.9642174
0.97	0.9034483	0.7800274	0.9074074	0.9552239	0.8955224	0.9644087
0.98	0.9172414	0.7668182	0.9137931	0.9552239	0.9117647	0.9616341
0.99	0.9172414	0.7817025	0.9137931	0.9552239	0.9117647	0.9695752
1	0.9103448	0.7779402	0.9137931	0.9552239	0.9037037	0.9668006
1.01	0.9103448	0.7671228	0.9062500	0.9253731	0.9037037	0.9649828
1.02	0.9103448	0.7737524	0.9206349	0.9253731	0.9051095	0.9677574
1.03	0.9103448	0.7671109	0.9047619	0.9701493	0.9090909	0.9653655
1.04	0.9103448	0.7588555	0.9047619	0.9402985	0.9051095	0.9636433
1.05	0.9172414	0.7704021	0.9032258	0.9552239	0.9142857	0.9543628
1.06	0.9172414	0.7610976	0.9047619	0.9701493	0.9154930	0.9522579
1.07	0.9241379	0.7759445	0.9206349	0.9552239	0.9208633	0.9483352
1.08	0.9172414	0.7619798	0.9193548	0.9552239	0.9142857	0.9525450
1.09	0.9241379	0.7930168	0.9076923	0.9701493	0.9219858	0.9633563
1.1	0.9310345	0.7937623	0.9193548	0.9701493	0.9285714	0.9624952
1.11	0.9241379	0.7808513	0.9076923	0.9850746	0.9230769	0.9591466
1.12	0.9103448	0.7866489	0.9090909	0.9402985	0.9022556	0.9560850
1.13	0.9103448	0.7849167	0.9206349	0.9402985	0.9022556	0.9478569
1.14	0.9034483	0.7569644	0.9107143	0.9402985	0.8955224	0.9474742
1.15	0.9103448	0.7586139	0.9245283	0.9402985	0.9064748	0.9505358
1.16	0.9103448	0.7581405	0.9218750	0.9701493	0.9090909	0.9460390
1.17	0.9034483	0.7595666	0.9818182	0.9104478	0.8854962	0.9430731
1.18	0.9034483	0.7575702	0.9818182	0.9253731	0.8872180	0.9410639
1.19	0.9034483	0.7525298	0.9818182	0.9104478	0.8852459	0.9498661
1.2	0.9103448	0.7740631	0.9821429	0.8955224	0.9007634	0.9439342
1.21	0.9103448	0.7805876	0.9821429	0.9104478	0.9007634	0.9499617
1.22	0.9241379	0.7910901	10.000.000	0.9104478	0.9105691	0.9556066
1.23	0.9103448	0.7504300	0.9411765	0.9402985	0.9064748	0.9528320
1.24	0.9241379	0.7667504	0.9600000	0.9552239	0.9185185	0.9490050
1.25	0.9241379	0.7738334	0.9666667	0.9402985	0.9185185	0.9530233
1.26	0.9103448	0.7592298	0.9607843	0.9253731	0.9037037	0.9564677
1.27	0.9172414	0.7823316	0.9583333	0.9253731	0.9104478	0.9589552
1.28	0.9034483	0.7773284	0.9591837	0.9253731	0.8985507	0.9566590
1.29	0.9103448	0.7690206	0.9333333	0.9104478	0.9022556	0.9576158
1.3	0.9241379	0.7987703	0.9666667	0.9402985	0.9133858	0.9615385
1.31	0.9241379	0.7921267	0.9818182	0.9402985	0.9133858	0.9597206

1.32	0.9172414	0.8103368	0.9500000	0.9402985	0.9076923	0.9537887
1.33	0.9172414	0.7823479	0.9661017	0.9253731	0.9051095	0.9582855
1.34	0.9103448	0.7675937	0.9482759	0.9253731	0.9051095	0.9450823
1.35	0.9172414	0.7725790	10.000.000	0.9253731	0.9117647	0.9510142
1.36	0.9103448	0.7800034	0.9814815	0.9701493	0.9064748	0.9503444
1.37	0.9310345	0.8025702	0.9811321	0.9402985	0.9264706	0.9512055
1.38	0.9172414	0.7929549	0.9636364	0.9701493	0.9154930	0.9525450
1.39	0.9103448	0.7812715	0.9444444	0.9850746	0.9103448	0.9602947
1.4	0.9241379	0.7912120	0.9454545	0.9850746	0.9230769	0.9577114
1.41	0.9103448	0.8212626	0.9615385	0.9701493	0.9090909	0.9536931
1.42	0.9103448	0.7746803	0.9615385	0.9552239	0.9078014	0.9516839
1.43	0.9103448	0.7711958	0.9411765	0.9701493	0.9078014	0.9514925
1.44	0.9034483	0.7786042	0.9387755	0.9701493	0.9027778	0.9513012
1.45	0.8965517	0.7430646	0.9574468	0.9253731	0.8905109	0.9467088
1.46	0.9034483	0.7632619	0.9591837	0.9701493	0.8970588	0.9537887
1.47	0.9034483	0.7358780	0.9600000	0.9701493	0.8970588	0.9530233
1.48	0.8896552	0.7164318	0.9583333	0.9552239	0.8888889	0.9493877
1.49	0.8965517	0.7362209	0.9583333	0.9402985	0.8888889	0.9426904
1.5	0.8965517	0.7271177	0.9555556	0.9402985	0.8888889	0.9424990
1.51	0.8965517	0.7270198	0.9795918	0.9552239	0.8951049	0.9513969
1.52	0.8896552	0.7392515	10.000.000	0.8805970	0.8805970	0.9371412
1.53	0.8689655	0.7489660	10.000.000	0.8656716	0.8387097	0.9556066
1.54	0.8620690	0.7302192	10.000.000	0.8656716	0.8437500	0.9314007
1.55	0.8620690	0.7428486	0.9782609	0.8656716	0.8372093	0.9350364
1.56	0.8551724	0.7325493	10.000.000	0.8507463	0.8321168	0.9341753
1.57	0.8551724	0.7305753	0.9761905	0.8358209	0.8346457	0.9280520
1.58	0.8413793	0.6898567	0.9722222	0.7611940	0.8160000	0.9208764
1.59	0.8206897	0.6710690	0.9743590	0.7611940	0.7906977	0.9146575
1.6	0.8551724	0.6830389	0.9761905	0.7462687	0.8235294	0.9200153
1.61	0.8482759	0.6937212	0.9772727	0.7164179	0.8135593	0.9157099
1.62	0.8344828	0.6921488	0.9767442	0.6865672	0.7931034	0.9085343
1.63	0.8413793	0.7327565	10.000.000	0.6716418	0.7927928	0.9228856
1.64	0.8206897	0.7113981	10.000.000	0.6417910	0.7636364	0.9185802
1.65	0.8068966	0.7148005	10.000.000	0.6119403	0.7454545	0.9175277
1.66	0.7931034	0.6750002	10.000.000	0.5970149	0.7115385	0.9144661
1.67	0.7862069	0.6892349	10.000.000	0.5970149	0.7142857	0.9071948
1.68	0.7862069	0.7292436	10.000.000	0.5820896	0.7102804	0.9272866
1.69	0.7793103	0.7289576	10.000.000	0.5820896	0.7090909	0.9178148
1.7	0.7793103	0.7512470	10.000.000	0.5522388	0.6981132	0.9213548
1.71	0.7724138	0.7323512	10.000.000	0.5522388	0.6857143	0.9133180
1.72	0.7586207	0.7295243	10.000.000	0.5373134	0.6666667	0.9111175
1.73	0.7586207	0.6798068	10.000.000	0.5373134	0.6666667	0.9075775
1.74	0.7655172	0.6935704	10.000.000	0.5522388	0.6730769	0.9083429
1.75	0.7655172	0.6378742	10.000.000	0.5522388	0.6730769	0.9083429

1.76	0.7586207	0.6117487	10.000.000	0.5671642	0.6666667	0.9095867
1.77	0.7586207	0.6138457	10.000.000	0.5671642	0.6666667	0.8952354
1.78	0.7517241	0.6123389	10.000.000	0.5522388	0.6607143	0.8926521
1.79	0.7379310	0.6426956	10.000.000	0.5522388	0.6607143	0.9019326

Πίνακας 7 Min τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Α

ΑΚΤΙΝΑ	Accuracy	H	Precision	Recall	F	AUC
0.4	0.6827586	0.3549686	0.7906977	0.4029851	0.5400000	0.7829124
0.41	0.6896552	0.3700415	0.8043478	0.4179104	0.5544554	0.7868351
0.42	0.6965517	0.3112698	0.7800000	0.4328358	0.5769231	0.7685610
0.43	0.7103448	0.3293159	0.7843137	0.4477612	0.5940594	0.7787026
0.44	0.7241379	0.3632287	0.7924528	0.4925373	0.6296296	0.7968810
0.45	0.7310345	0.4101498	0.7924528	0.5223881	0.6422018	0.8195561
0.46	0.7310345	0.3916167	0.8260870	0.5074627	0.6422018	0.8061615
0.47	0.7310345	0.4056607	0.8163265	0.5223881	0.6422018	0.8133372
0.48	0.7379310	0.4122535	0.8085106	0.5671642	0.6666667	0.8154420
0.49	0.7172414	0.4149742	0.7708333	0.5522388	0.6434783	0.8063529
0.5	0.7103448	0.3786512	0.7659574	0.5373134	0.6315789	0.7934367
0.51	0.7241379	0.4037827	0.7755102	0.5373134	0.6486486	0.8026215
0.52	0.7310345	0.4149917	0.7800000	0.5373134	0.6486486	0.8115193
0.53	0.7172414	0.4246044	0.7708333	0.5373134	0.6371681	0.8248182
0.54	0.7172414	0.4635714	0.7708333	0.5373134	0.6434783	0.8395522
0.55	0.7379310	0.4776201	0.7959184	0.5820896	0.6724138	0.8433793
0.56	0.7448276	0.4923402	0.8000000	0.5820896	0.6837607	0.8471106
0.57	0.7241379	0.4738933	0.7755102	0.5671642	0.6551724	0.8517987
0.58	0.7241379	0.4602333	0.7755102	0.5223881	0.6481481	0.8409874
0.59	0.7655172	0.5023576	0.8000000	0.5522388	0.6915888	0.8576349
0.6	0.7793103	0.5100384	0.8000000	0.5970149	0.7142857	0.8615576
0.61	0.7793103	0.5092575	0.8032787	0.6268657	0.7241379	0.8606965
0.62	0.7931034	0.5268836	0.8064516	0.6567164	0.7563025	0.8657673
0.63	0.7655172	0.5226275	0.7894737	0.6567164	0.7258065	0.8672981
0.64	0.7655172	0.5308880	0.7846154	0.6567164	0.7258065	0.8706468
0.65	0.7586207	0.5201547	0.7727273	0.6417910	0.7107438	0.8580176
0.66	0.7586207	0.5075305	0.7962963	0.6417910	0.7107438	0.8692116
0.67	0.7586207	0.5044178	0.7857143	0.6567164	0.7154472	0.8615576
0.68	0.7793103	0.5016812	0.7846154	0.6716418	0.7377049	0.8630884
0.69	0.7724138	0.5003326	0.7727273	0.6865672	0.7360000	0.8598354
0.7	0.7724138	0.5045790	0.7727273	0.6865672	0.7360000	0.8632798
0.71	0.7793103	0.5071678	0.7761194	0.7014925	0.7500000	0.8624187
0.72	0.7793103	0.5074406	0.7761194	0.6865672	0.7500000	0.8663414
0.73	0.7793103	0.5157317	0.7681159	0.6865672	0.7500000	0.8732300
0.74	0.7793103	0.5395684	0.7611940	0.7014925	0.7460317	0.8788749
0.75	0.7724138	0.5234609	0.7647059	0.6865672	0.7401575	0.8714122

0.76	0.7724138	0.5419089	0.7681159	0.6716418	0.7401575	0.8729430
0.77	0.7724138	0.5446446	0.7681159	0.6865672	0.7401575	0.8806927
0.78	0.7655172	0.5537456	0.7746479	0.6865672	0.7301587	0.8755262
0.79	0.7724138	0.5570128	0.7638889	0.6865672	0.7360000	0.8747608
0.8	0.7724138	0.5566208	0.7638889	0.6865672	0.7360000	0.8740911
0.81	0.7724138	0.5568847	0.7746479	0.6865672	0.7360000	0.8812667
0.82	0.7793103	0.5579736	0.7763158	0.7014925	0.7460317	0.8712208
0.83	0.7862069	0.5594859	0.7763158	0.7014925	0.7596899	0.8850938
0.84	0.8000000	0.5583078	0.7671233	0.7014925	0.7704918	0.8855721
0.85	0.8068966	0.5569196	0.7671233	0.7014925	0.7704918	0.8827976
0.86	0.8068966	0.5604853	0.7671233	0.7014925	0.7704918	0.8833716
0.87	0.8000000	0.5652974	0.7671233	0.6716418	0.7563025	0.8802143
0.88	0.7931034	0.5651376	0.7671233	0.6567164	0.7457627	0.8834673
0.89	0.7862069	0.5700731	0.7714286	0.6716418	0.7520000	0.8835630
0.9	0.7931034	0.5705838	0.7714286	0.6716418	0.7563025	0.8787792
0.91	0.8000000	0.5580913	0.7714286	0.6716418	0.7563025	0.8685419
0.92	0.8000000	0.5577204	0.7714286	0.6716418	0.7563025	0.8671068
0.93	0.8000000	0.5579605	0.7714286	0.6716418	0.7563025	0.8696900
0.94	0.7931034	0.5566761	0.7605634	0.6865672	0.7603306	0.8666284
0.95	0.7931034	0.5572123	0.7605634	0.6865672	0.7603306	0.8649062
0.96	0.7931034	0.5397454	0.7605634	0.6865672	0.7603306	0.8650976
0.97	0.8000000	0.5636129	0.7671233	0.6865672	0.7603306	0.8663414
0.98	0.8000000	0.5492769	0.7671233	0.6865672	0.7603306	0.8664370
0.99	0.8000000	0.5485374	0.7692308	0.6865672	0.7603306	0.8653846
1	0.8000000	0.5633932	0.7692308	0.6865672	0.7603306	0.8670111
1.01	0.8000000	0.5593782	0.7638889	0.6865672	0.7603306	0.8659587
1.02	0.8000000	0.5290856	0.7692308	0.7014925	0.7642276	0.8716992
1.03	0.8000000	0.5550429	0.7638889	0.7014925	0.7642276	0.8641408
1.04	0.8000000	0.5507033	0.7638889	0.7014925	0.7642276	0.8746651
1.05	0.8000000	0.5241619	0.7638889	0.7164179	0.7741935	0.8623230
1.06	0.8068966	0.5318139	0.7671233	0.7462687	0.7874016	0.8625144
1.07	0.7931034	0.5350577	0.7733333	0.7462687	0.7727273	0.8633754
1.08	0.8068966	0.4983386	0.7733333	0.7313433	0.7941176	0.8490241
1.09	0.8068966	0.5105207	0.7746479	0.7164179	0.7933884	0.8630884
1.1	0.8000000	0.5222887	0.7638889	0.7164179	0.7910448	0.8611749
1.11	0.8068966	0.5375422	0.7702703	0.7014925	0.7833333	0.8702641
1.12	0.7793103	0.5358751	0.7536232	0.7014925	0.7647059	0.8718905
1.13	0.8000000	0.5106258	0.7746479	0.7313433	0.7786260	0.8645235
1.14	0.7793103	0.4945300	0.7333333	0.6865672	0.7716535	0.8606008
1.15	0.7655172	0.5017809	0.7391304	0.7014925	0.7500000	0.8507463
1.16	0.7862069	0.5041009	0.7647059	0.6716418	0.7563025	0.8520857
1.17	0.8000000	0.5194903	0.7826087	0.6865672	0.7666667	0.8556257
1.18	0.8000000	0.5175090	0.7746479	0.6567164	0.7652174	0.8556257
1.19	0.7862069	0.5223497	0.7746479	0.6268657	0.7304348	0.8523728

1.2	0.7862069	0.5470868	0.7887324	0.6417910	0.7350427	0.8605052
1.21	0.7931034	0.5237299	0.7702703	0.6567164	0.7457627	0.8701684
1.22	0.8000000	0.4974569	0.7500000	0.6567164	0.7521368	0.8648106
1.23	0.7655172	0.4962379	0.7200000	0.6865672	0.7605634	0.8473020
1.24	0.7655172	0.4572085	0.7260274	0.7164179	0.7571429	0.8428052
1.25	0.7793103	0.4627036	0.7397260	0.7164179	0.7714286	0.8510333
1.26	0.7793103	0.4678896	0.7611940	0.7164179	0.7611940	0.8182166
1.27	0.7793103	0.4936909	0.7681159	0.6865672	0.7538462	0.8212782
1.28	0.7655172	0.4495188	0.7619048	0.7014925	0.7384615	0.8272101
1.29	0.7724138	0.4608288	0.7500000	0.7462687	0.7555556	0.8272101
1.3	0.8000000	0.5176306	0.7500000	0.7462687	0.7851852	0.8620360
1.31	0.8000000	0.5170718	0.7500000	0.7462687	0.7819549	0.8517030
1.32	0.8137931	0.4887965	0.7594937	0.7462687	0.7968750	0.8456755
1.33	0.7793103	0.4612969	0.7714286	0.6119403	0.7192982	0.8288366
1.34	0.7793103	0.4720467	0.7761194	0.6268657	0.7241379	0.8340987
1.35	0.7793103	0.4915138	0.7746479	0.6268657	0.7241379	0.8095101
1.36	0.7793103	0.4907093	0.7466667	0.6268657	0.7241379	0.8083620
1.37	0.7862069	0.4684265	0.7571429	0.6417910	0.7413793	0.8020475
1.38	0.7655172	0.4377714	0.7260274	0.6417910	0.7350427	0.7969767
1.39	0.7586207	0.4333696	0.7222222	0.6417910	0.7413793	0.7797551
1.4	0.7586207	0.4318608	0.7222222	0.6567164	0.7333333	0.7810945
1.41	0.7586207	0.4350808	0.7222222	0.6567164	0.7333333	0.8131458
1.42	0.7517241	0.4655529	0.7183099	0.6567164	0.7333333	0.8090318
1.43	0.7655172	0.4795119	0.7323944	0.6417910	0.7413793	0.8232874
1.44	0.7655172	0.4505172	0.7464789	0.6417910	0.7226891	0.8239571
1.45	0.7724138	0.4506118	0.7428571	0.6567164	0.7480916	0.8230004
1.46	0.7724138	0.4470363	0.7500000	0.6417910	0.7555556	0.8201301
1.47	0.7655172	0.4086958	0.7500000	0.6268657	0.7424242	0.7886529
1.48	0.7586207	0.3835315	0.7424242	0.6716418	0.7258065	0.7853042
1.49	0.7448276	0.4275007	0.7536232	0.6417910	0.6991870	0.8190777
1.5	0.7517241	0.4696036	0.7534247	0.6268657	0.7049180	0.8381171
1.51	0.7448276	0.4316345	0.7656250	0.5970149	0.6890756	0.8274971
1.52	0.7379310	0.4446119	0.7195122	0.5223881	0.6779661	0.7991772
1.53	0.7448276	0.4689560	0.7534247	0.5522388	0.6837607	0.8174512
1.54	0.7103448	0.4244309	0.7631579	0.3731343	0.5434783	0.8110409
1.55	0.7172414	0.4307865	0.7733333	0.4477612	0.6060606	0.8163031
1.56	0.6758621	0.4807166	0.7808219	0.2985075	0.4597701	0.8430922
1.57	0.6275862	0.4278836	0.7868852	0.2238806	0.3571429	0.8297933
1.58	0.6206897	0.4220898	0.7500000	0.1940299	0.3209877	0.8173555
1.59	0.5793103	0.3940592	0.6666667	0.1791045	0.2823529	0.8040566
1.6	0.6482759	0.4020559	0.7555556	0.2686567	0.4137931	0.8007080
1.61	0.6413793	0.3818990	0.7906977	0.2537313	0.3953488	0.8058745
1.62	0.6689655	0.3792508	0.7272727	0.2985075	0.4545455	0.8016648
1.63	0.6689655	0.3013486	0.7358491	0.2985075	0.4545455	0.7572713

1.64	0.6482759	0.3044988	0.7291667	0.2835821	0.4269663	0.7544011
1.65	0.6620690	0.2993571	0.7708333	0.2686567	0.4235294	0.7663605
1.66	0.6275862	0.3365913	0.7446809	0.2089552	0.3414634	0.7882702
1.67	0.6482759	0.3190747	0.7272727	0.3432836	0.4848485	0.7722924
1.68	0.6620690	0.3013813	0.7209302	0.3134328	0.4719101	0.7455989
1.69	0.6000000	0.2891808	0.7209302	0.1641791	0.2750000	0.7355530
1.7	0.5931034	0.2771013	0.7250000	0.1492537	0.2531646	0.7266552
1.71	0.6413793	0.2651489	0.7250000	0.2537313	0.4047619	0.6769996
1.72	0.6275862	0.2590095	0.7073171	0.3134328	0.4375000	0.6426521
1.73	0.6275862	0.1846643	0.7179487	0.2537313	0.4047619	0.6168197
1.74	0.6275862	0.1825594	0.7241379	0.2388060	0.3855422	0.6150976
1.75	0.6344828	0.1800780	0.7368421	0.2835821	0.4175824	0.6254305
1.76	0.6206897	0.1836807	0.6923077	0.2537313	0.3953488	0.6275354
1.77	0.6344828	0.1757416	0.6923077	0.2537313	0.3953488	0.6354765
1.78	0.6137931	0.1738958	0.6923077	0.2388060	0.3636364	0.6361462
1.79	0.5724138	0.1891627	0.6666667	0.1492537	0.2439024	0.6504975

Πίνακας 8: Min τιμές από την 1 επανάληψη για ακτίνες από 0,40-1,79 για την Μέθοδο Β

AKTINA	Accuracy	H	Precision	Recall	F	AUC
0.4	0.7395862	0.4551038	0.8835685	0.5032836	0.6395614	0.8457214
0.41	0.7470345	0.4680364	0.8848251	0.5211940	0.6543302	0.8512648
0.42	0.7546207	0.4714976	0.8766310	0.5471642	0.6721387	0.8520054
0.43	0.7602759	0.4856024	0.8807826	0.5576119	0.6812251	0.8586452
0.44	0.7714483	0.5079122	0.8809778	0.5853731	0.7018611	0.8667183
0.45	0.7808276	0.5237217	0.8855472	0.6047761	0.7173631	0.8722790
0.46	0.7855172	0.5251215	0.8848849	0.6173134	0.7258820	0.8715327
0.47	0.7942069	0.5358403	0.8845487	0.6394030	0.7406656	0.8743724
0.48	0.8002759	0.5440303	0.8863382	0.6525373	0.7504510	0.8781362
0.49	0.8013793	0.5514478	0.8847240	0.6567164	0.7526623	0.8785323
0.5	0.8033103	0.5546119	0.8841745	0.6617910	0.7555455	0.8811883
0.51	0.8034483	0.5612487	0.8829849	0.6635821	0.7562749	0.8846766
0.52	0.8108966	0.5700219	0.8824465	0.6820896	0.7680760	0.8856372
0.53	0.8136552	0.5825353	0.8889891	0.6820896	0.7705862	0.8916188
0.54	0.8182069	0.5819580	0.8893084	0.6934328	0.7776918	0.8910984
0.55	0.8245517	0.5977664	0.8970596	0.7014925	0.7860266	0.8960716
0.56	0.8233103	0.5989333	0.8935431	0.7017910	0.7845887	0.8980674
0.57	0.8257931	0.6025887	0.8961280	0.7053731	0.7877769	0.8982223
0.58	0.8268966	0.6095537	0.8931131	0.7113433	0.7901504	0.9015212
0.59	0.8365517	0.6256338	0.9000119	0.7286567	0.8035069	0.9077918
0.6	0.8375172	0.6261187	0.8969549	0.7340299	0.8058094	0.9089170
0.61	0.8343448	0.6252593	0.8865566	0.7370149	0.8035155	0.9114619
0.62	0.8435862	0.6384759	0.8852562	0.7611940	0.8173854	0.9161902
0.63	0.8402759	0.6385458	0.8808174	0.7582090	0.8137065	0.9131822
0.64	0.8362759	0.6390705	0.8713833	0.7591045	0.8101090	0.9142652

0.65	0.8383448	0.6415242	0.8713736	0.7644776	0.8132351	0.9152009
0.66	0.8387586	0.6469010	0.8701440	0.7665672	0.8140106	0.9179296
0.67	0.8430345	0.6536119	0.8706790	0.7767164	0.8201326	0.9178971
0.68	0.8409655	0.6519014	0.8696836	0.7725373	0.8174645	0.9175277
0.69	0.8402759	0.6477600	0.8648933	0.7770149	0.8177421	0.9186625
0.7	0.8417931	0.6482801	0.8622564	0.7841791	0.8206048	0.9177306
0.71	0.8405517	0.6509983	0.8598126	0.7847761	0.8195503	0.9193992
0.72	0.8422069	0.6436790	0.8593164	0.7895522	0.8220565	0.9186873
0.73	0.8382069	0.6444418	0.8496159	0.7919403	0.8187126	0.9188366
0.74	0.8364138	0.6428347	0.8460113	0.7922388	0.8170906	0.9148144
0.75	0.8353103	0.6409475	0.8418242	0.7946269	0.8161585	0.9158611
0.76	0.8366897	0.6418315	0.8395808	0.8011940	0.8184233	0.9163949
0.77	0.8404138	0.6509889	0.8429417	0.8062687	0.8229075	0.9200727
0.78	0.8459310	0.6577321	0.8453217	0.8176119	0.8299426	0.9227038
0.79	0.8455172	0.6613845	0.8450879	0.8170149	0.8294245	0.9223020
0.8	0.8463448	0.6617921	0.8432148	0.8217910	0.8310125	0.9211271
0.81	0.8463448	0.6629316	0.8427042	0.8223881	0.8309706	0.9207348
0.82	0.8460690	0.6593288	0.8395999	0.8262687	0.8314115	0.9182931
0.83	0.8471724	0.6647166	0.8398188	0.8292537	0.8328886	0.9198469
0.84	0.8456552	0.6651941	0.8375948	0.8289552	0.8315924	0.9209759
0.85	0.8464828	0.6658969	0.8385453	0.8298507	0.8325350	0.9205645
0.86	0.8440000	0.6643955	0.8359243	0.8271642	0.8298699	0.9197857
0.87	0.8442759	0.6648672	0.8356340	0.8283582	0.8302697	0.9202583
0.88	0.8474483	0.6664474	0.8363593	0.8358209	0.8342309	0.9203444
0.89	0.8500690	0.6653738	0.8376696	0.8408955	0.8375391	0.9198890
0.9	0.8535172	0.6688347	0.8408592	0.8453731	0.8413786	0.9214906
0.91	0.8572414	0.6637552	0.8435803	0.8510448	0.8457858	0.9194374
0.92	0.8547586	0.6633450	0.8411496	0.8480597	0.8429057	0.9187868
0.93	0.8532414	0.6620558	0.8381547	0.8486567	0.8415775	0.9166418
0.94	0.8561379	0.6658889	0.8423904	0.8498507	0.8442923	0.9173766
0.95	0.8517241	0.6615446	0.8385203	0.8441791	0.8393849	0.9149043
0.96	0.8526897	0.6609490	0.8374745	0.8483582	0.8410660	0.9148986
0.97	0.8524138	0.6601090	0.8372309	0.8480597	0.8408045	0.9142097
0.98	0.8575172	0.6596480	0.8398808	0.8573134	0.8470315	0.9135476
0.99	0.8583448	0.6644627	0.8390841	0.8608955	0.8483277	0.9159510
1	0.8582069	0.6643612	0.8388918	0.8608955	0.8482126	0.9160448
1.01	0.8588966	0.6629859	0.8392967	0.8617910	0.8490117	0.9161328
1.02	0.8565517	0.6590048	0.8367946	0.8594030	0.8465732	0.9138978
1.03	0.8551724	0.6592281	0.8320213	0.8629851	0.8457978	0.9139782
1.04	0.8529655	0.6536756	0.8273524	0.8644776	0.8442350	0.9135457
1.05	0.8543448	0.6535929	0.8294753	0.8647761	0.8455227	0.9107750
1.06	0.8560000	0.6495946	0.8277963	0.8716418	0.8481337	0.9085209
1.07	0.8539310	0.6458334	0.8282071	0.8653731	0.8453941	0.9086299
1.08	0.8536552	0.6448408	0.8266352	0.8671642	0.8454044	0.9082606

1.09	0.8503448	0.6440095	0.8282468	0.8552239	0.8404864	0.9087734
1.1	0.8515862	0.6454238	0.8287652	0.8579104	0.8421140	0.9085343
1.11	0.8520000	0.6468778	0.8298892	0.8570149	0.8422009	0.9097876
1.12	0.8506207	0.6418214	0.8301291	0.8531343	0.8405013	0.9092250
1.13	0.8493793	0.6418125	0.8303442	0.8495522	0.8388196	0.9072809
1.14	0.8442759	0.6344659	0.8322297	0.8340299	0.8316256	0.9071795
1.15	0.8441379	0.6357354	0.8378294	0.8250746	0.8298542	0.9063222
1.16	0.8413793	0.6361710	0.8417089	0.8131343	0.8249430	0.9041600
1.17	0.8411034	0.6355603	0.8479100	0.8029851	0.8229136	0.9024780
1.18	0.8408276	0.6365378	0.8447051	0.8074627	0.8234518	0.9036854
1.19	0.8448276	0.6371449	0.8462539	0.8152239	0.8284573	0.9056238
1.2	0.8471724	0.6481522	0.8520160	0.8128358	0.8302310	0.9098852
1.21	0.8486897	0.6470507	0.8504952	0.8191045	0.8328498	0.9110658
1.22	0.8444138	0.6387075	0.8518024	0.8074627	0.8269062	0.9098909
1.23	0.8446897	0.6354106	0.8456677	0.8158209	0.8287303	0.9054458
1.24	0.8456552	0.6356141	0.8465264	0.8173134	0.8299743	0.9041638
1.25	0.8449655	0.6332326	0.8458994	0.8158209	0.8292403	0.9057692
1.26	0.8560000	0.6493583	0.8547986	0.8322388	0.8420044	0.9079793
1.27	0.8577931	0.6566610	0.8575575	0.8328358	0.8434338	0.9070704
1.28	0.8520000	0.6481014	0.8515740	0.8256716	0.8370198	0.9062476
1.29	0.8520000	0.6465139	0.8505819	0.8271642	0.8377142	0.9081975
1.3	0.8617931	0.6593900	0.8598512	0.8411940	0.8489484	0.9115844
1.31	0.8612414	0.6558708	0.8621852	0.8364179	0.8474565	0.9098737
1.32	0.8634483	0.6575575	0.8686685	0.8331343	0.8492250	0.9097933
1.33	0.8587586	0.6423115	0.8606452	0.8313433	0.8442770	0.9047838
1.34	0.8580690	0.6449967	0.8611354	0.8283582	0.8430382	0.9049904
1.35	0.8571034	0.6423995	0.8592376	0.8289552	0.8423540	0.9059816
1.36	0.8514483	0.6416068	0.8451624	0.8343284	0.8376903	0.9022943
1.37	0.8550345	0.6471579	0.8543924	0.8301493	0.8401342	0.9074952
1.38	0.8528276	0.6407926	0.8509804	0.8301493	0.8381097	0.9073670
1.39	0.8514483	0.6401308	0.8499909	0.8280597	0.8363232	0.9074455
1.4	0.8520000	0.6403943	0.8507091	0.8283582	0.8370400	0.9081190
1.41	0.8526897	0.6423472	0.8586458	0.8208955	0.8364460	0.9091083
1.42	0.8484138	0.6379000	0.8538569	0.8170149	0.8318050	0.9065423
1.43	0.8471724	0.6444395	0.8488136	0.8202985	0.8310174	0.9097283
1.44	0.8398621	0.6416423	0.8393188	0.8131343	0.8229101	0.9083697
1.45	0.8365517	0.6246302	0.8397975	0.8044776	0.8187468	0.9024436
1.46	0.8405517	0.6255501	0.8406733	0.8146269	0.8237744	0.9023728
1.47	0.8375172	0.6211968	0.8438673	0.8014925	0.8185860	0.9007099
1.48	0.8402759	0.6160516	0.8452449	0.8053731	0.8221427	0.8995829
1.49	0.8360000	0.6103733	0.8400555	0.8011940	0.8174817	0.8938997
1.5	0.8271724	0.6017816	0.8447350	0.7719403	0.8034625	0.8917030
1.51	0.8217931	0.5947447	0.8452489	0.7564179	0.7945331	0.8870015
1.52	0.8080000	0.5852599	0.8470439	0.7208955	0.7743686	0.8794202

1.53	0.8140690	0.5991280	0.8611814	0.7188060	0.7793437	0.8818944
1.54	0.8064828	0.6101351	0.8862612	0.6734328	0.7596347	0.8887428
1.55	0.8051034	0.6028980	0.8889310	0.6677612	0.7563982	0.8892805
1.56	0.7968276	0.5905872	0.8778008	0.6585075	0.7442811	0.8867528
1.57	0.7539310	0.5845881	0.8824563	0.5429851	0.6579579	0.8862246
1.58	0.7562759	0.5705196	0.8763314	0.5564179	0.6664819	0.8801990
1.59	0.7452414	0.5607975	0.8663122	0.5331343	0.6458707	0.8765365
1.6	0.7628966	0.5436358	0.8661393	0.5791045	0.6882918	0.8709395
1.61	0.7608276	0.5396770	0.8732701	0.5662687	0.6819729	0.8694126
1.62	0.7500690	0.5309386	0.8740195	0.5397015	0.6617182	0.8659357
1.63	0.7554483	0.5265367	0.8847590	0.5444776	0.6694365	0.8588595
1.64	0.7427586	0.5189492	0.8816754	0.5146269	0.6454887	0.8567853
1.65	0.7347586	0.5212960	0.8918145	0.4877612	0.6267071	0.8586663
1.66	0.7209655	0.5135101	0.8951796	0.4525373	0.5938719	0.8550823
1.67	0.7288276	0.5012334	0.8940742	0.4710448	0.6138901	0.8508534
1.68	0.7245517	0.5157583	0.9047025	0.4528358	0.6007456	0.8567049
1.69	0.7148966	0.5047453	0.8995966	0.4334328	0.5806014	0.8512610
1.7	0.7113103	0.5074084	0.8928147	0.4298507	0.5756301	0.8514945
1.71	0.7071724	0.4976360	0.8894267	0.4223881	0.5687164	0.8433946
1.72	0.7080000	0.4839143	0.8855063	0.4268657	0.5727897	0.8381439
1.73	0.7063448	0.4778603	0.8933241	0.4188060	0.5658339	0.8371431
1.74	0.7078621	0.4777051	0.8990122	0.4191045	0.5667440	0.8392097
1.75	0.7084138	0.4700863	0.8891641	0.4238806	0.5714931	0.8347742
1.76	0.6966897	0.4480172	0.8364482	0.4313433	0.5643529	0.8174608
1.77	0.6957241	0.4483600	0.8385129	0.4277612	0.5615977	0.8148412
1.78	0.6877241	0.4357656	0.8317723	0.4116418	0.5449216	0.8057157
1.79	0.6748966	0.4332259	0.8341134	0.3731343	0.5075726	0.8069441

Κώδικας

```
install.packages("hmeasure")
install.packages("MASS")
install.packages("Metrics")
install.packages("caTools")
library(Metrics)
library(BallMapper)  #Ball Mapper graph
library(pROC)  #auc calculation
library(hmeasure)#h_measure
library(caret)#accurancy
library(MASS)
library(class) #h_measure

###step1
set.seed(333)#import dataset#
dataset<-australian_dat#data normalization
process <- preProcess(as.data.frame(dataset), method=c("range"))
norm_data <- predict(process, as.data.frame(dataset))
class<-norm_data$X15
norm_data_2 <- norm_data[-15]
train_indices <- sample(1:nrow(dataset), 0.7 * nrow(dataset))#mia sinartisi pou xwrizei sto 70% to dataset
training.set<- norm_data_2[train_indices, ]#xwrizei tixaia to 70% se training dedomena
class.training.set<-norm_data$X15[train_indices]#xwrizei tixaia to 70% apo to digma 0,1 athetisi,mi athetisi
training.set_15<- norm_data[train_indices, ]#xwrizei tixaia to 70% se training dedomena
class.training.set_15<-norm_data$X15[train_indices]
#class.training.set<-as.data.frame(class.training.set)#to kanoume dataframe.
testing.set<- norm_data_2[-train_indices, ]
class.testing.set<-norm_data$X15[-train_indices]
#class.testing.set<-as.data.frame(class.testing.set)
id.good2<-which(class.training.set==0)
id.good2
id.bad2<-which(class.training.set==1)
id.bad2
probbad<-222/482
probgood<-260/482
sample.good2<-sample(id.good2,260)# επιστρέφει δείγμα απο τους καλούς 260
sample.bad.2<-sample(id.bad2,222)# επιστρέφει δείγμα απο τους κακους 222
rownames(training.set)<-seq(1, length.out = nrow(training.set))
rownames(training.set)
rownames(testing.set) <- seq(483, length.out = nrow(testing.set))
rownames(testing.set)
#epsilon<-1.2  ## το βέλτιστο ε
#BMaustralian<-BallMapper(training.set,as.data.frame(class.training.set),epsilon)
#ColorIgraphPlot(BMaustralian)
#vertices<-c(BMaustralian[["landmarks"]]) #einai diktes twn paratirisewn.
```

```
#prepei na upologisoume tin apostasi meta3i kathe neas paratirisis me to kentro
#twn sfairwn kai stin sinexeia na episimanoume tis sfaires stis opoies aniki auti i paratirisi

###step2
PROB_BM_AB<-function(epsilon,training,testing,class.training,class.testing, prob)
{BM<-BallMapper(training,as.data.frame(class.training),epsilon)
vertices<-c(BM[["landmarks"]])  # indices of the observaions- ball centers
#we must calculate the distance of every new obs from the ball centers and then decide
the balls that this new obs belongs to
default.probabilityA<-c()
default.probabilityB<-c()
p<-list()
numbers_of_landmarks<-c()
landmarks<-list()
##probability<-list()
for (i in 1:nrow(testing)) {new.obs<-testing[i,]
distance<-as.matrix(dist(rbind(new.obs,training[vertices,])))[1,][-1]
distance_new<-as.matrix(dist(rbind(new.obs,training[,])))[1,][-1]
ball.centers<-which(distance<epsilon)  #distances smaller than epsilon
landmarks[[i]]<-as.numeric(names(ball.centers))
numbers_of_landmarks[i]<-length(landmarks[[i]])
# default probability A
if (numbers_of_landmarks[i]>0)
{p[[i]]<-
points_covered_by_landmarks(BM,unique(unlist(BM$coverage[landmarks[[i]]])))
default.probabilityA[i]<-mean(class.training[p[[i]]])}
else
{neighbor<-as.numeric(which.min(distance_new))
centers<-as.numeric(BM$coverage[[neighbor]])
p[[i]]<-points_covered_by_landmarks(BM,centers)
default.probabilityA[i]<-mean(class.training[p[[i]]])}
# default probability B
if (numbers_of_landmarks[i]>0)
{p[[i]]<-
points_covered_by_landmarks(BM,unique(unlist(BM$coverage[landmarks[[i]]])))
default.probabilityB[i]<-mean(class.training[p[[i]]])}
else {
  default.probabilityB[i]<-prob}}
actual_labels <- class.testing
predicted_labels_A <- ifelse(default.probabilityA > 0.5 , 1, 0)
predicted_labels_B <- ifelse(default.probabilityB > 0.5 , 1, 0)
# metrics
accurancy_A <-Metrics::accuracy(actual_labels ,predicted_labels_A)
AUC_method_A<-auc(class.testing,default.probabilityA)
HMeasure_A<-HMeasure(class.testing,default.probabilityA)
accurancy_B <-Metrics::accuracy(actual_labels ,predicted_labels_B)
HMeasure_B<-HMeasure(class.testing,default.probabilityB)
```

```
 AUC_method_B<-auc(class.testing,default.probabilityB)
 result <- list(default.probabilityA = default.probabilityA,
              default.probabilityB = default.probabilityB,
              HMeasure_A = HMeasure_A ,
              HMeasure_B = HMeasure_B,
              AUC_method_A = AUC_method_A,
              AUC_method_B=AUC_method_B,
              accurancy_A=accurancy_A,
              accurancy_B=accurancy_B,
              BM=BM)

 return(result)
 }
 result_59<-PROB_BM_AB(0.59,training.set,          testing.set,          class.training.set,
class.testing.set, 0.46)
 result_59$default.probabilityA
 result_59$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
 result_59$AUC_method_A
 result_59$accurancy_A
 result_59$BM
 ColorIgraphPlot(result_59$BM)

 result_67<-PROB_BM_AB(0.67,training.set,          testing.set,          class.training.set,
class.testing.set, 0.46)
 result_67$default.probabilityA
 result_67$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
 result_67$accurancy_A
 ColorIgraphPlot(result_67$BM)

 result_78<-PROB_BM_AB(0.78,training.set,          testing.set,          class.training.set,
class.testing.set, 0.46)
 result_78$default.probabilityA
 result_78$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
 result_78$accurancy_A
 ColorIgraphPlot(result_78$BM)

 result_90<-PROB_BM_AB(0.90,training.set,          testing.set,          class.training.set,
class.testing.set, 0.46)
 result_90$default.probabilityA
 result_90$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
 result_90$accurancy_A
 ColorIgraphPlot(result_90$BM)

 result_106<-PROB_BM_AB(1.06,training.set,          testing.set,          class.training.set,
class.testing.set, 0.46)
 result_106$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
 result_106$accurancy_A
```

```
ColorIgraphPlot(result_106$BM)

result_132<-PROB_BM_AB(1.32,training.set, testing.set, class.training.set,
class.testing.set, 0.46)
result_132$default.probabilityA
result_132$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
result_132$accurancy_A
ColorIgraphPlot(result_132$BM)

result_168<-PROB_BM_AB(1.68,training.set, testing.set, class.training.set,
class.testing.set, 0.46)
result_168$default.probabilityA
result_168$HMeasure_A$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
result_168$accurancy_A
ColorIgraphPlot(result_168$BM)

###step 3
dataset_2<-norm_data[train_indices, ]
second.training.set<-dataset_2
library(dplyr)
set.seed(333)
h_measure.list_A <- list()
h_measure.list_B <- list()
for (k in 1:50) {
  set.seed(k)
  id.good2 <- which(second.training.set$X15 == 0)
  id.bad2 <- which(second.training.set$X15 == 1)
  sample.good2 <- sample(id.good2, 182)
  sample.bad.2 <- sample(id.bad2, 155)
  training.sample <- sort(union(sample.bad.2, sample.good2))
  new.training.set <- second.training.set[,-15][training.sample, ]
  new.class.train.set <- dataset_2$X15[training.sample]
  rownames(new.training.set) <- seq(1, 337)
  new.testing.set <- second.training.set[,-15][-training.sample, ]
  rownames(new.testing.set) <- seq(338, 482)
  new.class.test.set <- dataset_2$X15[-training.sample]
  # Ball Mapper construction for various epsilons
  measure_A <- matrix(nrow=140,ncol=6)
  measure_B <- matrix(nrow=140,ncol=6)
  for (i in 1:140) {
    epsilon <- 0.39 + i/100
    result <- PROB_BM_AB(epsilon, new.training.set, new.testing.set,
                         new.class.train.set, new.class.test.set, 0.46)
    measure_A[i, 1] <- result$accurancy_A
    measure_A[i, 2:6] <- as.numeric(result$HMeasure_A$metrics[c('H', 'Precision',
'Recall', 'F','AUC')])
    colnames(measure_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
```

```
    rownames(measure_A) <- c(seq(0.40, length.out = nrow(measure_A), by = 0.01))
    measure_B[i, 1] <- result$accurancy_B
    measure_B[i, 2:6] <- as.numeric(result$HMeasure_B$metrics[c('H', 'Precision',
'Recall', 'F', 'AUC')])
    colnames(measure_B) <- c('Accurancy','H', 'Precision', 'Recall', 'F','AUC')
    rownames(measure_B) <- c(seq(0.40, length.out = nrow(measure_B), by = 0.01))
  }
  h_measure.list_A[[k]] <- as.matrix(measure_A)
  h_measure.list_B[[k]] <- as.matrix(measure_B)
 }

 h_measure.list_A
 h_measure.list_B

 ##step 5 MAX_MIN_MEAN METHODS A
 {max_elements_A <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      max_val <- h_measure.list_A[[1]][i, j]
      for (k in 2:50) {
        current_val <- h_measure.list_A[[k]][i, j]
        if (!is.na(current_val) && current_val >= max_val) {
          max_val <- current_val
        }
      }
      max_elements_A[i, j] <- max_val
      colnames(max_elements_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
      rownames(max_elements_A) <- c(seq(0.40, length.out = nrow(measure_A), by =
0.01))
    }
  }
  max_elements_A

  #step 5a to max ton max
  max_max_values_A <- matrix(NA, nrow = 1, ncol = 6)
  colnames(max_max_values_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  max_max_values_A  <- apply(max_elements_A, 2, max) #to 2 ypodilwni  kata mikos
twn stilwn, 3ekarei ana stili to max arxeio
  max_max_values_A
  max_max_values_A  <-  rownames(max_elements_A)[apply(max_elements_A,  2,
which.max)]
  max_max_values_A

  min_elements_A <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      min_val <- h_measure.list_A[[1]][i, j]
```

```
      for (k in 2:50) {
        current_val <- h_measure.list_A[[k]][i, j]
        if (!is.na(current_val) && current_val < min_val) {
          min_val <- current_val
        }
      }
      min_elements_A[i, j] <- min_val
      colnames(min_elements_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(min_elements_A) <- c(seq(0.40, length.out = nrow(measure_A), by =
0.01))
    }
  }
  min_elements_B

  #step 5a to max ton max
  min_min_values_A <- matrix(NA, nrow = 1, ncol = 6)
  colnames(min_min_values_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  min_min_values_A  <- apply(min_elements_A, 2, min) #to 2 ypodilwni  kata mikos
twn stilwn, 3ekarei ana stili to max arxeio
  min_min_values_A
  min_min_values_A  <-  rownames(min_elements_A)[apply(min_elements_A,  2,
which.min)]
  min_min_values_A

  mean_elements_A <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      # Get values from h_measure.list_A
      values_A <- sapply(h_measure.list_A, function(mat) mat[i, j])
      # Find the mean value for h_measure.list_A
      mean_elements_A[i, j] <- mean(values_A, na.rm = TRUE)
      colnames(mean_elements_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
      rownames(mean_elements_A) <- c(seq(0.40, length.out = nrow(mean_elements_A),
by = 0.01)
    }
  }
  mean_elements_A

  #step 3a to max ton max
  max_mean_values_A <- matrix(NA, nrow = 1, ncol = 6)
  colnames(max_mean_values_A) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  max_mean_values_A <- apply(mean_elements_A, 2, max) #to 2 ypodilwni  kata mikos
twn stilwn, 3ekarei ana stili to max arxeio
  max_mean_values_A
  max_row_indices_A  <-  rownames(mean_elements_A)[apply(mean_elements_A,  2,
which.max)]
  max_row_indices_A
```

```
  #graphs_a
  {
   library(ggplot2)
   #accurancy_a
   # metatropi tou pinaka se data frames
   max_df_acc_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(max_elements_A), by = 0.01),
                              Value = max_elements_A[, 1],
                              Type = "Max")

   min_df_acc_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A), by = 0.01),
                              Value = min_elements_A[, 1],
                              Type = "Min")

   mean_df_acc_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(mean_elements_A), by = 0.01),
                               Value = mean_elements_A[, 1],
                               Type = "Mean")
   # sinxwnefsi twn data frames
   combined_df_acc <- rbind(max_df_acc_a, min_df_acc_a, mean_df_acc_a)
   # dimiourgia grafimatos
   library(ggplot2)

   ggplot(combined_df_acc, aes(x = Epsilon, y = Value, color = Type)) +
    geom_line() +
    labs(title = "Max, Min, and Mean Values for the Accuracy_A",
         x = "Epsilon", y = "Value") +
    theme_minimal()

   #h_measure
   # metatropi tou pinaka se data frames
   max_df_h_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(max_elements_A), by = 0.01),
                            Value = max_elements_A[, 2],
                            Type = "Max")

   min_df_h_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A), by = 0.01),
                            Value = min_elements_A[, 2],
                            Type = "Min")

   mean_df_h_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(mean_elements_A), by = 0.01),
                             Value = mean_elements_A[, 2],
                             Type = "Mean")
```

```
    # sinxwnefsi twn data frames
    combined_df_h <- rbind(max_df_h_a, min_df_h_a, mean_df_h_a)

    # dimiourgia grafimatos
    ggplot(combined_df_h, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the HMeasure_A",
           x = "Epsilon", y = "Value") +
      theme_minimal()

    #percision
    # metatropi tou pinaka se data frames
    max_df_per_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_A), by = 0.01),
                               Value = max_elements_A[, 3],
                               Type = "Max")

    min_df_per_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A),
by = 0.01),
                               Value = min_elements_A[, 3],
                               Type = "Min")

    mean_df_per_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_A), by = 0.01),
                                Value = mean_elements_A[, 3],
                                Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_per <- rbind(max_df_per_a, min_df_per_a, mean_df_per_a)

    # dimiourgia grafimatos
    ggplot(combined_df_per, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the Percision_A",
           x = "Epsilon", y = "Value") +
      theme_minimal()

 #recall
    # metatropi tou pinaka se data frames
    max_df_rec_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_A), by = 0.01),
                               Value = max_elements_A[, 4],
                               Type = "Max")

    min_df_rec_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A),
by = 0.01),
                               Value = min_elements_A[, 4],
                               Type = "Min")
```

```
    mean_df_rec_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_A), by = 0.01),
                          Value = mean_elements_A[, 4],
                          Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_rec <- rbind(max_df_rec_a, min_df_rec_a, mean_df_rec_a)

    # dimiourgia grafimatos
    ggplot(combined_df_rec, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the Recall_A",
           x = "Epsilon", y = "Value") +
      theme_minimal()

    #f1_score
    # metatropi tou pinaka se data frames
    max_df_f1_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(max_elements_A),
by = 0.01),
                          Value = max_elements_A[, 5],
                          Type = "Max")

    min_df_f1_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A),
by = 0.01),
                          Value = min_elements_A[, 5],
                          Type = "Min")

    mean_df_f1_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_A), by = 0.01),
                          Value = mean_elements_A[, 5],
                          Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_f1 <- rbind(max_df_f1_a, min_df_f1_a, mean_df_f1_a)
    # dimiourgia grafimatos
    ggplot(combined_df_f1, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the F1_score_A",
           x = "Epsilon", y = "Value") +
      theme_minimal()

    #AUC
    # metatropi tou pinaka se data frames
    max_df_auc_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_A), by = 0.01),
                          Value = max_elements_A[, 6],
                          Type = "Max")
```

```
    min_df_auc_a <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_A),
by = 0.01),
                               Value = min_elements_A[, 6],
                               Type = "Min")

    mean_df_auc_a <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_A), by = 0.01),
                                Value = mean_elements_A[, 6],
                                Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_auc <- rbind(max_df_auc_a, min_df_auc_a, mean_df_auc_a)

    # dimiourgia grafimatos
    ggplot(combined_df_auc, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the AUC_A",
           x = "Epsilon", y = "Value") +
      theme_minimal()
  }
 }

 ##step 5 MAX_MIN_MEAN METHODS B
 {max_elements_B <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      max_val <- h_measure.list_B[[1]][i, j]
      for (k in 2:50) {
        current_val <- h_measure.list_B[[k]][i, j]
        if (!is.na(current_val) && current_val >= max_val) {
          max_val <- current_val
        }
      }
      max_elements_B[i, j] <- max_val
      colnames(max_elements_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
      rownames(max_elements_B) <- c(seq(0.40, length.out = nrow(measure_B), by =
0.01))
    }
  }
  max_elements_B

  #step 3a to max ton max
  max_max_values_B <- matrix(NA, nrow = 1, ncol = 6)
  colnames(max_max_values_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  max_max_values_B  <- apply(max_elements_B, 2, max) #to 2 ypodilwni  kata mikos
twn stilwn, 3ekarei ana stili to max arxeio
  max_max_values_B
```

```
  max_max_values_B <- rownames(max_elements_B)[apply(max_elements_B, 2, which.max)]
  max_max_values_B

  min_elements_B <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      min_val <- h_measure.list_B[[1]][i, j]
      for (k in 2:40) {
        current_val <- h_measure.list_B[[k]][i, j]
        if (!is.na(current_val) && current_val < min_val) {
          min_val <- current_val
        }
      }
      min_elements_B[i, j] <- min_val
      colnames(min_elements_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(min_elements_B) <- c(seq(0.40, length.out = nrow(measure_B), by = 0.01))
    }
  }
  min_elements_B

  #step 3a to max ton max
  min_min_values_B <- matrix(NA, nrow = 1, ncol = 6)
  colnames(min_min_values_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  min_min_values_B  <- apply(min_elements_B, 2, min) #to 2 ypodilwni  kata mikos twn stilwn, 3ekarei ana stili to max arxeio
  min_min_values_B
  min_min_values_B <- rownames(min_elements_B)[apply(min_elements_B, 2, which.min)]
  min_min_values_B

  mean_elements_B <- matrix(NA, nrow = 140, ncol = 6)
  for (i in 1:140) {
    for (j in 1:6) {
      # Get values from h_measure.list_B
      values_B <- sapply(h_measure.list_B, function(mat) mat[i, j])
      # Find the mean value for h_measure.list_B
      mean_elements_B[i, j] <- mean(values_B, na.rm = TRUE)
      colnames(mean_elements_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
      rownames(mean_elements_B) <- c(seq(0.40, length.out = nrow(mean_elements_B), by = 0.01))
    }
  }
  mean_elements_B
```

```
  #step 5a to max ton max
  max_mean_values_B <- matrix(NA, nrow = 1, ncol = 6)
  colnames(max_mean_values_B) <- c('Accuracy','H', 'Precision', 'Recall', 'F','AUC')
  max_mean_values_B  <- apply(mean_elements_B, 2, max) #to 2 ypodilwni  kata mikos twn stilwn, 3ekarei ana stili to max arxeio
  max_mean_values_B

  max_row_indices_B  <-  rownames(mean_elements_B)[apply(mean_elements_B,  2, which.max)]
  max_row_indices_B

  #graphs B
  {
   library(ggplot2)
   #accurancy_B
   # metatropi tou pinaka se data frames
   max_df_Acc_B  <-  data.frame(Epsilon  =  seq(0.40,  length.out  =  nrow(max_elements_B), by = 0.01),
                        Value = max_elements_B[, 1],
                        Type = "Max")

   min_df_Acc_B  <-  data.frame(Epsilon  =  seq(0.40,  length.out  =  nrow(min_elements_B), by = 0.01),
                        Value = min_elements_B[, 1],
                        Type = "Min")

   mean_df_Acc_B  <-  data.frame(Epsilon  =  seq(0.40,  length.out  =  nrow(mean_elements_B), by = 0.01),
                         Value = mean_elements_B[, 1],
                         Type = "Mean")
   # sinxwnefsi twn data frames
   combined_df_Acc_b <- rbind(max_df_Acc_B, min_df_Acc_B, mean_df_Acc_B)
   # dimiourgia grafimatos
   ggplot(combined_df_Acc_b, aes(x = Epsilon, y = Value, color = Type)) +
    geom_line() +
    labs(title = "Max, Min, and Mean Values for the Accuracy_B",
         x = "Epsilon", y = "Value") +
    theme_minimal()

   #h_measure
   # metatropi tou pinaka se data frames
   max_df_h_B <- data.frame(Epsilon = seq(0.40, length.out = nrow(max_elements_B), by = 0.01),
                        Value = max_elements_B[, 2],
                        Type = "Max")
```

```
    min_df_h_B <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_B),
by = 0.01),
                             Value = min_elements_B[, 2],
                             Type = "Min")

    mean_df_h_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_B), by = 0.01),
                              Value = mean_elements_B[, 2],
                              Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_h <- rbind(max_df_h_B, min_df_h_B, mean_df_h_B)

    # dimiourgia grafimatos
    ggplot(combined_df_h, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the HMeasure_B",
           x = "Epsilon", y = "Value") +
      theme_minimal()


    #percision
    # metatropi tou pinaka se data frames
    max_df_per_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_B), by = 0.01),
                               Value = max_elements_B[, 3],
                               Type = "Max")

    min_df_per_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(min_elements_B), by = 0.01),
                               Value = min_elements_B[, 3],
                               Type = "Min")

    mean_df_per_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_B), by = 0.01),
                                Value = mean_elements_B[, 3],
                                Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_per <- rbind(max_df_per_B, min_df_per_B, mean_df_per_B)
    # dimiourgia grafimatos
    ggplot(combined_df_per, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the Percision_B",
           x = "Epsilon", y = "Value") +
      theme_minimal()
    #recall
    # metatropi tou pinaka se data frames
```

```
    max_df_rec_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_B), by = 0.01),
                           Value = max_elements_B[, 4],
                           Type = "Max")

    min_df_rec_B <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_B),
by = 0.01),
                           Value = min_elements_B[, 4],
                           Type = "Min")

    mean_df_rec_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_B), by = 0.01),
                            Value = mean_elements_B[, 4],
                            Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_rec <- rbind(max_df_rec_B, min_df_rec_B, mean_df_rec_B)

    # dimiourgia grafimatos
    ggplot(combined_df_rec, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the Recall_B",
           x = "Epsilon", y = "Value") +
      theme_minimal()
    #f1_score
    # metatropi tou pinaka se data frames
    max_df_f1_B <- data.frame(Epsilon = seq(0.40, length.out = nrow(max_elements_B),
by = 0.01),
                          Value = max_elements_B[, 5],
                          Type = "Max")

    min_df_f1_B <- data.frame(Epsilon = seq(0.40, length.out = nrow(min_elements_B),
by = 0.01),
                          Value = min_elements_B[, 5],
                          Type = "Min")

    mean_df_f1_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_B), by = 0.01),
                            Value = mean_elements_B[, 5],
                            Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_f1 <- rbind(max_df_f1_B, min_df_f1_B, mean_df_f1_B)
    # dimiourgia grafimatos
    ggplot(combined_df_f1, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the F1_score_B",
           x = "Epsilon", y = "Value") +
      theme_minimal()
```

```
      #AUC
    # metatropi tou pinaka se data frames
    max_df_Auc_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(max_elements_B), by = 0.01),
                          Value = max_elements_B[, 6],
                          Type = "Max")

    min_df_Auc_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(min_elements_B), by = 0.01),
                          Value = min_elements_B[, 6],
                          Type = "Min")

    mean_df_Auc_B <- data.frame(Epsilon = seq(0.40, length.out =
nrow(mean_elements_B), by = 0.01),
                           Value = mean_elements_B[, 6],
                           Type = "Mean")
    # sinxwnefsi twn data frames
    combined_df_Auc <- rbind(max_df_Auc_B, min_df_Auc_B, mean_df_Auc_B)

    # dimiourgia grafimatos
    ggplot(combined_df_Auc, aes(x = Epsilon, y = Value, color = Type)) +
      geom_line() +
      labs(title = "Max, Min, and Mean Values for the AUC_B",
           x = "Epsilon", y = "Value") +
      theme_minimal()
  }
 }

 max_result_1_A
 max_result_B

 #Step 4 CROSS VALIDATION
 ## Ball Mapper cross validation ##
 id.good <- which(class == 0)
 id.bad <- which(class == 1)
 # Δημιουργία των πινάκων
 epsilon.methodA <- matrix(NA, nrow = 40, ncol = 6)
 epsilon.methodB <- matrix(NA, nrow = 40, ncol = 6)
 max_values_mean_A <- matrix(NA, nrow = 40, ncol = 6)
 max_values_mean_B <- matrix(NA, nrow = 40, ncol = 6)

 for (n in 37:40) {
   set.seed(n)
   sample.good <- sample(id.good, 260)
   sample.bad <- sample(id.bad, 222)
   sample <- sort(union(sample.bad, sample.good))
```

```
  training.cv <- norm_data[, -15][sample, ]
  class.train.cv <- class[sample]
  rownames(training.cv) <- seq(1, 482)
  testing.cv <- norm_data[, -15][-sample, ]
  rownames(testing.cv) <- seq(483, 690)
  class.test.cv <- class[-sample]

  measure.list.cv.A <- list()
  measure.list.cv.B <- list()

  for (k in 1:10) {
    set.seed(k + 100)
    id.good2 <- which(class.train.cv == 0)
    id.bad2 <- which(class.train.cv == 1)
    sample.good.cv <- sample(id.good2, 182)
    sample.bad.cv <- sample(id.bad2, 155)
    training.sample.cv <- sort(union(sample.bad.cv, sample.good.cv))
    new.training.set.cv <- training.cv[training.sample.cv, ]
    new.class.train.set.cv <- class.train.cv[training.sample.cv]
    rownames(new.training.set.cv) <- seq(1, 337)
    new.testing.set.cv <- training.cv[-training.sample.cv, ]
    rownames(new.testing.set.cv) <- seq(338, 482)
    new.class.test.set.cv <- class.train.cv[-training.sample.cv]

    measure_cv_A <- matrix(nrow = 25, ncol = 6)
    measure_cv_B <- matrix(nrow = 25, ncol = 6)

    for (m in 1:25) {
      epsilon <- 0.55 + m / 20
      x <- PROB_BM_AB(epsilon, new.training.set.cv, new.testing.set.cv,
new.class.train.set.cv, new.class.test.set.cv, 0.46)
      measure_cv_A[m, 1] <- x$accurancy_A
      measure_cv_A[m, 2:6] <- as.numeric(x$HMeasure_A$metrics[c('H', 'Precision',
'Recall', 'F', 'AUC')])
      colnames(measure_cv_A) <- c('Accuracy', 'H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(measure_cv_A) <- seq(0.55, length.out = nrow(measure_cv_A), by =
0.05)

      measure_cv_B[m, 1] <- x$accurancy_B
      measure_cv_B[m, 2:6] <- as.numeric(x$HMeasure_B$metrics[c('H', 'Precision',
'Recall', 'F', 'AUC')])
      colnames(measure_cv_B) <- c('Accurancy', 'H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(measure_cv_B) <- seq(0.55, length.out = nrow(measure_cv_B), by =
0.05)
    }

    measure.list.cv.A[[k]] <- as.matrix(measure_cv_A)
```

```
    measure.list.cv.B[[k]] <- as.matrix(measure_cv_B)
  }

  mean_elements_cv_A <- matrix(NA, nrow = 25, ncol = 6)
  mean_elements_cv_B <- matrix(NA, nrow = 25, ncol = 6)

  for (i in 1:25) {
    for (j in 1:6) {
      values_cv_A <- sapply(measure.list.cv.A, function(mat) mat[i, j])
      values_cv_B <- sapply(measure.list.cv.B, function(mat) mat[i, j])

      mean_elements_cv_A[i, j] <- mean(values_cv_A, na.rm = TRUE)
      mean_elements_cv_B[i, j] <- mean(values_cv_B, na.rm = TRUE)

      colnames(mean_elements_cv_A) <- c('Accuracy', 'H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(mean_elements_cv_A) <- seq(0.55, length.out =
nrow(mean_elements_cv_A), by = 0.05)

      colnames(mean_elements_cv_B) <- c('Accuracy', 'H', 'Precision', 'Recall', 'F', 'AUC')
      rownames(mean_elements_cv_B) <- seq(0.55, length.out =
nrow(mean_elements_cv_B), by = 0.05)
    }

    max_mean_values_A <- apply(mean_elements_cv_A, 2, max)
    max_values_mean_A[n, ] <- as.numeric(max_mean_values_A)

    max_mean_values_B <- apply(mean_elements_cv_B, 2, max)
    max_values_mean_B[n, ] <- as.numeric(max_mean_values_B)
  }

  epsilon.methodA[n, ] <-
rownames(mean_elements_cv_A)[apply(mean_elements_cv_A, 2, which.max)]
  epsilon.methodB[n, ] <-
rownames(mean_elements_cv_B)[apply(mean_elements_cv_B, 2, which.max)]
 }
 max_values_mean_A
 max_values_mean_B
 epsilon.methodA
 epsilon.methodB
 #histogram
 hist(as.numeric(max_values_mean_B[, 1]),
      breaks = 10,
      main = "Ιστόγραμμα Συχνοτήτων Accuracy_B",
      xlab = "Epsilon",
      col = "skyblue",  # Χρώμα των μπάρων
      border = "black",  # Χρώμα του περιγράμματος των μπάρων
```

```
    xlim = c(min(as.numeric(max_values_mean_B[, 1])),
max(as.numeric(max_values_mean_B[, 1]))) # Ορισμός ορίων του άξονα x
 )
 hist(as.numeric(max_values_mean_B[, 2]),
    breaks = 10,
    main = "Ιστόγραμμα Συχνοτήτων Hmeasure_B",
    xlab = "Epsilon",
    col = "skyblue", # Χρώμα των μπάρων
    border = "black", # Χρώμα του περιγράμματος των μπάρων
    xlim = c(min(as.numeric(max_values_mean_B[, 2])),
max(as.numeric(max_values_mean_B[, 2]))) # Ορισμός ορίων του άξονα x
 )
 hist(as.numeric(max_values_mean_B[, 3]),
    breaks = 10,
    main = "Ιστόγραμμα Συχνοτήτων Percision_B",
    xlab = "Epsilon",
    col = "skyblue", # Χρώμα των μπάρων
    border = "black", # Χρώμα του περιγράμματος των μπάρων
    xlim = c(min(as.numeric(max_values_mean_B[, 3])),
max(as.numeric(max_values_mean_B[, 3]))) # Ορισμός ορίων του άξονα x
 )
 hist(as.numeric(max_values_mean_B[, 4]),
    breaks = 10,
    main = "Ιστόγραμμα Συχνοτήτων Recall_B",
    xlab = "Epsilon",
    col = "skyblue", # Χρώμα των μπάρων
    border = "black", # Χρώμα του περιγράμματος των μπάρων
    xlim = c(min(as.numeric(max_values_mean_B[, 4])),
max(as.numeric(max_values_mean_B[, 4]))) # Ορισμός ορίων του άξονα x
 )
 hist(as.numeric(max_values_mean_B[, 5]),
    breaks = 10,
    main = "Ιστόγραμμα Συχνοτήτων F1_score_B",
    xlab = "Epsilon",
    col = "skyblue", # Χρώμα των μπάρων
    border = "black", # Χρώμα του περιγράμματος των μπάρων
    xlim = c(min(as.numeric(max_values_mean_B[, 5])),
max(as.numeric(max_values_mean_B[, 5]))) # Ορισμός ορίων του άξονα x
 )
 hist(as.numeric(max_values_mean_B[, 6]),
    breaks = 10,
    main = "Ιστόγραμμα Συχνοτήτων AUC_B",
    xlab = "Epsilon",
    col = "skyblue", # Χρώμα των μπάρων
    border = "black", # Χρώμα του περιγράμματος των μπάρων
    xlim = c(min(as.numeric(max_values_mean_B[, 6])),
max(as.numeric(max_values_mean_B[, 6]))) # Ορισμός ορίων του άξονα x
```

```
)
##Step 5
#LOGISTIC REGRESSION##
library(caTools)
training_set<- norm_data[train_indices, ]
logistic_model<-
glm(training_set$X15~training_set$X1+training_set$X2+training_set$X3+training_set
$X4+training_set$X5+training_set$X6+training_set$X7+training_set$X8+training_set
$X9+training_set$X10+training_set$X11+training_set$X12+training_set$X13+training
_set$X14,data=training_set,family="binomial")
summary(logistic_model)
#predict test data
test.set<- norm_data_2[-train_indices, ]
class.test<- class[-train_indices]
p<-
logistic_model$coefficients[1]+as.matrix(test.set)%*%logistic_model$coefficients[-1]
predicted<-exp(p)/(1+exp(p))
predicted_labels_A <- ifelse(as.numeric(predicted) > 0.5 , 1, 0)
#HMeasure logistic regression
HMeasure_lg <-HMeasure(class.test,as.numeric(predicted),severity.ratio = 0.5)
HMeasure_lg$metrics[c('H', 'Precision', 'Recall', 'F','AUC')]
accurancy_A_lg <-Metrics::accuracy(predicted_labels_A,class.test)
accurancy_A_lg

#neural_networks
library(tidyverse)
library(neuralnet)
train_data <- norm_data[train_indices, ]
test_data <- norm_data[-train_indices,]
classification_nn <- neuralnet(X15 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
X10 + X11 + X12 + X13 + X14,
                               data = train_data,
                               linear.output = FALSE,
                               likelihood = TRUE,
                               err.fct = 'ce')
predicted_probs_nn <- predict(classification_nn, newdata = test_data, type = "response")
predicted_labels_nn <- ifelse(predicted_probs_nn > 0.5, 1, 0)
HMeasure_nn <- HMeasure(class.test, predicted_probs_nn, severity.ratio = 0.5)
metrics_nn <- HMeasure_nn$metrics[c('H', 'Precision', 'Recall', 'F', 'AUC')]
print(metrics_nn)
accuracy_nn <- sum(predicted_labels_nn == class.test) / length(class.test)
print(paste("Accuracy: ", round(accuracy_nn, 3)))

# Support Vector Machines (SVM)
library(e1071)
svm_model <- svm(X15 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
X11 + X12 + X13 + X14, data = training_set, kernel = "linear", cost = 1)
```

```
predicted_svm <- predict(svm_model, newdata = test.set)
predicted_svm_ <-ifelse(predicted_svm > 0.5, 1, 0)
accurancy_svm <- Metrics::accuracy(class.test, as.numeric(predicted_svm_))
accurancy_svm
HMeasure_svm <-HMeasure(class.test, as.numeric(predicted_svm_))
HMeasure_svm$metrics[c('H', 'Precision', 'Recall', 'F','AUC')].
```