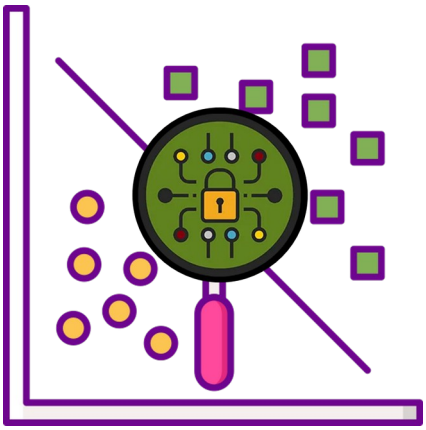
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΩΝ ΟΜΑΔΟΠΟΙΗΣΗΣ ΜΕ ΣΚΟΠΟ ΤΗΝ ΕΠΙΤΕΥΞΗ ΑΝΩΝΥΜΙΑΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ ΤΩΝ ΠΡΑΓΜΑΤΩΝ



ΚΛΕΒΕΣΤ ΠΑΛΟΥΤΣΑΙ, 888

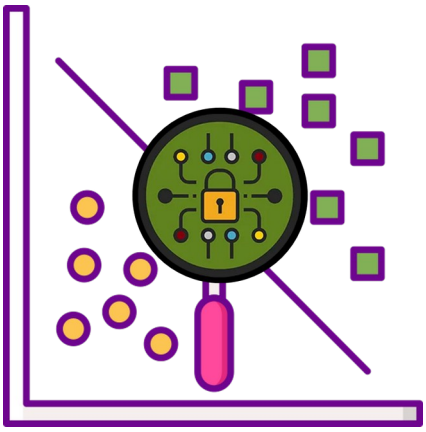
ΕΠΙΒΛΕΠΩΝ: ΕΠΙΚ. ΚΑΘΗΓΗΤΗΣ, ΠΑΝΑΓΙΩΤΗΣ ΣΑΡΗΓΙΑΝΝΙΔΗΣ

ΚΟΖΑΝΗ, ΙΟΥΛΙΟΣ 2020

UNIVERSITY OF WESTERN MACEDONIA
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DIPLOMA DISSERTATION

IMPLEMENTATION AND COMPARATIVE STUDY OF CLUSTERING ALGORITHMS ON PROVISIONING ANONYMITY IN THE INTERNET OF THINGS



KLEVEST PALUCAJ, 888

SUPERVISOR: ASS. PROFESSOR, PANAGIOTIS SARIGIANNIDIS

KOZANI, JULY 2020

Περίληψη

Στη παρούσα διπλωματική εργασία αναλύεται η ανάγκη για τη διασφάλιση της ιδιωτικότητας και της ανωνυμίας των χρηστών στο Διαδίκτυο των Πραγμάτων (IoT) και μελετώνται τέσσερις αλγόριθμοι συσταδοποίησης οι οποίοι συμβάλλουν στη διασφάλιση της ιδιωτικότητας των δεδομένων.

Χρησιμοποιώντας το προγραμματιστικό εργαλείο της Matlab, υλοποιήθηκαν οι αλγόριθμοι K-means, DBSCAN (Density Based Spatial Clustering of Applications with Noise) και PADC (Privacy and Availability Data Clustering), ενώ στα πλαίσια της διπλωματικής εργασίας αναπτύχθηκε ένα νέο αλγοριθμικό μοντέλο (Utilized Outlier Effect-Density Based Spatial Clustering of Applications with Noise - UOE-DBSCAN) που συμβάλλει στην παροχή ανωνυμίας των δεδομένων των χρηστών.

Αρχικά, παρουσιάζεται ο χώρος του Διαδικτύου των Πραγμάτων και περιγράφονται τα χαρακτηριστικά και οι ιδιότητές του. Αναλύονται τα επίπεδα αρχιτεκτονικής και οι τεχνολογίες που χρησιμοποιούνται στο Διαδίκτυο των Πραγμάτων και αναφέρονται ορισμένα θέματα ασφάλειας που υπάρχουν σε αυτό.

Στη συνέχεια, περιγράφονται οι βασικές τεχνικές διασφάλισης της ιδιωτικότητας στο Διαδίκτυο των Πραγμάτων k-Ανωνυμία, l-diversity, t-closeness και ε-Διαφορική Ιδιωτικότητα.

Στα επόμενα δύο κεφάλαια, γίνεται μια αναλυτική περιγραφή των αλγορίθμων συσταδοποίησης με τους οποίους αποκρύπτονται τα δεδομένα και η ταυτότητα των χρηστών από κακόβουλους παρατηρητές. Οι αλγόριθμοι που αναλύονται είναι ο K-means, ο DBSCAN, ο αλγόριθμος ιεραρχικής συσταδοποίησης και ο PADC. Επίσης, παρουσιάζεται ένα νέο, προτεινόμενο αλγοριθμικό σχήμα που ονομάζεται UOE-DBSCAN και γίνεται αναφορά στα ιδιαίτερα χαρακτηριστικά του.

Τέλος, πραγματοποιείται σύγκριση των παραπάνω αλγορίθμων με τον νέο αλγόριθμο ως προς τις ακόλουθες μετρικές σύγκρισης της αποδοτικότητας: Precision, Recall και F-measure.

Λέξεις Κλειδιά: Διαδίκτυο των Πραγμάτων; IoT; Ανωνυμία; Αλγόριθμοι Συσταδοποίησης; k-Ανωνυμία; ε-Διαφορική Ιδιωτικότητα; K-means; DBSCAN.

Abstract

This diploma dissertation focuses on privacy and security issues regarding the Internet of Things (IoT) paradigm and evaluates the performance of four clustering algorithms towards achieving data anonymization.

By utilizing Matlab, K-means, DBSCAN (Density Based Spatial Clustering of Applications with Noise) and PADC (Privacy and Availability Data Clustering) algorithms were implemented while a new algorithmic model (Utilized Outlier Effect-Density Based Spatial Clustering of Applications with Noise - UOE-DBSCAN) has also been developed as part of the dissertation, which helps to increase the security of user data.

At first, the Internet of Things paradigm was analyzed based on its specific characteristics. Then, an architectural point of view was presented, as well as according technologies for IoT deployment. Moreover, focus was given on existing IoT security issues and vulnerabilities.

Continuing, a description is provided regarding the most well-known privacy-preservation techniques, namely as k-anonymity, l-diversity, t-closeness and ε-differential privacy.

In the next two chapters, privacy-preserving algorithms are presented and discussed focusing on identity disclosure and safeguarding user data from cyber-attackers. The evaluation process considers the following algorithms: K-means, DBSCAN, Hierarchical clustering and PADC algorithm. Additionally, a novel algorithmic model, UOE-DBSCAN is proposed and described.

Finally, a comparison is provided between the novel algorithm scheme and the other models in terms of the following metrics regarding performance efficiency: Precision, Recall and F-measure.

Keywords: Internet of Things; IoT; Anonymity; Clustering Algorithms; k-anonymity; ε-Differential Privacy; K-means; DBSCAN.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω την οικογένειά μου για τη στήριξη που μου πρόσφερε καθ' όλη τη διάρκεια εκπόνησης της εργασίας. Επίσης τους φίλους που έκανα και που με έκαναν να περάσω αξέχαστες στιγμές κατά τη διάρκεια των σπουδών μου. Ιδιαίτερες ευχαριστίες στην κοπέλα μου η οποία ήταν πάντα εκεί να με συντροφεύσει και να με ωθήσει ψυχολογικά σε κάθε στάδιο της εκπόνησης της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου κύριο Παναγιώτη Σαρηγιαννίδη και την Άννα Τριανταφύλλου, υποψήφια διδάκτορα στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών για τις συμβουλές τους και τη καθοδήγησή τους.

Πίνακας περιεχομένων

1. Εισαγωγή

1.1 Εισαγωγή στο Διαδίκτυο των Πραγμάτων

Η σύγχρονη εποχή χαρακτηρίζεται από πλήθος τεχνολογικών ανακαλύψεων που επιβάλλουν την άμεση και γρήγορη ανταλλαγή δεδομένων μέσω του Διαδικτύου. Αξιοσημείωτη και αναγκαία κρίνεται πλέον η χρήση έξυπνων ηλεκτρονικών συσκευών στην καθημερινότητα. Βασικό ρόλο κατέχουν τα έξυπνα κινητά τηλέφωνα, οι έξυπνες τηλεοράσεις και άλλες οικιακές ηλεκτρονικές συσκευές, τα ηλεκτρονικά περικάρπια για παρακολούθηση φυσικής κατάστασης (smartwatches), όπως επίσης και ηλεκτρονικές συσκευές που ορίζουν την κυκλοφορία (smart city), το παρκάρισμα των αυτοκινήτων (smart parking), τη φωταγώγηση της πόλης, την διακίνηση ηλεκτρικού ρεύματος (smart grid) και άλλες πτυχές της καθημερινότητας για τη διευκόλυνση και την αναβάθμιση της ποιότητας ζωής των πολιτών.

Οι σύγχρονες αυτές ανάγκες οδήγησαν στον σχεδιασμό νέων τεχνολογιών για τη διασύνδεση και επικοινωνία των έξυπνων συσκευών, στοχεύοντας στη διατήρηση της ποιότητας υπηρεσιών και της αξιοπιστίας στην ανταλλαγή πληροφοριών. Το Διαδίκτυο των Πραγμάτων (Internet of Things - IoT) ως όρος διατυπώθηκε πρώτη φορά απο τον Kevin Aston στα τέλη της δεκαετίας του 1990 [1]. Η ιδέα του στηριζόταν στη διασύνδεση των συσκευών που θα αποτελούσαν το Διαδίκτυο των Πραγμάτων μέσω της αναγνώρισης με ραδιοσυχνότητες (Radio Frequency Identification - RFID), όπου κάθε συσκευή θα είχε ένα μοναδικό αναγνωριστικό.

Το Διαδίκτυο των Πραγμάτων είναι μία από τις πιο ανερχόμενες τεχνολογικές εξελίξεις [2]. Το Διαδίκτυο των Πραγμάτων αποτελεί ένα σύστημα-δίκτυο στο οποίο μπορούν να συνδεθούν διαφόρων ειδών συσκευές, όπως οικιακές συσκευές, συσκευές ελέγχου άσκησης και υγείας (πχ. smartwatches), κινητά, φορητοί υπολογιστές, αυτοκίνητα και γενικά αντικείμενα που διαθέτουν ηλεκτρονικά εξαρτήματα με τα οποία μπορούν να μεταφέρουν πληροφορίες και δεδομένα μέσω ενός δικτύου χωρίς να υπάρχει η παρέμβαση του ανθρώπινου παράγοντα. Δηλαδή, οποιαδήποτε επικοινωνία και ανταλλαγή δεδομένων μεταξύ των συσκευών δε θα γίνεται με τη παρέμβαση του ανθρώπου-χρήστη αυτή καθαυτή. Εν αντιθέσει, θα γίνεται μέσω ορισμένων τεχνολογιών οι οποίες θα επιτρέπουν την επικοινωνία μεταξύ των συσκευών με τη χρήση ασύρματης τεχνολογίας, όπως είναι το BlueTooth, το WiFi, το RFID αλλά και με χρήση ενσύρματης τεχνολογίας όπως είναι οι συσκευές οι οποίες συνδέονται μεταξύ τους ή μέσω τρίτων χρησιμοποιώντας καλώδια.

1.2 Αρχιτεκτονική του Διαδικτύου των Πραγμάτων

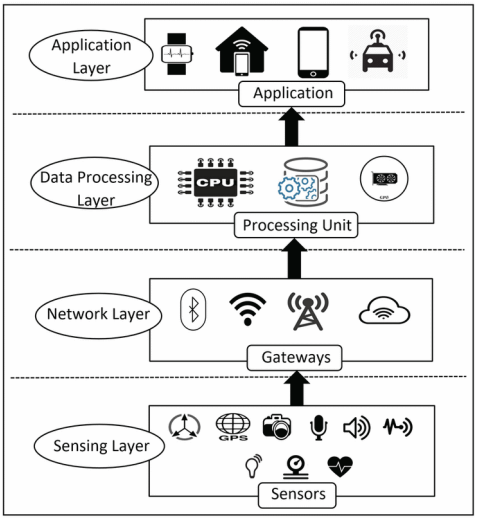
Το Διαδίκτυο των Πραγμάτων βασίζεται σε έξυπνες συσκευές και αισθητήρες για να συλλέξει δεδομένα και στη συνέχεια ορίζει την επεξεργασία, αλλά και την ανάλυσή τους για την εξαγωγή υπηρεσιών και συμπερασμάτων που θα είναι συμβατά με τις επιθυμίες των χρηστών. Βασικό συστατικό του Διαδικτύου των Πραγμάτων αποτελεί το Νέφος (Cloud), για την προσωρινή αποθήκευση των δεδομένων ενδιαμέσω διεργασιών [3]. Η επικοινωνία των έξυπνων συσκευών ορίζεται μέσω διαφόρων μοτίβων και τεχνολογιών στο Διαδίκτυο των Πραγμάτων, έχοντας ως στόχο την επίτευξη συμβατότητας και αξιοπιστίας στη μετάδοση πληροφοριών. Η ανταλλαγή πληροφοριών στηρίζεται στις ακόλουθες μεθόδους και τεχνολογίες:

- Μέσω BlueTooth.
- Μέσω WiFi.
- Με χρήση δορυφορικής σύνδεσης.
- Via ethernet σύνδεσης.

Η κάθε τεχνολογία χαρακτηρίζεται από συγκεκριμένα πλεονεκτήματα και μειονεκτήματα και η απόδοσή της εξαρτάται από το είδος και το μέγεθος της εφαρμογής που θέλει να εξυπηρετήσει. Μετά την αποστολή των συλλεγόμενων στοιχείων στην ανάλογη online πλατφόρμα ή εξυπηρετητή, τα δεδομένα δέχονται κάποιου είδους επεξεργασία ανάλογα με τις πληροφορίες που απαιτούνται να εξαχθούν. Βέβαια, υπάρχουν και κάποιες περιπτώσεις στις οποίες πρέπει να ενημερωθεί ο χρήστης, ανάλογα με τα αποτελέσματα της επεξεργασίας αυτής. Ένα παράδειγμα είναι η περίπτωση στην οποία ένας αισθητήρας στο σπίτι ενός χρήστη στέλνει δεδομένα τα οποία υποδηλώνουν ότι κάποιος έχει μπει στο σπίτι του χρήστη, οπότε και πρέπει να του σταλεί ειδοποίηση.

1.2.1 Επίπεδα Αρχιτεκτονικής

Στην Εικόνα 1 απεικονίζεται η αρχιτεκτονική δομή του Διαδικτύου των Πραγμάτων [4]. Τα αρχιτεκτονικά επίπεδα που ορίζονται για την ανταλλαγή δεδομένων και την επεξεργασία των συλλεγόμενων πληροφοριών μέχρι τον τελικό χρήστη είναι τα ακόλουθα [5] :



Εικόνα 1 - Αρχιτεκτονική του Διαδικτύου των Πραγμάτων [6].

- **<u>1ο Επίπεδο – Επίπεδο Συλλογής Δεδομένων:</u>** Το επίπεδο αυτό αποτελείται από όλα τα όργανα με τα οποία γίνεται εφικτή η καταγραφή πληροφοριών και η συλλογή δεδομένων, οποιαδήποτε είδους έξυπνη συσκευή όπως είναι για παράδειγμα το παγκόσμιο σύστημα τοποθεσίας (Global Positioning System - GPS), το μικρόφωνο, η κάμερα. Οι έξυπνες συσκευές βασίζονται σε διαφορετικών ειδών αισθητήρες για την συλλογή πληροφοριών απο το περιβάλλον. Οι αισθητήρες έχουν συνήθως χαμηλή υπολογιστική ισχύ και μέγεθος. Η λειτουργία τους βασίζεται σε μπαταρίες και σε τεχνικές εξοικονόμησης ενέργειας για την αποτελεσματικότητά τους.

- **<u>2ο Επίπεδο – Επίπεδο Δικτύου:</u>** Το συγκεκριμένο επίπεδο έχει τις παρακάτω λειτουργίες:

    1. Συλλογή δεδομένων από τις αισθητήριες συσκευές του επιπέδου συλλογής δεδομένων.
    2. Δρομολόγηση δεδομένων μεταξύ κομβικών σημείων του δικτύου.
    3. Μετάδοση πληροφοριών μέσω πρωτοκόλλων μεταφοράς και ενσωμάτωση διευθύνσεων πρωτοκόλλου διαδικτύου (Internet Protocol - IP).
    4. Αποθήκευση δεδομένων στον επιθυμητό εξυπηρετητή του δικτύου ή απευθείας αποθήκευση σε μια βάση δεδομένων.

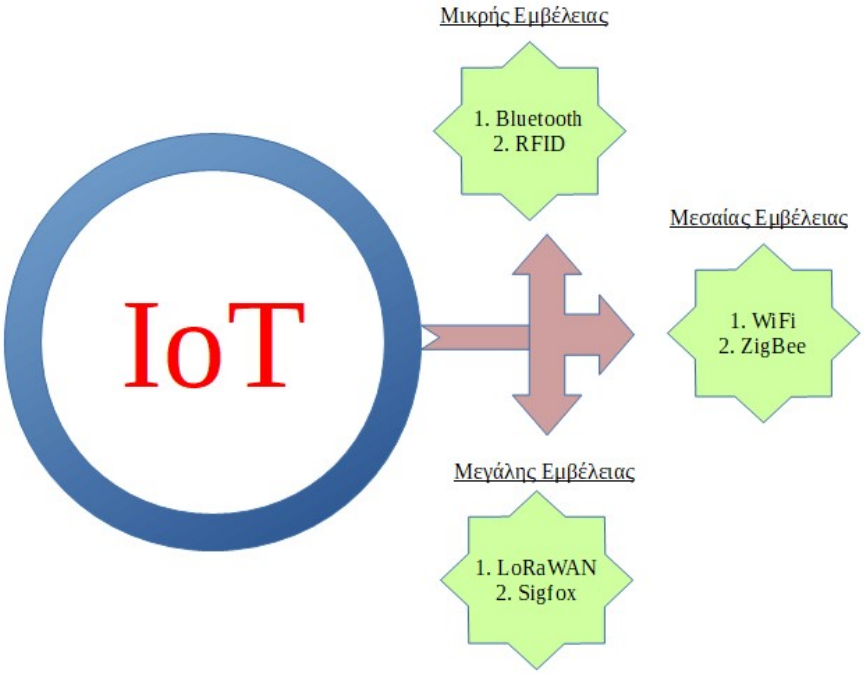
- **3ο Επίπεδο – Επίπεδο Επεξεργασίας Δεδομένων:** Το επίπεδο επεξεργασίας δεδομένων περιλαμβάνει όλες τις διαθέσιμες πλατφόρμες στις οποίες μπορούν να σταλούν τα δεδομένα από το επίπεδο δικτύου και στις οποίες πλατφόρμες γίνεται η κατάλληλη επεξεργασία των δεδομένων. Επίσης, στο συγκεκριμένο επίπεδο εμπεριέχονται διάφορα μοτίβα και τρόποι μορφοποίησης δεδομένων, όπως και εξειδικευμένοι αλγόριθμοι για κατηγοριοποίηση των συλλεγόμενων πληροφοριών στοχεύοντας στην παραγωγή μοτίβων συμπεριφοράς.

- **4ο Επίπεδο – Επίπεδο Εφαρμογής:** Το επίπεδο αυτό είναι το τελευταίο, είναι το επίπεδο που αλληλεπιδρά με το χρήστη και η χρησιμότητά του έγκειται στο γεγονός ότι χρησιμοποιεί τα επεξεργασμένα δεδομένα που λαμβάνει από το επίπεδο επεξεργασίας δεδομένων για να καθοδηγήσει και να ενημερώσει το χρήστη σχετικά με τα αποτελέσματα των δεδομένων.

1.3 Βασικές Τεχνολογίες του Διαδικτύου των Πραγμάτων

Το Διαδίκτυο των Πραγμάτων απαιτεί μια ποικιλομορφία τεχνολογιών και μεθόδων επικοινωνίας μεταξύ των έξυπνων συσκευών. Για την εξυπηρέτηση των αναγκών του έχει αποδειχθεί ότι μπορεί να χρησιμοποιηθούν υπάρχουσες τεχνολογίες διαφόρων εμβελειών. Παρόλα αυτά έχουν ήδη αναπτυχθεί νέα επικοινωνιακά πρωτόκολλα με συγκεκριμένες προϋποθέσεις και χαρακτηριστικά ώστε να συμβάλλουν στην υλοποίησή του.

Στο παρακάτω σχεδιάγραμμα, παρατηρούνται κάποιες απ' αυτές τις τεχνολογίες και κάποιες οι οποίες είναι ακόμη καινούργιες στο χώρο των δικτύων και επικοινωνιών [7].



Εικόνα 2 - Σχεδιάγραμμα τεχνολογιών που χρησιμοποιούνται στο Διαδίκτυο των Πραγμάτων.

Οι μικρής εμβέλειας τεχνολογίες περιλαμβάνουν τις ακόλουθες :

1. Το BlueTooth, τεχνολογία η οποία θεωρείται αρκετά σημαντική, κυρίως σε είδη τεχνολογίας που αφορούν τα κινητά τηλέφωνα (smartphones) εξαιτίας της ευρείας ενσωμάτωσής του σε αυτά. Το πρωτόκολλο BlueTooth χαμηλής ενέργειας (BlueTooth Low-Energy - BLE), παρά την χρησιμότητά του εξαιτίας του μικρού κόστους και της μικρής κατανάλωσης ενέργειας, δεν αποδίδει αποτελεσματικά όταν μεταφέρει συνεχώς μεγάλες ποσότητες δεδομένων.

2. Το RFID θεωρείται μία από τις πρώτες τεχνολογίες του Διαδικτύου των Πραγμάτων που υλοποιήθηκαν. Η τωρινή χρήση του εστιάζεται κυρίως σε εφαρμογές που αφορούν υπηρεσίες εντοπισμού και καταγραφής αντικειμένων σε κάποια επιχείρηση. Στο μέλλον, βέβαια, η χρήση του θα είναι πιο σημαντική καθώς, για παράδειγμα, θα είναι δυνατή η αξιοποίηση αυτής της τεχνολογίας για να εντοπίζονται και να ακολουθούνται οι κινήσεις των ασθενών ή για να καταγράφεται συνεχώς το απόθεμα ενός καταστήματος για να μην τελειώσει το εμπόρευμα.

Στις μεσαίας εμβέλειας τεχνολογίες συμπεριλαμβάνονται οι παρακάτω:

1. Το WiFi είναι ίσως το πιο γνωστό και διαδεδομένο είδος ασύρματης τεχνολογίας που βασίζεται στο πρωτόκολλο IEEE 802.11. Η χρήση του WiFi στο Διαδίκτυο των Πραγμάτων περιορίζεται εξαιτίας της αρκετά υψηλής κατανάλωσης ενέργειας που απαιτείται για την γρήγορη μεταφορά δεδομένων με φόντο την επίτευξη αξιοπιστίας και καλύτερης συνδεσιμότητας. Όντας μία σημαντική τεχνολογία στην ανάπτυξη του Διαδικτύου των Πραγμάτων, η εμβέλεια του WiFi παρέχει λύσεις σε αρκετά ζητήματα του Διαδικτύου του Πραγμάτων αλλά θα πρέπει να χρησιμοποιηθεί με τρόπους τέτοιους ώστε να επωφελούνται όλοι όσοι θα συμμετέχουν σε αυτό.

2. Το ZigBee είναι μια αρκετά γνωστή τεχνολογία στο χώρο του Διαδικτύου των Πραγμάτων, καθώς λειτουργεί χρησιμοποιώντας μια τοπολογία δικτύου που ονομάζεται mesh network. Αυτή η τοπολογία εκφράζει επακριβώς αυτό που ζητείται να συμβαίνει σε ένα έξυπνο σπίτι. Πιο συγκεκριμένα, σε ένα έξυπνο σπίτι, οι συσκευές πρέπει να επικοινωνούν μεταξύ τους ασύρματα, διαδικασία η οποία είναι η βασική λειτουργία ενός mesh δικτύου. Η γρήγορη μεταφορά δεδομένων και η αξιοπιστία είναι αυτό που κάνουν το ZigBee να είναι στις κορυφαίες επιλογές για την εφαρμογή του στο Διαδίκτυο των Πραγμάτων.

Τέλος, στις τεχνολογίες μεγάλης εμβέλειας μπορούν να συμπεριληφθούν οι εξής :

1. Το μεγάλης εμβέλειας ευρείας περιοχής δίκτυο (Long Range Wide-Area Network - LoRaWAN) αποτελεί ένα πρωτόκολλο χαμηλής κατανάλωσης ενέργειας, το οποίο υλοποιήθηκε με σκοπό την υποστήριξη μεγάλων δικτύων στα οποία συμμετέχει ένας μεγάλος αριθμός από συσκευές. Το LoRaWAN στοχεύει σε εφαρμογές δικτύων μεγάλης εμβέλειας (Wide Area Networks - WANs) και είναι σχεδιασμένο με τέτοιο τρόπο ώστε να διαθέτει τα απαραίτητα χαρακτηριστικά για να υποστηρίξει χαμηλού κόστους, κινητή και ασφαλής επικοινωνία.

2. Το Sigfox δημιουργήθηκε με σκοπό να παρέχει μια αποτελεσματική λύση στη διασύνδεση μεταξύ χαμηλής κατανάλωσης μηχανής σε μηχανή (machine to machine - M2M) εφαρμογών για τις οποίες η εμβέλεια του WiFi δεν είναι αρκετά μεγάλη. Το Sigfox

χρησιμοποιεί την UNB (Ultra NarrowBand) τεχνολογία η οποία επιτρέπει τη διαχείριση ταχυτήτων μεταφοράς δεδομένων της τάξεως των 10-1000 bps. Η κατανάλωση ενέργειας στο Sigfox είναι πολύ χαμηλή (σχεδόν 100 φορές χαμηλότερη από την κατανάλωση στις κινητές επικοινωνίες). Ταυτόχρονα προσφέρει ένα ισχυρό, αποδοτικό και ευμετάβλητο δίκτυο το οποίο μπορεί να υποστηρίξει επικοινωνία μεταξύ συσκευών τα οποία απέχουν μεταξύ τους πολλά τετραγωνικά χιλιόμετρα.

1.4 Προβλήματα ασφαλείας και προστασίας ιδιωτικότητας στο Διαδίκτυο των Πραγμάτων

Με τη συνεχόμενη ανάπτυξη και δημιουργία έξυπνων συσκευών που μπορούν να συλλέξουν και να ανταλλάξουν αυτόματα προσωπικά δεδομένα για την αποτελεσματικότερη και πιο γρήγορη εξυπηρέτηση των χρηστών, αυξάνονται ταυτόχρονα και οι ανησυχίες για το αν αυτές οι συσκευές είναι αρκετά ασφαλείς ώστε να αποτρέψουν κακόβουλους χρήστες από το να αποκτήσουν αυτά τα δεδομένα χωρίς την απαιτούμενη εξουσιοδότηση από τους χρήστες. Πράγματι, καθημερινά ο αριθμός των συσκευών που μπορούν να συνδεθούν στο Διαδίκτυο των Πραγμάτων αυξάνεται με γοργούς ρυθμούς και θεωρείται ότι στα τέλη του 2020 θα φτάσει μέχρι τα 30 δισεκατομμύρια ενώ εκτιμάται ότι μέχρι τα τέλη του 2025 ο αριθμός αυτός θα έχει υπερδιπλασιαστεί [1].

Παρότι ο αριθμός των συσκευών αυτών αυξάνεται συνεχώς, δεν έχει δοθεί η απαιτούμενη προσοχή τόσο στην κατασκευή των έξυπνων συσκευών, ώστε να διαθέτουν ένα ενσωματωμένο μηχανισμό ασφαλείας, όσο και στην ασφάλεια και στη σωστή διαχείριση των δεδομένων που συλλέγονται.

Στη συνέχεια, γίνεται αναφορά στους κυριότερους λόγους και προβλήματα για την έλλειψη ασφαλείας στο Διαδίκτυο των Πραγμάτων αλλά και στο ζήτημα της προστασίας της ιδιωτικότητας [8][9].

1.4.1. Έλλειψη επίγνωσης του χρήστη σχετικά με το Διαδίκτυο των Πραγμάτων

Οι χρήστες είναι πλέον αρκετά ενημερωμένοι και εξοικειωμένοι με τις βασικές αρχές και τους κανόνες ασφάλειας που πρέπει να ακολουθήσουν προκειμένου να διασφαλίσουν την προστασία και ακεραιότητα των προσωπικών τους δεδομένων. Αυτή η εξέλιξη είναι θετική αλλά καθώς η τεχνολογία εξελίσσεται, οι χρήστες θα πρέπει και αυτοί να προσαρμόζονται στα νέα δεδομένα.

Το Διαδίκτυο των Πραγμάτων είναι μια νέα τεχνολογία η οποία είναι ακόμη σε πρώιμο στάδιο ως προς την εισχώρηση και τη διάδοσή της στην καθημερινότητα των απλών χρηστών. Αυτό έχει ως αποτέλεσμα οι καθημερινοί χρήστες να είναι πιο ευάλωτοι σε επιθέσεις κακόβουλων παρατηρητών. Γι' αυτό το λόγο, είναι απαραίτητη η ενημέρωση των χρηστών σχετικά με τους κινδύνους που υπάρχουν στο Διαδίκτυο των Πραγμάτων.

1.4.2. Έλλειψη ενημερώσεων των συσκευών

Η ενημέρωση του λογισμικού των συσκευών είναι πολύ σημαντική για να προστατευτούν οι συσκευές και κατ' επέκταση οι χρήστες που τις διαθέτουν. Οι συσκευές του Διαδικτύου των Πραγμάτων που διατίθενται στο εμπόριο διαθέτουν όλες τις ενημερώσεις για αδυναμίες στο λογισμικό τους. Παρόλα αυτά, οι κατασκευαστικές εταιρίες δίνουν περισσότερη έμφαση στην αύξηση της παραγωγής των συσκευών παρά στο να τα κάνουν περισσότερο ασφαλή. Έτσι, είναι απαραίτητη η συνεχής και άμεση ενημέρωση των συσκευών, καθώς και η προστασία των δεδομένων που έχουν συλλέξει όταν αυτά κάνουν τις απαραίτητες ενημερώσεις.

1.4.3. Μετάδοση δεδομένων με μη ασφαλή τρόπο

Για να μπορέσει το Διαδίκτυο των Πραγμάτων να λειτουργήσει με ασφάλεια και διαφυλάσσοντας την ιδιωτικότητα των χρηστών, θα πρέπει οι επιμέρους οντότητές του, οι οποίες ορίζουν τη μετάδοση δεδομένων από τις συσκευές, να διαθέτουν τους καλύτερους μηχανισμούς ασφαλείας. Είναι απαραίτητη η διασφάλιση της ακεραιότητας αλλά και της εγκυρότητας των ευαίσθητων δεδομένων που ανταλλάσσονται μεταξύ των έξυπνων συσκευών στο Διαδίκτυο των

Πραγμάτων, τόσο για την εξασφάλιση της ποιότητας των παρεχόμενων υπηρεσιών, όσο και για την διατήρηση της εμπιστοσύνης στο τεχνολογικό αυτό επίτευγμα. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί το πρότυπο TLS ή η μετάδοση να γίνεται μέσω κρυπτογραφημένου καναλιού.

1.4.4. Φυσική προστασία των συσκευών

Εκτός από την ανάπτυξη λογισμικών για τη διατήρηση της ασφάλειας κατά τη μετάδοση δεδομένων στο Διαδίκτυο των Πραγμάτων, υπάρχει και η ανάγκη για την προστασία των συσκευών αυτών. Η εξασφάλιση της φυσικής ασφάλειας των συσκευών κρίνεται απαραίτητη για να υπάρχει ομαλή και σωστή λειτουργία μεταξύ των συσκευών στο Διαδίκτυο των Πραγμάτων. Ένα παράδειγμα παραβίασης της ασφάλειας είναι η εισαγωγή κακόβουλου λογισμικού μέσω ενός ενιαίου σειριακού διαύλου (Universal Serial Bus - USB) σε κάποια απόμερη και απομονωμένη συσκευή δίνοντας την δυνατότητα σε κάποιον κακόβουλο χρήστη να πάρει τα δεδομένα της συσκευής ή ακόμη να προσβάλλει το δίκτυο με κάποιον ιό.

1.4.5. Συσκευές με εργοστασιακές ρυθμίσεις

Η διατήρηση των εργοστασιακών ρυθμίσεων και κωδικών στις έξυπνες συσκευές είναι ένας ακόμη παράγοντας που μπορεί να οδηγήσει στην εύκολη παραβίασή τους. Πιο συγκεκριμένα, οι συσκευές διαθέτουν προεπιλεγμένους κωδικούς ασφαλείας τους οποίους είναι εύκολο κανείς να αναπαράγει κάνοντας, για παράδειγμα, χρήση ενός ειδικού αλγορίθμου (bruteforce) που στοχεύει στην ανίχνευση πιθανών συνδυασμών. Έτσι, εάν ο κωδικός είναι ένας από τους γνωστούς και δημόσιους προεπιλεγμένους κωδικούς, τότε είναι εύκολο κανείς να αποκτήσει πρόσβαση στη συγκεκριμένη συσκευή.

Αναφορικά, έχει υπάρξει επίθεση η οποία πραγματοποιήθηκε εκμεταλλεύοντας τη παραπάνω αδυναμία σε συσκευές. Το 2016, έγινε επίθεση από Mirai bots, η οποία "έριξε" τους σέρβερ γνωστών πλατφόρμων όπως είναι το Twitter, το Netflix, κλπ και η οποία έστρεψε την προσοχή στα κενά ασφαλείας που προκαλούν οι συσκευές με εργοστασιακές ρυθμίσεις [10].

1.4.6. Κακόβουλο λογισμικό

Η συνεχής αύξηση των συσκευών του Διαδικτύου των Πραγμάτων έχει ως αποτέλεσμα την ανάλογη αύξηση των κακόβουλων λογισμικών τα οποία αναπτύσσονται για την εκμετάλλευση των συσκευών αυτών. Το κακόβουλο λογισμικό ή όπως αλλιώς ονομάζεται malware (malicious software) είναι ένα κομμάτι κώδικα ή ένα πρόγραμμα το οποίο έχει ως κύριο στόχο να βλάψει κάποιο υπολογιστικό σύστημα. Το κακόβουλο λογισμικό, ανάλογα με το σκοπό του, μπορεί να παρακολουθεί, να επιτεθεί ή και να καταστρέψει οποιαδήποτε συσκευή στην οποία θα εισχωρήσει. Το κακόβουλο λογισμικό είναι μια γενική κατηγορία και χωρίζεται σε υποκατηγορίες ανάλογα με τις λειτουργίες και τον σκοπό του. Παρακάτω αναλύονται ορισμένες κατηγορίες στις οποίες κατηγοριοποιείται το malware [11].

- **Ιός (Virus)**

  Ο ιός είναι ίσως το πιο γνωστό είδος κακόβουλου λογισμικού στους απλούς χρήστες. Έχει τη δυνατότητα να βλάψει ή και να διαγράψει ολόκληρα αρχεία και δεδομένα του υπολογιστικού συστήματος και μεταδίδεται πολύ εύκολα με τη βοήθεια κάποιας εξωτερικής πηγής όπως είναι το USB.

- **Ransomware**

  Το ransomware είναι ένα είδος λογισμικού το οποίο όταν εισέλθει στο σύστημα, κρυπτογραφεί όλα τα δεδομένα. Στη συνέχεια, για να ανακτηθούν τα δεδομένα θα πρέπει να πληρωθούν κάποια λύτρα (ransom) στον επιτιθέμενο.

- **Trojan Horse**

  Αυτό το είδος λογισμικού έχει ως σκοπό τη μόλυνση του υπολογιστή, αλλά κύριο χαρακτηριστικό του είναι ότι εμφανίζεται ως ένα λογισμικό το οποίο είναι χρήσιμο για τον υπολογιστή αλλά στην πραγματικότητα μπορεί να υποκλέψει αρχεία ή να αποκτήσει τον έλεγχο του συστήματος.

- **Ζόμπι (Zombie)**

  Το λογισμικό αυτό προσβάλλει τους υπολογιστές και τους καθιστά μέλη ενός δικτύου το οποίο ελέγχεται απομακρυσμένα από κακόβουλους χρήστες. Στη συνέχεια, το σύνολο

των μολυσμένων υπολογιστών θα προσπαθήσουν να επιτεθούν στο δίκτυο-στόχο με αποτέλεσμα να μην μπορεί το δίκτυο να παρέχει τις υπηρεσίες του. Η επίθεση αυτή ονομάζεται κατανεμημένη επίθεση άρνησης υπηρεσιών (Distributed Denial of Service attack – DDoS attack).

- **Σκουλήκι (Worm)**

  Το σκουλήκι είναι ένα είδος λογισμικού το οποίο αφού μολύνει κάποιον υπολογιστή, μεταδίδεται σε κάποιον άλλον που βρίσκεται στο ίδιο δίκτυο. Η γρήγορη εξάπλωσή του προκαλεί μεγάλη επιβάρυνση στο δίκτυο.

- **Rootkit**

  Το λογισμικό αυτό έχει την ιδιαιτερότητα ότι μπορεί να περιέχει οποιοδήποτε από τα παραπάνω λογισμικά και ότι μπορεί να τα αποκρύψει από τα λογισμικά ασφαλείας του υπολογιστικού συστήματος. Μία συνήθης λειτουργία του είναι ότι δημιουργεί πίσω πόρτες (backdoors), τις οποίες μπορεί να εκμεταλλευτεί κάποιος απομακρυσμένος κακόβουλος χρήστης.

- **Spyware**

  Το λογισμικό αυτό έχει τη λειτουργία να παρακολουθεί κρυφά τις δραστηριότητες του χρήστη και να τις αναφέρει στον προγραμματιστή που το δημιούργησε. Μπορεί να μολύνει πολύ εύκολα κάποιο υπολογιστικό σύστημα καθώς διαδίδεται στους υπολογιστές με εγκατάσταση κάποιου προγράμματος ή την επίσκεψη σε κάποια σελίδα που είναι μολυσμένη και είναι πολύ δύσκολος ο εντοπισμός του. Συνήθως υποκλέβουν σημαντικές πληροφορίες όπως στοιχεία πιστωτικών καρτών, προσωπικά δεδομένα αλλά μερικές φορές μπορεί να εγκαταστήσουν άλλα προγράμματα ή να αλλάξουν ρυθμίσεις του συστήματος.

- **Adware**

  Το adware είναι ένα ανεπιθύμητο λογισμικό το οποίο σχεδιάστηκε για να εμφανίζει διαφημίσεις στην οθόνη του υπολογιστή και συνήθως ενσωματώνεται σε ένα άλλο πρόγραμμα για να ξεγελάσει το χρήστη να το εγκαταστήσει στο υπολογιστικό σύστημα.

Ένας τρόπος για να γίνει αντιληπτό ότι υπάρχει στο σύστημα είναι όταν παρατηρηθεί ότι το σύστημα είναι αργό και αυτό συμβαίνει διότι το λογισμικό αυτό καταναλώνει αρκετή υπολογιστική ισχύ για να φορτώσει και να εμφανίσει τις διαφημίσεις.

- **Λογικές Βόμβες (Logic Bombs)**

  Οι λογικές βόμβες ή αλλιώς logic bombs είναι κομμάτια κώδικα ή ολόκληρα προγράμματα τα οποία εκτελούνται όταν συμβεί κάποιο προκαθορισμένο γεγονός, όπως η έλευση κάποιας συγκεκριμένης ημερομηνίας. Μπορεί να μείνουν αδρανή για χρόνια μέχρι να ενεργοποιηθούν και έχουν τη δυνατότητα να τροποποιήσουν ή να διαγράψουν αρχεία του συστήματος, να διαγράψουν εγγραφές από βάσεις δεδομένων ή και να απενεργοποιήσουν ολόκληρα συστήματα.

2. Τεχνικές προστασίας και ιδιωτικότητας δεδομένων στο Διαδίκτυο των Πραγμάτων

Η αυτοματοποίηση των έξυπνων συσκευών είναι ένας από τους βασικούς στόχους του Διαδικτύου των Πραγμάτων. Παρόλα αυτά, η αξιοπιστία και η ασφάλεια των πληροφοριών που μεταδίδονται μέσω των πολλαπλών έξυπνων συσκευών δεν μπορεί να εξασφαλιστεί με βάση τις υπάρχουσες τεχνολογίες και τα επικοινωνιακά πρωτόκολλα.

Τα δεδομένα που παρέχουν με την έγκρισή τους οι χρήστες είναι πάρα πολλά και είναι απαραίτητα για την ομαλή και σωστή ανάπτυξη της τεχνολογίας στον τομέα του Διαδικτύου των Πραγμάτων. Η χρησιμότητα των δεδομένων αυτών είναι αρκετά υψηλή, καθώς μπορούν να αξιοποιηθούν σε ερευνητικές, επιστημονικές ή στατιστικές μελέτες.

Η δημοσίευση των δεδομένων γίνεται κυρίως μέσω μεγάλων βάσεων δεδομένων [12]. Παρόλα αυτά, θα πρέπει να υπάρχουν μηχανισμοί και τρόποι με τους οποίους θα μπορούν να τα προφυλάσσουν από κακόβουλους χρήστες. Το πρόβλημα της ασφάλειας έγκειται στο γεγονός ότι η ασφάλεια έχει πολλές πτυχές και πρέπει να καλυφθούν όλες. Συγκεκριμένα, όταν δημοσιοποιούνται δεδομένα για συγκεκριμένες μελέτες, απαιτείται η εκμετάλλευση όσο το δυνατόν περισσότερων δεδομένων, χωρίς όμως να αποκαλυφθεί η ταυτότητα των χρηστών.

Στο παρελθόν έχουν υπάρξει περιστατικά στα οποία διέρρευσαν προσωπικές πληροφορίες σχετικά με την ταυτότητα των χρηστών χωρίς, βέβαια, την εσκεμμένη πρόθεση των εταιριών που τις δημοσίευσαν. Ένα παράδειγμα αποτελεί η διαρροή πληροφοριών από την εταιρία παροχής online υπηρεσιών που ονομάζεται AOL (America Online) [13]. Το 2006, η εταιρία δημοσιοποίησε στην ιστοσελίδα της ένα αρχείο το οποίο περιείχε 22.000.000 αναζητήσεις λέξεων από περισσότερους από 650.000 χρήστες σε περίοδο 3 μηνών. Στο αρχείο αυτό δεν υπήρχε αναφορά σε προσωπικές πληροφορίες αναφορικά με τους χρήστες. Παρόλα αυτά, μέσα στις αναζητήσεις υπήρχαν λέξεις με προσωπικές πληροφορίες των χρηστών, και αυτό έδινε τη δυνατότητα σε οποιονδήποτε κακόβουλο χρήστη να πάρει αυτά τα δεδομένα, να τα συγκρίνει με κάποια άλλα δημόσια αρχεία (πχ. δημοσιονομικά αρχεία) και να εξάγει αποτελέσματα τα οποία θα αποκάλυπταν την ταυτότητα των χρηστών. Ένα ακόμη παράδειγμα, είναι η υπόθεση με τη πλατφόρμα του Netflix. Το 2006, το Netflix θέλησε να βελτιστοποιήσει τον τρόπο αξιολόγησης των ταινιών του, χρησιμοποιώντας έναν αλγόριθμο ο οποίος δεχόταν σαν είσοδο τις παλιές κριτικές [13][14]. Για το σκοπό αυτό, το Netflix δημοσίευσε κριτικές από περισσότερους από 500.000 συνδρομητές του. Παρά το γεγονός ότι τα ονόματα και οι IP διευθύνσεις είχαν αντικατασταθεί με έναν αριθμό, μπορούσε κάποιος να συγκρίνει τις κριτικές αυτές με τις κριτικές στη διαδικτυακή βάση δεδομένων των ταινιών (Internet Movie Database - IMDb) και να μπορέσει να ταυτοποιήσει τους χρήστες.

Παρατηρείται, λοιπόν, ότι η ασφάλεια είναι μία έννοια πολύ εύθραυστη και γι'αυτόν τον λόγο έχει απασχολήσει σε μεγάλο βαθμό την επιστημονική κοινότητα. Βέβαια, με την όλο και αυξανόμενη τάση της χρήσης νέων τεχνολογιών, τα δεδομένα αυξάνονται σε τεράστιο βαθμό, με αποτέλεσμα την ώθηση για συνεχή αναζήτηση νέων τρόπων διασφάλισης της προστασίας των δεδομένων. Η επιστημονική κοινότητα δημιούργησε ορισμούς και ανακάλυψε τεχνικές προστασίας των δεδομένων. Οι ορισμοί χωρίζονται σε 2 μεγάλες κατηγορίες [14].

1. Συντακτική προστασία των δεδομένων, η οποία περιλαμβάνει τεχνικές στις οποίες τα δεδομένα στη βάση πρέπει να είναι παρόμοια μεταξύ τους και ο βαθμός ομοιότητας πρέπει να υπερβαίνει έναν συγκεκριμένο αριθμό γραμμών των οποίων το περιεχόμενό τους είναι ίδιο.

2. Σημασιολογική προστασία των δεδομένων, η οποία αναφέρεται σε τεχνικές στις οποίες τα δεδομένα προς δημοσίευση δεν θα μεταβληθούν εάν προστεθούν νέα δεδομένα ή αφαιρεθούν κάποια ήδη υπάρχοντα.

Όπως αναφέρθηκε και νωρίτερα, τα δεδομένα αναπαρίστανται και δημοσιεύονται σε βάσεις δεδομένων. Τα δεδομένα αυτά έχουν ιδιότητες (attributes) οι οποίες κατηγοριοποιούνται ως εξής :

- **Identifiers**, τα οποία είναι ιδιότητες που μπορούν ταυτοποιήσουν μοναδικά ένα πρόσωπο (πχ. ΑΦΜ).
- **Quasi-identifiers**, τα οποία είναι ένα σύνολο από ιδιότητες που μπορούν να χρησιμοποιηθούν για να ταυτοποιήσουν κάποια ή όλα τα πρόσωπα της βάσης (πχ. ταχυδρομικός κώδικας (Τ.Κ) και φύλο). Σε αυτό το σημείο αξίζει να σημειωθεί ότι το σύνολο των γραμμών των οποίων τα quasi-identifiers τους είναι ίδια αποτελούν μια κλάση ισοτιμίας (equivalence class).
- **Confidential attributes**, τα οποία είναι ιδιότητες οι οποίες είναι προσωπικές και απόρρητες πέραν του ατόμου που σχετίζονται και είναι απαραίτητη η απόκρυψή τους (πχ. κατάσταση υγείας).
- **Non-confidential attributes**, ιδιότητες οι οποίες θεωρούνται ασφαλείς για δημοσίευση (πχ. όνομα κατοικίδιου).

Στον παρακάτω Πίνακα 1 παρατηρούνται οι τεχνικές προστασίας δεδομένων για κάθε έναν ορισμό και στις οποίες θα γίνει αναλυτικότερη αναφορά παρακάτω.

Πίνακας 1 - Πίνακας τεχνικών προστασίας δεδομένων.

	Συντακτική Προστασία Δεδομένων	**Σημασιολογική Προστασία Δεδομένων**
Τεχνικές	1. k-anonymity 2. l-diversity 3. t-closeness	1. Differential privacy

2.1 Συντακτική προστασία των δεδομένων

Οι τεχνικές της συντακτικής προστασίας δεδομένων βασίζονται στη παραδοχή ότι η δημοσίευση του πίνακα δεδομένων μπορεί να θέσει σε κίνδυνο την αποκάλυψη των ταυτοτήτων μόνο των ατόμων που περιλαμβάνονται στον πίνακα [15]. Οι τεχνικές αυτές, αρχικά, λειτουργούν κρύβοντας τα στοιχεία τα οποία χαρακτηρίζουν μοναδικά τα πρόσωπα της βάσης. Αυτός, όμως, δεν

είναι ασφαλής τρόπος καθώς κάποιος μπορεί να χρησιμοποιήσει τους quasi-identifiers για να ταυτοποιήσει κάποιο πρόσωπο. Μια έρευνα που πραγματοποιήθηκε το 2000 στο Κάνσας της Αμερικής έδειξε ότι το 63% του πληθυσμού μπορεί να ταυτοποιηθεί χρησιμοποιώντας σε συνδυασμό την ημερομηνία γέννησης, τον ταχυδρομικό κώδικα και το φύλο.

Πίνακας 2 - Πίνακας ασθενών ενός νοσοκομείου.

Όνομα	**ΑΦΜ**	**Φύλο**	**Τ.Κ.**	**Ημ. Γέννησης**	**Ασθένεια**
Νίκος Κ.	2698268730	Α	66102	23/6/1996	Χλαμύδια
Κώστας Σ.	6987820123	Α	66100	10/6/1996	AIDS
Μαρία Π.	1365204589	Γ	66196	29/7/1996	Πνευμονία
Γεωργία Μ.	2551230147	Γ	66191	13/7/1996	Λευχαιμία
Κάτια Ν.	4112659785	Γ	66152	12/10/2000	Πνευμονία
Λάζαρος Σ.	3131206989	Α	66110	3/10/2000	Πνευμονία
Ξένια Γ.	7556458963	Γ	66150	1/12/2000	Εγκεφαλικό
Άγγελος Π.	5620103056	Α	66118	6/12/2000	Εγκεφαλικό
Ιωάννα Μ.	2102305684	Α	66114	30/12/2000	Εγκεφαλικό

Στον Πίνακα 2 παρατηρείται ο πίνακας της βάσης ενός νοσοκομείου για κάποιους ασθενείς και στον Πίνακα 3 παρατηρείται ο πίνακας που πρόκειται να δημοσιευθεί.

Πίνακας 3 - Παράδειγμα απόκρυψης προσωπικών στοιχείων.

Όνομα	**ΑΦΜ**	**Φύλο**	**Τ.Κ.**	**Ημ. Γέννησης**	**Ασθένεια**
*	*	Α	66102	23/6/1996	Χλαμύδια
*	*	Α	66100	10/6/1996	AIDS
*	*	Γ	66196	29/7/1996	Πνευμονία
*	*	Γ	66191	13/7/1996	Λευχαιμία
*	*	Γ	66152	12/10/2000	Πνευμονία
*	*	Α	66110	3/10/2000	Πνευμονία
*	*	Γ	66150	1/12/2000	Εγκεφαλικό
*	*	Α	66118	6/12/2000	Εγκεφαλικό
*	*	Α	66114	30/12/2000	Εγκεφαλικό

Η δημοσίευση του Πίνακα 3 δεν είναι ασφαλής διότι θα μπορούσε κάποιος κακόβουλος χρήστης να συγκρίνει τα στοιχεία αυτά με κάποια δημόσια αρχεία (πχ. αρχεία ψηφοφόρων) και να βγάλει ασφαλή συμπεράσματα για την ασθένεια που έχει κάποιο άτομο της βάσης. Συνεπώς, οι τεχνικές της συντακτικής προστασίας δεδομένων έχουν επικεντρωθεί στην προστασία των quasi-identifiers.

2.1.1 Η τεχνική k-anonymity

Η τεχνική του k-anonymity είναι η πιο γνωστή τεχνική ασφάλειας και παροχής ανωνυμίας και επικεντρώνεται στην προστασία της ταυτότητας των ατόμων της βάσης δεδομένων [16]. Η ιδέα πάνω στην οποία βασίζεται η τεχνική του k-anonymity είναι ότι στον πίνακα της βάσης που θα δημοσιευθεί, τα quasi-identifiers πρέπει είναι παρόμοια για τουλάχιστον k γραμμές του πίνακα [14] [17]. Με άλλα λόγια, ο πίνακας της βάσης ικανοποιεί τη τεχνική προστασίας k-anonymity εάν κάθε γραμμή του πίνακα σχετίζεται με τουλάχιστον k πρόσωπα και εάν κάθε πρόσωπο του πίνακα σχετίζεται με τουλάχιστον k γραμμές της βάσης. Άρα, οι γραμμές του πίνακα θα πρέπει να ικανοποιούν την παρακάτω σχέση:

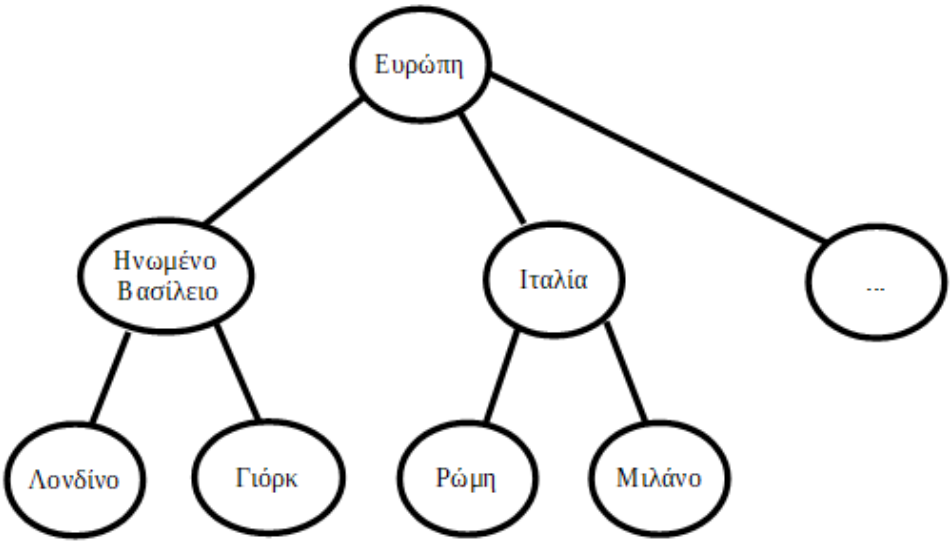
$$Pr(x) \leq \frac{1}{k} \quad (1)$$

Στη σχέση 1, με x συμβολίζεται κάθε γραμμή του πίνακα και με Pr(x) συμβολίζεται η πιθανότητα να ταυτοποιηθεί κάθε γραμμή του πίνακα.

Για να μπορέσει να δημοσιευθεί ο πίνακας και να εξασφαλιστεί ότι δεν θα μπορέσει κανείς να εξάγει συμπεράσματα για τα άτομα που υπάρχουν σε αυτόν, θα πρέπει να υπάρξει πλήρης γνώση, εκ των προτέρων, όλων των τρόπων τους οποίους θα μπορούσε κανείς να χρησιμοποιήσει για να κάνει ταυτοποίηση των ατόμων της βάσης. Αυτή, όμως, η μέθοδος είναι πρακτικά αδύνατη με αποτέλεσμα η τεχνική του k-anonymity να είναι δομημένη με τρόπο τέτοιο ώστε να εξασφαλίζει ότι τουλάχιστον k πρόσωπα του πίνακα δεν θα είναι παρόμοια μεταξύ τους.

Η μεθοδολογία που εφαρμόζει το k-anonymity είναι να χρησιμοποιεί τις τεχνικές της γενίκευσης (generalization) και της συμπίεσης (suppression) [18][19]. Η τεχνική της γενίκευσης εφαρμόζεται με διαφορετικούς τρόπους, ανάλογα τον τύπο των δεδομένων που θα εφαρμοστεί η

γενίκευση. Η γενίκευση μιας αριθμητικής ιδιότητας (numeric attribute), όπως είναι για παράδειγμα η ηλικία (πχ. 23), γίνεται με τη μετατροπή της σε ένα διάστημα τιμών (πχ. [20 – 30]). Από την άλλη, η γενίκευση μιας κατηγοριακής ιδιότητας (category attribute), όπως είναι για παράδειγμα η χώρας γέννησης (πχ. Λονδίνο), πραγματοποιείται με τη χρήση ιεραρχικού δένδρου (πχ. Ηνωμένο Βασίλειο) όπως φαίνεται και στην Εικόνα 3.



Εικόνα 3 - Παράδειγμα Ιεραρχικού Δένδρου.

Η τεχνική της συμπίεσης είναι το ανώτατο επίπεδο γενίκευσης και συνήθως, όταν εφαρμόζεται, η τιμή της ιδιότητας αντικαθίσταται ολόκληρη από το σύμβολο *. Το σύμβολο αυτό αναπαριστά όλο το εύρος τιμών που μπορεί να πάρει η αντίστοιχη ιδιότητα του πίνακα. Παρέχεται, όμως, η δυνατότητα εφαρμογής αυτού του συμβόλου σε ένα τμήμα και όχι υποχρεωτικά σε ολόκληρη τη τιμή.

Θα χρησιμοποιηθούν ως παράδειγμα τα στοιχεία του Πίνακα 3 και θα εφαρμοστούν σε αυτά οι παραπάνω τεχνικές. Ο Πίνακας 4 δείχνει ότι η τεχνική του k-anonymity μπορεί να εφαρμοστεί με επιτυχία για k=2 και αν θεωρηθούν ως quasi-identifiers ο συνδυασμός του φύλου, του ταχυδρομικού κώδικα και της ημερομηνίας γέννησης.

Πίνακας 4 - Παράδειγμα k-anonymity πίνακα για k=2.

Φύλο	**Τ.Κ.**	**Ημ. Γέννησης**	**Ασθένεια**
*	661**	**/6/1996	Χλαμύδια
*	661**	**/6/1996	AIDS
*	661**	**/7/1996	Πνευμονία
*	661**	**/7/1996	Λευχαιμία
*	661**	**/10/2000	Πνευμονία
*	661**	**/10/2000	Πνευμονία
*	661**	**/12/2000	Εγκεφαλικό
*	661**	**/12/2000	Εγκεφαλικό
*	661**	**/12/2000	Εγκεφαλικό

Στον Πίνακα 4 παρατηρείται ότι για την ιδιότητα του φύλου έχει χρησιμοποιηθεί η τεχνική της συμπίεσης, η οποία μπορεί να εφαρμοστεί σε ολόκληρη τη γραμμή αν είναι απαραίτητο, αλλά στους άλλους δύο quasi-identifiers έχει χρησιμοποιηθεί η γενίκευση σε τμήματα των τιμών. Επίσης, παρατηρείται ότι στην ευαίσθητη ιδιότητα (sensitive attribute) δεν εφαρμόζονται οι παραπάνω τεχνικές.

Βέβαια, χρησιμοποιώντας τις τεχνικές της γενίκευσης και της συμπίεσης, τα δεδομένα του πίνακα αλλοιώνονται. Από τη μία πλευρά, είναι απαραίτητη η προστασία των δεδομένων από τους κακόβουλους παρατηρητές και αναπόφευκτα είναι επιτακτικής σημασίας η χρήση των παραπάνω τεχνικών. Από την άλλη μεριά, όμως, όσο περισσότερο γενικευθεί μια τιμή, τόσο μεγαλύτερες θα είναι και οι απώλειες στη ποσότητα της πληροφορίας που πρόκειται να δημοσιευθεί. Συνεπώς, εδώ τίθεται ένα ερώτημα στο πως θα πραγματοποιηθεί η μέτρηση των απωλειών των πληροφοριών και μέχρι τι ποσοστό από τις πληροφορίες μπορεί να διατεθεί για να επιτευχθεί ασφαλής δημοσίευση του πίνακα αλλά και οι πληροφορίες που θα υπάρχουν στον πίνακα να είναι επαρκείς για να αξιοποιηθούν για μελέτες και έρευνες.

Για τη μέτρηση των απωλειών των πληροφοριών, χρησιμοποιείται η μεταβλητή Information Loss (IL). Η μεταβλητή αυτή ορίζεται διαφορετικά ανάλογα με τι τύπου είναι η ιδιότητα της οποίας θα μετρηθεί η απώλειά της. Πιο συγκεκριμένα, εάν πραγματοποιηθεί μέτρηση της απώλειας σε μια

αριθμητική ιδιότητα, τότε η αρχική τιμή x γενικεύεται σε ένα εύρος τιμών [x_{min} x_{max}], όπου το x_{min} είναι η ελάχιστη τιμή που παίρνει η μεταβλητή σε όλο τον πίνακα και x_{max} είναι η αντίστοιχη μέγιστη τιμή. Σε αυτήν την περίπτωση, ο τύπος που εκφράζει τις απώλειες είναι ο παρακάτω:

$$IL = \frac{x_{max} - x_{min}}{X_{Max} - \chi_{Min}} \tag{2}$$

Στον τύπο (2), οι μεταβλητές X_{Max} και X_{Min} συμβολίζουν τη μέγιστη και την ελάχιστη, αντίστοιχα, δυνατή τιμή που μπορεί να πάρει η συγκεκριμένη ιδιότητα. Στην περίπτωση των ιδιοτήτων που ανήκουν στην κατηγορία των κατηγοριακών ιδιοτήτων, η μέτρηση των απωλειών γίνεται με τον παρακάτω τύπο:

$$IL = \frac{dnode(x)}{Tnodes} \tag{3}$$

Στον τύπο (3), η μεταβλητή dnode εκφράζει τον αριθμό των φύλλων που έχει η τιμή της ιδιότητας στο ιεραρχικό δένδρο, ενώ η μεταβλητή Tnodes εκφράζει των αριθμό των συνολικών φύλλων. Ας αναφερθούμε για παράδειγμα στο ιεραρχικό δένδρο της Εικόνας 3. Ας υποθέσουμε ότι στο εικονιζόμενο δένδρο δεν υπάρχουν άλλες πόλεις πέρα από το Ηνωμένο Βασίλειο και την Ιταλία. Εάν γενικευθεί η τιμή "Λονδίνο" στη τιμή "Ηνωμένο Βασίλειο", τότε, σύμφωνα με τον παραπάνω τύπο, η μεταβλητή dnode (Ηνωμένο Βασίλειο) παίρνει τη τιμή 2 και η μεταβλητή Tnodes παίρνει τη τιμή 4. Άρα οι απώλειες είναι 50%. Όλα αυτά, όμως, γίνονται σε επίπεδο κελιού, δηλαδή υπολογίζεται σε κάθε κελί του πίνακα η τιμή των απωλειών. Σε ένα πιο γενικευμένο πλαίσιο, θα γινόταν ο υπολογισμός των απωλειών σε όλη τη γραμμή κάνοντας χρήση του παρακάτω τύπου:

$$IL = \frac{\sum_{i=1}^{n} IL_i}{n} \tag{4}$$

Γενικά, γίνεται προσπάθεια να μειωθεί όσο το δυνατόν περισσότερο η τιμή IL ή με άλλα λόγια θεσπίζεται ως στόχος η μεγιστοποίηση της εξαγωγής ωφέλιμης πληροφορίας από τον πίνακα που θα δημοσιευθεί.

2.1.2 Η τεχνική l-diversity

Η τεχνική του k-anonymity είναι αρκετά αποτελεσματική στην απόκρυψη της ταυτότητας των προσώπων του πίνακα. Όμως, η τεχνική αυτή υστερεί όταν κάποιος κακόβουλος παρατηρητής προσπαθήσει να ταυτοποιήσει ένα άτομο του πίνακα εκμεταλλευόμενος τις τιμές των ευαίσθητων πληροφοριών που υπάρχουν σε αυτόν. Οι επιθέσεις τέτοιου τύπου καταφέρνουν να διαπεράσουν την ασφάλεια της τεχνικής του k-anonymity, με αποτέλεσμα την εφαρμογή μιας νέας τεχνικής για την προστασία του πίνακα, η οποία ονομάζεται l-diversity [16].

Η τεχνική αυτή εφαρμόζεται παράλληλα με τη τεχνική του k-anonymity και κατακερματίζει τον πίνακα σε τμήματα, στα οποία είναι υποχρεωτική η ύπαρξη τουλάχιστον l διαφορετικών τιμών για τις τιμές των ευαίσθητων ιδιοτήτων. Με άλλα λόγια, κάθε τμήμα το οποίο δημιουργείται θα πρέπει να περιέχει l διαφορετικές τιμές για να αποτραπεί η σύνδεση μεταξύ της ταυτότητας και των ευαίσθητων πληροφοριών του ατόμου [20][21].

Στη περίπτωση αυτή, θα χρησιμοποιηθεί ως παράδειγμα ο Πίνακας 4. Στον Πίνακα 4, η τεχνική του k-anonymity εξάγει μη ασφαλή αποτελέσματα και ο λόγος που συμβαίνει αυτό είναι γιατί στον πίνακα υπάρχουν ίδιες τιμές για την ευαίσθητη ιδιότητα και ταυτόχρονα τα quasi-identifiers για αυτές τις τιμές είναι ίδιες.

Πίνακας 5 - Παράδειγμα για 2-anonymity πίνακα με 2-diversity.

Φύλο	**Τ.Κ.**	**Ημ. Γέννησης**	**Ασθένεια**
Α	6610*	**/**/1996	Χλαμύδια
Α	6610*	**/**/1996	AIDS
Γ	6619*	**/**/1996	Πνευμονία
Γ	6619*	**/**/1996	Λευχαιμία
Γ	6615*	**/**/2000	Πνευμονία
Γ	6615*	**/**/2000	Εγκεφαλικό
Α	6611*	**/**/2000	Πνευμονία
Α	6611*	**/**/2000	Εγκεφαλικό
Α	6611*	**/**/2000	Εγκεφαλικό

Με αυτό το σκεπτικό, θα μπορούσε κάποιος κακόβουλος παρατηρητής, ο οποίος είναι γνωστός της Κάτιας που έχει γεννηθεί στις 12/10/2000 και μένει σε μια περιοχή με ταχυδρομικό κώδικα 66152, να χρησιμοποιήσει τις πληροφορίες από τον Πίνακα 4 και να εξάγει ως σίγουρο συμπέρασμα ότι η Κάτια νοσεί από πνευμονία. Αυτό το πρόβλημα είναι ένα είδος επίθεσης το οποίο ονομάζεται επίθεση ομοιογένειας (homogeneity attack). Η επίθεση αυτή λειτουργεί με την υπόθεση ότι αν ο παρατηρητής γνωρίζει τις τιμές των quasi-identifiers του ατόμου, τότε μπορεί να υποκλέψει τις ευαίσθητες πληροφορίες του.

Η τεχνική του l-diversity καταφέρνει να αποτρέψει τέτοιου είδους επιθέσεις. Όπως φαίνεται και στον Πίνακα 5, εφαρμόζοντας στον πίνακα της βάσης της τεχνική του l-diversity για l=2, δεν είναι δυνατόν για κάποιον εξωτερικό παρατηρητή, ακόμη και να γνωρίζει τα στοιχεία των ατόμων, να εξάγει ασφαλή συμπεράσματα. Στον Πίνακα 5, έχοντας κάνει ορισμένες τροποποιήσεις σχετικά με τη γενίκευση των τιμών των quasi-identifiers, ο παρατηρητής δεν μπορεί να είναι σίγουρος για το αν η Κάτια νοσεί από πνευμονία ή έχει υποστεί εγκεφαλικό. Με αυτόν τον τρόπο, πραγματοποιείται απόκρυψη των ευαίσθητων πληροφοριών ή με άλλα λόγια, δεν δίνεται η δυνατότητα σε κανέναν εξωτερικό παρατηρητή να μπορεί να υποκλέψει ή να υποθέσει με 100% σιγουριά τις ευαίσθητες πληροφορίες.

2.1.3 Η τεχνική t-closeness

Παρά το γεγονός ότι η τεχνική του l-diversity λειτουργεί πολύ καλά στο να αποτρέπει σε έναν κακόβουλο παρατηρητή την αποκάλυψη των ευαίσθητων πληροφοριών του πίνακα, υπάρχουν περιπτώσεις στις οποίες ο παρατηρητής μπορεί να εξάγει συμπεράσματα βασιζόμενος σε στατιστικές πιθανότητες ανάλογα με την κατανομή των τιμών των ιδιοτήτων στον πίνακα και με αυτόν το τρόπο μπορεί να κάνει κάποιες υποθέσεις οι οποίες πιθανόν να έχουν μεγάλα ποσοστά βεβαιότητας. Ας χρησιμοποιηθεί ως παράδειγμα ο Πίνακας 5. Στον πίνακα 5, κάποιος κακόβουλος παρατηρητής μπορεί να γνωρίζει τον Νίκο Κ., που ζει σε περιοχή με ταχυδρομικό κώδικα 66102 και είναι γεννημένος το 1996. Ο παρατηρητής μπορεί να εξάγει με 50% πιθανότητα ότι ο Νίκος πάσχει από χλαμύδια. Ενώ στον γενικό πίνακα, η πιθανότητα αυτή είναι μικρότερη και πιο συγκεκριμένα 11,1% (1/9). Αυτός είναι ένας τρόπος επίθεσης βασιζόμενος σε κατανομές συχνοτήτων των τιμών και ονομάζεται επίθεση λοξότητας (skewness attack).

Πίνακας 6 - Παράδειγμα για 2-anonymity πίνακα με 2-diversity και 0.5-closeness.

Φύλο	**Τ.Κ.**	**Ημ. Γέννησης**	**Ασθένεια**
Α	6610*	**/**/1996	Χλαμύδια
Α	6610*	**/**/1996	AIDS
Γ	6619*	**/**/1996	Πνευμονία
Γ	6619*	**/**/1996	Λευχαιμία
Γ	6615*	**/**/2000	Πνευμονία
Γ	6615*	**/**/2000	Εγκεφαλικό
Α	6611*	**/**/2000	Πνευμονία
Α	6611*	**/**/2000	Εγκεφαλικό
Α	6611*	**/**/2000	Γρίπη

Επίσης, ένας παρατηρητής θα μπορούσε να βγάλει ένα συμπέρασμα με 100% σιγουριά στο γεγονός ότι ο Νίκος πάσχει από κάποιο σεξουαλικώς μεταδιδόμενο νόσημα, εξαιτίας του τρόπου με τον οποίο έχουν χωριστεί τα μπλόκ εφαρμόζοντας την τεχνική του l-diversity. Αυτού του είδους οι επιθέσεις συμβαίνουν όταν οι τιμές των ιδιοτήτων είναι σημασιολογικά ίδιες και ονομάζονται επιθέσεις ομοιότητας (similarity attacks).

Για την αντιμετώπιση τέτοιου είδους επιθέσεων, εφαρμόζεται μια νέα τεχνική που είναι επέκταση της τεχνικής του l-diversity και η οποία ονομάζεται t-closeness [16][22]. Η τεχνική αυτή απαιτεί η συχνότητα εμφάνισης των στοιχείων που ανήκουν στην ίδια κλάση ισοτιμίας να μην υπερβαίνει ένα άνω όριο t σε σχέση με τη συχνότητα εμφάνισης των στοιχείων σε ολόκληρο τον πίνακα. Έτσι, παρατηρείται στον Πίνακα 6 ότι η συχνότητα εμφάνισης των στοιχείων σε κάθε κλάση ισοτιμίας είναι το πολύ 50% και ότι σε ολόκληρο τον πίνακα η συχνότητα εμφάνισης των στοιχείων είναι το πολύ 22,2%. Οπότε, για τον Πίνακα 6, η τιμή του t είναι 0.5. Γενικά, όμως, γίνεται προσπάθεια να κρατηθεί όσο το δυνατόν μικρότερη η τιμή του t για να ελαχιστοποιηθεί η βεβαιότητα κάποιου παρατηρητή.

2.2 Σημασιολογική προστασία των δεδομένων

Οι τεχνικές της σημασιολογικής προστασίας δεδομένων εστιάζονται σε τρόπους αντιμετώπισης των ευπαθειών των πινάκων προς δημοσίευση και έχουν την ιδιαιτερότητα ότι δεν επηρεάζονται από την εγγραφή ή τη διαγραφή κάποιου προσώπου από τον πίνακα. Ταυτόχρονα, οι τεχνικές αυτές δεν επιτρέπουν την αποκάλυψη των ταυτοτήτων των ατόμων, είτε αυτά τα άτομα υπάρχουν μέσα στον πίνακα είτε όχι. Ο Dalenius αναφέρει ότι οποιαδήποτε πληροφορία χρειάζεται να μάθει κάποιος για ένα πρόσωπο της βάσης θα πρέπει να μπορεί να το κάνει αυτό χωρίς να χρησιμοποιήσει τη βάση [23]. Στη συνέχεια, θα μελετηθεί η πιο βασική και διαδεδομένη τεχνική στη σημασιολογική προστασία των δεδομένων.

2.2.1 Η τεχνική ε-Differential Privacy

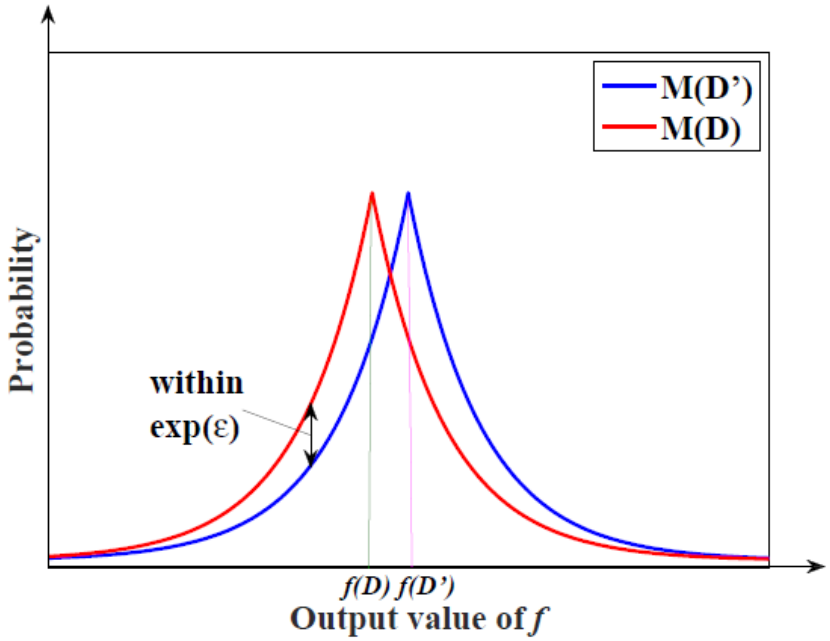
Η πιο γνωστή και ευρέως χρησιμοποιούμενη τεχνική προστασίας των δεδομένων με τέτοιο τρόπο ώστε να μην μπορεί κάποιος να χρησιμοποιήσει εξωτερικές πληροφορίες για να ταυτοποιήσει κάποιο άτομο της βάσης είναι η τεχνική του differential privacy [24] [25].

Η τεχνική του differential privacy προτάθηκε για πρώτη φορά το 2006 από τον Dwork [26]. Το differential privacy βασίζεται στην τεχνική της παραμόρφωσης των δεδομένων, με την οποία οι ευαίσθητες ιδιότητες παραμορφώνονται προσθέτοντας θόρυβο στις τιμές τους ενώ κάποιες ιδιότητες μπορεί να παραμείνουν αναλλοίωτες [27][28]. Η τεχνική αυτή αποσκοπεί στο να προστατέψει τα άτομα της βάσης με το να εμφανίζει πληροφορίες σχετικά με ένα σύνολο από άτομα και όχι για κάθε άτομο ξεχωριστά. Με άλλα λόγια, το άτομο προστατεύεται από ένα σύνολο ατόμων με αποτέλεσμα ένας κακόβουλος παρατηρητής να μην μπορεί να ταυτοποιήσει συγκεκριμένα κάποιο άτομο αφού έχει πληροφορίες μόνο για το σύνολο .

Ο ορισμός του differential privacy αναφέρει ότι αν θεωρήσουμε δύο γκρουπ από δεδομένα (datasets) D1 και D2, τα οποία διαφέρουν μεταξύ τους κατά μια γραμμή, τότε μια αυθαίρετη τυχαία συνάρτηση M επιτυγχάνει ε-differential privacy, εάν και μόνο εάν, ισχύει ότι [29]:

$$Pr(M(D1)) \leq e^{\varepsilon} * Pr(M(D2)) \quad (5)$$

Ο τύπος (5) αναφέρεται στο γεγονός ότι η ύπαρξη ή όχι ενός ατόμου στον πίνακα δεν επηρεάζει κατά πολύ τα αποτελέσματα της συνάρτησης M. Στον παραπάνω τύπο, υπάρχει η μεταβλητή ε, η οποία αποτελεί τη παράμετρο προστασίας (privacy budget parameter) και σχετίζεται με τη ποσότητα του θορύβου που θα εισαχθεί στα δεδομένα.



Εικόνα 4: Καμπύλες πιθανοτήτων αποκάλυψης των δεδομένων στην τεχνική του differential privacy [29].

Γίνεται αντιληπτό ότι ο τύπος (5) δεν έχει μεταβλητές ή παραμέτρους που να εμπλέκουν τις εξωτερικές πληροφορίες που μπορεί να έχει κάποιος κακόβουλος παρατηρητής, με αποτέλεσμα να θωρακίζονται τα δεδομένα από τέτοιου είδους επιθέσεις.

Νωρίτερα αναφερθήκαμε στη βασική λειτουργία της τεχνικής του differential privacy η οποία είναι η παραμόρφωση των δεδομένων. Για την επίτευξη αυτής της παραμόρφωσης, εφαρμόζεται τυχαίος θόρυβος στα αποτελέσματα της αναζήτησης (query) πάνω στα δεδομένα και με αυτόν τον τρόπο επιτυγχάνεται η προστασία.

Ο μηχανισμός που αναλαμβάνει την παραγωγή του τυχαίου θορύβου ακολουθεί τη κατανομή Laplace και συμβολίζεται ως Lap(b). Για το b ισχύει ότι :

$$b=\frac{\Delta f}{\varepsilon} \tag{6}$$

Ο παραπάνω τύπος δείχνει ότι ο θόρυβος που εισάγεται στις τιμές των δεδομένων είναι αντιστρόφως ανάλογος με τη παράμετρο προστασίας ε. Δηλαδή, όσο μικρότερη είναι η τιμή της παραμέτρου ε, τόσο μεγαλύτερη παραμόρφωση και κατά συνέπεια τόσο μεγαλύτερη ασφάλεια παρέχονται στα δεδομένα. Στον τύπο (6), παρατηρείται η μεταβλητή Δf, η οποία ονομάζεται ευαισθησία (sensitivity). Σαν ευαισθησία ορίζεται η μέγιστη διαφορά δύο γειτονικών datasets D1, D2 όσον αφορά τα αποτελέσματα της συνάρτησης αναζήτησης [30]. Πιο συγκεκριμένα, για μια συνάρτηση f : D → R^k , με R συμβολίζεται το εύρος των τιμών και k ορίζεται η διάσταση του R, τότε ο τύπος της ευαισθησίας της f ορίζεται ως εξής:

$$\Delta f=max_{D1,D2}|f(D1)-f(D2)| \tag{7}$$

Όπως φαίνεται στο παραπάνω τύπο, η ευαισθησία Δf εξαρτάται από τα αποτελέσματα της συνάρτησης f και δεν επηρεάζεται από το μέγεθος του dataset. Όπως αναφερθήκαμε παραπάνω, ο θόρυβος ο οποίος εισάγεται στα δεδομένα συμβολίζεται ως Lap(b) και η τιμή του είναι η εξής:

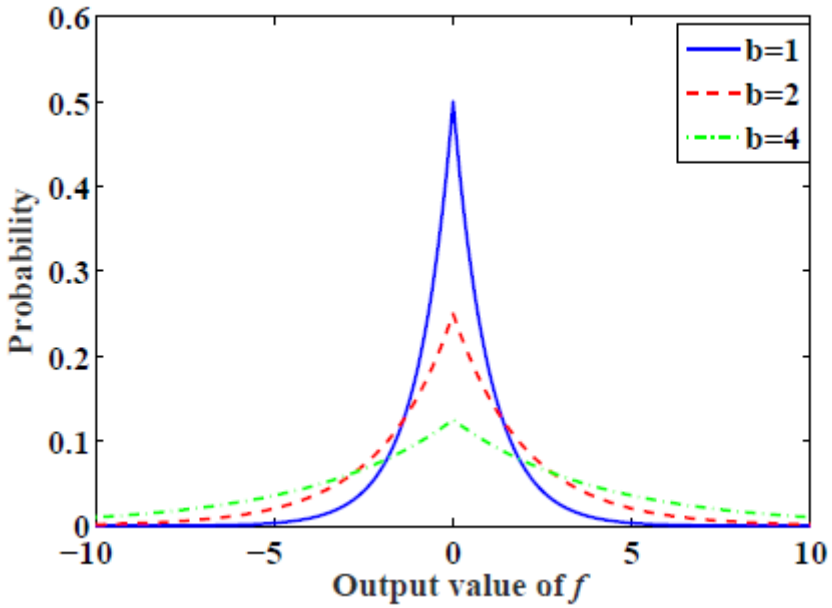
$$Lap(b)=\exp(\frac{-|x|}{b}) \tag{8}$$

Επίσης, ισχύει ότι για μια τυχαία μεταβλητή x, θεωρείται ότι η μεταβλητή αυτή ακολουθεί την κατανομή Laplace αν η συνάρτηση πυκνότητας πιθανότητας της μεταβλητής αυτής είναι η εξής:

$$p(x)=\frac{1}{2b}\cdot\exp(\frac{-|x-u|}{b}) \tag{9}$$

Καταλήγοντας, αν θεωρηθεί ως dataset η μεταβλητή D, ως συνάρτηση αναζήτησης η μεταβλητή f και ως τυχαία συνάρτηση η μεταβλητή M, τότε η παραμόρφωση που δέχονται τα δεδομένα καθορίζονται στον παρακάτω τύπο, ο οποίος ικανοποιεί ε-differential privacy :

$$M(D)=f(D)+Lap(b)^{k} \tag{10}$$



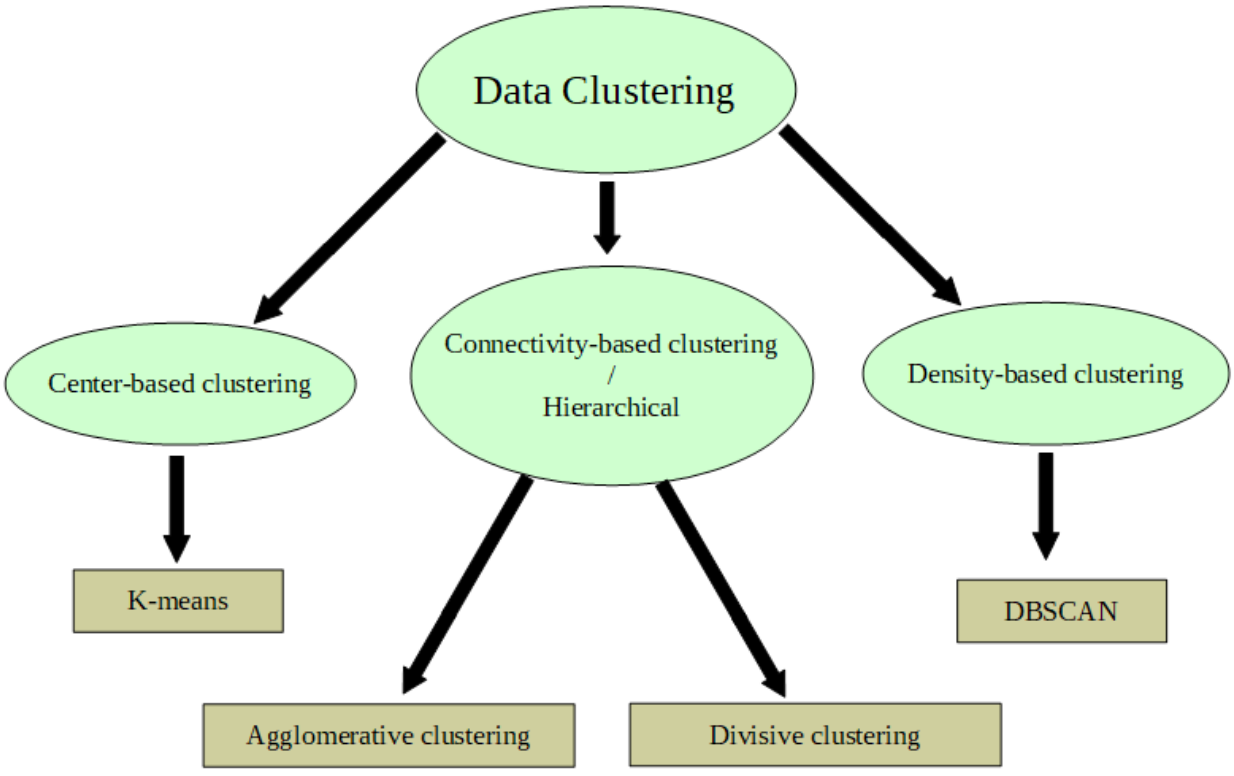
Εικόνα 5 - Συνάρτηση πυκνότητας πιθανότητας του θορύβου Laplace για διαφορετικές τιμές του b [29].

3. Αλγόριθμοι Ομαδοποίησης

Η συνεχής και αυξανόμενη παραγωγή δεδομένων από τις συσκευές δημιουργεί τεράστιες ποσότητες δεδομένων. Τα δεδομένα αυτά μπορούν να πάρουν διάφορες μορφές, όπως για παράδειγμα, μπορεί να αφορούν την υγεία μας, την καθημερινότητά μας, τη δουλειά μας ή οτιδήποτε άλλο θα μπορούσε να βοηθήσει στο να εξελιχθεί και να αναπτυχθεί το Διαδίκτυο των Πραγμάτων. Βέβαια, θα πρέπει γίνεται διαχείριση όλου αυτού του όγκου δεδομένων με τέτοιο τρόπο ώστε να αξιοποιούνται μόνο οι πληροφορίες που χρειάζονται και να απορρίπτονται οι υπόλοιπες.

Για να επιτευχθεί η αποτελεσματικότερη διαχείριση του τεράστιου όγκου δεδομένων, αναπτύχθηκαν, από την επιστημονική κοινότητα, ορισμένες τεχνικές εξόρυξης δεδομένων [31]. Μια από αυτές τις τεχνικές είναι η ομαδοποίηση των δεδομένων ή αλλιώς συσταδοποίηση (data clustering) [32][33].

Με τον όρο συσταδοποίηση ορίζεται η διαδικασία με την οποία διαχωρίζονται και ομαδοποιούνται τα δεδομένα σε συστάδες (clusters) ανάλογα με την ομοιότητα που έχουν τα αντικείμενα. Με άλλα λόγια, γίνεται προσπάθεια δημιουργίας ομάδων των οποίων τα δεδομένα τους να μοιάζουν περισσότερο μεταξύ τους (intra-cluster) και να μοιάζουν λιγότερο με τα δεδομένα των άλλων ομάδων (inter-cluster). Η τεχνική της συσταδοποίησης επάγεται στη γενικότερη κατηγορία τεχνικών μηχανικής μάθησης και πιο συγκεκριμένα στις τεχνικές μη επιβλεπόμενης μάθησης (unsupervised learning techniques), οι οποίες λειτουργούν αυτόνομα χωρίς να χρησιμοποιούν εξωτερικές πληροφορίες για να εξάγουν αποτελέσματα [34]. Η τεχνική της συσταδοποίησης έχει μελετηθεί και εφαρμοστεί σε αρκετά πεδία της επιστήμης, όπως είναι οι αλγόριθμοι μηχανικής μάθησης, οι στατιστικές μελέτες και τα αποτελέσματά τους, η επεξεργασία εικόνας, η αναγνώριση μοτίβων, η διαχείριση και η επεξεργασία πληροφοριών, τα γραφικά υπολογιστών, η βιοπληροφορική και πολλά άλλα.

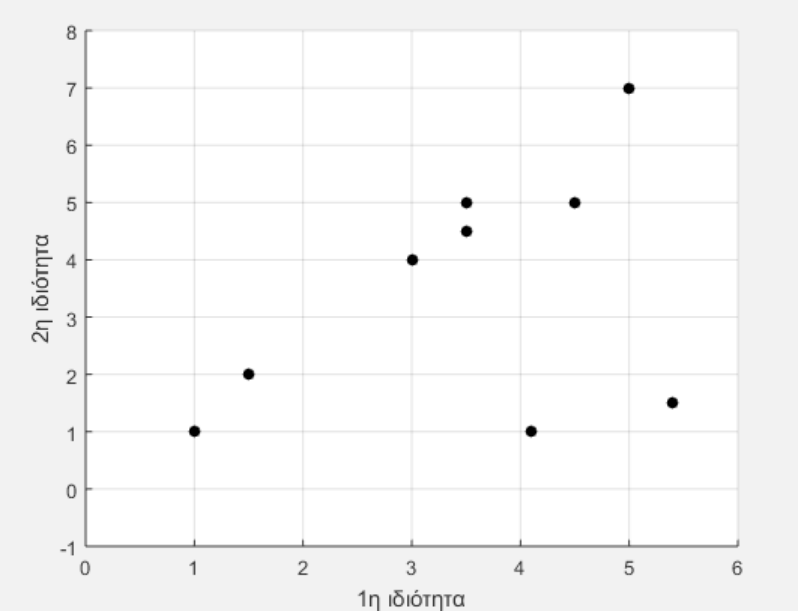


Εικόνα 6 - Μερικές κατηγορίες συσταδοποίησης που θα μελετηθούν παρακάτω.

Όταν εφαρμόζεται η τεχνική της συσταδοποίησης θα πρέπει να χρησιμοποιείται μια μετρική μονάδα για να υπολογιστεί η ομοιότητα μεταξύ των αντικειμένων. Η μονάδα που χρησιμοποιείται είναι μια συνάρτηση απόστασης μεταξύ των αντικειμένων. Όπως παρατηρείται στην Εικόνα 6, κάποιες κύριες τεχνικές συσταδοποίησης χωρίζονται σε κατηγορίες οι οποίες είναι βασισμένες σε πρότυπα (K-means), σε συνδέσεις (Hierarchical) και στην πυκνότητα των αντικειμένων (DBSCAN).

3.1 Ο αλγόριθμος K-means

Ο αλγόριθμος K-means είναι ένας αλγόριθμος ο οποίος προτάθηκε για πρώτη φορά από τον Mac-Queen [35] και είναι παρόμοιος με τον αλγόριθμο του Lloyd [36]. Ο αλγόριθμος του K-means είναι ένας επαναληπτικός αλγόριθμος, ο οποίος χωρίζει τα αντικείμενα του συνόλου σε συστάδες. Η τεχνική του K-means είναι μια τεχνική συσταδοποίησης βασισμένη σε πρότυπο, δηλαδή όλα τα αντικείμενα της συστάδας είναι περισσότερο κοντά στο κέντρο της συγκεκριμένης συστάδας σε σύγκριση με την απόσταση από τα κέντρα των άλλων συστάδων [37].



Εικόνα 7 - Παράδειγμα δισδιάστατου πίνακα κάνοντας χρήση του εργαλείου Matlab.

Ιδιότητες	1η	2η
A	1.0	1.0
B	1.5	2.0
Γ	3.0	4.0
Δ	5.0	7.0
E	3.5	5.0
Z	4.5	5.0
H	3.5	4.5
Θ	4.1	1
I	5.4	1.5

Εικόνα 8 - Τα περιεχόμενα του πίνακα 7.

Στην Εικόνα 7 παρατηρείται πως αναπαρίστανται, σε διάγραμμα, τα αντικείμενα του πίνακα της Εικόνας 8, χρησιμοποιώντας το εργαλείο της Matlab. Στη συνέχεια, παρατηρούνται τα αποτελέσματα του αλγορίθμου για διαφορετικά αρχικά κέντρα.

Υπάρχουν περιπτώσεις στις οποίες ο αλγόριθμος K-means θα εξάγει διαφορετικά αποτελέσματα εξαιτίας της τυχαιότητας της επιλογής των αρχικών σημείων. Τα βήματα του αλγορίθμου K-means φαίνονται παρακάτω σε μορφή ψευδοκώδικα [24].

K-means Αλγόριθμος

Είσοδος: ο αριθμός των συστάδων **k**,
το dataset **D** = { x_1 , x_2 , … , x_n }
Έξοδος: **k** συστάδες

Βήμα 1ο : Επέλεξε K σημεία ως αρχικά κέντρα βάρους (centroid).

Βήμα 2ο : Υπολόγισε την Ευκλείδεια απόσταση μεταξύ όλων των σημείων και των κέντρων βάρους, σύμφωνα με τον παρακάτω τύπο:

$$d(x_q, x_j) = \sqrt{((x_{q1} - x_{j1})^2 + \ldots + (x_{qn} - x_{jn})^2)} \tag{11}$$

Βήμα 3ο : Ανάλογα με τις αποστάσεις που προέκυψαν από το βήμα 2 , τα κέντρα βάρους υπολογίζονται εκ νέου χρησιμοποιώντας τον εξής τύπο:

$$g_i = \frac{1}{m} \cdot \sum_{x \in C_i} x \tag{12}$$

όπου $\mathbf{g_i}$ το νέο κέντρο βάρους για την i-οστή συστάδα, **m** το σύνολο των αντικειμένων στην i-οστή συστάδα και $\mathbf{C_i}$ είναι η i-οστή συστάδα.

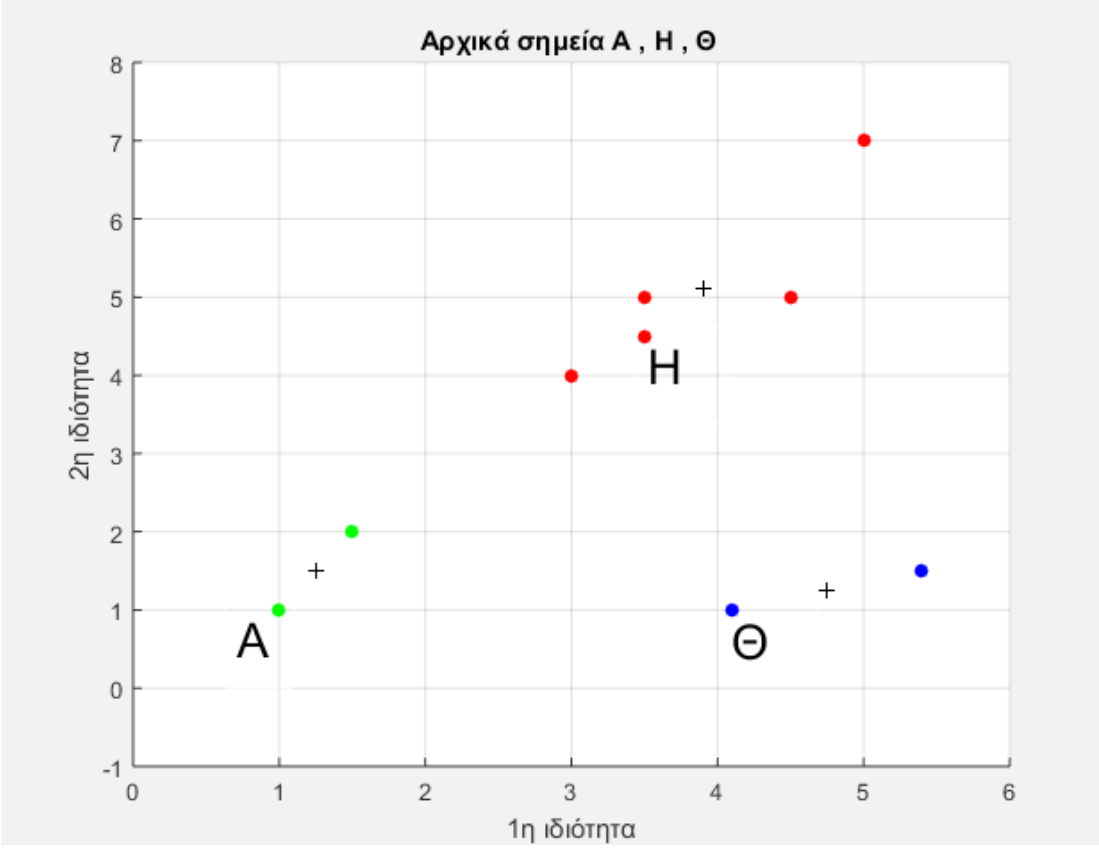
Βήμα 4ο : Υπολογίζουμε το άθροισμα του τετραγωνικού σφάλματος (SSE) :

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(g_i, x) \tag{13}$$

όπου x είναι ένα σημείο στη συστάδα C_i και g_i είναι το κέντρο βάρους της συστάδας C.

Βήμα 5ο : Σε περίπτωση που το SSE πληρεί τις προϋποθέσεις τερματισμού (πχ. η απόκλιση των νέων κέντρων από τα κέντρα της προηγούμενης επανάληψης να είναι μικρότερη της τάξεως του 10^{-3}), τότε ο αλγόριθμος σταματάει. Διαφορετικά επιστρέφουμε στο βήμα 2.

Η μετρική μονάδα που χρησιμοποιείται σε αυτόν τον αλγόριθμο είναι η Ευκλείδεια απόσταση. Η επιλογή των αρχικών κέντρων γίνεται με τυχαίο τρόπο από τον αλγόριθμο με αποτέλεσμα να υπάρχουν περιπτώσεις στις οποίες οι συστάδες που δημιουργούνται να διαφέρουν ανάλογα με την επιλογή των αρχικών κέντρων, γεγονός που γίνεται αντιληπτό με το παρακάτω παράδειγμα.

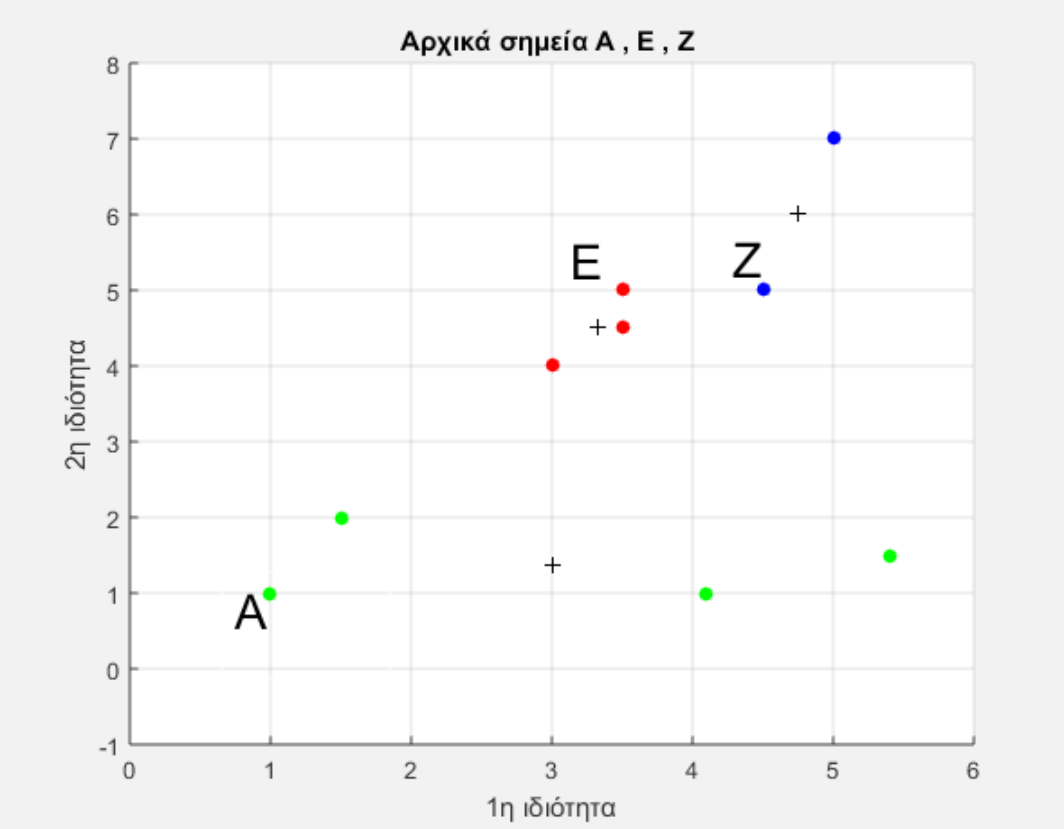


Εικόνα 9 - Αποτελέσματα K-means για τα αρχικά σημεία Α, Η, Θ.

```
ans =

    5.0000    7.0000
    3.6250    4.6250
    3.0000    1.3750
```

Εικόνα 10 - Συντεταγμένες των κέντρων των συστάδων της Εικόνας 9.



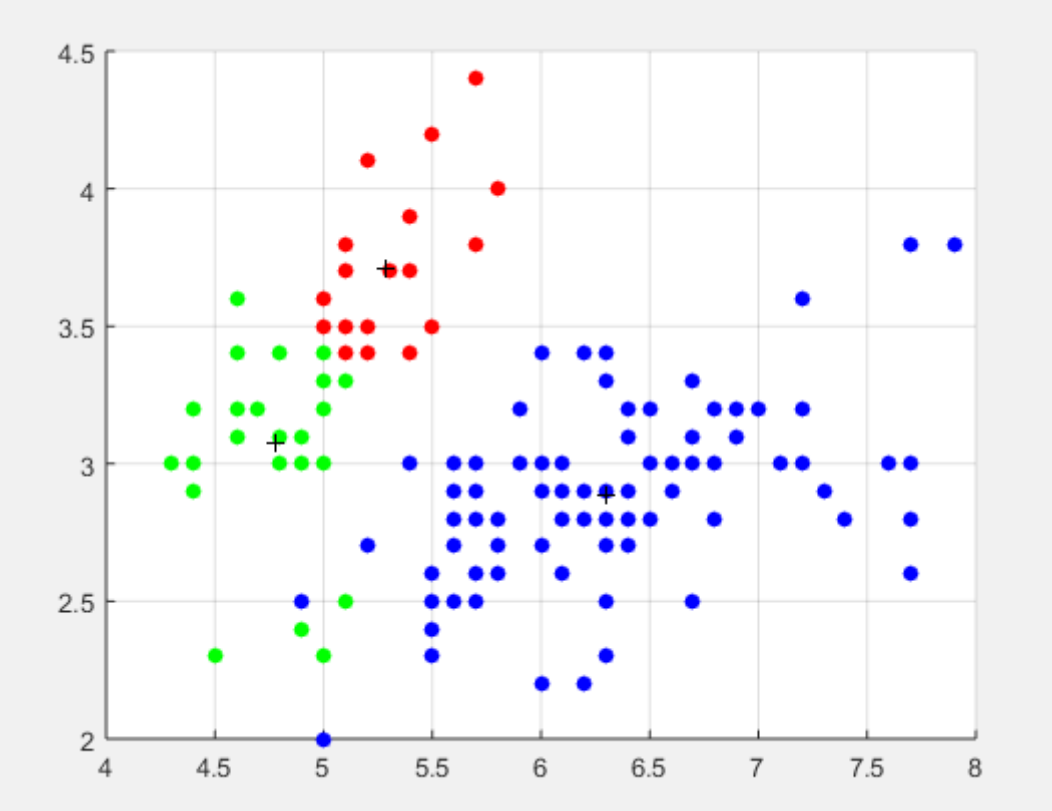
Εικόνα 11 - Αποτελέσματα K-means για τα αρχικά σημεία A, E, Z.

```
ans =

    3.0000    1.3750
    3.3333    4.5000
    4.7500    6.0000
```

Εικόνα 12 - Συντεταγμένες των κέντρων των συστάδων της Εικόνας 11.

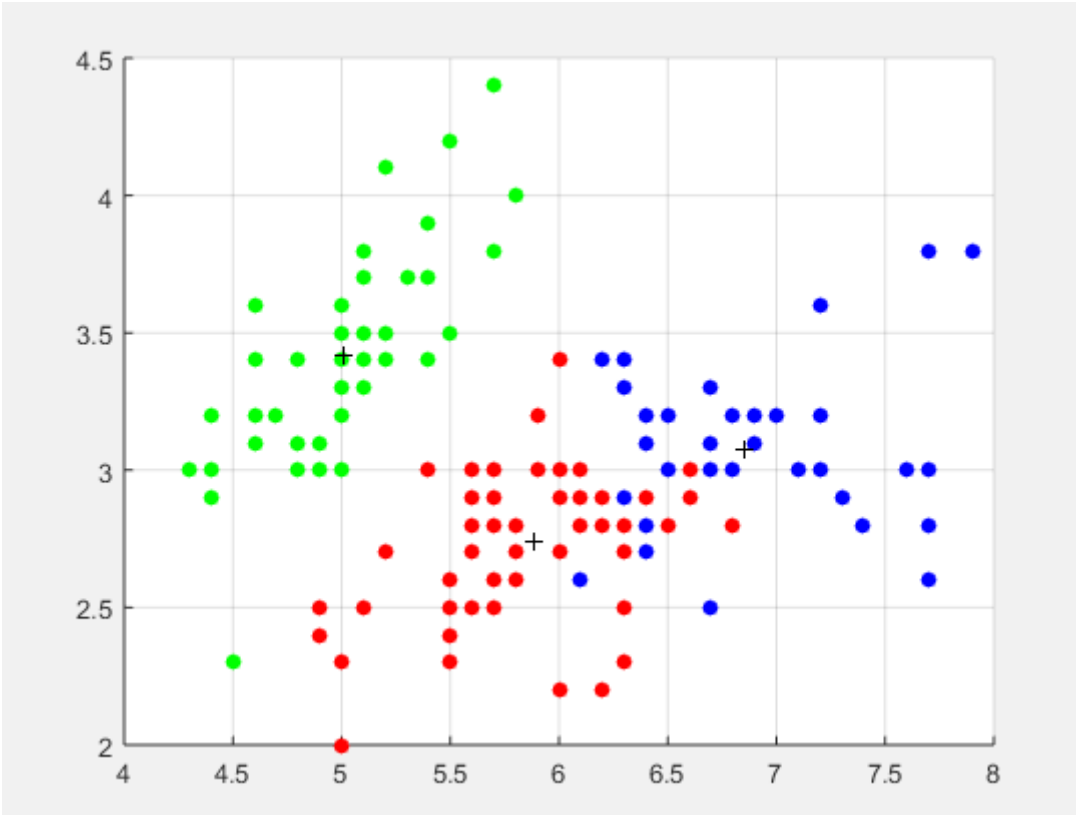
Οι Εικόνες 9 και 11 δείχνουν μια από τις αδυναμίες του αλγορίθμου K-means όσον αφορά την τυχαιότητα της επιλογής των αρχικών κέντρων. Στην Εικόνα 9 παρατηρείται ότι επιλέγοντας τα σημεία Α, Η, Θ σαν αρχικά κέντρα, ο αλγόριθμος εξάγει ως αποτέλεσμα τρεις συστάδας από αντικείμενα των οποίων τα κέντρα είναι οι τιμές που εμφανίζονται στην Εικόνα 10. Τα κέντρα κάθε συστάδας εμφανίζονται στο σχεδιάγραμμα ως ένας μαύρος σταυρός '+', ενώ τα αντικείμενα κάθε συστάδας αναπαρίστανται με διαφορετικό χρώμα. Στην πραγματικότητα, ο αλγόριθμος εφαρμόζεται σε δεδομένα πολλών διαστάσεων και με πολλές περισσότερες γραμμές. Παρακάτω, χρησιμοποιώντας το Iris dataset από τη βάση δεδομένων του πανεπιστημίου της Καλιφόρνιας (University of California, Irvine - UCI) UCI Knowledge Discovery Archive [38], ο αλγόριθμος εξάγει τα αποτελέσματα της Εικόνας 13.



Εικόνα 13 - Αποτελέσματα του K-means στο Iris dataset.

Το Iris dataset είναι ένα από τα πιο γνωστά και απλά dataset όλης της βάσης δεδομένων του UCI παρά το γεγονός ότι περιέχει αντικείμενα τεσσάρων διαστάσεων, αριθμός σχετικά μικρός σε

σχέση με τις διαστάσεις των υπολοίπων dataset. Το περιεχόμενο των αντικειμένων είναι πραγματικοί αριθμοί (REAL) ενώ ο συνολικός αριθμός των γραμμών (tuples) είναι 150. Βέβαια, το κάθε dataset περιέχει και κάποιο αναγνωριστικό (identifier), το οποίο βοηθάει στο να εξαχθούν κάποια στατιστικά αποτελέσματα. Τα στατιστικά αποτελέσματα θα μελετηθούν και θα σχολιαστούν στο κεφάλαιο 5. Στην Εικόνα 14 παρατηρούνται τα αποτελέσματα του μειονεκτήματος που έχει ο αλγόριθμος του K-means σε ένα αρκετά μεγάλο dataset.

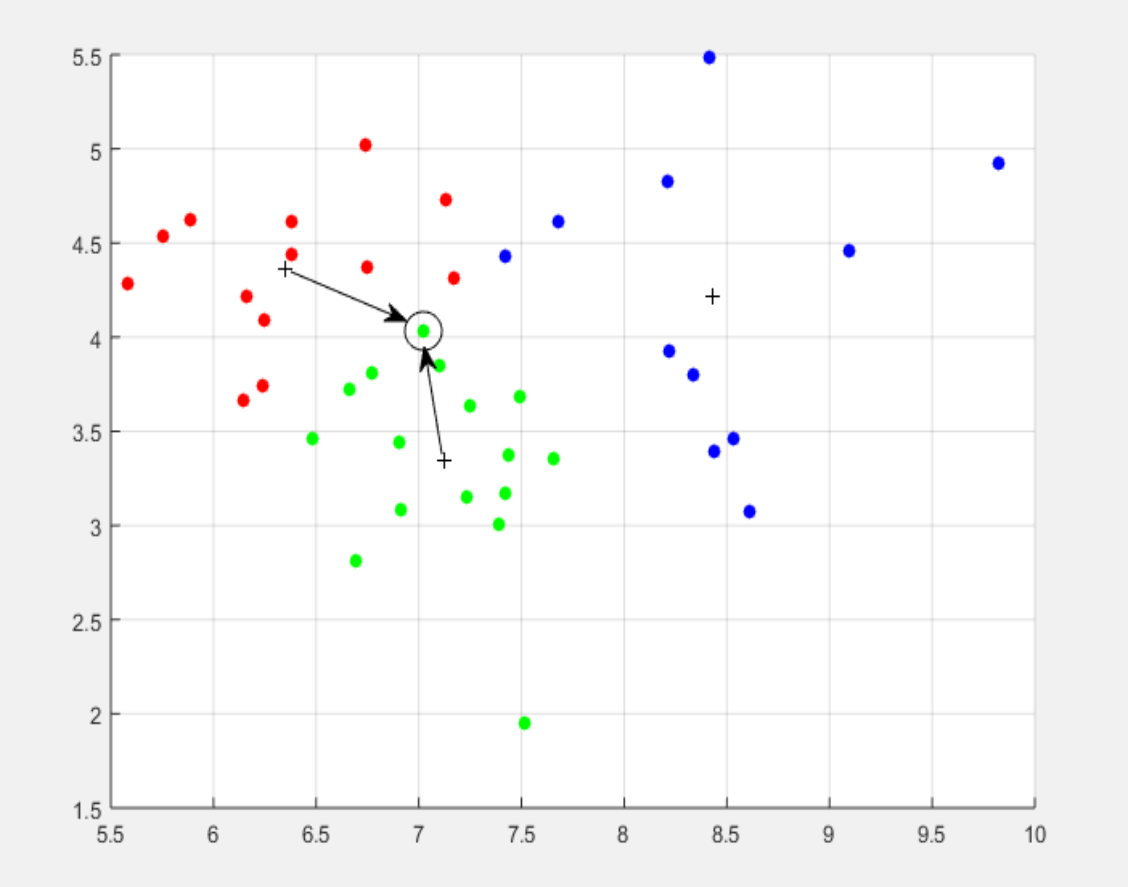


Εικόνα 14 - Αποτελέσματα του K-means στο Iris dataset για διαφορετικά αρχικά κέντρα.

Στις εικόνες 13, 14 παρατηρείται ότι υπάρχουν σημεία που φαινομενικά είναι πλησιέστερα σε κέντρα διαφορετικών συστάδων σε σχέση με αυτά που δείχνουν το χρώμα τους και κατά συνέπεια από αυτά που πραγματικά ανήκουν. Αυτό είναι αποτέλεσμα του γεγονότος ότι τα αντικείμενα, τα οποία είναι τεσσάρων διαστάσεων, αναπαρίστανται σε ένα σχεδιάγραμμα δύο διαστάσεων.

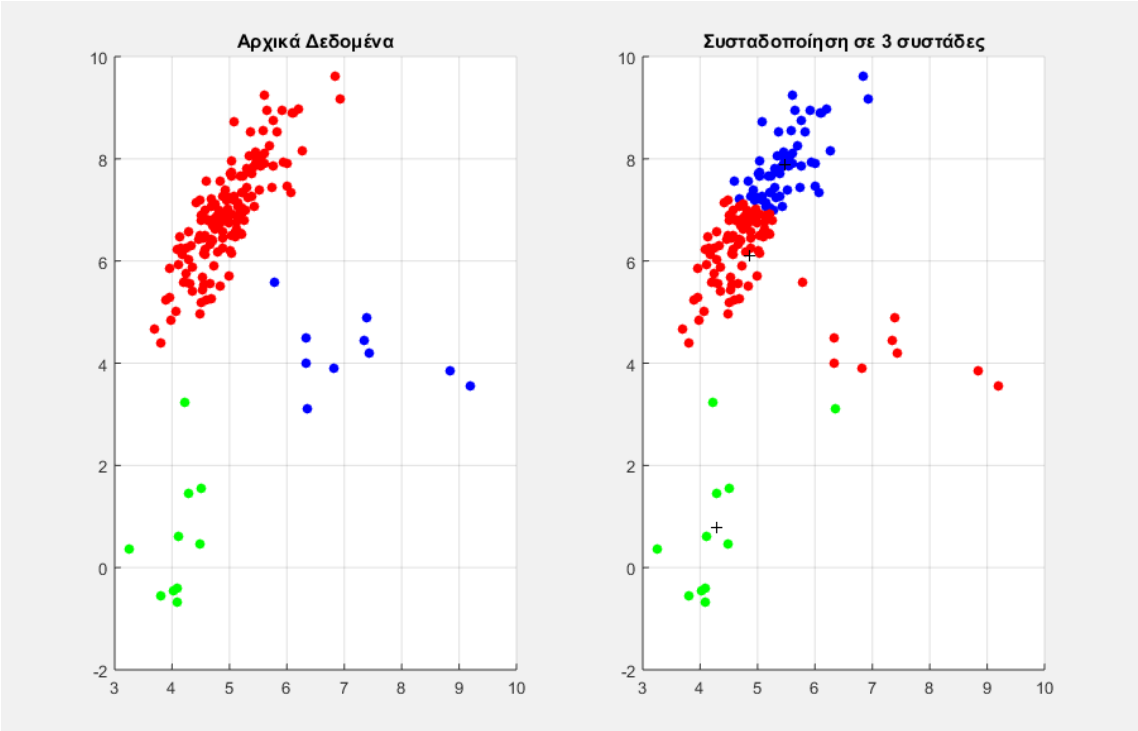
Επίσης, πέρα από το μειονέκτημα που έχει ο αλγόριθμος εξαιτίας της τυχαίας επιλογής αρχικών σημείων, ο K-means είναι αρκετά προβληματικός όταν τα δεδομένα περιέχουν ακραίες τιμές ή τους λεγόμενους outliers. Οι outliers είναι αντικείμενα τα οποία ανήκουν σε κάποια συστάδα αλλά έχουν αρνητική επίδραση στον σωστό υπολογισμό του κέντρου της συστάδας [39]. Όπως είδαμε και στον αλγόριθμο του K-means, για να υπολογιστεί το κέντρο μια συστάδας πρέπει να βρεθεί το μέσο όλων των αντικειμένων που περιέχονται στη συστάδα. Όταν, όμως, η συστάδα περιέχει ακραίες τιμές, αυτό έχει ως αποτέλεσμα την μετακίνηση του κέντρου της συστάδας προς τις ακραίες τιμές.

Ακόμη, υπάρχουν περιπτώσεις στις οποίες οι outliers έχουν κατανεμηθεί από τον αλγόριθμο σε λάθος συστάδα. Ένα τέτοιο παράδειγμα απεικονίζεται στην Εικόνα 15. Το σημείο που σημειώνεται με κύκλο έχει τοποθετηθεί σε λάθος συστάδα καθώς η Ευκλείδεια απόσταση του σημείου από τη συστάδα με το κόκκινο χρώμα είναι μικρότερη από αυτήν με το πράσινο χρώμα.



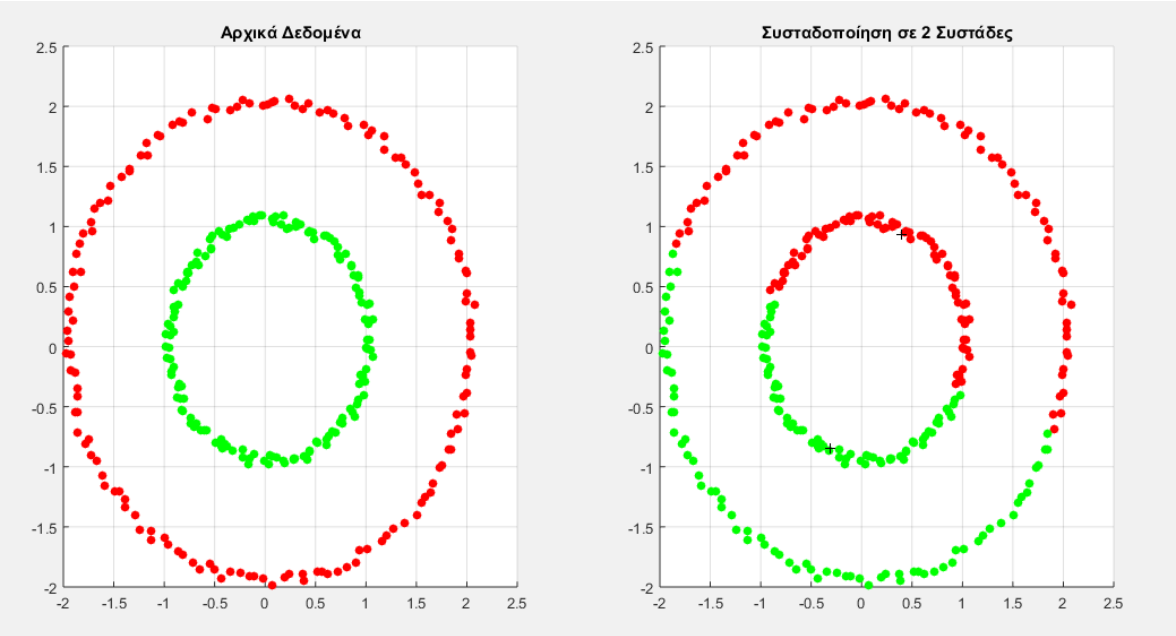
Εικόνα 15 - Παράδειγμα του προβλήματος των ακραίων τιμών (outliers) στον αλγόριθμο K-means.

Ακόμη ένα πρόβλημα που εμφανίζει ο αλγόριθμος K-means είναι η αδυναμία του να ξεχωρίσει συστάδες οι οποίες έχουν διαφορετικά μεγέθη. Στην Εικόνα 16, τα αποτελέσματα αναδεικνύουν την συγκεκριμένη αδυναμία καθώς, όπως φαίνεται στο δεξιό σχεδιάγραμμα, ο αλγόριθμος διασπά την μεγάλη συστάδα σε δύο μικρότερες καταλήγοντας σε δημιουργία τριών λανθασμένων συστάδων σε σχέση με τα αρχικά δεδομένα.



Εικόνα 16 - Πρόβλημα του αλγορίθμου για συστάδες διαφορετικών μεγεθών.

Ένα ακόμη πρόβλημα του αλγορίθμου παρατηρείται στην Εικόνα 17. Το πρόβλημα βασίζεται στο γεγονός ότι ο αλγόριθμος δεν παρουσιάζει τα επιθυμητά αποτελέσματα όταν τα δεδομένα αναπαριστάνουν σχηματικές δομές.



Εικόνα 17 - Πρόβλημα του αλγορίθμου K-means για δεδομένα που αναπαριστούν σχηματικές δομές.

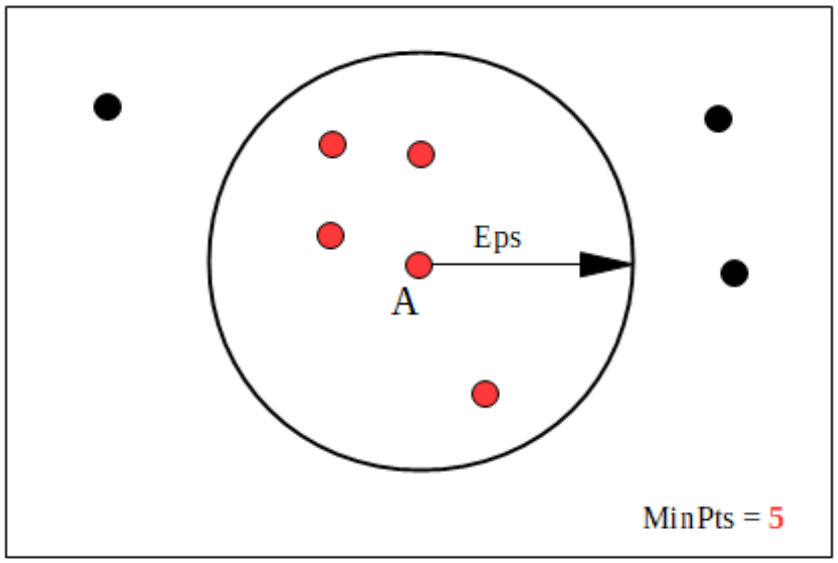
3.2 Ο αλγόριθμος DBSCAN

Ο αλγόριθμος DBSCAN αποτελεί, όπως αναφέρεται και στην ονομασία του, έναν αλγόριθμο συσταδοποίησης ο οποίος βασίζεται στη πυκνότητα των δεδομένων για να εξάγει συμπεράσματα [40]. Προτάθηκε για πρώτη φορά το 1996 από τους Martin Ester, Jörd Sander, Hans-Peter Kriegel και Xiaowei Xu [41]. Ο αλγόριθμος παρέχει το πλεονέκτημα ότι αναγνωρίζει συστάδες από αντικείμενα που σχηματίζουν σχηματικές δομές, σε αντίθεση με τις μεθόδους διαμεριστικής και ιεραρχικής συσταδοποίησης.

Ο συγκεκριμένος αλγόριθμος είναι πολύ αποτελεσματικός στην ανίχνευση και στην περιθωριοποίηση των outliers, γεγονός που συμβαίνει λόγο της φιλοσοφίας που έχει ο αλγόριθμος DBSCAN [42]. Πιο συγκεκριμένα, ο αλγόριθμος DBSCAN λειτουργεί χρησιμοποιώντας δύο βασικές παραμέτρους:

- **Eps**, παράμετρος η οποία αναφέρεται στην ακτίνα με την οποία σχηματίζεται μια γειτονιά γύρω από κάθε σημείο.

- **MinPts**, παράμετρος η οποία αναφέρεται στον ελάχιστο αριθμό από σημεία τα οποία πρέπει να περιέχονται μέσα στη γειτονιά που σχηματίζεται από την τιμή Eps.

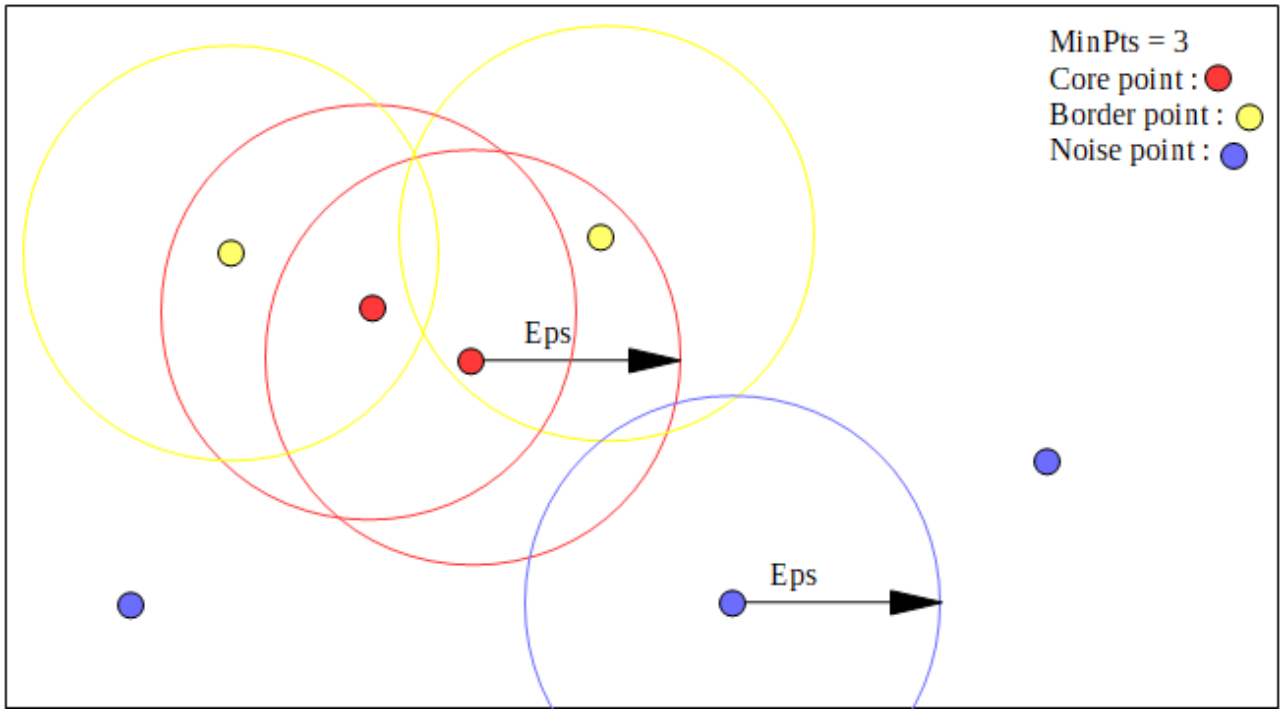


Εικόνα 18 - Παράδειγμα δημιουργίας γειτονιάς στον DBSCAN.

Αξίζει να σημειωθεί ότι το σημείο γύρω από το οποίο σχηματίζεται η κάθε γειτονιά συμπεριλαμβάνεται στη παράμετρο MinPts. Όπως φαίνεται και στην Εικόνα 18, γύρω από το σημείο Α δημιουργείται μία γειτονιά με ακτίνα **Eps** και με **MinPts** = 5.

Με βάση τις παραμέτρους που αναφέρθηκαν προηγουμένως, τα αντικείμενα χωρίζονται στις παρακάτω τρεις κατηγορίες:

- **Σημεία πυρήνα (core points)**, ονομάζουμε τα σημεία τα οποία περιέχουν μέσα στη γειτονιά τους αριθμό μεγαλύτερο ή ίσο από MinPts σημεία.
- **Σημεία ορίου (border points)**, ονομάζουμε τα σημεία τα οποία έχουν μικρότερο από MinPts αριθμό σημείων στην γειτονιά τους, αλλά έχουν το χαρακτηριστικό ότι περιέχονται στη γειτονιά ενός σημείου πυρήνα.
- **Σημεία θορύβου (noise points)**, ονομάζουμε τα σημεία τα οποία δεν είναι ούτε σημεία πυρήνα ούτε σημεία ορίου.



Εικόνα 19 - Παράδειγμα κατηγοριοποίησης των σημείων ανάλογα με τις παραμέτρους Eps,MinPts.

Όπως αναφέρθηκε προηγουμένως, ο αλγόριθμος DBSCAN είναι αρκετά αποτελεσματικός απέναντι σε outliers αλλά και απέναντι σε δεδομένα που δημιουργούν αυθαίρετες σχηματικές δομές [43]. Ένα επιπλέον πλεονέκτημα αυτού του αλγορίθμου είναι ότι ο αριθμός των συστάδων δεν είναι προκαθορισμένος αλλά οι συστάδες δημιουργούνται κατά την εκτέλεση του αλγορίθμου. Βέβαια ο αλγόριθμος έχει και τα μειονεκτήματά του. Καταρχάς, ο αλγόριθμος δεν λειτουργεί αποτελεσματικά όταν τα δεδομένα έχουν μεγάλες διαφορές στις πυκνότητες που δημιουργούν επειδή η τιμή της παραμέτρου MinPts που σχετίζεται με τη πυκνότητα είναι προκαθορισμένη σε όλη τη διάρκεια εκτέλεσης του αλγορίθμου. Επίσης, υπάρχει η λεγόμενη "κατάρα των διαστάσεων", η οποία αναφέρεται στο πρόβλημα που δημιουργείται όταν ο αριθμός των διαστάσεων ενός dataset είναι πολύ μεγάλος και η ύπαρξη αυτού του προβλήματος οφείλεται στη συνάρτηση απόστασης, στη περίπτωσή αυτή είναι η Ευκλείδεια απόσταση, με αποτέλεσμα να αλλοιώνεται η ποιότητα των αποτελεσμάτων του αλγορίθμου. Ένα ακόμη μειονέκτημα είναι απόρροια του τελευταίου βήματος του ψευδοκώδικα του αλγορίθμου DBSCAN [44].

DBSCAN Αλγόριθμος

Είσοδος: η ακτίνα της γειτονιάς **Eps**,
η παράμετρος που ορίζει τη πυκνότητα των σημείων **MinPts**,
το dataset **D = { x_1 , x_2 , … , x_n }**
Έξοδος: συσταδοποιημένο **D**

Βήμα 1ο : Υπολόγισε την Ευκλείδεια απόσταση μεταξύ όλων των σημείων και των κέντρων βάρους, σύμφωνα με τον παρακάτω τύπο:

$$d(x_q, x_j) = \sqrt{\left((x_{q1} - x_{j1})^2 + \ldots + (x_{qn} - x_{jn})^2\right)} \tag{11}$$

Βήμα 2ο : Κατηγοριοποίησε κάθε σημείο σε σημείο πυρήνα, σημείο ορίου και σημείο θορύβου.

Βήμα 3ο : Διέγραψε τα σημεία θορύβου.

Βήμα 4ο : Βρες τα σημεία πυρήνα που είναι σε απόσταση μικρότερη ή ίση από Eps μεταξύ τους.

Βήμα 5ο : Για κάθε σύνολο από σημεία πυρήνα που δημιουργούνται στο βήμα 4 δημιούργησε μια ξεχωριστή συστάδα.

Βήμα 6ο : Ανάθεσε κάθε σημείο ορίου στη συστάδα που ανήκει το σημείο πυρήνα το οποίο περιέχει στη γειτονιά του το σημείο ορίου.

Το πρόβλημα που δημιουργείται στο τελευταίο βήμα του αλγορίθμου είναι ότι υπάρχει πιθανότητα τα σημεία ορίου να ανατίθενται σε διαφορετικές συστάδες μετά από πολλές εκτελέσεις του αλγορίθμου, εάν αυτά τα σημεία ανήκουν στη γειτονιά δύο ή περισσότερων σημείων πυρήνα. Αυτό εξαρτάται από τη σειρά με την οποία θα επεξεργαστεί τα δεδομένα ο αλγόριθμος.

Βέβαια, όπως παρατηρείται στον παραπάνω αλγόριθμο, οι παράμετροι δίνονται έτοιμοι σαν είσοδο στον αλγόριθμο. Επειδή, όμως, ο αλγόριθμος βασίζεται στην πυκνότητα, αυτό έχει σαν αποτέλεσμα η παραμικρή μεταβολή στις τιμές των παραμέτρων να παράγει εντελώς διαφορετικά αποτελέσματα στις συστάδες που θα δημιουργηθούν από τον αλγόριθμο. Η ιδιαιτερότητα, αυτή, του αλγορίθμου μπορεί να αντιμετωπιστεί χρησιμοποιώντας μια συνάρτηση που θα βελτιστοποιεί τις τιμές που θα εισαχθούν στον αλγόριθμο DBSCAN σε σχέση με τα δοθέντα δεδομένα. Η συνάρτηση αυτή είναι η λεγόμενη k-distance συνάρτηση και ο τρόπος λειτουργίας του φαίνεται στο ψευδοκώδικα παρακάτω.

k-distance

Είσοδος: το dataset **D = { x_1 , x_2 , … , x_n }**
Έξοδος: βελτιστοποιημένες τιμές για τις παραμέτρους **Eps** και **MinPts**

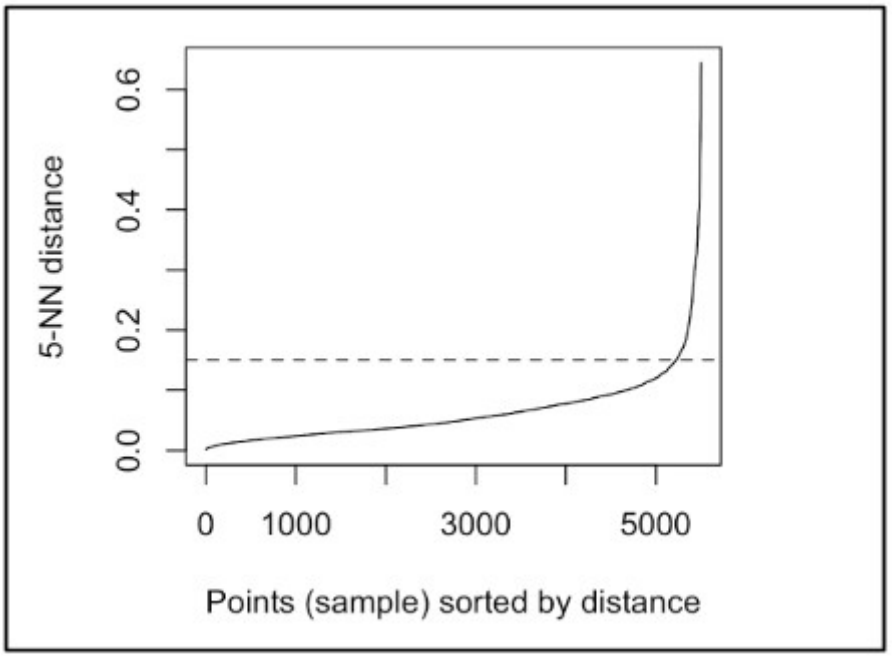
Βήμα 1ο : Υπολόγισε την Ευκλείδεια απόσταση k-distance όλων των σημείων για κάποια τιμή k :

$$d(x_q, x_j) = \sqrt{\left((x_{q1} - x_{j1})^2 + \ldots + (x_{qn} - x_{jn})^2\right)} \quad (11)$$

Βήμα 2ο : Ταξινόμησε τις τιμές των αποστάσεων σε αύξουσα σειρά.

Βήμα 3ο : Κάνε τη γραφική παράσταση των παραπάνω τιμών.

Βήμα 4ο : Παρατηρώντας τη γραφική παράσταση, επέλεξε ως τιμή της παραμέτρου **Eps** την τιμή στην οποία φαίνεται μια απότομη αλλαγή στη καμπύλη της γραφικής παράστασης. Η παράμετρος **MinPts** παίρνει την τιμή του k.



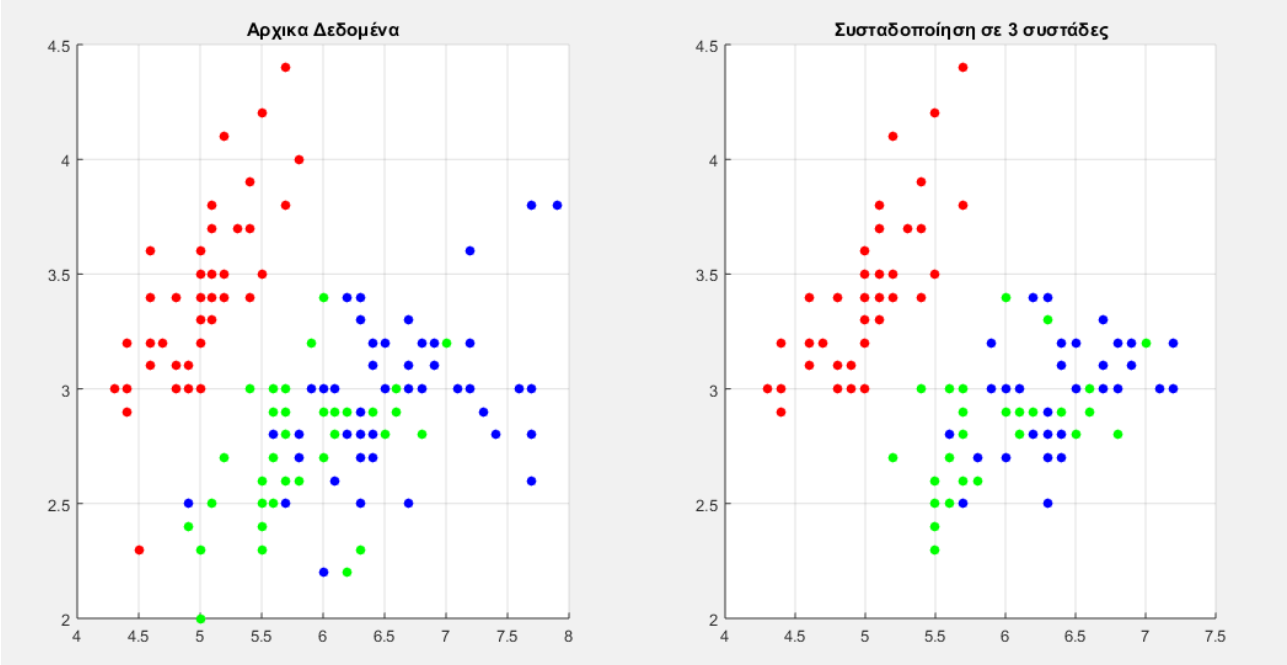
Εικόνα 20 - Καμπύλη επιλογής της παραμέτρου Eps για MinPts = 5 [45].

Στον ψευδοκώδικα της συνάρτησης k-distance γίνεται η επιλογή μιας αυθαίρετης τιμής για το k. Αυτή η τιμή, όμως, πρέπει να μεταβληθεί αρκετές φορές και να δοκιμαστεί για να επιλεχθεί η τιμή που εξυπηρετεί καλύτερα τον αλγόριθμο.

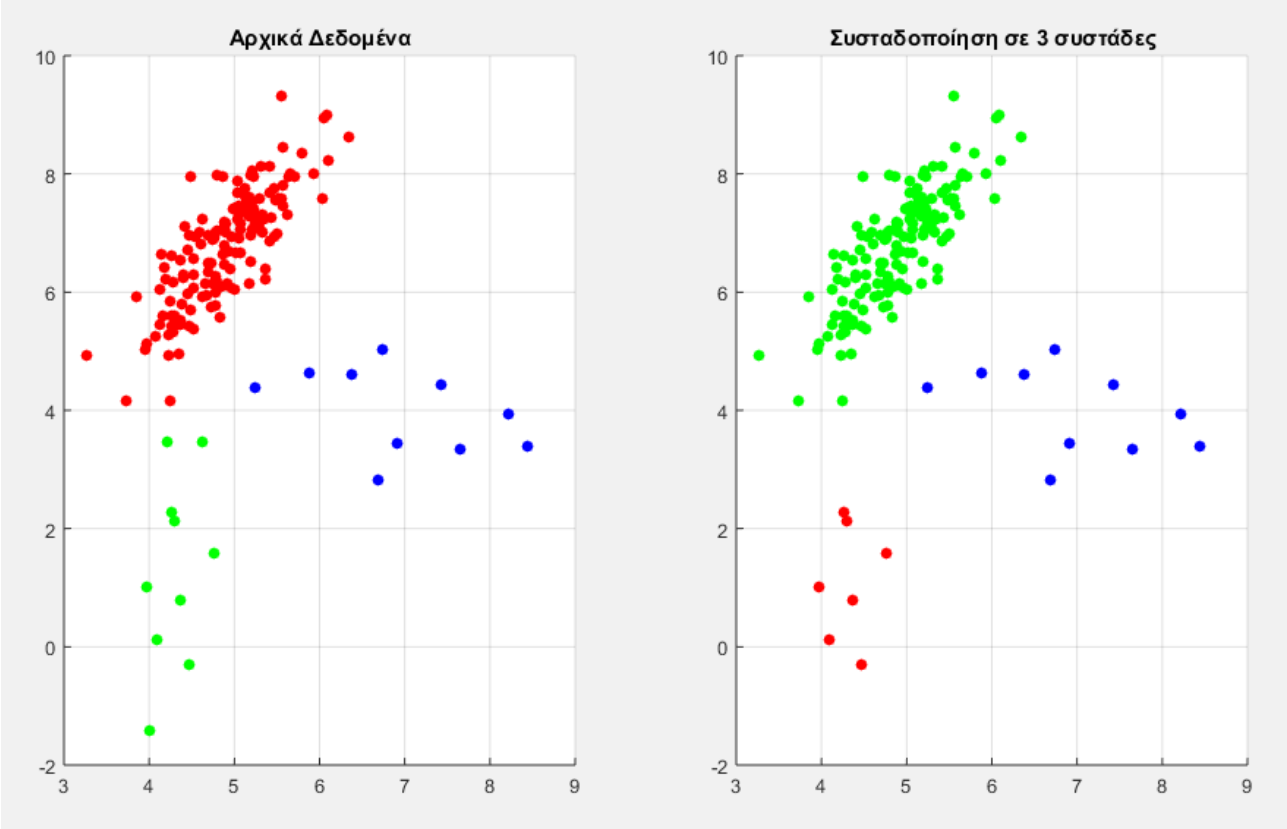
Στην Εικόνα 20, παρατηρείται η γραφική παράσταση που δημιουργείται από τη συνάρτηση k-distance για τιμή του k=5 και για τη συγκεκριμένη τιμή επιλέγεται Eps = 0.15, δηλαδή το σημείο εκείνο που υπάρχει η απότομη αλλαγή στη καμπύλη.

Λαμβάνοντας υπόψη όσα προαναφέρθηκαν, εφαρμόζοντας τον αλγόριθμο DBSCAN στο Iris dataset, τα αποτελέσματα του αλγορίθμου εμφανίζονται στην Εικόνα 21. Η επιλογή των παραμέτρων έγινε με τη μέθοδο του k-distance, με αποτέλεσμα οι τιμές των παραμέτρων εισόδου να είναι οι εξής: **Eps=0.4 , MinPts=5**.

Επίσης, στην Εικόνα 21 παρατηρείται ότι τα αρχικά δεδομένα περιέχουν σημεία τα οποία δεν υπάρχουν στα αποτελέσματα του αλγορίθμου στη δεξιά εικόνα. Τα σημεία που δεν εμφανίζονται είναι οι outliers αποδεικνύοντας την αποτελεσματικότητα του αλγορίθμου απέναντι σε ακραίες τιμές που προκαλούν θόρυβο και επηρεάζουν αρνητικά τα αποτελέσματα.



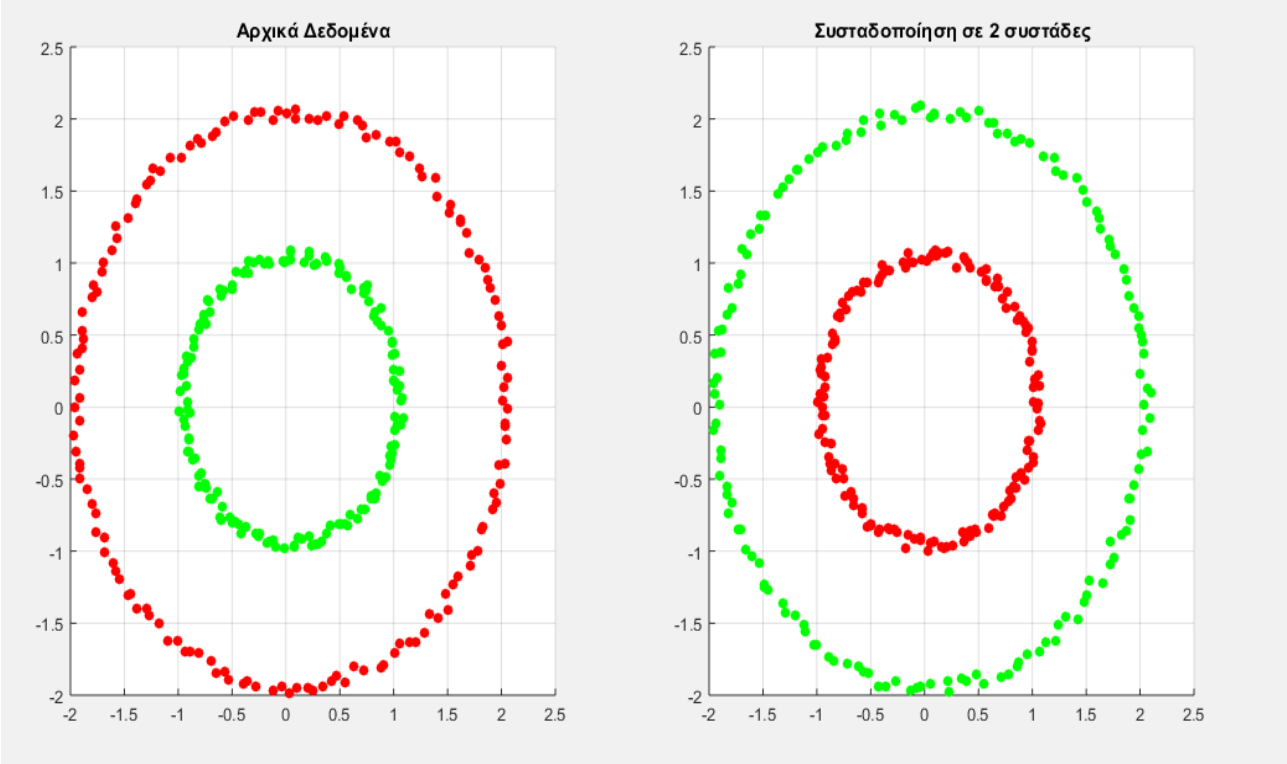
Εικόνα 21 - Παράδειγμα εφαρμογής του αλγορίθμου DBSCAN στο Iris dataset για Eps=0.4 και MinPts = 5.



Εικόνα 22 - Παράδειγμα αντιμετώπισης του προβλήματος για δεδομένα με διαφορετικές πυκνότητες με κατάλληλη επιλογή των παραμέτρων εισόδου.

Ακόμη, στην Εικόνα 22 παρατηρείται ότι ο αλγόριθμος αναγνωρίζει τη διαφορά στη πυκνότητα των δεδομένων και εξάγει σωστά αποτελέσματα. Βέβαια, όσο μεγαλώνει ο αριθμός των διαστάσεων, τόσο αυξάνεται και η δυσκολία του αλγορίθμου να διαχωρίζει δεδομένα με σημαντικά μεγάλες διαφορές στις πυκνότητες.

Τέλος, στην Εικόνα 23 παρατηρείται η ικανότητα του αλγορίθμου να αναγνωρίζει σχηματικές δομές. Ασφαλώς, όμως, θα υπάρχουν και σε αυτή την περίπτωση αρκετοί outliers ανάλογα με τις παραμέτρους εισόδου που θα εισαχθούν, πράγμα που γίνεται αντιληπτό στο δεξιό κομμάτι της Εικόνας 23, στην οποία υπάρχουν μικρές ασυνέχειες των δύο σχηματικών δομών.



Εικόνα 23 - Παράδειγμα επιτυχούς αναγνώρισης σχηματικών δομών από τον αλγόριθμο DBSCAN.

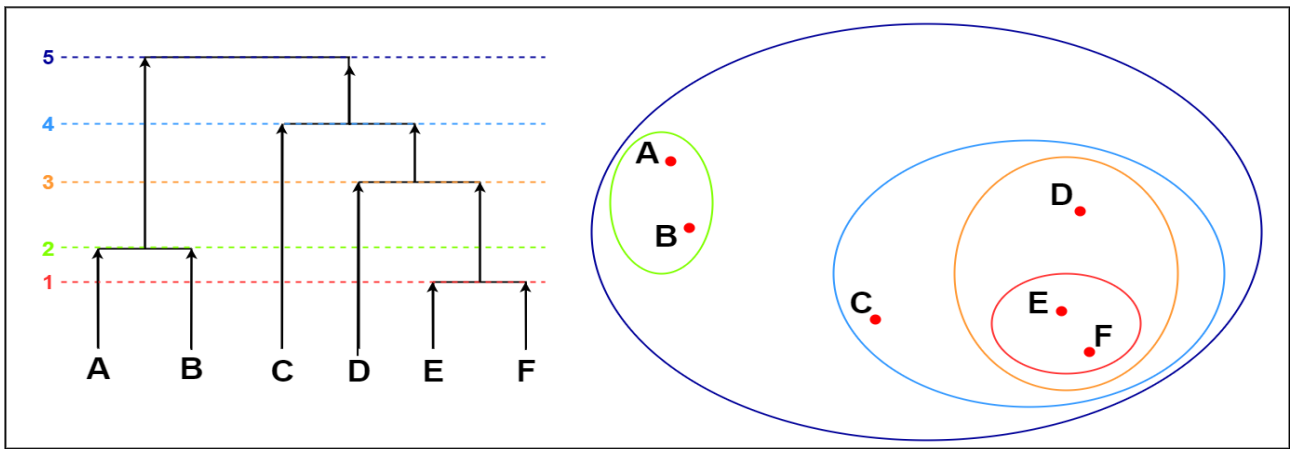
3.3 Ο αλγόριθμος Ιεραρχικής Συσταδοποίησης

Η ιεραρχική συσταδοποίηση είναι μια ευρέως γνωστή μέθοδος συσταδοποίησης εξαιτίας του γεγονότος ότι μπορεί να παράξει αρκετά ποιοτικές συστάδες. Το κύριο χαρακτηριστικό αυτής της μεθόδου είναι το γεγονός ότι οι συστάδες που δημιουργούνται μπορούν να αναπαρασταθούν σαν ένα δένδρο ιεραρχίας ή δενδρόγραμμα και με αυτόν τον τρόπο δημιουργείται μια ιεραρχία από συστάδες [46][47]. Κάθε κόμβος στο δενδρόγραμμα αναπαριστά μία συστάδα. Η ιεραρχική συσταδοποίηση μπορεί να χωριστεί σε δύο κατηγορίες ανάλογα με τον τρόπο που διαχειρίζονται τα δεδομένα. Οι δυο κατηγορίες είναι οι εξής:

- **Συσσωρευτική (Agglomerative) Συσταδοποίηση**
- **Διαχωριστική (Divisive) Συσταδοποίηση**

Η συσσωρευτική συσταδοποίηση λειτουργεί τη μέθοδο "bottom-up", δηλαδή κάθε αντικείμενο από τα αρχικά δεδομένα θεωρείται ως μία ξεχωριστή συστάδα και καθώς ανεβαίνουμε επίπεδα στο δένδρο ιεραρχίας, τα αντικείμενα αυτά ενώνονται σε μια ενιαία συστάδα. Η ένωση αυτή γίνεται με το κριτήριο που χρησιμοποιείται σε όλους τους αλγόριθμους ομαδοποίησης και αυτό είναι η συνάρτηση απόστασης [48]. Με άλλα λόγια, τα αντικείμενα που βρίσκονται πιο κοντά ή διαφορετικά τα αντικείμενα που είναι πιο όμοια μεταξύ τους, ενώνονται σε κάθε επανάληψη. Κάθε φορά η ένωση γίνεται μεταξύ των δύο πιο όμοιων αντικειμένων και έτσι σε κάθε επίπεδο του δενδρογράμματος υπάρχει μόνο μία συστάδα.

Στην Εικόνα 24 παρατηρείται η εφαρμογή του αλγορίθμου της ιεραρχικής συσταδοποίησης σε ένα απλό παράδειγμα με έξι αντικείμενα. Ο αλγόριθμος υπολογίζει τις αποστάσεις μεταξύ των σημείων και επιλέγει να ενώσει τα δύο ομοιότερα αντικείμενα κάθε φορά μέχρι να ενωθούν όλα τα σημεία σε μια ενιαία συστάδα. Βέβαια, αυτή η μέθοδος υστερεί στο γεγονός ότι θέλει περισσότερο αποθηκευτικό χώρο και απαιτεί περισσότερο υπολογιστικό χρόνο, ενώ παράγει και περισσότερο θόρυβο όσο μεγαλύτερες είναι οι διαστάσεις των δεδομένων.



Εικόνα 24 - Παράδειγμα συσσωρευτικής ιεραρχικής συσταδοποίησης.

Η συσσωρευτική ιεραρχική συσταδοποίηση μπορεί να χωριστεί σε υποκατηγορίες ανάλογα με το πως θα οριστεί ο ορισμός της εγγύτητας των συστάδων. Οι υποκατηγορίες είναι οι εξής:

- Απλός σύνδεσμος ή Ελάχιστης απόστασης (Simple-linkage or Minimum distance) ονομάζεται η κατηγορία στην οποία η εγγύτητα μεταξύ δύο συστάδων ορίζεται ως η ελάχιστη απόσταση ανάμεσα σε οποιαδήποτε σημεία των συστάδων [49]. Ξεκινώντας με όλα τα σημεία ως μια ξεχωριστή συστάδα, στη συνέχεια ενώνονται τα δύο κοντινότερα σημεία σε μια συστάδα και σε κάθε επανάληψη γίνεται η σύγκριση κάθε εξωτερικού σημείου με όλα τα σημεία που περιέχονται στη συστάδα. Εάν η μικρότερη απόσταση του εκάστοτε σημείου είναι με μια συγκεκριμένη συστάδα, τότε το σημείο ενώνεται με τη συστάδα αυτή, ενώ σε αντίθετη περίπτωση το σημείο δημιουργεί μια άλλη συστάδα με το κοντινότερο σημείο του. Αυτή η μέθοδος είναι αρκετά καλή στη διαχείριση μη ελλειπτικών σχημάτων αλλά είναι εξίσου ευαίσθητη σε θόρυβο και ακραίες τιμές.

- Πλήρης σύνδεσμος ή Μέγιστης απόστασης (Complete-linkage or Maximum distance) ονομάζεται η κατηγορία στην οποία η εγγύτητα μεταξύ δύο συστάδων ορίζεται ως η μέγιστη απόσταση ανάμεσα σε οποιαδήποτε σημεία των συστάδων [50][51]. Η μέθοδος αυτή ξεκινάει διαλέγοντας τα δύο κοντινότερα σημεία και ενώνοντάς τα σε μια συστάδα. Στη συνέχεια όμως, αφού υπολογιστούν οι αποστάσεις, το σημείο που θα ενωθεί με τη συστάδα θα είναι αυτό το οποίο έχει τη μεγαλύτερη απόσταση με κάποιο από τα σημεία της συστάδας. Εάν, όμως, το σημείο απέχει περισσότερο με κάποιο άλλο σημείο που είναι μόνο του, τότε δημιουργείται μια άλλη συστάδα με τα δύο αυτά σημεία. Η συγκεκριμένη μέθοδος έχει την δυνατότητα να ευνοεί σφαιρικά σχήματα, είναι λιγότερο ευαίσθητη στον θόρυβο από την προηγούμενη μέθοδο αλλά μπορεί να διασπάσει και μεγάλες συστάδες.

- Μέθοδος Ward ονομάζεται η κατηγορία στην οποία η εγγύτητα μεταξύ των συστάδων ορίζεται ως η αύξηση του τετραγωνικού σφάλματος που προκύπτει από τη συγχώνευση δύο συστάδων. Έχοντας προταθεί από τον Ward το 1963, η μέθοδος αυτή, σε κάθε βήμα, υπολογίζει όλους τους πιθανούς συνδυασμούς από συστάδες και επιλέγει εκείνον τον συνδυασμό για τον οποίο η απώλεια πληροφοριών ελαχιστοποιείται [52]. Ως απώλεια πληροφοριών ορίζεται το τετραγωνικό σφάλμα που περιγράφεται από τον εξής τύπο:

$$D_w(C_i, C_j) = \sum_{x \in C_i} (x - r_i)^2 + \sum_{x \in C_j} (x - r_j)^2 + \sum_{x \in C_{ij}} (x - r_{ij})^2 \quad (14)$$

συμβολίζοντας με r_i το κέντρο της συστάδας C_i , με r_j το κέντρο της συστάδας C_j και με r_{ij} το κέντρο της συστάδας C_{ij}. Όπως γίνεται φανερό και από τον παραπάνω τύπο, η μέθοδος Ward υπολογίζει την απόσταση των σημείων από το κέντρο των συστάδων. Με άλλα λόγια, υπολογίζει τη διακύμανση των σημείων και επιλέγει τα σημεία τα οποία έχουν την ελάχιστη διακύμανση.

Παρακάτω αναλύεται ο αλγόριθμος της συσσωρευτικής ιεραρχικής συσταδοποίησης.

Αλγόριθμος Συσσωρευτικής Ιεραρχικής Συσταδοποίησης

Είσοδος: το dataset **D = { x_1 , x_2 , … , x_n }**
Έξοδος: μια συστάδα αποτελούμενη από όλα τα σημεία του **D**

Βήμα 1ο : Θεώρησε κάθε σημείο ως μια ξεχωριστή συστάδα.

Βήμα 2ο : Υπολόγισε τον πίνακα εγγύτητας (διαφέρει ανάλογα με το αν χρησιμοποιήσουμε τη μέθοδο απλού συνδέσμου, πλήρους συνδέσμου ή Ward).

Βήμα 3ο : Συγχώνευσε τις δύο κοντινότερες συστάδες.

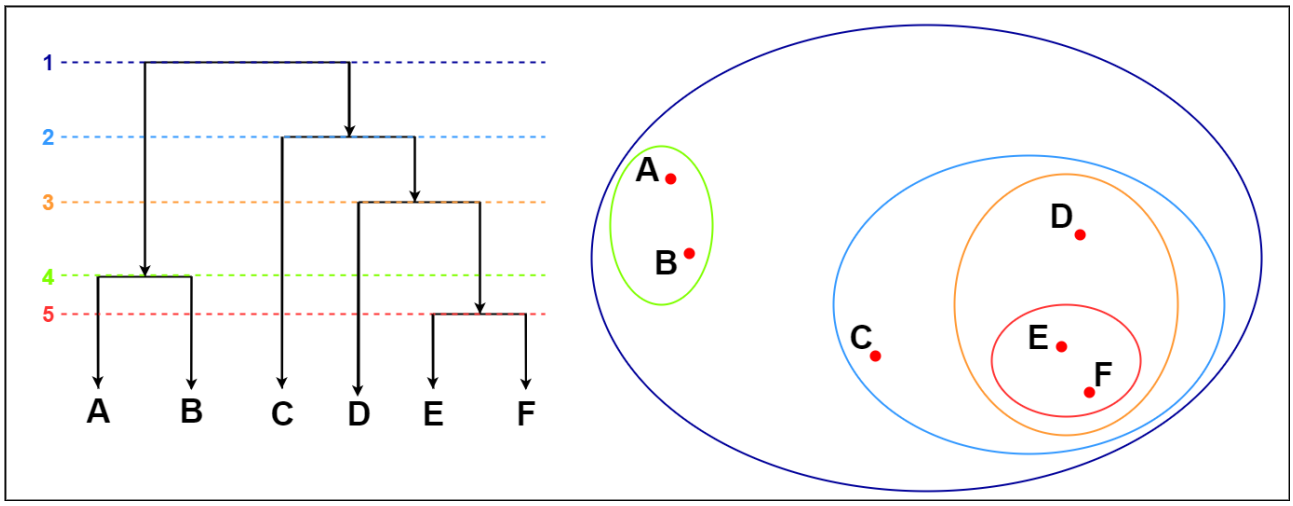
Βήμα 4ο : Ενημέρωσε τον πίνακα εγγύτητας ούτως ώστε να περιλαμβάνει τη νέα συστάδα.

Βήμα 5ο : Εάν ο αριθμός των συστάδων είναι μεγαλύτερος από ένα, τότε πήγαινε στο βήμα 3, αλλιώς τερμάτισε τερμάτισε τον αλγόριθμο.

Η διαχωριστική ιεραρχική συσταδοποίηση λειτουργεί με τη μέθοδο "top-down" και κάνει την αντίστροφη διαδικασία από την συσσωρευτική συσταδοποίηση [53]. Από την ονομασία γίνεται αντιληπτό ότι σε κάθε βήμα της μεθόδου πραγματοποιείται διάσπαση μιας συστάδας σε δύο μικρότερες. Ο αλγόριθμος ξεκινάει θεωρώντας όλα τα σημεία ως μια ενιαία συστάδα και σε κάθε επανάληψη διασπάται το σημείο ή η συστάδα που είναι λιγότερο όμοιο με το αντίστοιχο σημείο ή συστάδα που ορίσαμε ως αναφορά. Η διαδικασία αυτή συνεχίζεται μέχρι να διασπαστούν όλες οι συστάδες και κάθε σημείο της αρχικής συστάδας να αποτελεί μια συστάδα από μόνο του.

Η μέθοδος αυτή, παρόλο που είναι αρκετά περίπλοκη από την αντίστοιχη συσσωρευτική, είναι πιο γρήγορη και έχει τις πληροφορίες όλων των σημείων στην αρχή του αλγορίθμου, κάτι το οποίο δεν ισχύει στη συσσωρευτική συσταδοποίηση.

Βλέποντας την Εικόνα 25, παρατηρείται ότι εφαρμόζεται η ακριβώς αντίστροφη διαδικασία από την συσσωρευτική. Στην αρχή, ο αλγόριθμος συνθέτει όλα τα σημεία σε μια συστάδα και σε κάθε βήμα πραγματοποιείται διαχωρισμός. Παρακάτω, αναλύεται ο αλγόριθμος της διαχωριστικής συσταδοποίησης.



Εικόνα 25: Παράδειγμα διαχωριστικής ιεραρχικής συσταδοποίησης.

Αλγόριθμος Διαχωριστικής Ιεραρχικής Συσταδοποίησης

Είσοδος: το dataset $\mathbf{D} = \{ x_1 , x_2 , \dots , x_n \}$
Έξοδος: κάθε σημείο αποτελεί μια ξεχωριστή συστάδα

Βήμα 1ο : Θεώρησε όλα τα σημεία ως μια ενιαία συστάδα.

Βήμα 2ο : Επέλεξε και χώρισε το σημείο ή τη συστάδα που διαφέρει περισσότερο.

Βήμα 3ο : Εάν ο αριθμός των συστάδων είναι ίσος με τον αριθμό των σημείων που περιέχονταν στην αρχική συστάδα, τότε τερμάτισε τερμάτισε τον αλγόριθμο, αλλιώς πήγαινε στο βήμα 2.

3.4 Ο αλγόριθμος PADC

Η τεχνική PADC αποτελεί μια νέα τεχνική συσταδοποίησης, η οποία χρησιμοποιεί τον συνδυασμό της τεχνικής K-means και του differential privacy [54]. Η τεχνική αυτή βελτιώνει το μείζον πρόβλημα του K-means που αφορά την επιλογή των αρχικών σημείων καθώς ο αλγόριθμος λειτουργεί με τέτοιο τρόπο ώστε τα κέντρα που θα επιλεχθούν να είναι σχετικά με τις διαφορετικές πυκνότητες και κατανομές των δεδομένων. Επίσης, υπάρχει σημαντική βελτίωση στην ανίχνευση των ακραίων τιμών (outliers) κατά την διάρκεια εκτέλεσης του αλγορίθμου. Πιο συγκεκριμένα, ο αλγόριθμος λειτουργεί σε δύο φάσεις. Στη πρώτη φάση γίνεται η επιλογή των αρχικών κέντρων και στη δεύτερη φάση ο αλγόριθμος χρησιμοποιεί τα αρχικά κέντρα που επιλέχθηκαν στη πρώτη φάση για να εξάγει όσο το δυνατόν πιο ρεαλιστικά αποτελέσματα.

Όπως προαναφέρθηκε, η πρώτη φάση του αλγορίθμου PADC έχει ως στόχο τη βελτιστοποιημένη επιλογή των αρχικών κέντρων. Για να επιτευχθεί αυτό, βέβαια, θα πρέπει να υπάρξει περιορισμός και αν είναι δυνατόν, πλήρης απομάκρυνση όλων των outliers. Όπως φαίνεται και από τον παρακάτω ψευδοκώδικα της πρώτης φάσης, λαμβάνεται υπόψη μια νέα μεταβλητή r που καθορίζει το ποσοστό των σημείων που θα θεωρηθούν ως outliers. Εάν για παράδειγμα η τιμή της παραμέτρου είναι 0.8, τότε μόνο το 80% των σημείων συμβάλλουν στον υπολογισμό των κέντρων ενώ το υπόλοιπο 20% θεωρούνται ακραίες τιμές. Για κάθε σημείο, υπολογίζεται η μεταβλητή πυκνότητάς του, γίνεται μια ταξινόμηση των σημείων σε φθίνουσα σειρά και τα σημεία που βρίσκονται πιο χαμηλά θεωρούνται, πάντα με βάση και τη μεταβλητή r, ακραίες τιμές. Με αυτόν τον τρόπο, επιλέγονται τα αρχικά κέντρα χωρίς να επηρεάζουν καθόλου οι outliers την έκβαση των αποτελεσμάτων.

PADC Αλγόριθμος – 1η Φάση

Είσοδος: ο αριθμός των συστάδων **k**,
η παράμετρος των ακραίων τιμών **r**,
το dataset $\mathbf{D = \{ x_1 , x_2 , \dots , x_n \}}$
Έξοδος: βελτιστοποιημένα **k** αρχικά κέντρα

Βήμα 1ο : Υπολόγισε τη πυκνότητα κάθε σημείου του **D** με βάση τον παρακάτω τύπο:

$$density(x) = \frac{n}{\sum_{i=1}^{n} (dist^2(x, y_i))} \tag{15}$$

Βήμα 2ο : Ταξινόμησε τις πυκνότητες σε φθίνουσα σειρά.

Βήμα 3ο : Με βάση τη παράμετρο **r**, οι **n** * (**r** – 1) τελευταίες ταξινομημένες πυκνότητες θεωρούνται ακραίες τιμές και διαγράφονται.

Βήμα 4ο : Με βάση τη παράμετρο **k**, τα ταξινομημένα σημεία που δεν είναι ακραίες τιμές, χωρίζονται σε **k** ίσες συστάδες.

Βήμα 5ο : Τα κέντρα των συστάδων αποτελούν τα **k** αρχικά κέντρα.

Η δεύτερη φάση του αλγορίθμου PADC εκμεταλλεύεται τη βελτιστοποιημένη επιλογή των αρχικών κέντρων της πρώτης φάσης και χρησιμοποιώντας τα πλεονεκτήματα της μεθόδου του differential privacy, καταφέρνει να εξάγει αποτελέσματα τα οποία είναι βελτιστοποιημένα και παράλληλα είναι ασφαλή απέναντι σε ενέργειες υποκλοπής πληροφοριών σχετικά με τις ταυτότητες των συμμετεχόντων. Πιο συγκεκριμένα, όπως παρατηρείται και στον παρακάτω αλγόριθμο της δεύτερης φάσης, μια σημαντική διαφορά σε σχέση με άλλες μεθόδους είναι η αλλαγή της συνάρτησης απόστασης από Ευκλείδεια σε Σχετική απόσταση, το οποίο βοηθάει στο να δημιουργηθούν πιο ακριβείς συστάδες. Αυτό επιτυγχάνεται με το να εκχωρείται σε κάθε συστάδα ένα βάρος ανάλογα με το πόσο όμοια είναι τα σημεία της κάθε συστάδας σε κάθε επανάληψη.

Επίσης, ο αλγόριθμος χρησιμοποιεί τη συνάρτηση ευαισθησίας Δf και τη παράμετρο ε για να επιτύχει ε-differential privacy, εξασφαλίζοντας με αυτόν τον τρόπο την ασφάλεια των δεδομένων. Αναλυτικότερα, σε κάθε επανάληψη, υπολογίζονται οι μεταβλητές sum και num και προστίθεται θόρυβος με τον μηχανισμό Laplace για να αποκρυφτούν οι πραγματικές πληροφορίες, και στην συγκεκριμένη περίπτωση να αποκρυφτεί το κέντρο της συστάδας.

PADC Αλγόριθμος – 2η Φάση

Είσοδος: ο αριθμός των συστάδων **k**,
βελτιστοποιημένα **k** αρχικά κέντρα,
η παράμετρος των ακραίων τιμών **r**,
το dataset **D** = { $\mathbf{x_1}$, $\mathbf{x_2}$, … , $\mathbf{x_n}$ },
η συνάρτηση ευαισθησίας **Δf**,
η παράμετρος προστασίας **ε**
Έξοδος: τα συσταδοποιημένα δεδομένα

Βήμα 1ο : Με βάση τα αρχικά κέντρα, συσταδοποιούμε τα δεδομένα σε **k** συστάδες.

Βήμα 2ο : Υπολόγισε τη διακύμανση των σημείων χωρίς να λαμβάνουμε υπόψη τους outliers με βάση τον παρακάτω τύπο:

$$s_i^2 = \frac{\sum_{x \in C_i} \sum_{i-1}^{n*r} (x - c_i)^2}{n_{c_i}} \tag{16}$$

και το βάρος της κάθε συστάδας είναι:

$$w_i = \frac{1}{s_i^2} \tag{17}$$

με C_i ορίζουμε τη i-οστή συστάδα, c_i ορίζουμε το κέντρο της i συστάδας.

Βήμα 3ο : Υπολόγισε τη σχετική απόσταση όλων των σημείων από το κέντρο των συστάδων εφαρμόζοντας τον παρακάτω τύπο:

$$dist^2(x, c_i) = w_i * \sum_{i}^{d} (x_i - c_i)^2 \tag{18}$$

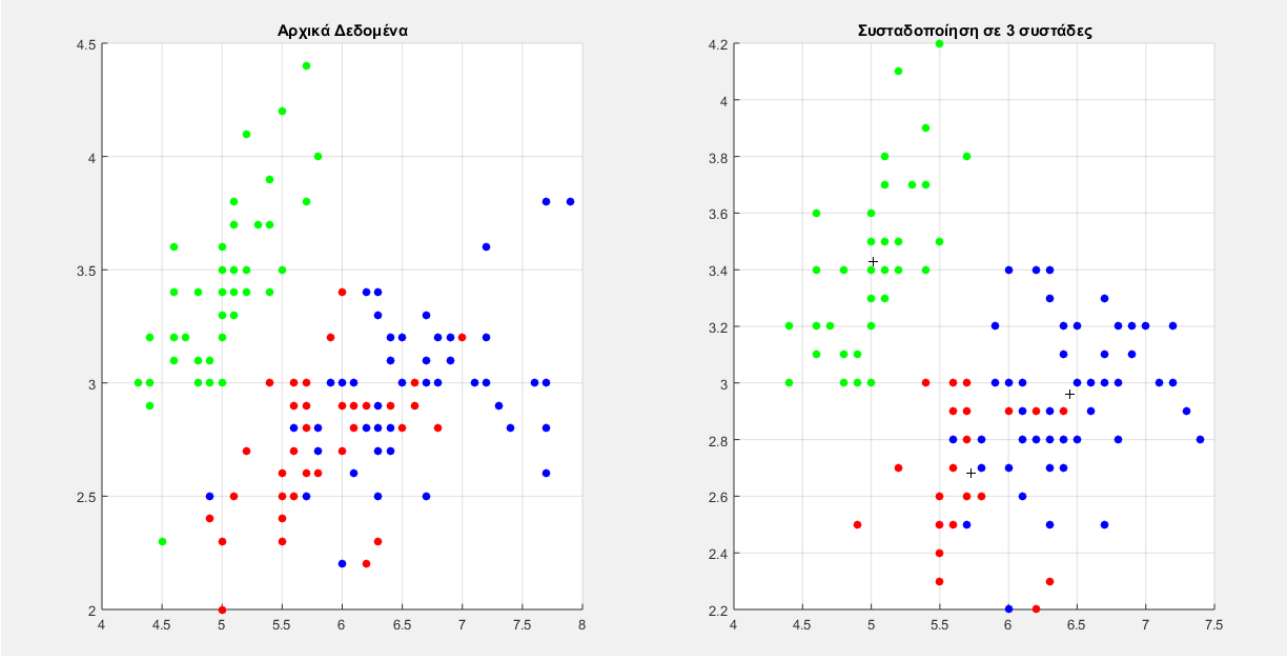
Βήμα 4ο : Με βάση τις σχετικές αποστάσεις που υπολογίστηκαν πιο πάνω, επανατοποθέτησε τα σημεία στις κοντινότερες συστάδες.

Βήμα 5ο : Με βάση τους παρακάτω τύπους υπολόγισε τα νέα κέντρα για κάθε συστάδα, κάνοντας χρήση και της μεθόδου του differential privacy :

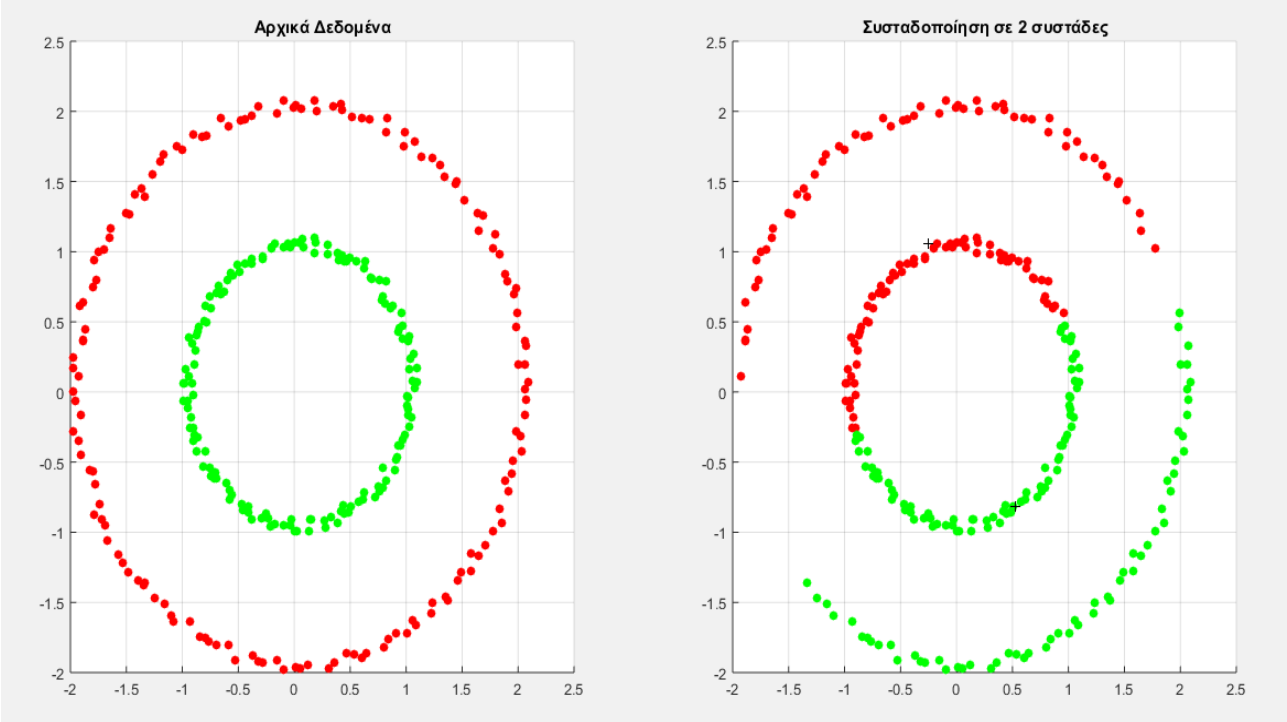
$$sum' = sum + Lap(\Delta f / \varepsilon) \quad (19)$$
$$num' = num + Lap(\Delta f / \varepsilon) \quad (20)$$
$$new\_center = sum' / num' \quad (21)$$

με sum ορίζουμε το άθροισμα των σημείων ξεχωριστά για κάθε συστάδα και με num ορίζουμε το σύνολο των σημείων για κάθε συστάδα.

Βήμα 6ο : Εάν ο αλγόριθμος δεν συγκλίνει επέστρεψε στο βήμα 2, αλλιώς τερμάτισε τη διαδικασία.



Εικόνα 26 - Αποτελέσματα του αλγορίθμου PADC στο Iris dataset.



Εικόνα 27 - Αποτελέσματα αλγορίθμου PADC σε σφαιρική σχηματική δομή.

Η Εικόνα 26 εμφανίζει τα αποτελέσματα του αλγορίθμου PADC όταν εφαρμοστεί στο Iris dataset. Τα αποτελέσματα δείχνουν την ικανότητα του αλγορίθμου να εντοπίζει και να εκμηδενίζει τις ακραίες τιμές καθώς, όπως παρατηρείται, κάποια αρχικά δεδομένα, τα οποία βρίσκονται απομακρυσμένα από τα κέντρα των συστάδων, δεν εμφανίζονται καθόλου στο τελικό αποτέλεσμα. Επίσης, παρατηρώντας την Εικόνα 26 γίνεται αντιληπτό και η αποτελεσματικότητα του αλγορίθμου όσον αφορά την ορθότητα των αποτελεσμάτων καθώς η πλειονότητα από τα τελικά αποτελέσματα είναι στη σωστή συστάδα, με βάση πάντα τα αρχικά δεδομένα.

Η Εικόνα 27 αποκαλύπτει ότι ο αλγόριθμος PADC δεν εξάγει ορθά αποτελέσματα και επηρεάζεται από δεδομένα που σχηματίζουν περίπλοκες σχηματικές δομές. Αν και παρατηρείται να εξαλείφονται οι ακραίες τιμές, ο αλγόριθμος, όντας επέκταση του K-means και περιέχοντας τη φιλοσοφία του, δεν μπορεί να αναγνωρίσει σωστά τα δεδομένα.

4. Ανάπτυξη νέου μοντέλου για την ομαδοποίηση δεδομένων

Η παρούσα διπλωματική στοχεύει στην ανάπτυξη ενός καινούργιου μοντέλου όσον αφορά την συσταδοποίηση των δεδομένων. Το νέο μοντέλο αναπτύχθηκε αντλώντας ως βασική φιλοσοφία την διαδικασία που χρησιμοποιεί ο αλγόριθμος DBSCAN. Ο νέος αλγόριθμος UOE-DBSCAN στοχεύει στην βελτιστοποίηση των αποτελεσμάτων σε σχέση με τον απλό DBSCAN, εκμεταλλεύοντας τη ιδιαιτερότητα που έχει ο αλγόριθμος να απομονώνει και να εξαλείφει τις ακραίες τιμές.

Πιο συγκεκριμένα, ο νέος αλγόριθμος δεν εξουδετερώνει όλες τις ακραίες τιμές αλλά τις εισάγει εκ νέου στα αποτελέσματα αφού, όμως, πρώτα υποστούν κάποια επεξεργασία με σκοπό να μην αλλοιωθεί η ποιότητα των αποτελεσμάτων. Σε αυτή την φάση, ο αλγόριθμος αξιοποιεί και τα πλεονεκτήματα της μεθόδου του differential privacy, εφαρμόζοντας τη τεχνική αυτή στις ακραίες τιμές. Ο μηχανισμός Laplace δημιουργεί ένα θόρυβο τον οποίο και προσθέτει στις ακραίες τιμές. Έτσι, οι τιμές των outliers επαναπροσδιορίζονται και αυτό έχει ως αποτέλεσμα κάποιες ακραίες τιμές, που σε κανονικές συνθήκες απλά θα διαγραφόντουσαν, να συμπεριληφθούν σε κάποια από τις τελικές συστάδες, βελτιώνοντας τη ποιότητα των αποτελεσμάτων.

Όπως γίνεται αντιληπτό και από τον παρακάτω ψευδοκώδικα, η διαδικασία του αλγορίθμου UOE-DBSCAN πραγματοποιεί επανένταξη των ακραίων τιμών στις συστάδες που βρίσκονται κοντινότερα, με προϋπόθεση η ακραία τιμή να είναι στη γειτονιά κάποιου σημείου πυρήνα ή σημείου ορίου.

UOE-DBSCAN Αλγόριθμος

Είσοδος: η ακτίνα της γειτονιάς **Eps**,
η παράμετρος που ορίζει τη πυκνότητα των σημείων **MinPts**,
το dataset **D = { x_1 , x_2 , … , x_n }**,
η παράμετρος προστασίας **ε**,
η ευαισθησία **Δf**
Έξοδος: συσταδοποιημένο **D**

Βήμα 1ο : Υπολόγισε την Ευκλείδεια απόσταση μεταξύ όλων των σημείων και των κέντρων βάρους, σύμφωνα με τον παρακάτω τύπο:

$$d(x_q, x_j) = \sqrt{((x_{q1} - x_{j1})^2 + ... + (x_{qn} - x_{jn})^2)} \quad (11)$$

Βήμα 2ο : Κατηγοριοποίησε κάθε σημείο σε σημείο πυρήνα, σημείο ορίου και σημείο θορύβου.

Βήμα 3ο : Αποθήκευσε τα σημεία θορύβου σε έναν ξεχωριστό πίνακα.

Βήμα 4ο : Βρες τα σημεία πυρήνα που είναι σε απόσταση μικρότερη ή ίση από Eps μεταξύ τους
.

Βήμα 5ο : Για κάθε σύνολο από σημεία πυρήνα που δημιουργούνται στο βήμα 4 δημιούργησε μια ξεχωριστή συστάδα.

Βήμα 6ο : Ανάθεσε κάθε σημείο ορίου στη συστάδα που ανήκει το σημείο πυρήνα το οποίο περιέχει στη γειτονιά του το σημείο ορίου.

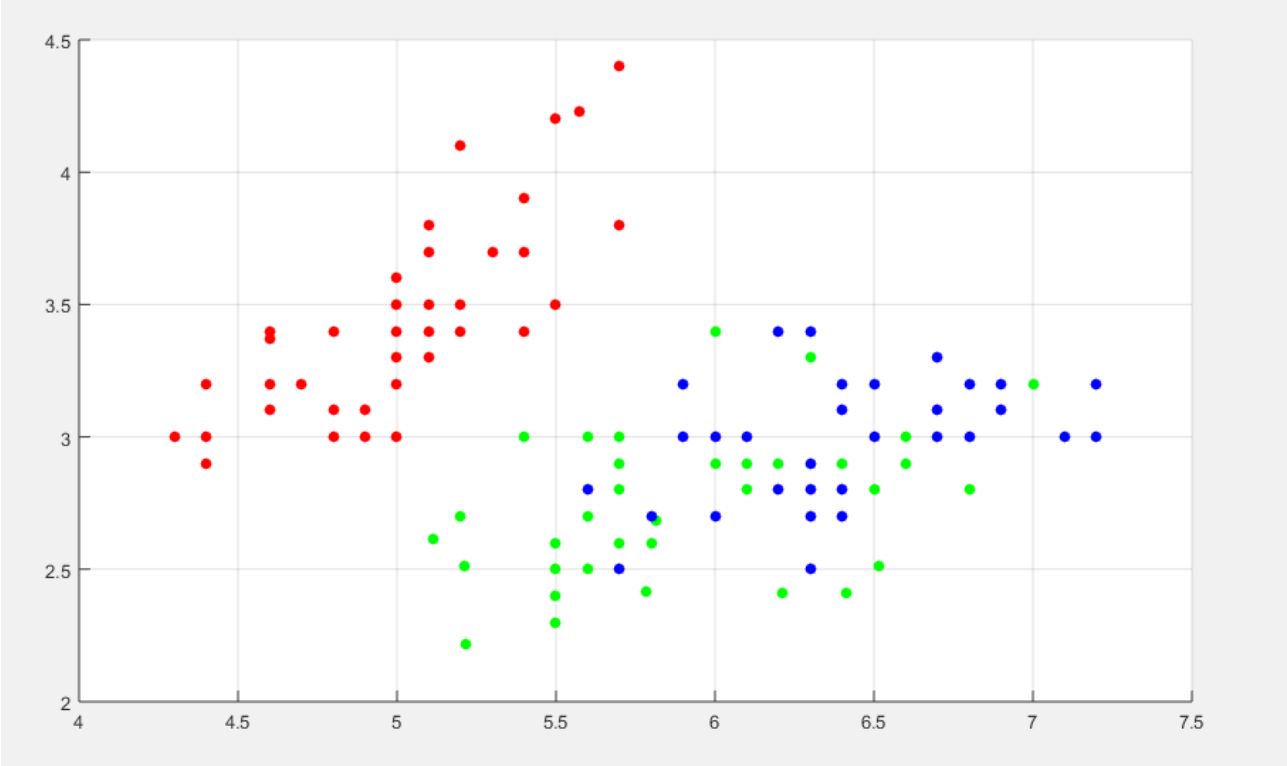
Βήμα 7ο : Για κάθε σημείο θορύβου, βρες το κοντινότερο σημείο πυρήνα ή σημείο ορίου με βάση τη σχέση (11) και αποθήκευσε τον αριθμό συστάδας του.

Βήμα 8ο : Πρόσθεσε ή αφαίρεσε το θόρυβο του μηχανισμού Laplace σε κάθε σημείο θορύβου σύμφωνα με τον εξής τύπο :

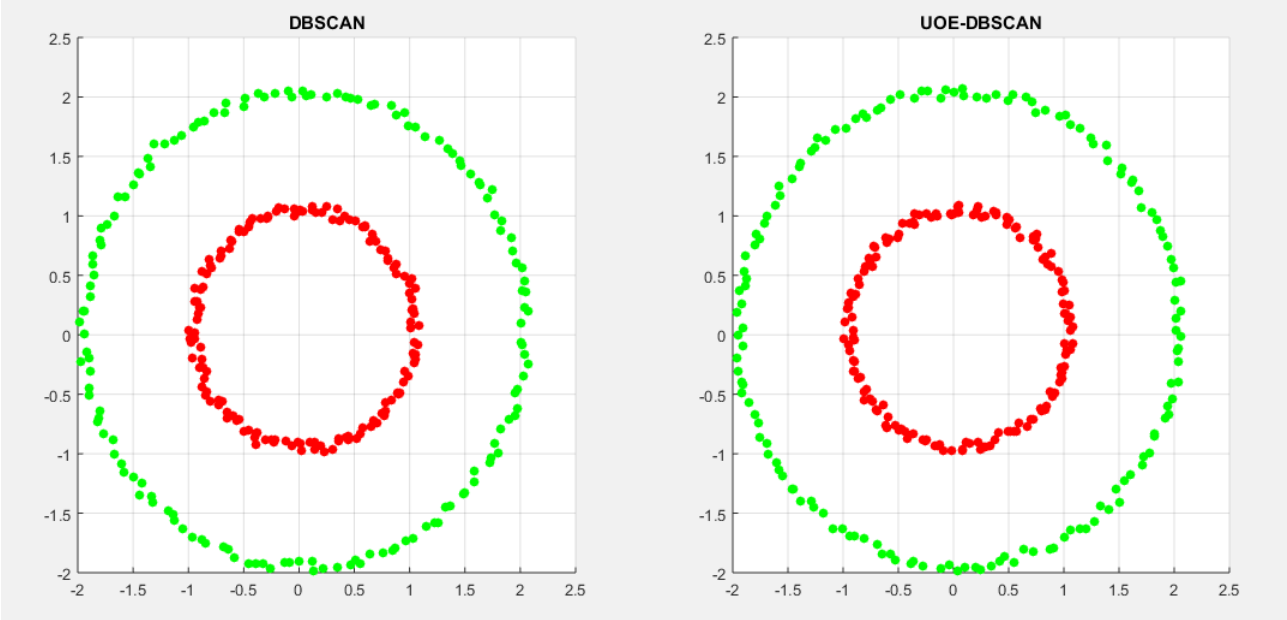
$$result_i = x_i \pm lap / dim \quad (22)$$

με $result_i$ συμβολίζουμε τη νέα τιμή για κάθε ακραία τιμή, με x_i συμβολίζουμε τη τρέχουσα τιμή του κάθε outlier, με τη μεταβλητή lap ορίζουμε το μηχανισμό που παράγει το θόρυβο Laplace και με dim συμβολίζουμε τις διαστάσεις του εκάστοτε dataset.

Βήμα 9ο : Σύγκρινε τις νέες τιμές των outliers με τη παράμετρο εισόδου **Eps** και εάν η νέα τιμή είναι μέσα στη γειτονιά του κοντινότερου σημείου που βρέθηκε στο βήμα 7, εκχώρησε το συγκεκριμένο σημείο στη αντίστοιχη συστάδα.



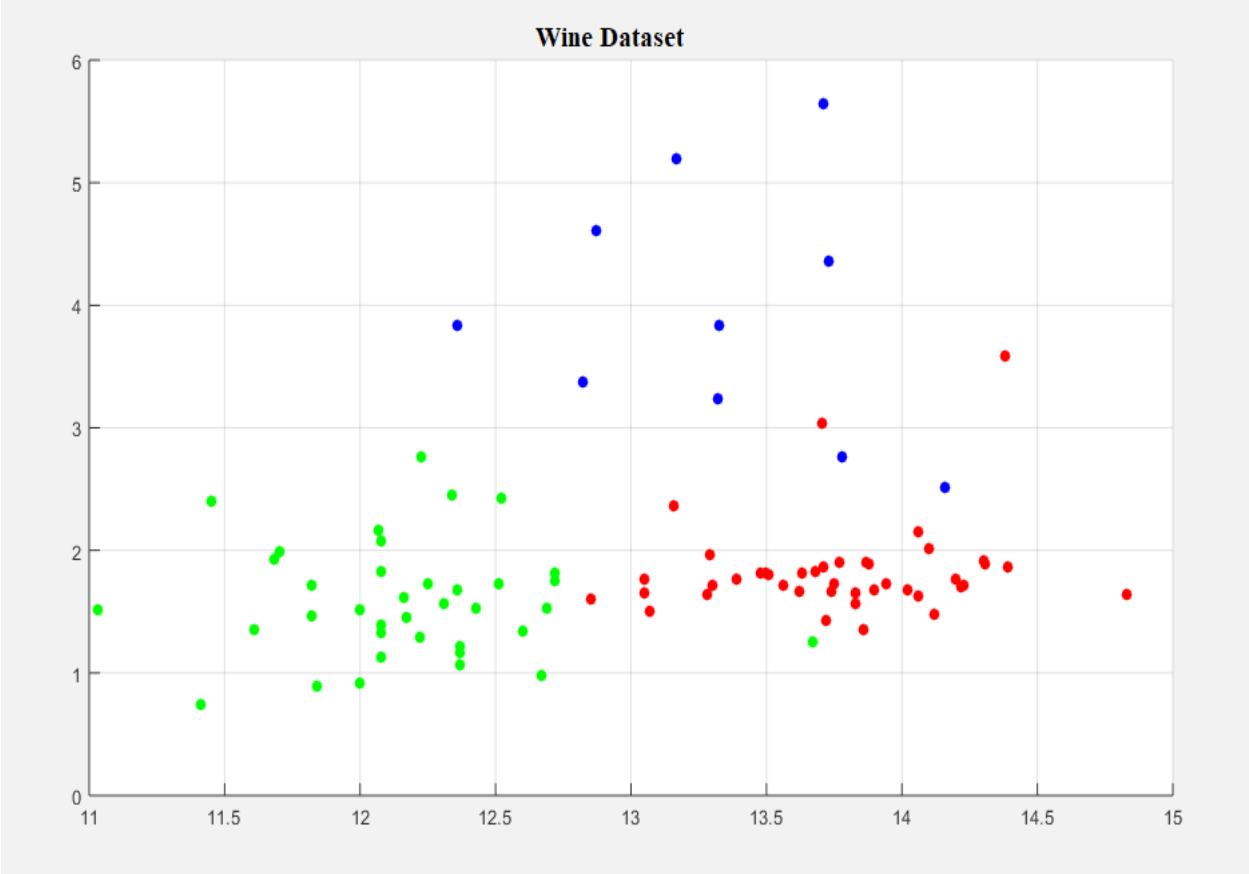
Εικόνα 28 - Αποτελέσματα του νέου αλγορίθμου UOE-DBSCAN στο Iris dataset.



Εικόνα 29 - Αποτελέσματα των δύο αλγορίθμων σε σχηματικές δομές.

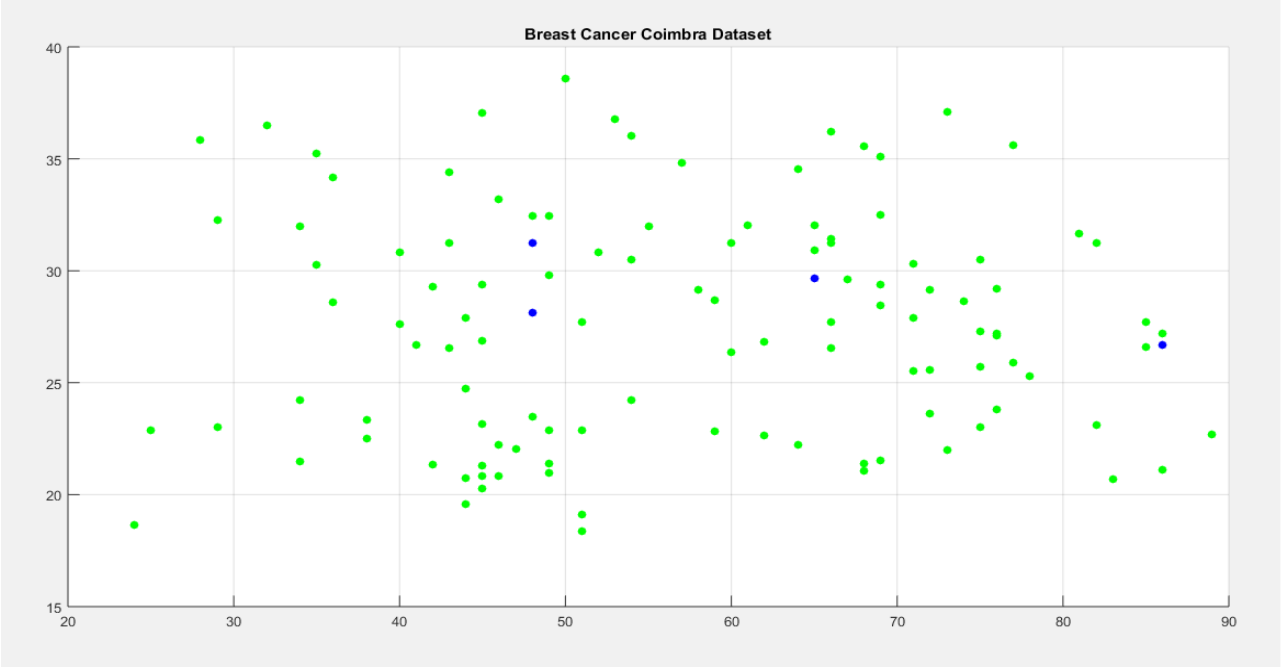
Εφαρμόζοντας τον αλγόριθμο UOE-DBSCAN στο Iris dataset, παρουσιάζεται στην Εικόνα 28 η αποδοτικότητα του νέου μοντέλου. Συγκρίνοντας, μάλιστα, τα αποτελέσματα με τα αντίστοιχα της Εικόνας 21, γίνεται αντιληπτό ότι τα αποτελέσματα της Εικόνας 28 είναι καλύτερα όσον αφορά την συμμετέχει όσο το δυνατόν περισσότερων αντικειμένων σε συστάδες. Αυτό συμβαίνει διότι ακραίες τιμές που στον απλό αλγόριθμο DBSCAN θα διαγραφόντουσαν, με το νέο αλγοριθμικό μοντέλο δίνεται η δυνατότητα συμμετοχής των αντικειμένων αυτών στα τελικά αποτελέσματα.

Επίσης, η Εικόνα 29 παρουσιάζει τα αποτελέσματα του απλού και του βελτιστοποιημένου DBSCAN. Παρατηρείται η ομοιότητα των αποτελεσμάτων και το γεγονός ότι ο νέος αλγόριθμος λειτουργεί εξίσου ορθά και σωστά απέναντι σε δεδομένα που δημιουργούν σχηματικές δομές.



Εικόνα 30 - Αποτελέσματα του νέου αλγοριθμικού μοντέλου στο Wine dataset.

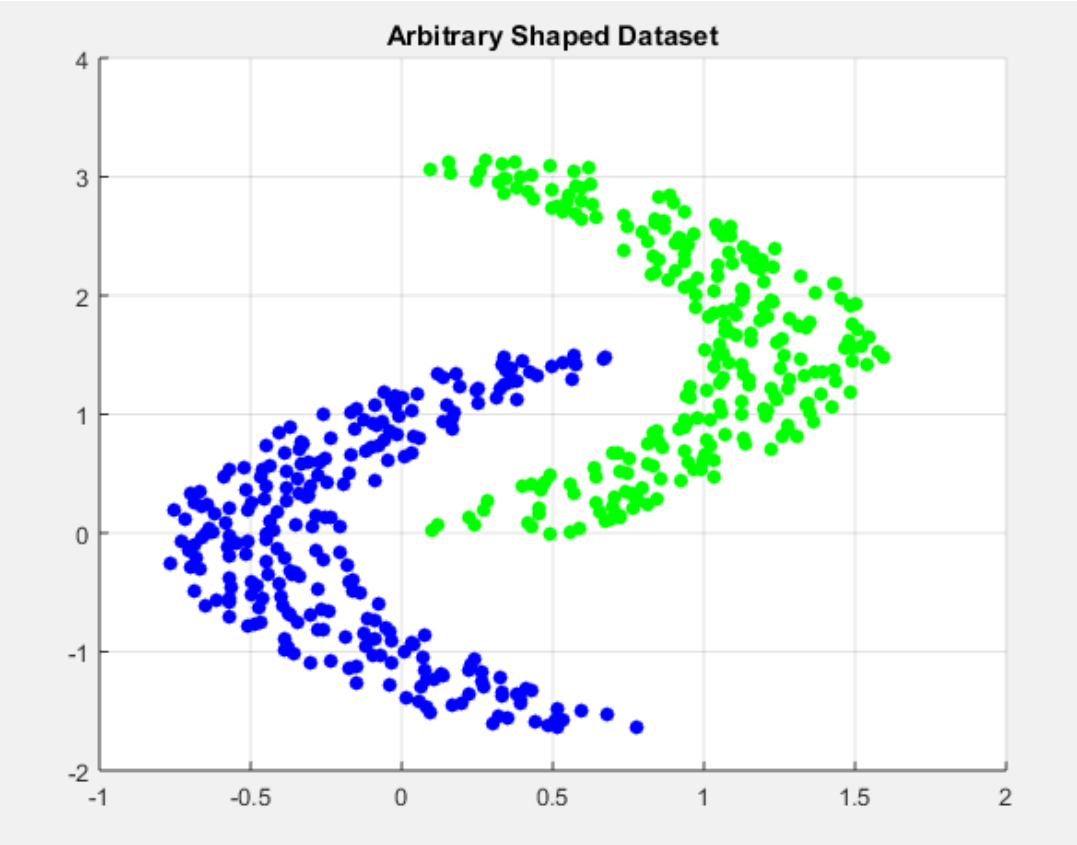
Χρησιμοποιώντας ως μια νέα βάση δεδομένων από το UCI το Wine dataset, η Εικόνα 30 δείχνει τα αποτελέσματα του νέου μοντέλου στο Wine dataset. Η απόδοση του νέου αλγοριθμικού μοντέλου στο συγκεκριμένο dataset είναι παρόμοια με την απόδοση του απλού DBSCAN και αυτό συμβαίνει διότι με τις ίδιες παραμέτρους εισόδου, η απόδοση των αλγορίθμων δεν αλλάζει. Οι τιμές των παραμέτρων που χρησιμοποιούνται είναι οι εξής : **Eps = 2.35, MinPts = 4**. Αναφορικά με το Wine dataset, η συγκεκριμένη βάση δεδομένων περιέχει μικρό αριθμό από γραμμές (178 για την ακρίβεια) αλλά οι διαστάσεις των δεδομένων είναι 13, σχετικά μεγάλος αριθμός. Η συσταδοποίηση των δεδομένων πραγματοποιείται σε 3 συστάδες, ενώ οι τιμές που παίρνουν τα δεδομένα είναι πραγματικές (REAL).



Εικόνα 31 - Αποτελέσματα του αλγορίθμου UOE-DBSCAN στο Breast Cancer Coimbra dataset.

Επιπλέον, στην Εικόνα 31 παρουσιάζονται τα αποτελέσματα του νέου μοντέλου που αναπτύχθηκε πάνω στο Breast Cancer Coimbra dataset του UCI. Το συγκεκριμένο dataset περιλαμβάνει δεδομένα ακέραιων τιμών (INT) και οι διαστάσεις των τιμών αυτών αγγίζουν τις 9. Το Breast Cancer Coimbra dataset περιέχει, επίσης, 116 γραμμές (tuples) τιμών, ενώ τα δεδομένα ομαδοποιούνται σε 2 κατηγορίες. Σε αυτό το dataset, οι παράμετροι εισόδου είναι οι εξής: **Eps = 140, MinPts = 4**.

Παρά το γεγονός ότι οι διαστάσεις είναι αρκετά μεγάλες, η απόδοση του νέου αλγορίθμου UOE-DBSCAN στο Breast Cancer Coimbra dataset φτάνει σε ποσοστό σχεδόν 97%.



Εικόνα 32 - Το Arbitrary Shaped dataset με αυθαίρετες τιμές.

Η Εικόνα 32 παρουσιάζει ένα dataset το οποίο περιέχει αυθαίρετες τιμές, οι οποίες όπως φαίνεται και στην εικόνα, δημιουργούν μια σχηματική δομή τύπου σπιράλ. Το Arbitrary Shaped dataset δημιουργήθηκε στα πλαίσια του συγγράμματος και αποτελείται από 500 γραμμές (tuples) και τα δεδομένα χωρίζονται σε 2 κατηγορίες που λαμβάνουν τιμές χαρακτήρων (CHAR). Τα δεδομένα είναι δύο διαστάσεων και οι τιμές που λαμβάνουν τα δεδομένα είναι πραγματικές (REAL).

Όλα οι εικόνες και τα αποτελέσματα που προηγήθηκαν και που θα επακολουθήσουν έχουν αναπτυχθεί και αναπαρασταθεί χρησιμοποιώντας το προγραμματιστικό εργαλείο και περιβάλλον της Matlab.

Στο συγκεκριμένο σημείο, είναι απαραίτητη η ανάλυση και επεξήγηση ενός μεγέθους το οποίο χρησιμοποιείται στον επιστημονικό τομέα με στόχο την σύγκριση των αλγορίθμων ως προς την αποδοτικότητα των αποτελεσμάτων τους. Το μέγεθος αυτό ονομάζεται F-measure και αποτελεί συνδυασμό δύο άλλων μετρικών μεγεθών, του Precision και του Recall. Το Precision και το Recall χρησιμοποιούνται ευρέως στον τομέα της ανάκτησης πληροφοριών. Το Precision ορίζεται ως ο λόγος των σημείων που έχουν συσταδοποιηθεί σωστά δια τον συνολικό αριθμό των σημείων που έχουν συσταδοποιηθεί. Το Recall ορίζεται ως ο λόγος των σημείων που έχουν συσταδοποιηθεί στη συστάδα στην οποία ανήκουν εξαρχής δια τον συνολικό αριθμό των συσταδοποιημένων σημείων.

Το F-measure για οποιαδήποτε συστάδα δίνεται από τον παρακάτω τύπο:

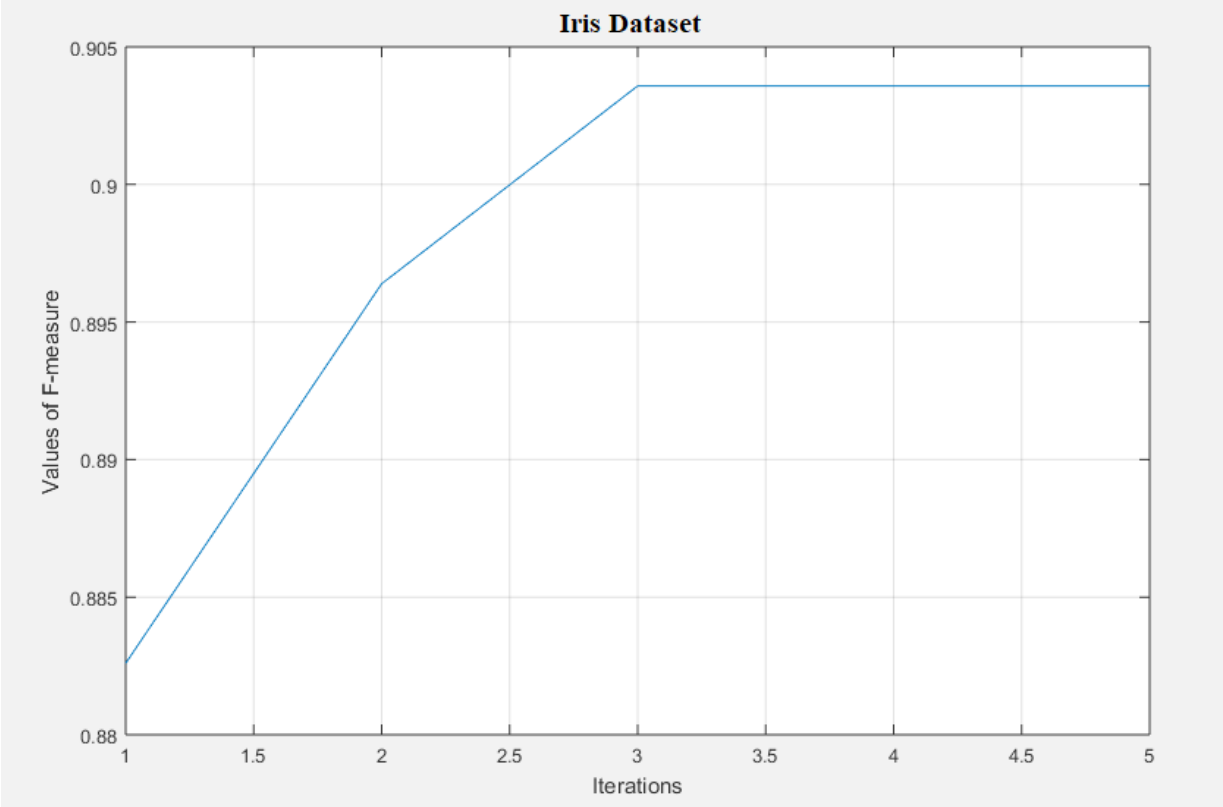
$$F(Cluster_x) = \frac{2 * Precision(Cluster_x) * Recall(Cluster_x)}{Precision(Cluster_x) + Recall(Cluster_x)} \tag{23}$$

με $Cluster_x$ ορίζεται οποιαδήποτε συστάδα.

Για να προσδιοριστεί η τιμή του F-measure για ολόκληρο το dataset χρησιμοποιείται ο παρακάτω τύπος:

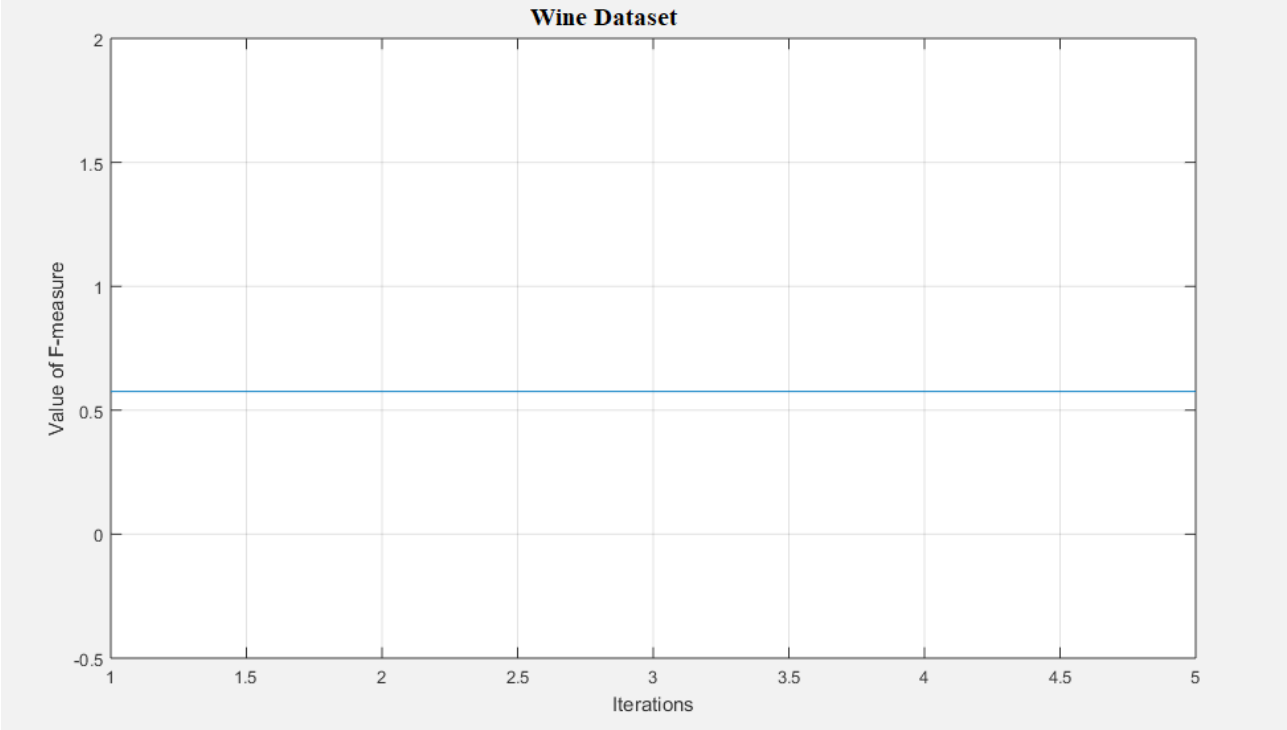
$$F = \frac{\sum_{i=1}^{k} (|c_i| * F(i))}{\sum_{i=1}^{k} |c_i|} \tag{24}$$

με i συμβολίζεται η τρέχων συστάδα, με F(i) συμβολίζεται το F-measure για την i συστάδα και με c_i συμβολίζεται ο αριθμός των σημείων για την i συστάδα.

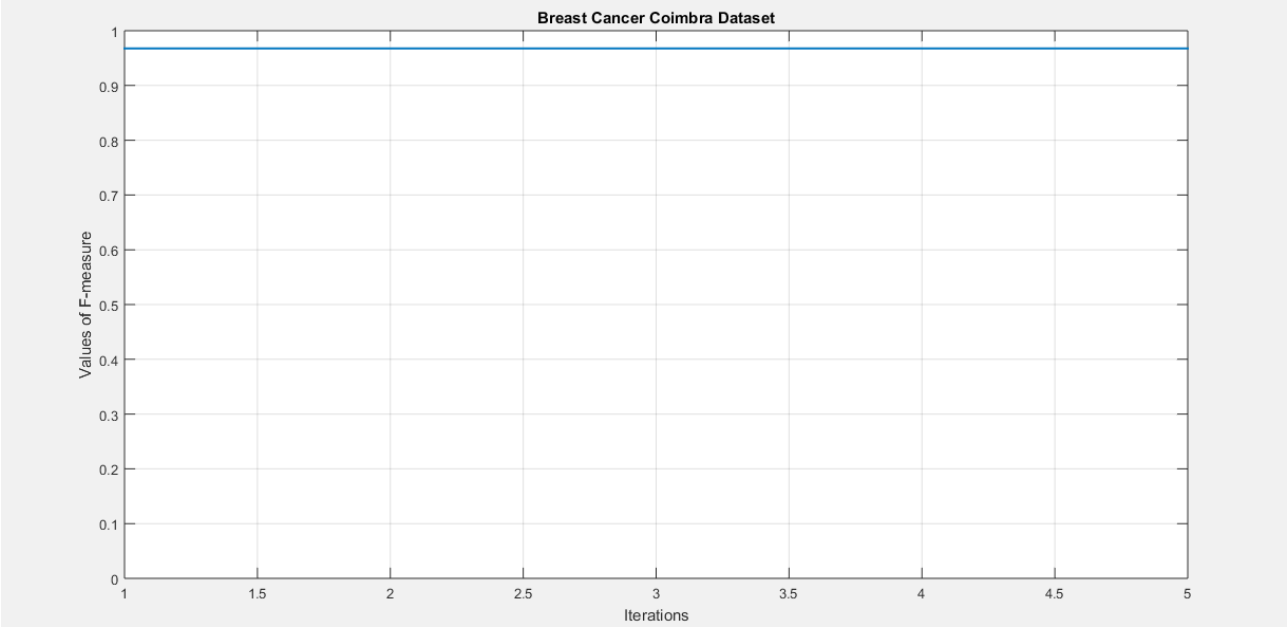


Εικόνα 33 - Βελτιστοποίηση του F-measure μετά τις τρεις πρώτες επαναλήψεις των βημάτων 7,8,9 του ψευδοκώδικα του UOE-DBSCAN στο Iris dataset.

Μια ενδιαφέρουσα εξέλιξη κατά την αξιολόγηση του αλγορίθμου αποτελεί η καμπύλη που φαίνεται στην Εικόνα 33. Η εξέλιξη αυτή αφορά την βελτιστοποίηση των αποτελεσμάτων του νέου μοντέλου UOE-DBSCAN που αναπτύχθηκε. Πιο συγκεκριμένα, η Εικόνα 33 αποκαλύπτει την ικανότητα του αλγορίθμου να βελτιστοποιεί την τιμή του F-measure μετά από κάποιον συγκεκριμένο αριθμό επαναλήψεων. Οι επαναλήψεις αυτές αφορούν τα τελευταία τρία βήματα του αλγορίθμου και παρατηρείται ότι σε κάθε επανάληψη των συγκεκριμένων βημάτων υπάρχει βελτιστοποίηση στην τιμή του F-measure. Βέβαια, παρατηρώντας τις Εικόνες 34, 35 γίνεται αντιληπτό ότι η ικανότητα αυτή δεν λειτουργεί εξίσου αποτελεσματικά σε άλλα dataset, γεγονός που είναι λογικό να συμβαίνει καθώς ο αλγόριθμος μπορεί από τη πρώτη επανάληψη να βρει το μεγαλύτερο F-measure και να μην υπάρχει η δυνατότητα περαιτέρω βελτίωσης.



Εικόνα 34 - Μη βελτιστοποίηση του F-measure μετά τις πέντε πρώτες επαναλήψεις των βημάτων 7,8,9 του ψευδοκώδικα του UOE-DBSCAN στο Wine dataset.



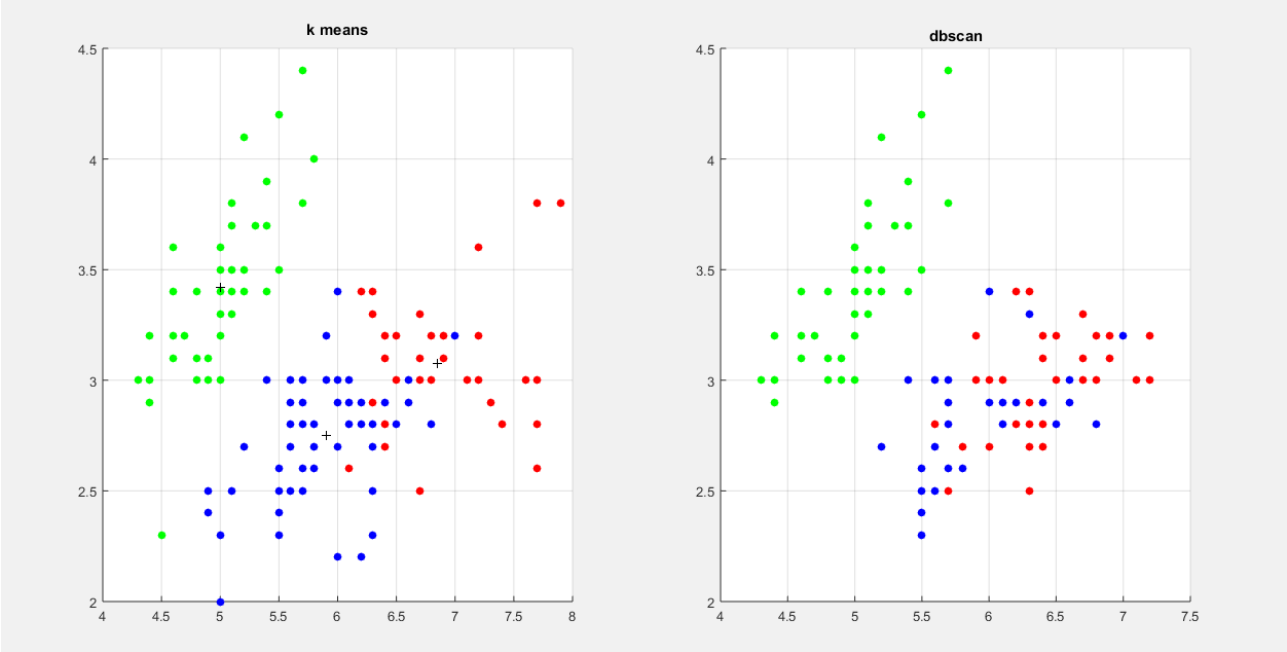
Εικόνα 35 - Μη βελτιστοποίηση του F-measure μετά τις πέντε πρώτες επαναλήψεις των βημάτων 7,8,9 του ψευδοκώδικα του UOE-DBSCAN στο Breast Cancer Coimbra dataset.

5. Σύγκριση αλγορίθμων ομαδοποίησης

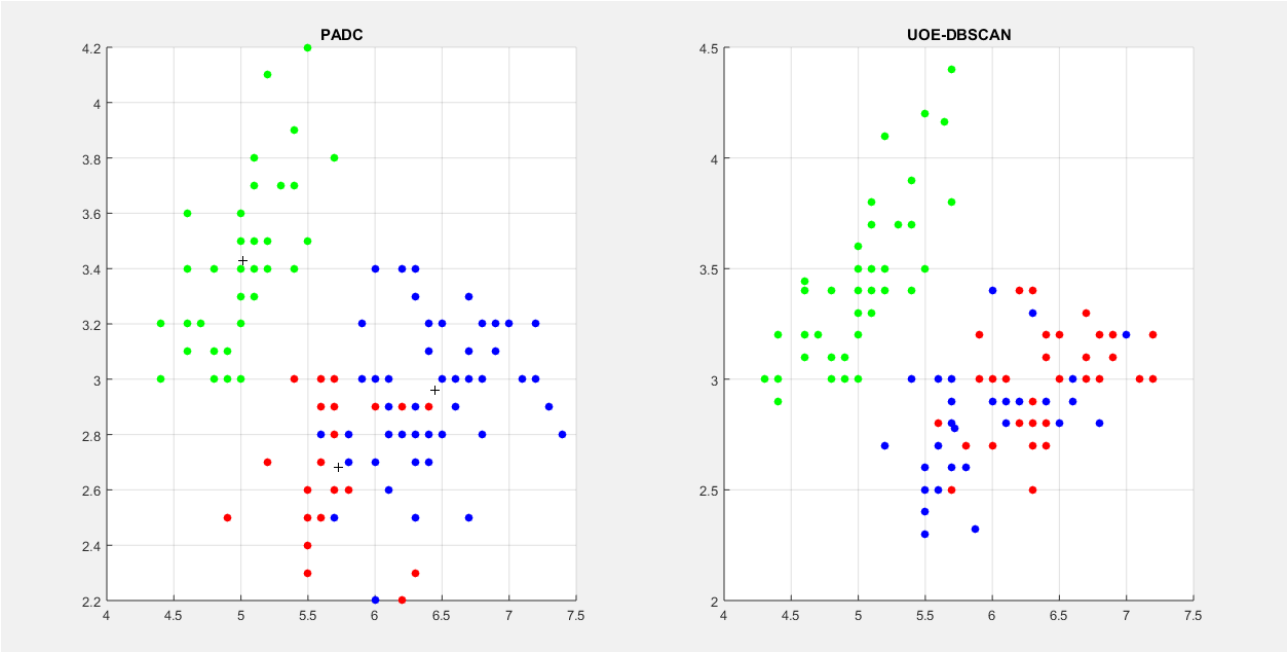
Για να κατανοηθεί βαθύτερα η σημασία, η λειτουργικότητα και η χρησιμότητα των αλγορίθμων που επεξηγήθηκαν και αναλύθηκαν στις προηγούμενες ενότητες, θα χρειαστεί να πραγματοποιηθεί κάποιου είδους σύγκριση μεταξύ των αλγορίθμων αυτών. Τα αποτελέσματα της σύγκρισης αυτής, αντικατοπτρίζουν την υπεροχή ή μη κάποιων αλγορίθμων έναντι άλλων.

5.1 Απόδοση αλγορίθμων στην δημιουργία συστάδων

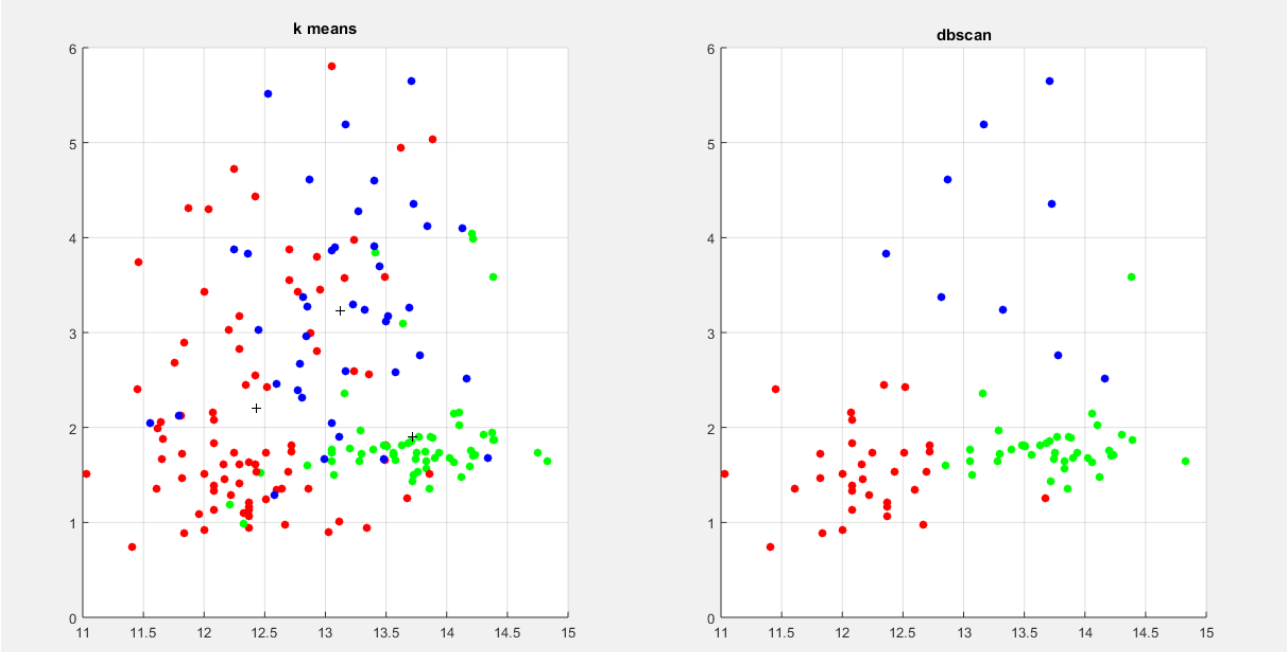
Οι Εικόνες 36 και 37 αποτελούν μια σύνοψη όλων των αλγορίθμων που αναλύθηκαν στο προηγούμενο κεφάλαιο και εμφανίζουν τα αποτελέσματα των αλγορίθμων αυτών, δηλαδή του K-means, του DBSCAN, του PADC και του UOE-DBSCAN. Το dataset που χρησιμοποιήθηκε είναι το Iris dataset και με πρώτη ματιά φαίνεται η αποτελεσματικότητα των αλγορίθμων PADC και UOE-DBSCAN.



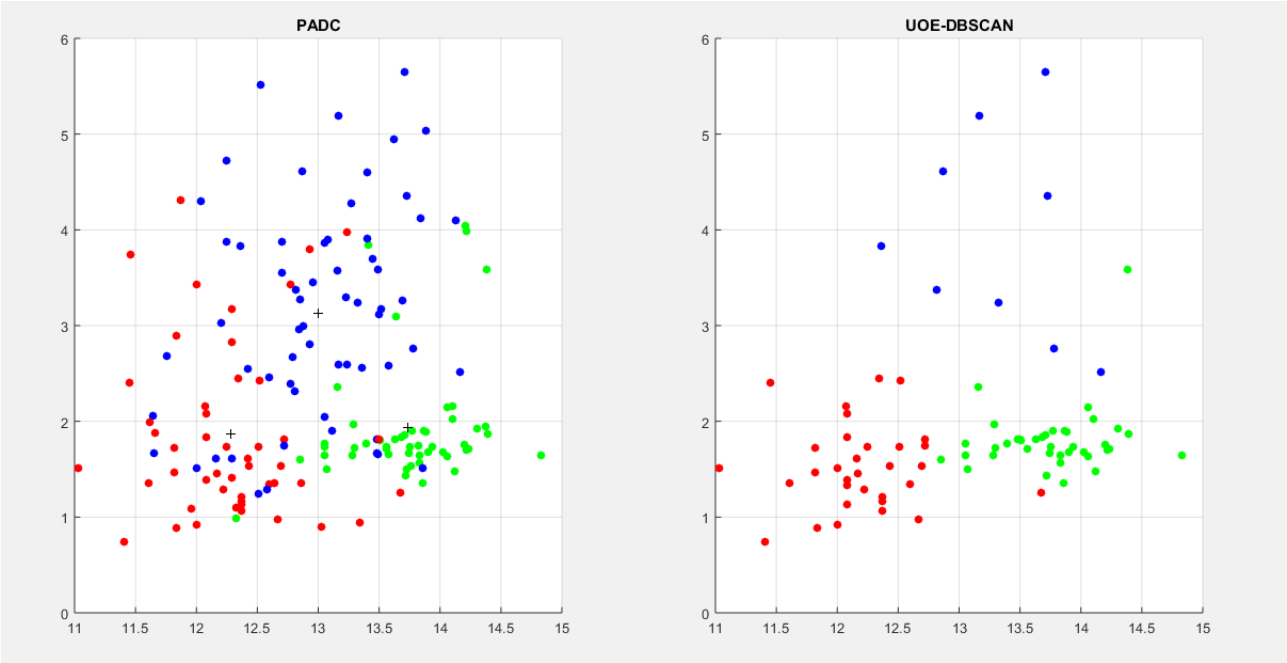
Εικόνα 36: Συγκριτικά αποτελέσματα των αλγορίθμων K-means και DBSCAN στο Iris dataset.



Εικόνα 37 - Συγκριτικά αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN στο Iris dataset.



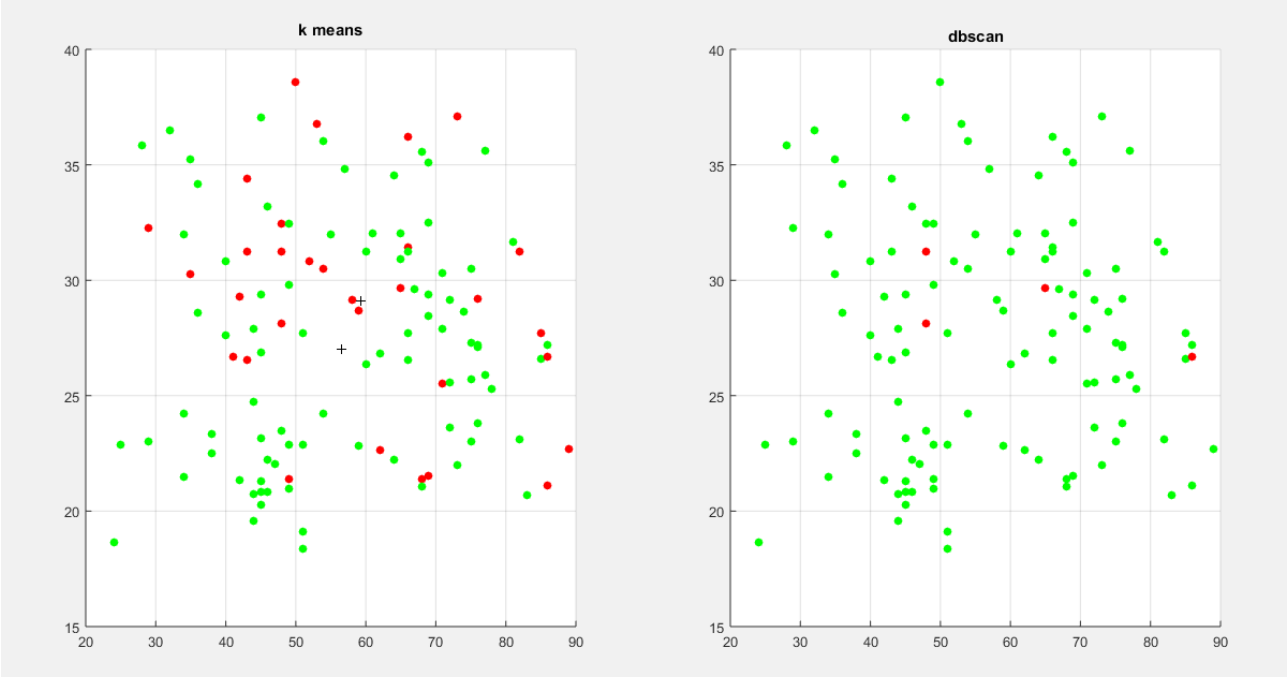
Εικόνα 38 - Συγκριτικά αποτελέσματα των αλγορίθμων K-means και DBSCAN στο Wine dataset.



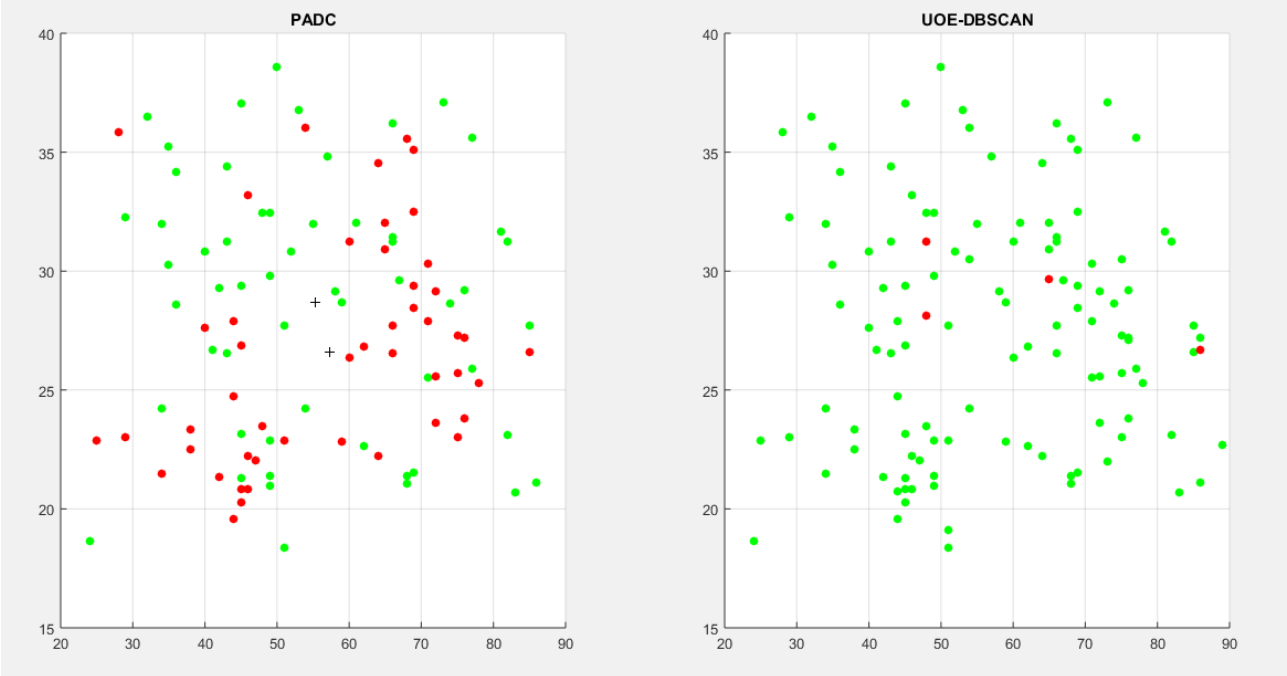
Εικόνα 39 - Συγκριτικά αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset.

Αντίστοιχα, στις Εικόνες 38, 39 εμφανίζονται τα αποτελέσματα των αλγορίθμων ομαδοποίησης που προαναφέρθηκαν στο Wine dataset. Είναι φανερό ότι οι αλγόριθμοι K-means και PADC είναι αυτοί που αποδίδουν καλύτερα, ενώ οι αλγόριθμοι DBSCAN και UOE-DBSCAN υστερούν αρκετά σε ποσοστό που φτάνει σχεδόν το 50%. Είναι φανερό, επίσης, στις Εικόνες 38, 39 ότι οι αλγόριθμοι DBSCAN και UOE-DBSCAN εξάγουν τις ίδιες συστάδες, γεγονός που σημαίνει ότι για τις συγκεκριμένες τιμές των παραμέτρων εισόδου, τα αποτελέσματα δεν επιδέχονται περαιτέρω βελτίωση. Όσον αφορά τους αλγορίθμους K-means και PADC, ποσοτικά παράγουν και οι δύο σχεδόν το ίδιο αποτέλεσμα αλλά ποιοτικά τα αποτελέσματά τους διαφέρουν καθώς η ανάθεση των δεδομένων στις συστάδες είναι διαφορετική στους δύο αλγορίθμους.

Αντίστοιχα, στις παρακάτω Εικόνες 40, 41 παρουσιάζονται τα αποτελέσματα των αλγορίθμων K-means, DBSCAN, PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset. Με πρώτη ματιά φαίνεται ότι ο αλγόριθμος PADC είναι αποδοτικότερος, αλλά όπως φαίνεται στην Εικόνα 46, οι αλγόριθμοι DBSCAN και UOE-DBSCAN παράγουν καλύτερο Precision, ενώ στην Εικόνα 47, πλησιάζουν αρκετά το ποσοστό του Recall του PADC.

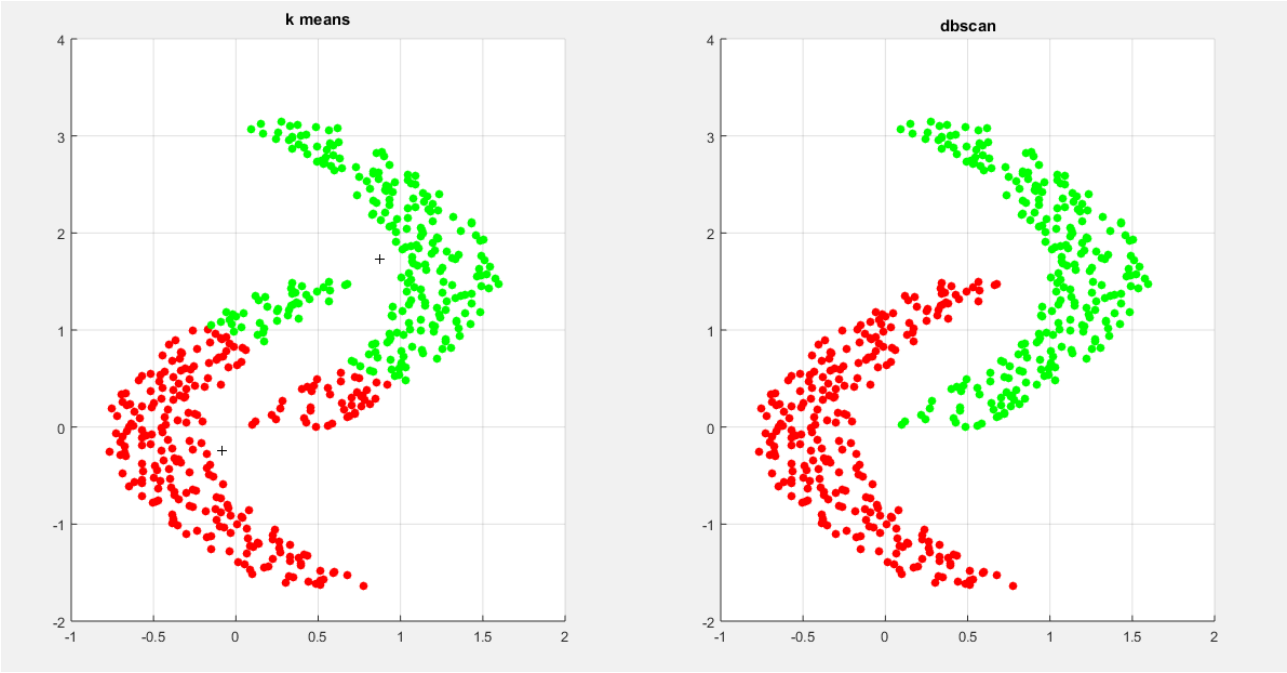


Εικόνα 40 - Συγκριτικά αποτελέσματα των αλγορίθμων K-means και DBSCAN στο Breast Cancer Coimbra dataset.

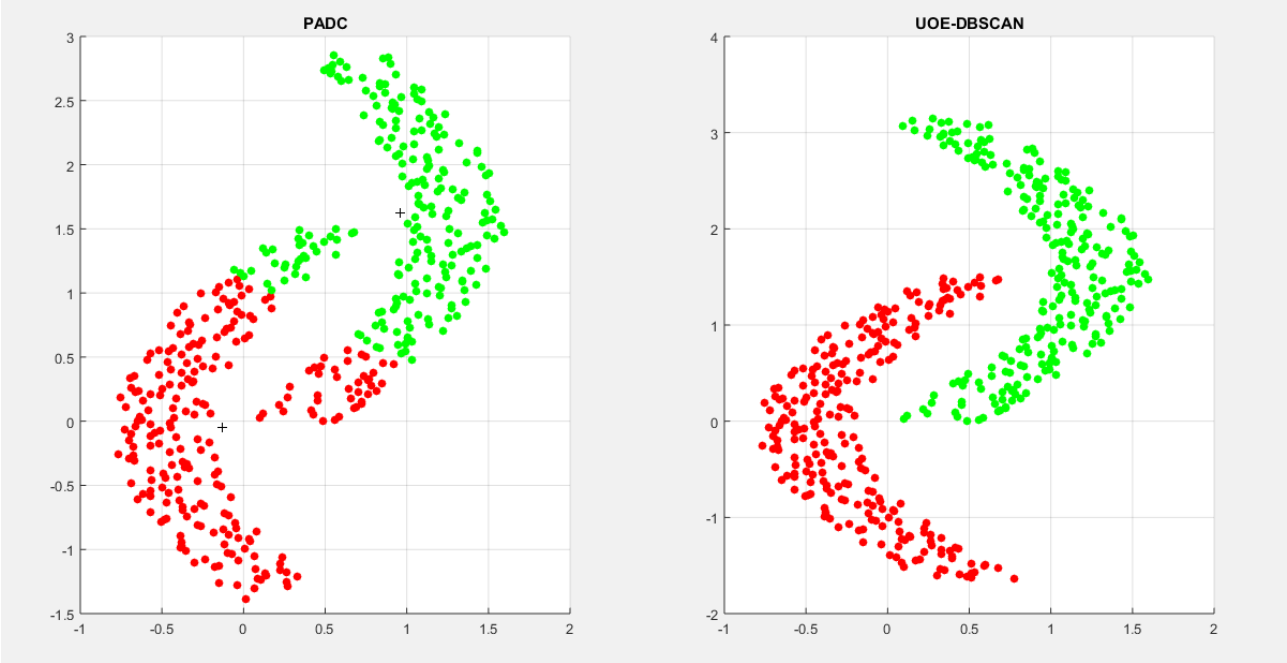


Εικόνα 41 - Συγκριτικά αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset.

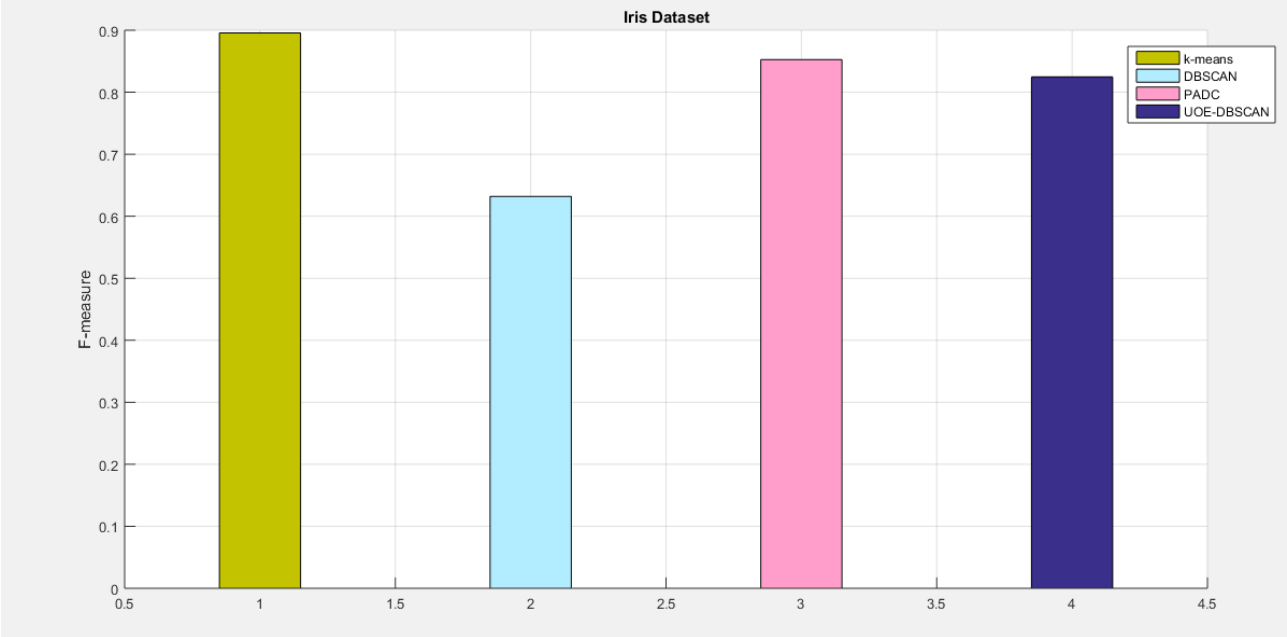
Επιπλέον, οι Εικόνες 42, 43 παρουσιάζουν τα αποτελέσματα των αλγορίθμων ομαδοποίησης που αναφερθήκαμε προηγουμένως στο Arbitrary Shaped dataset. Από τα αποτελέσματα γίνεται αντιληπτό ότι στο συγκεκριμένο dataset, υπερτερούν οι αλγόριθμοι DBSCAN και UOE-DBSCAN. Η υπεροχή αυτή είναι λογική καθώς οφείλεται στα πλεονεκτήματα των δύο αλγορίθμων να παράγουν καλύτερα και πιο ποιοτικά αποτελέσματα σε δεδομένα που δημιουργούν σχηματικές δομές. Αντιθέτως, οι αλγόριθμοι K-means και PADC, παράγουν αδύναμα και μη ποιοτικά αποτελέσματα όταν εφαρμόζονται σε δεδομένα σχηματικών δομών όπως είναι το συγκεκριμένο dataset.



Εικόνα 42 - Συγκριτικά αποτελέσματα των αλγορίθμων K-means και DBSCAN στο Arbitrary Shaped dataset.

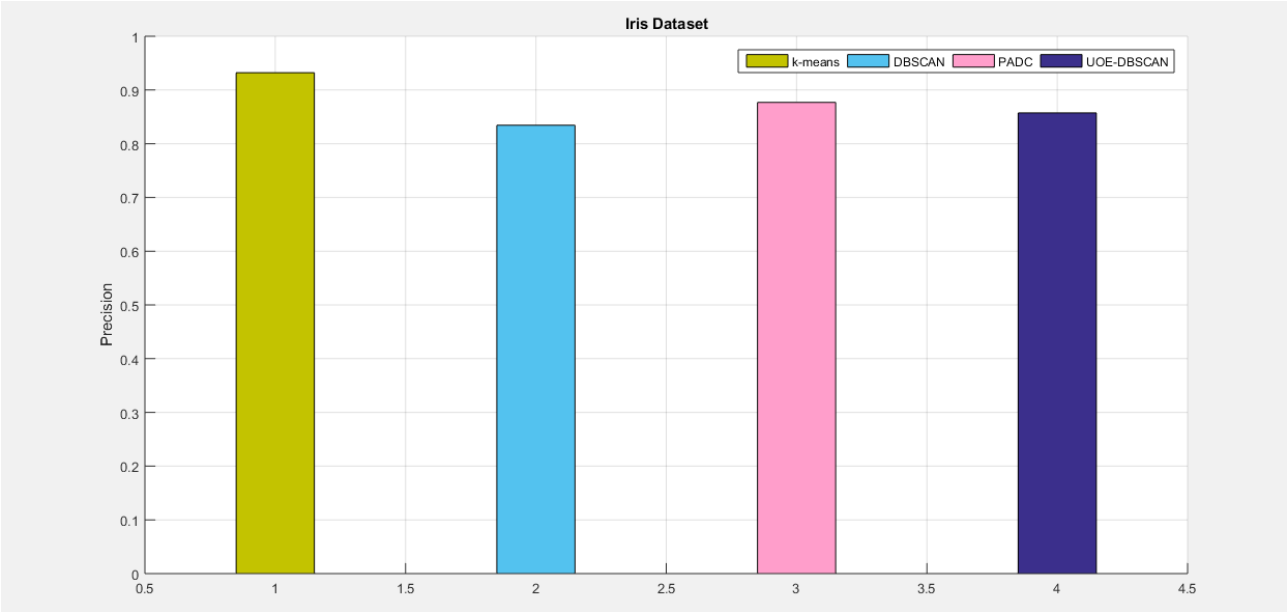


Εικόνα 43 - Συγκριτικά αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN στο Arbitrary Shaped dataset.



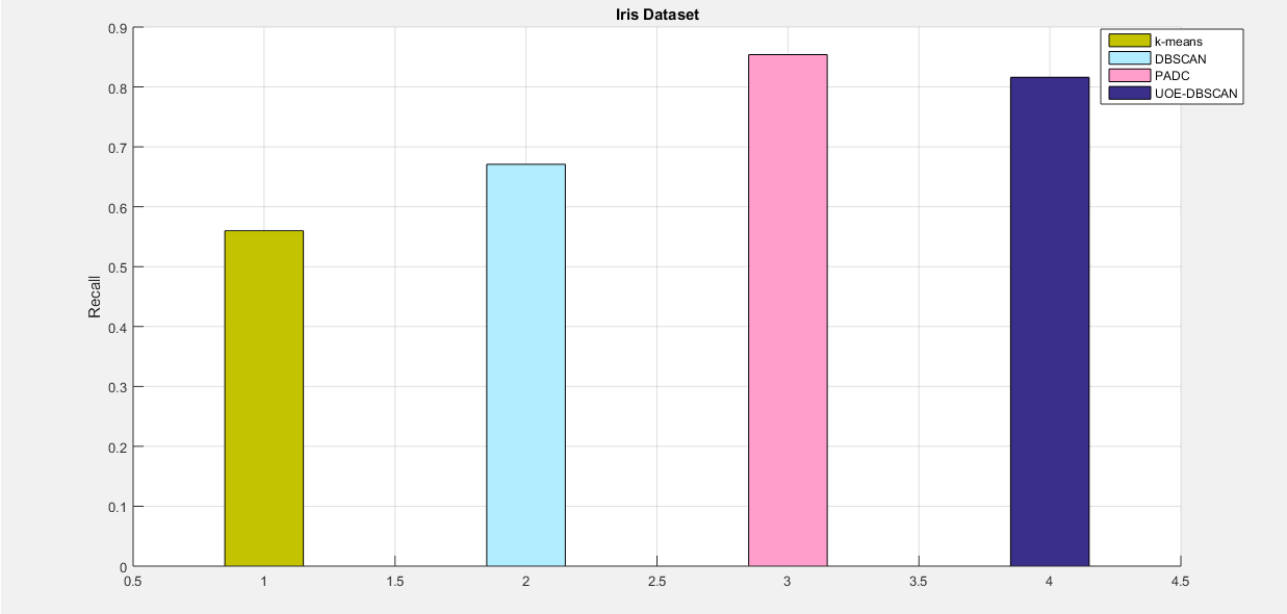
Εικόνα 44 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure όλων των αλγορίθμων στο Iris dataset.

Στην Εικόνα 44 παρατηρείται η απόδοση της μετρικής F--measure όλων των αλγορίθμων συσταδοποίησης στο Iris dataset. Από τα αποτελέσματα της Εικόνας 44, προκύπτει το συμπέρασμα ότι ο αλγόριθμος K-means είναι ο πιο αποδοτικός ως προς το F-measure φτάνοντας σε ποσοστό 99% με τον PADC και το νέο μοντέλο UOE-DBSCAN να έχουν αρκετά καλή απόδοση (95% και 92% αντίστοιχα).



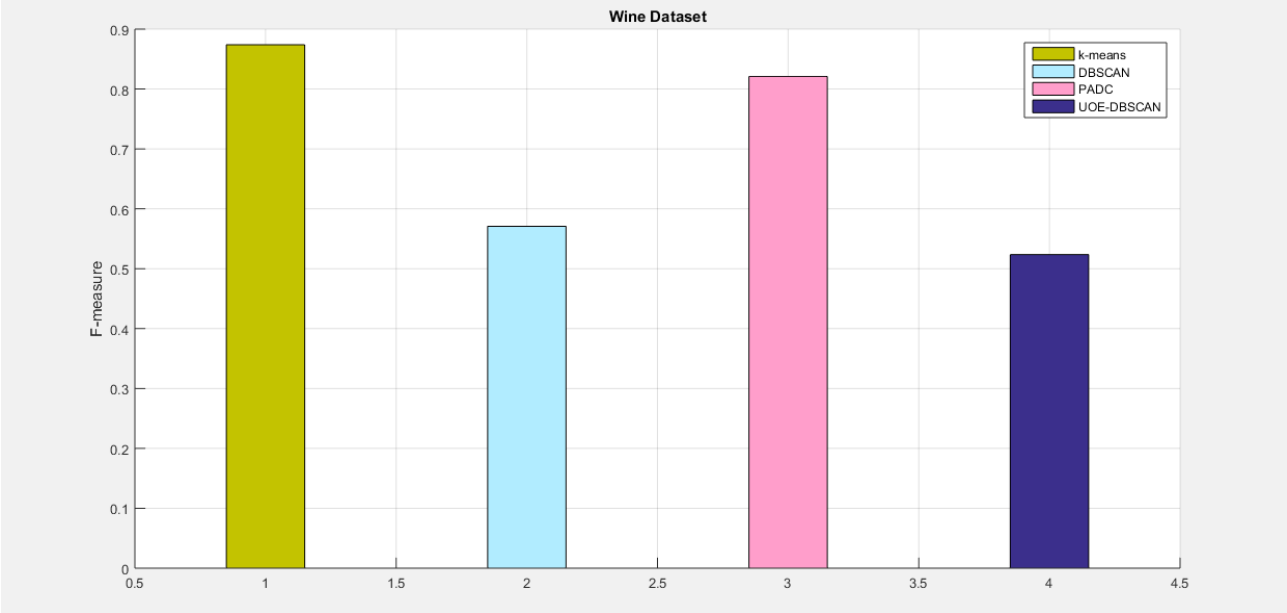
Εικόνα 45 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision όλων των αλγορίθμων στο Iris dataset.

Στις Εικόνες 45, 46 παρατηρείται η απόδοση του Precision και του Recall όλων των αλγορίθμων συσταδοποίησης στο Iris dataset. Αναλύοντας τα αποτελέσματα της Εικόνας 45, προκύπτει το συμπέρασμα ότι ο αλγόριθμος K-means είναι ο πιο αποδοτικός ως προς το Precision φτάνοντας σε ποσοστό 93%.

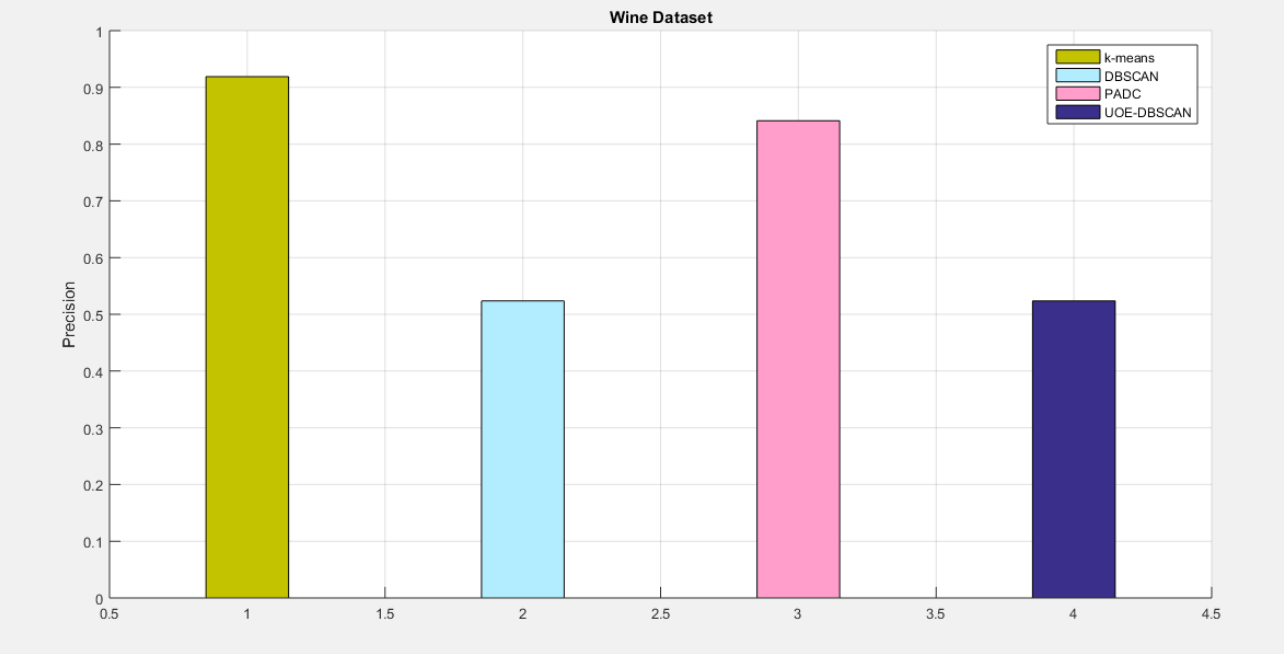


Εικόνα 46 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Recall όλων των αλγορίθμων στο Iris dataset.

Ο νέος αλγόριθμος UOE-DBSCAN πλησιάζει αρκετά την απόδοση του PADC (85% και 88% αντίστοιχα) ενώ είναι καλύτερος σε Precision από τον απλό DBSCAN (85% και 83% αντίστοιχα). Σε αντίθεση με το Precision, το Recall του K-means που παρατηρείται στην Εικόνα 46 είναι πολύ χαμηλό (56%) ενώ οι αλγόριθμοι PADC και UOE-DBSCAN είναι αρκετά αποδοτικοί ως προς το Recall (95% και 92% αντίστοιχα).

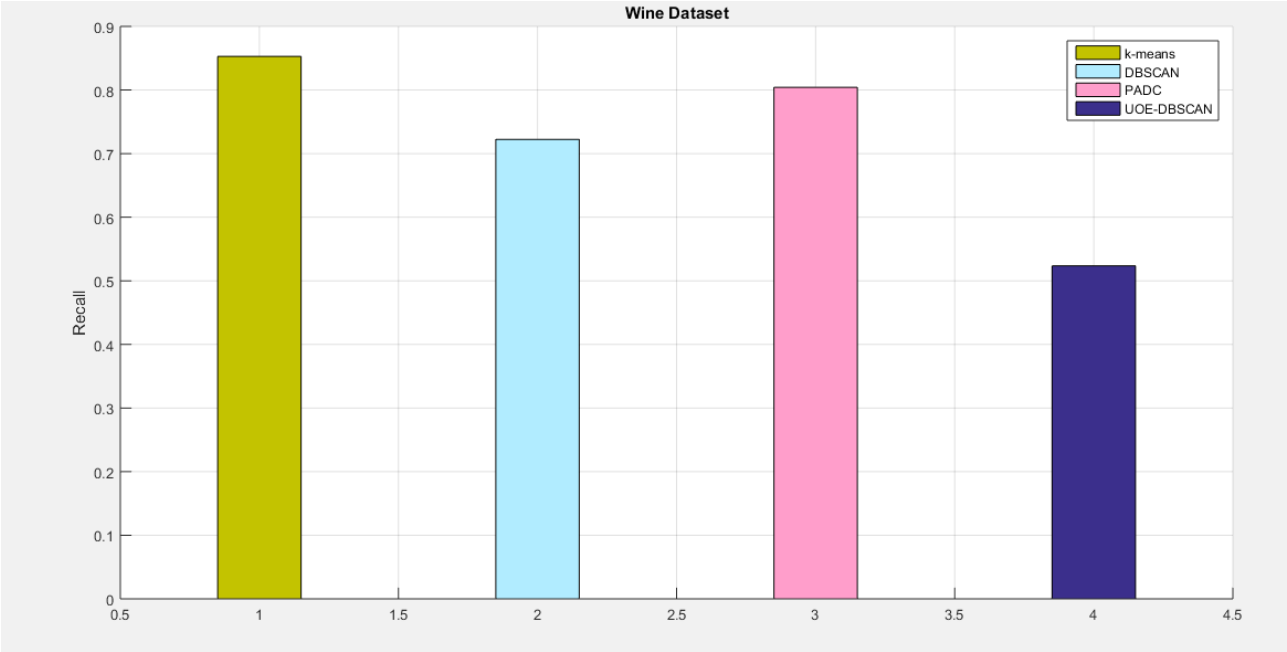


Εικόνα 47 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure όλων των αλγορίθμων στο Wine dataset.



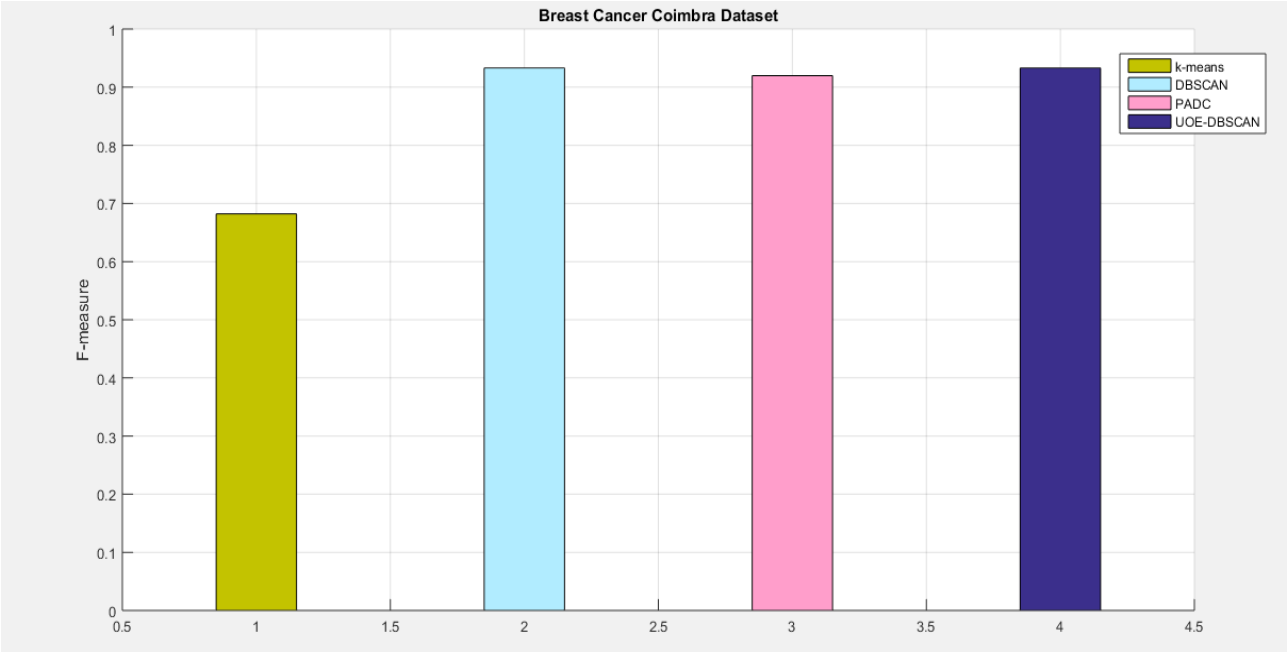
Εικόνα 48 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision όλων των αλγορίθμων στο Wine dataset.

Στις Εικόνες 47, 48, 49 εμφανίζονται τα αποτελέσματα όλων των αλγορίθμων στο Wine dataset συγκρίνοντάς τους ως προς τις μετρικές F-measure, Precision και Recall αντίστοιχα. Παρατηρώντας την Εικόνα 47, γίνεται αντιληπτό η υπεροχή του αλγορίθμου K-means έναντι του νέου μοντέλου κατά 45% (97% και 52% αντίστοιχα). Σε αυτό το σημείο είναι σημαντικό να τονίσουμε ότι το dataset αυτό, είναι το μόνο το dataset στον οποίο ο απλός DBSCAN υπερέχει σε κάποια από τις μετρικές του σε σχέση με τον UOE-DBSCAN. Παρακάτω, στην Εικόνα 48 παρατηρείται η μεγαλύτερη απόδοση του K-means σε σύγκριση με τον UOE-DBSCAN (92% και 52% αντίστοιχα) στην μετρική του Precision, ο PADC αποδίδει εξίσου καλό Precision (84%) ενώ ο αλγόριθμος ο DBSCAN αποδίδει Precision σε ποσοστό 52%.

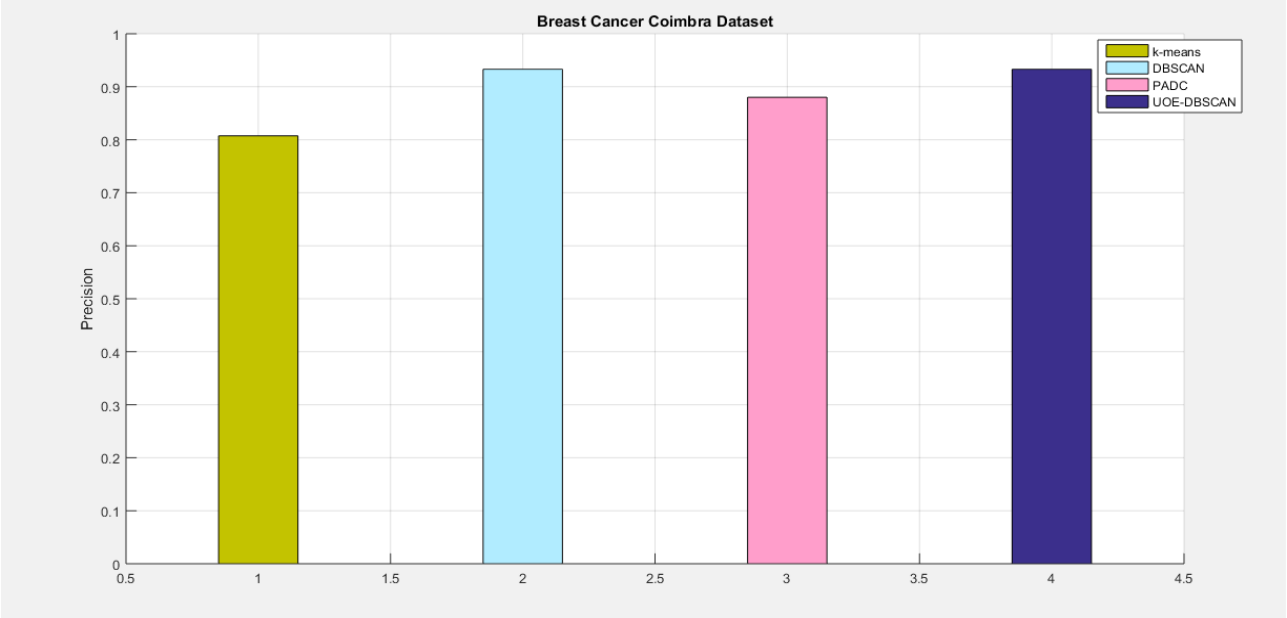


Εικόνα 49 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Recall όλων των αλγορίθμων στο Wine dataset.

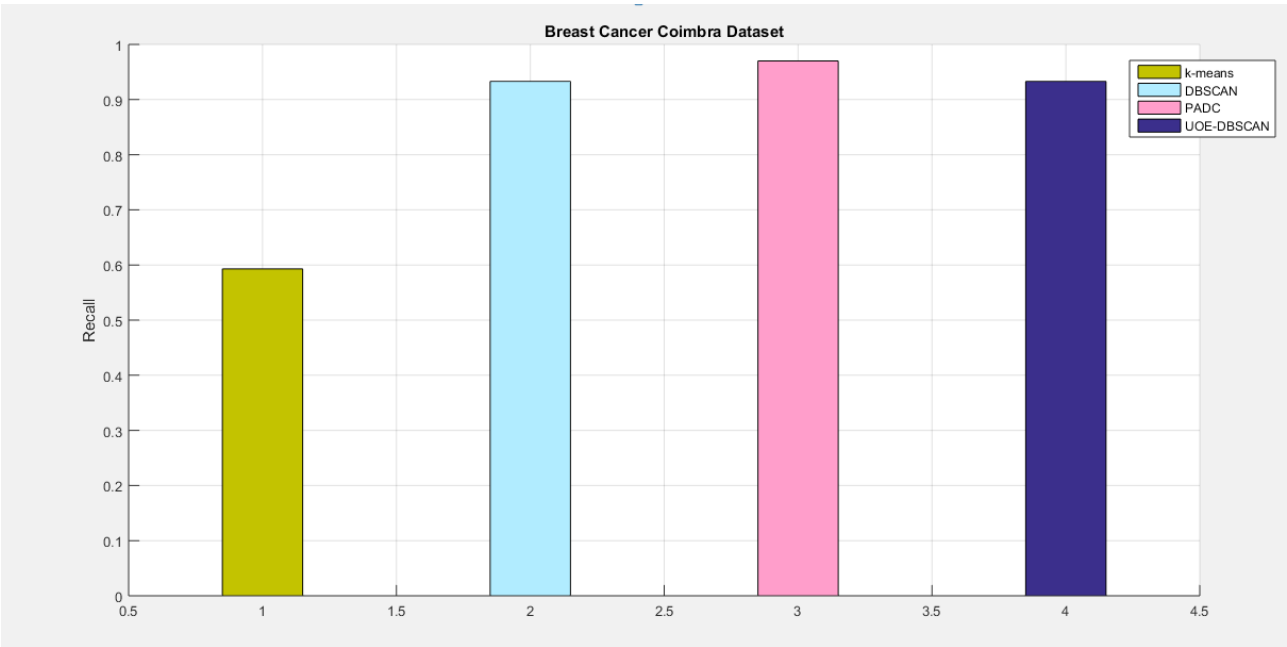
Στις Εικόνες 50, 51, 52 εμφανίζονται τα αποτελέσματα όλων των αλγορίθμων στο Breast Cancer Coimbra dataset συγκρίνοντάς τους ως προς το F-measure, Precision και το Recall αντίστοιχα. Από τα αποτελέσματα της Εικόνας 50, εξάγεται το συμπέρασμα ότι ο αλγόριθμος UOE-DBSCAN παράγει συνολικά καλύτερα αποτελέσματα σε σύγκριση με τον PADC (94% και 92% αντίστοιχα). Στην Εικόνα 51 παρατηρείται η μεγαλύτερη απόδοση σε Precision του νέου αλγοριθμικού μοντέλου UOE-DBSCAN σε σύγκριση με τον PADC (93% και 88% αντίστοιχα). Ο απλός DBSCAN αποδίδει εξίσου καλό Precision (93%) ενώ ο αλγόριθμος K-means αποδίδει Precision σε ποσοστό 81%. Παρατηρώντας την Εικόνα 52, προκύπτει το συμπέρασμα ότι το Recall των αλγορίθμων DBSCAN και UOE-DBSCAN κυμαίνεται στις ίδιες τιμές με το Precision τους, για τον PADC παρατηρείται μεγάλη απόδοση στο Recall σε ποσοστό 97% ενώ η τιμή του Recall για τον K-means είναι αρκετά χαμηλά (59%).



Εικόνα 50 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure όλων των αλγορίθμων στο Breast Cancer Coimbra dataset.

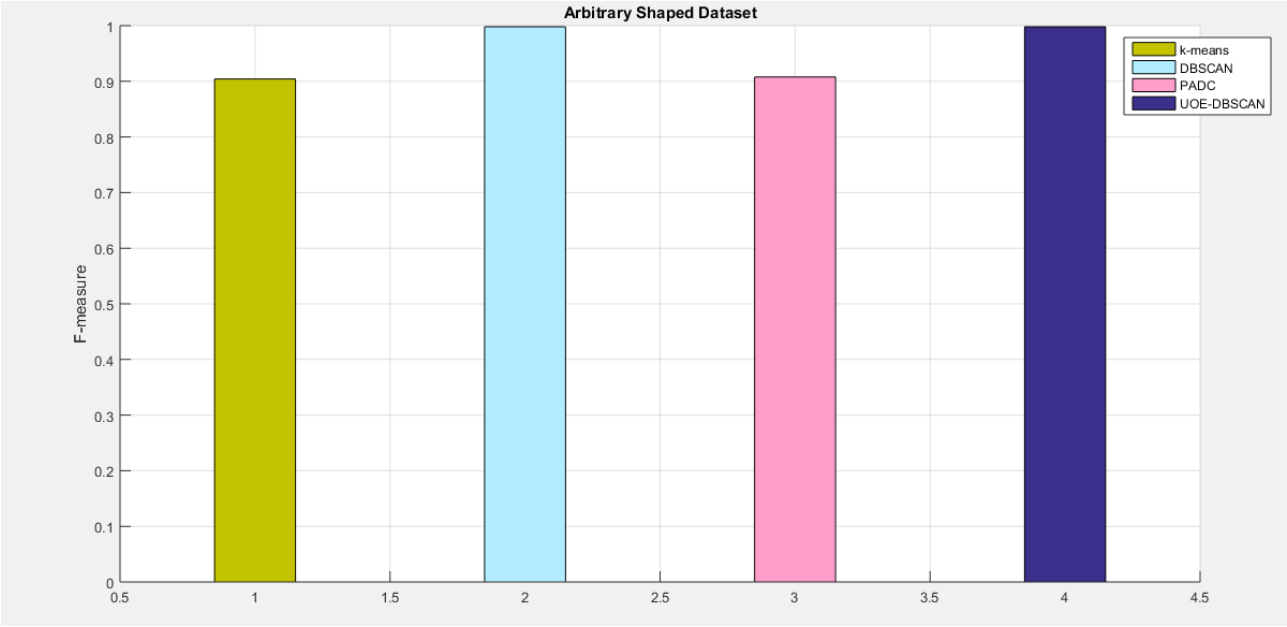


Εικόνα 51 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision όλων των αλγορίθμων στο Breast Cancer Coimbra dataset.



Εικόνα 52 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Recall όλων των αλγορίθμων στο Breast Cancer Coimbra dataset.

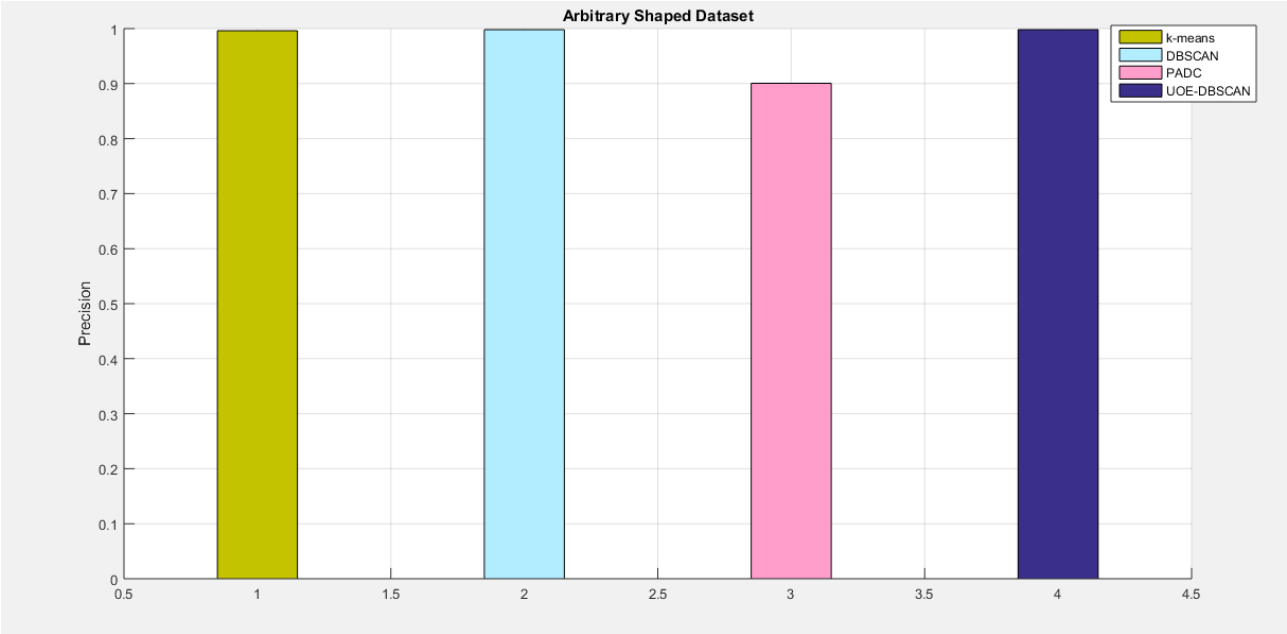
Στην Εικόνα 53 παρατηρείται η άριστη απόδοση στη μετρική του F-measure των αλγορίθμων DBSCAN και UOE-DBSCAN στο Arbitrary Shaped dataset ενώ οι αλγόριθμοι



Εικόνα 53 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure όλων των αλγορίθμων στο Arbitrary Shaped dataset.

K-means και PADC υστερούν σε αυτό το κομμάτι. Πιο συγκεκριμένα, τα αποτελέσματα του F-measure των δύο αλγορίθμων βασισμένων στη πυκνότητα αγγίζουν το ποσοστό του 99%, αποτέλεσμα λογικό αφού το dataset δημιουργεί σχηματική δομή, αντικείμενο στο οποίο είναι αρκετά αποδοτικοί οι συγκεκριμένοι αλγόριθμοι. Εν αντιθέσει, οι αλγόριθμοι K-means και PADC αποδίδουν χαμηλότερο ποσοστό F-measure (90% αμφότεροι).

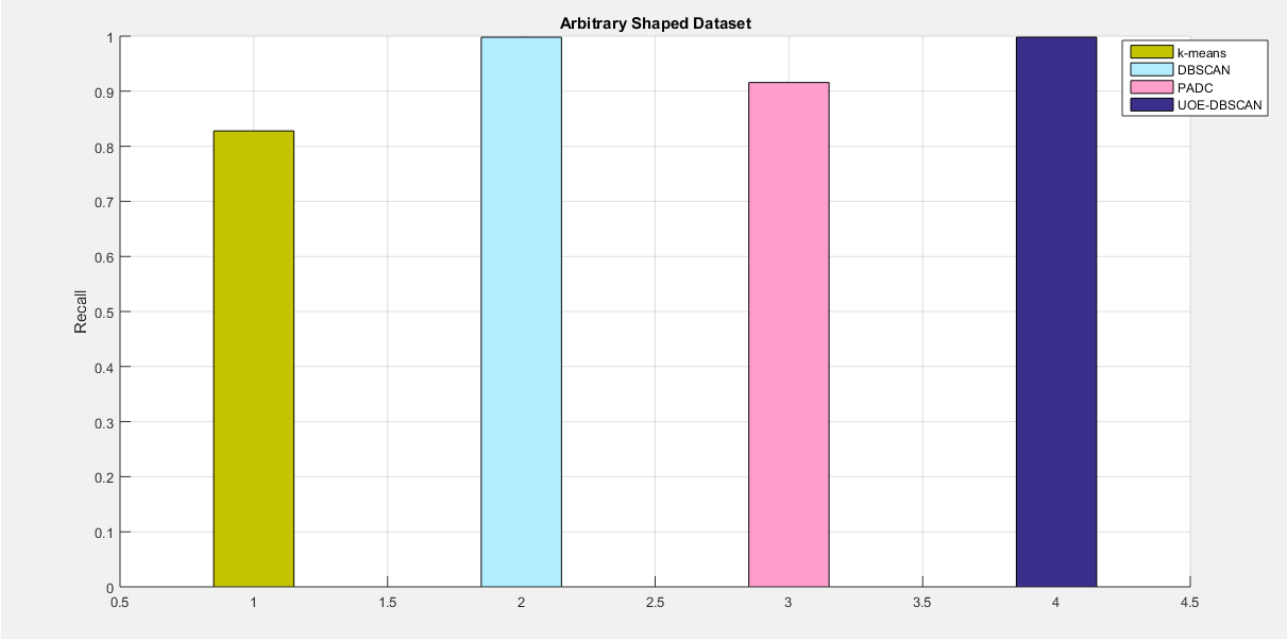
Στην Εικόνα 54 παρατηρείται η άριστη απόδοση στη μετρική του Precision των αλγορίθμων K-means, DBSCAN και UOE-DBSCAN στο Arbitrary Shaped dataset ενώ ο αλγόριθμος PADC υστερεί σε αυτό το κομμάτι. Πιο συγκεκριμένα, τα αποτελέσματα του Precision όλων των αλγορίθμων πέραν του PADC αγγίζουν το ποσοστό του 99% ενώ ο PADC αποδίδει Precision σε ποσοστό 90%.



Εικόνα 54 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision όλων των αλγορίθμων στο Arbitrary Shaped dataset.

Παρατηρώντας την Εικόνα 55, προκύπτει το συμπέρασμα ότι το Recall των αλγορίθμων DBSCAN και UOE-DBSCAN κυμαίνεται στις ίδιες τιμές με το Precision τους, για τον PADC παρατηρείται μεγάλη απόδοση στο Recall σε ποσοστό 91% ενώ η τιμή του Recall για τον K-means

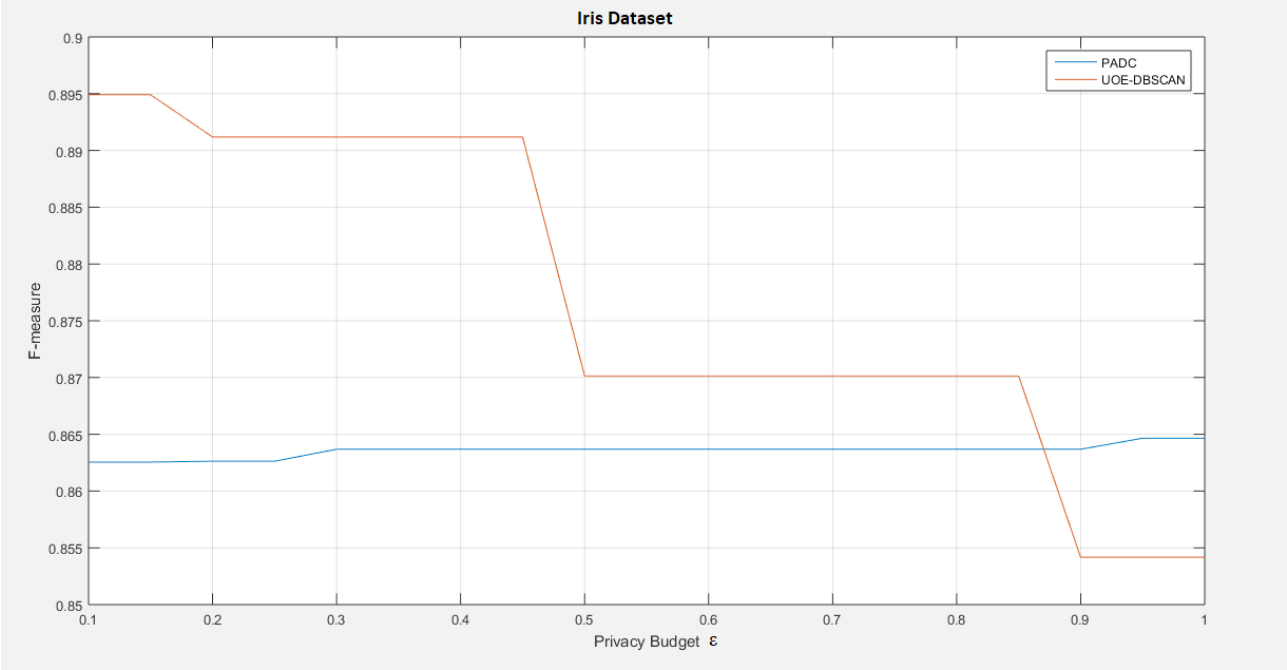
είναι αρκετά υψηλή σε ποσοστό 82%. Τα αποτελέσματα που παρατηρούνται στις δύο εικόνες επιβεβαιώνουν τις αναφορές που έχουν γίνει στο γεγονός ότι οι αλγόριθμοι DBSCAN και UOE-DBSCAN παράγουν ποιοτικότερα αποτελέσματα έναντι άλλων αλγορίθμων όταν τα δεδομένα αντικατοπτρίζονται ως σχηματικές δομές.



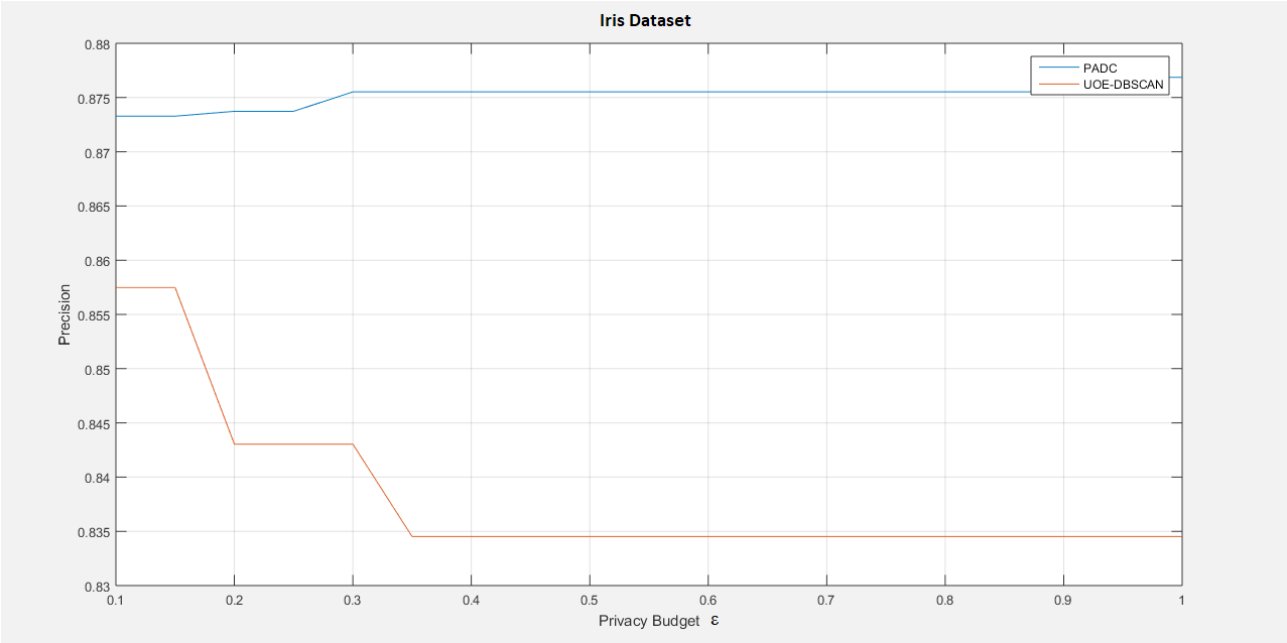
Εικόνα 55 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Recall όλων των αλγορίθμων στο Arbitrary Shaped dataset.

5.2 Σύγκριση αλγορίθμων με βάση το Iris dataset

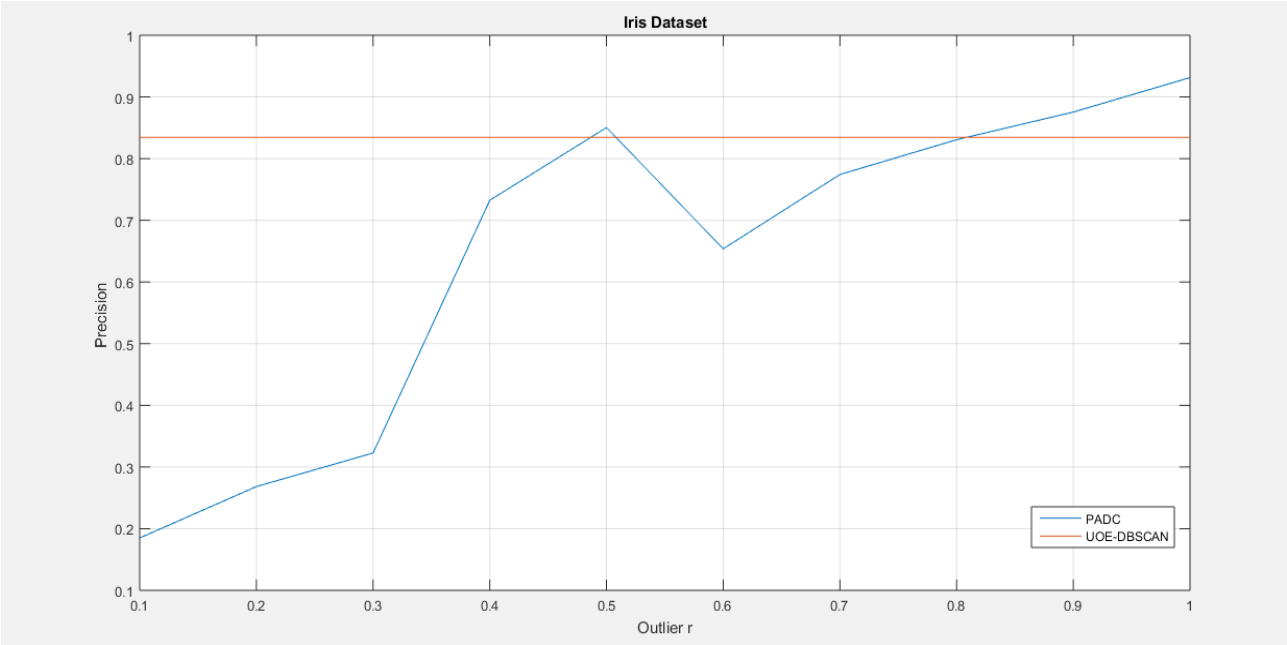
Στην Εικόνα 56, παρατηρείται η σύγκριση των αλγορίθμων PADC και UOE-DBSCAN. Η σύγκριση των αποτελεσμάτων γίνεται ως προς τη μετρική του F-measure και πραγματοποιείται με βάση την παράμετρο προστασίας (privacy budget) ε. Τα αποτελέσματα δείχνουν την υπεροχή του νέου αλγορίθμου που αναπτύχθηκε στα πλαίσια του συγγράμματος συγκριτικά με τον αλγόριθμο PADC. Πιο συγκεκριμένα, από τα αποτελέσματα είναι φανερό ότι όταν η παράμετρος προστασίας παίρνει τιμές στο διάστημα [0.1 - 0.87] , ο αλγόριθμος είναι πιο αποτελεσματικός σε σχέση με τον PADC κατά [0.7% – 3.2%] αντίστοιχα. Βέβαια, στο διάστημα [0.88 – 1] ο PADC παράγει καλύτερα αποτελέσματα αλλά η διαφορά από τον νέο αλγόριθμο είναι μικρότερη της τάξης 1%.



Εικόνα 56: Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure των αλγορίθμων PADC και UOE-DBSCAN στο Iris dataset ως προς τη παράμετρο προστασίας ε.



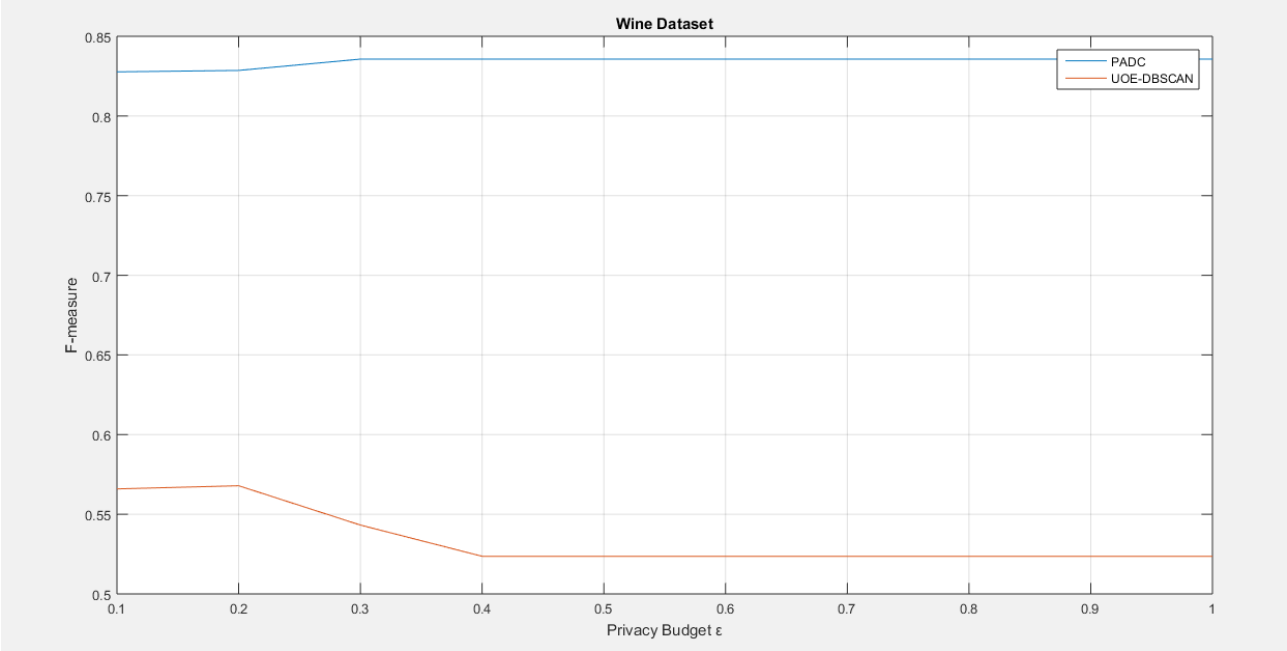
Εικόνα 57 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Iris dataset ως προς τη παράμετρο προστασίας ε.



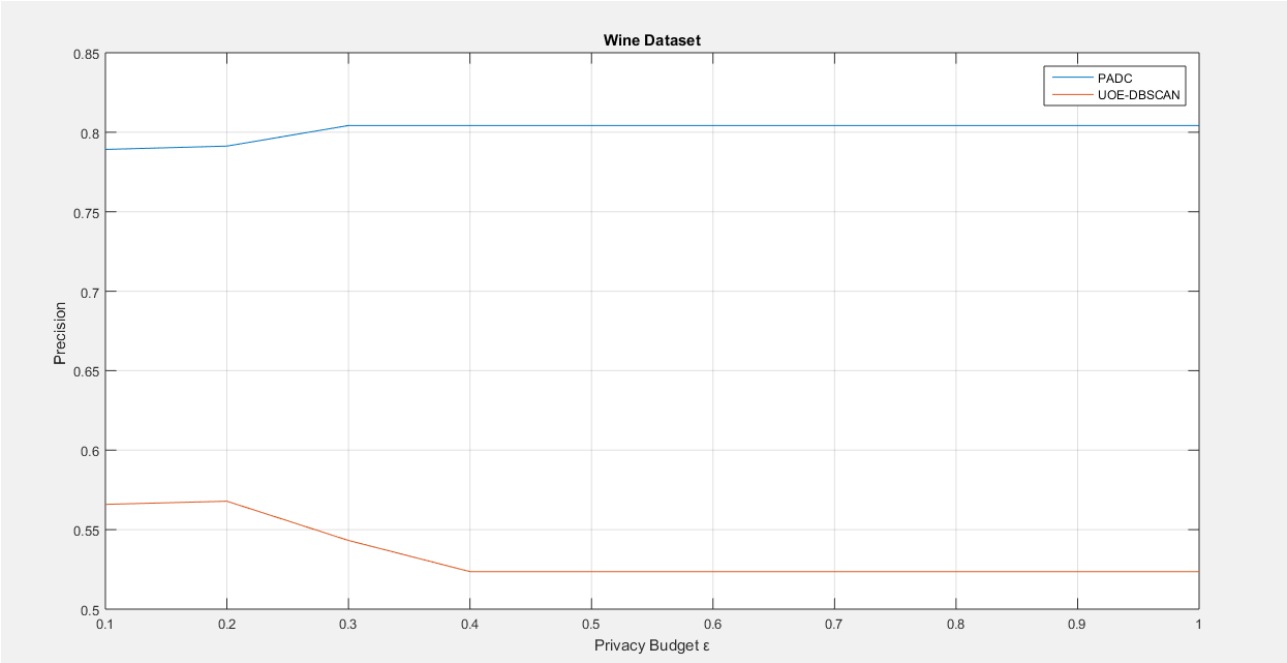
Εικόνα 58 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Iris dataset ως προς τη παράμετρο των ακραίων τιμών r.

Στις παραπάνω Εικόνες 57, 58 εμφανίζονται τα αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN με τη σύγκριση να γίνεται στη μετρική Precision ως προς το privacy budget και τη παράμετρο r (outlier parameter) αντίστοιχα. Η Εικόνα 57 αναδεικνύει τη μικρή διαφορά της μετρικής Precision για τους δύο αλγορίθμους. Ο αλγόριθμος PADC υπερτερεί σε σύγκριση με τον νέο αλγόριθμο για όλες τις δυνατές τιμές της παραμέτρου προστασίας, αλλά η διαφορά είναι αρκετά μικρή, της τάξεως του 1.9%. Από την άλλη, στην Εικόνα 58 απεικονίζεται η υπεροχή του νέου αλγορίθμου UOE-DBSCAN σε σχέση με τον PADC στην μετρική του Precision για διαφορετικές τιμές του r. Πιο συγκεκριμένα, ο UOE-DBSCAN είναι πιο αποδοτικός κατά [50%-60%] στο διάστημα [0.1 – 0.4] αντίστοιχα και κατά 12% στο διάστημα [0.6 – 0.7] ενώ στο υπόλοιπο διάστημα ο PADC υπερτερεί σχεδόν κατά 1% - 10%.

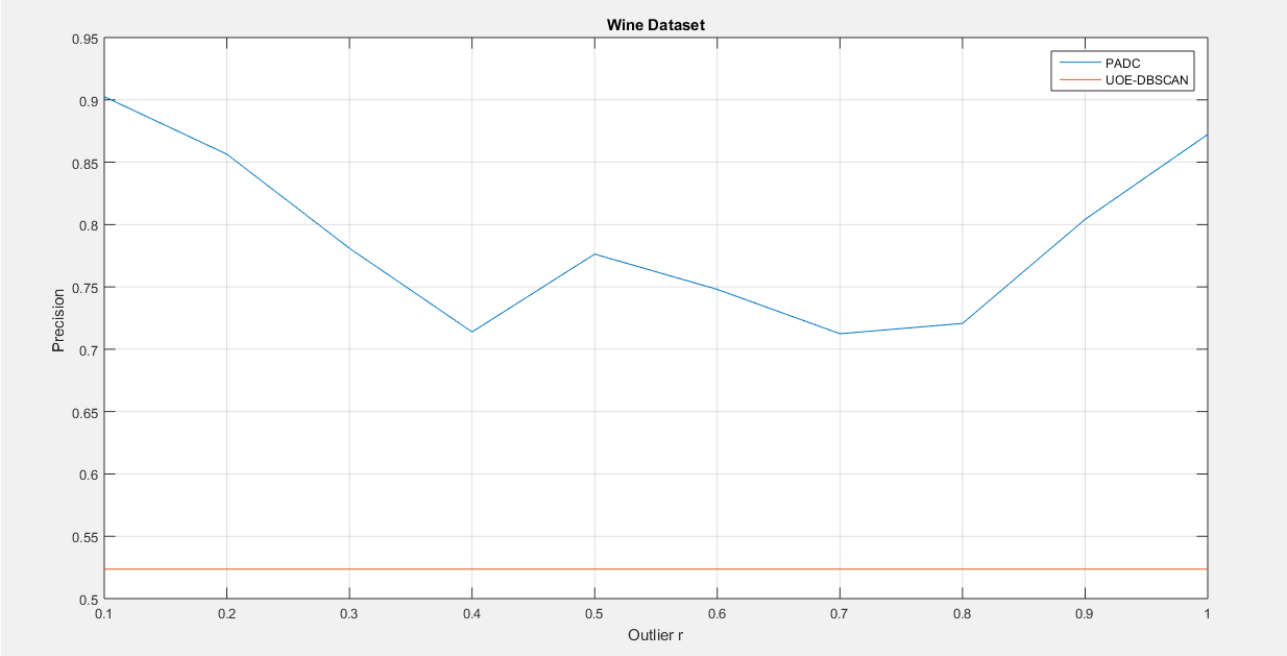
5.3 Σύγκριση αλγορίθμων με βάση το Wine dataset



Εικόνα 59 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset ως προς τη παράμετρο προστασίας ε.



Εικόνα 60 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset ως προς τη παράμετρο προστασίας ε.

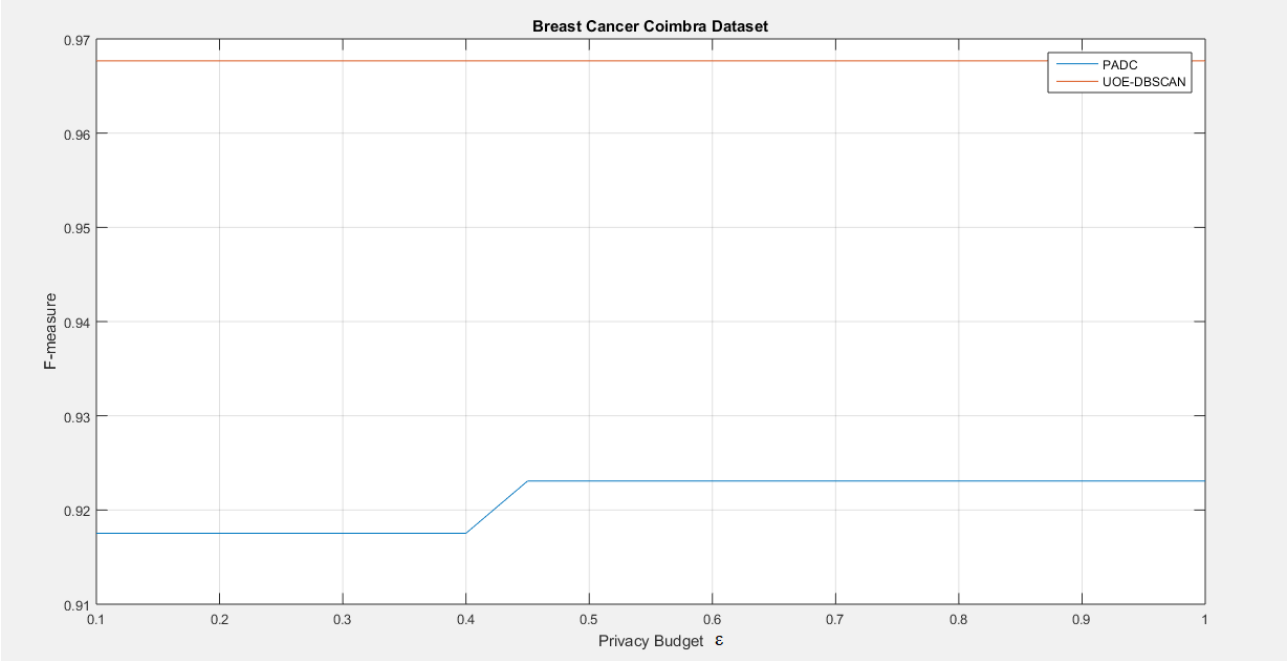


Εικόνα 61 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset ως προς τη παράμετρο ακραίων τιμών r.

Στην Εικόνα 59, παρατηρείται η σύγκριση των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset. Η σύγκριση των αποτελεσμάτων γίνεται ως προς τη μετρική του F-measure και πραγματοποιείται με βάση την παράμετρο προστασίας (privacy budget) ε. Από τα αποτελέσματα βγαίνει το συμπέρασμα ότι ο αλγόριθμος PADC αποδίδει καλύτερα από τον UOE-DBSCAN. Πιο συγκεκριμένα, από τα αποτελέσματα είναι φανερό ότι για οποιαδήποτε τιμή της παραμέτρου προστασίας ε, ο PADC είναι αποδοτικότερος σε ποσοστό από 26.5% (για ε < 0.2) έως 31% (για ε >0.4).

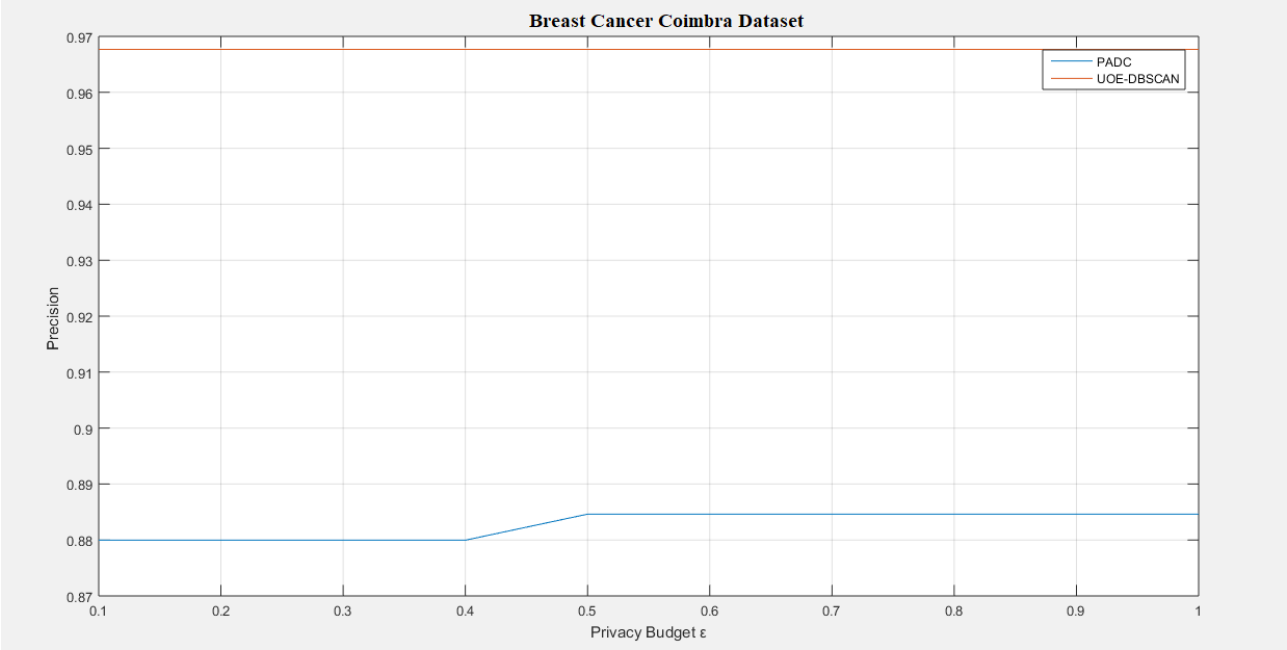
Στις παραπάνω Εικόνες 60, 61 εμφανίζονται τα αποτελέσματα των αλγορίθμων PADC και UOE-DBSCAN με τη σύγκριση να γίνεται στη μετρική Precision ως προς το privacy budget και τη παράμετρο r (outlier parameter) στο Wine dataset. Η Εικόνα 60 εμφανίζει τα ίδια αποτελέσματα με την Εικόνα 59, πράγμα που σημαίνει ότι στο συγκεκριμένο dataset όλες οι μετρικές F-measure, Precision και Recall είναι ίσες μεταξύ τους. Από την άλλη, στην Εικόνα 61 απεικονίζεται η υπεροχή του αλγορίθμου PADC έναντι του νέου μοντέλου UOE-DBSCAN στην μετρική του Precision για διαφορετικές τιμές του r. Πιο συγκεκριμένα, ο PADC είναι πιο αποδοτικός κατά [19%-38%] σε όλο το διάστημα [0.1 – 1]. Στο συγκεκριμένο dataset, παρατηρείται ότι ο PADC είναι σε όλες τις μετρικές ανώτερος σε απόδοση από τον UOE-DBSCAN.

5.4 Σύγκριση αλγορίθμων με βάση το Breast Cancer Coimbra dataset

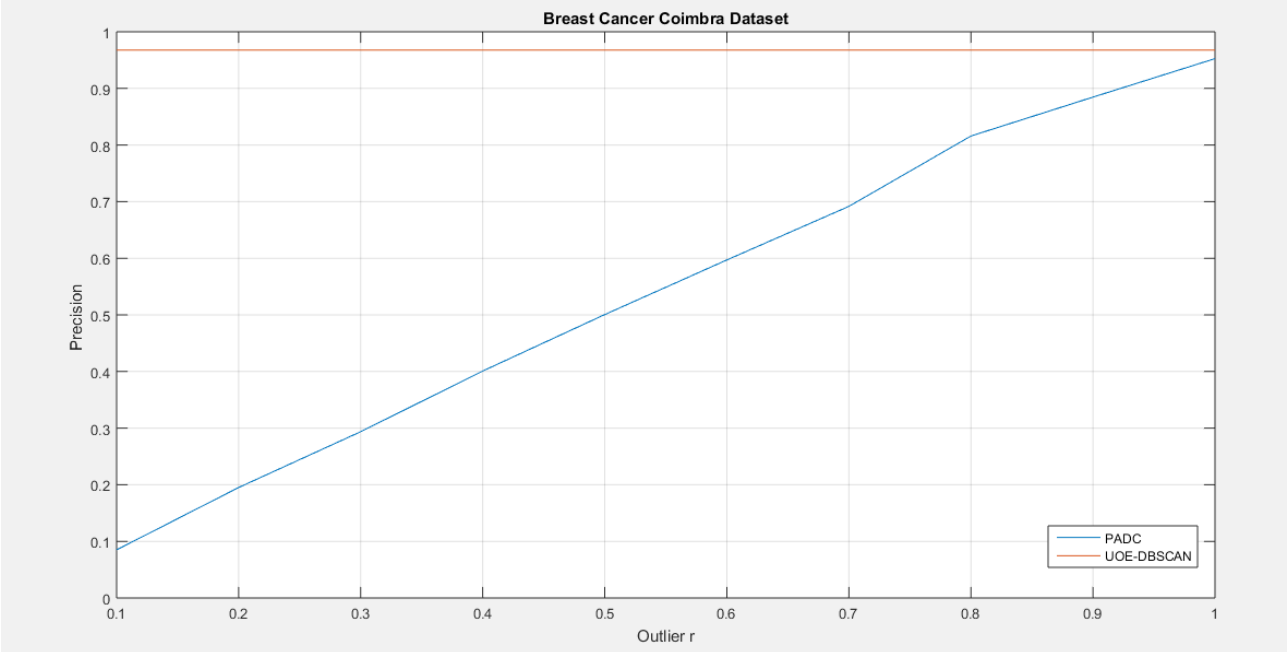


Εικόνα 62 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure των αλγορίθμων PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset ως προς τη παράμετρο προστασίας ε.

Στις Εικόνες 62, 63, 64 παρατηρούνται τα αποτελέσματα των μετρήσεων της αποδοτικότητας των αλγορίθμων PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset. Πιο συγκεκριμένα, η Εικόνα 62 παρουσιάζει τα αποτελέσματα του F-measure ως προς τη παράμετρο προστασίας ε. Παρατηρώντας τα αποτελέσματα γίνεται αντιληπτό ότι ο νέος αλγόριθμος που αναπτύχθηκε υπερέχει σταθερά εναντίον του PADC κατά 5% (96.8%/91.8% αντίστοιχα) για τιμές του ε από [0.1 – 0.4] και κατά 4.5% (96.8%/92.3%) για τιμές του ε από [0.45 – 1]. Άρα προκύπτει το συμπέρασμα ότι ο νέος αλγόριθμος είναι περισσότερο αποδοτικός και αποτελεσματικός στη ποιότητα των αποτελεσμάτων που παράγει.



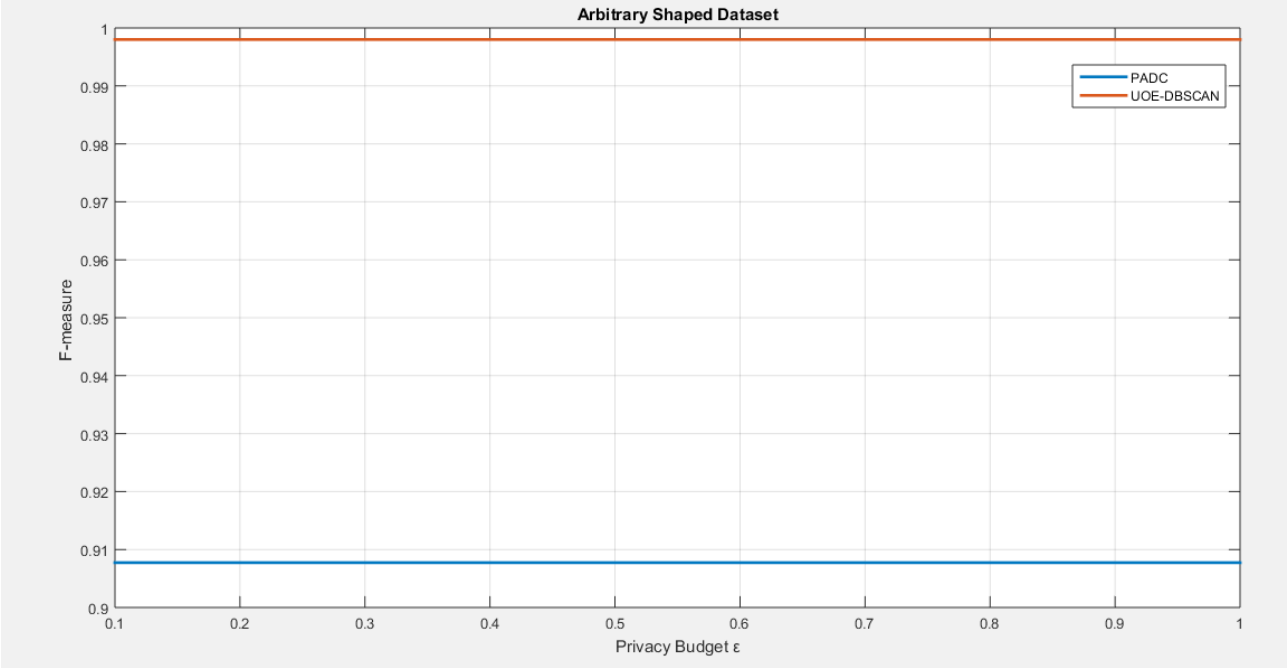
Εικόνα 63 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset ως προς τη παράμετρο προστασίας ε.



Εικόνα 64 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Breast Cancer Coimbra dataset ως προς τη παράμετρο r.

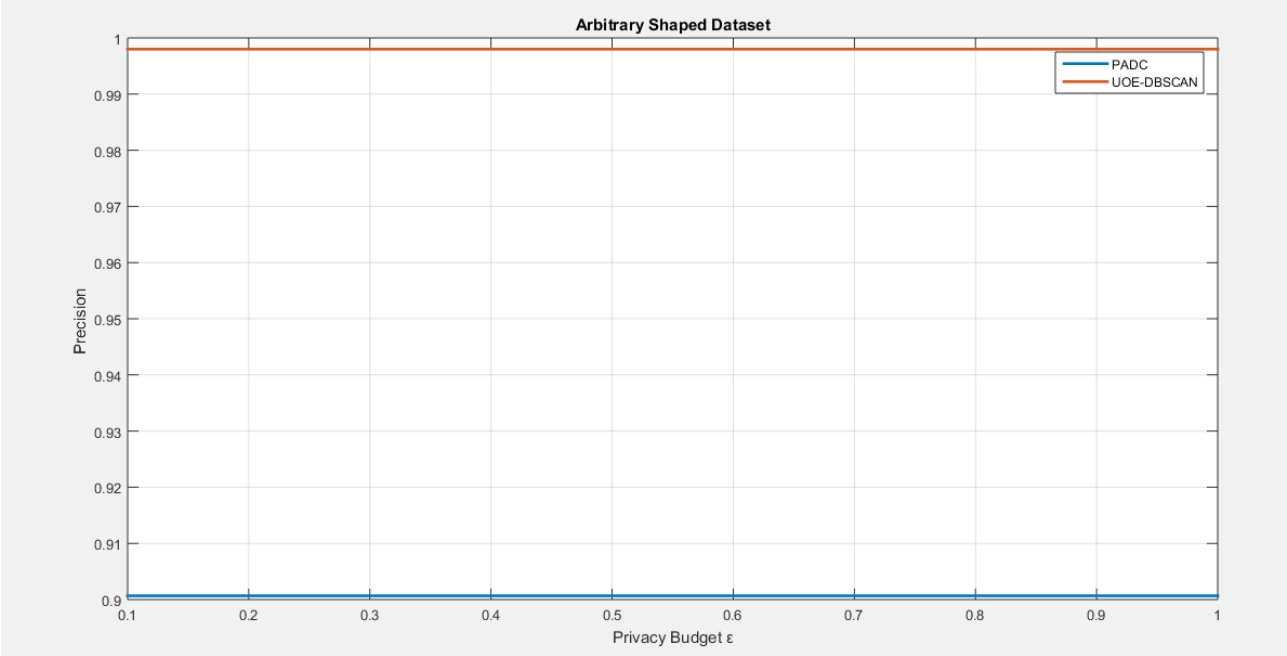
Το παραπάνω συμπέρασμα επιβεβαιώνεται και από τις Εικόνες 63 και 64. Η Εικόνα 63 δείχνει ότι ο νέος αλγόριθμος είναι αποδοτικότερος κατά 8,8% (96.8%/88% αντίστοιχα) για τιμές της παραμέτρου προστασίας ε από [0.1 – 0.4] ενώ για τιμές από [0.45 – 1] ο UOE-DBSCAN αποδίδει κατά 8.35% (96.8%/88.45% αντίστοιχα) καλύτερα. Επίσης, η Εικόνα 64 παρουσιάζει την απόδοση του Precision ως προς τη παράμετρο των outliers r και παρατηρείται ότι ο νέος αλγόριθμος συνεχίζει να είναι αποδοτικότερος της τάξεως από 1% (για r=1) έως 88,7% (για r=0.1) ανεξάρτητα από τις διαφορετικές τιμές του r. Είναι φανερό ότι ο νέος αλγόριθμος UOE-DBSCAN είναι πολύ ποιοτικός καθώς, στο συγκεκριμένο dataset, τα αποτελέσματα που αποδίδει βρίσκονται σε πολύ υψηλό επίπεδο.

5.5 Σύγκριση αλγορίθμων με βάση το Arbitrary dataset

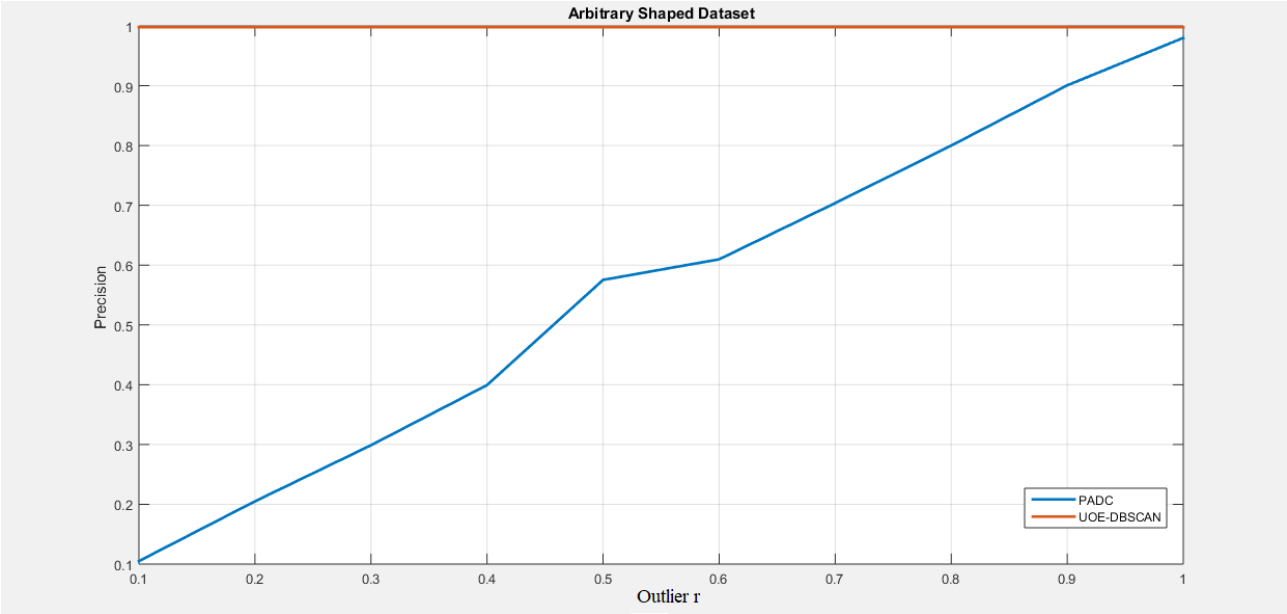


Εικόνα 65 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους F-measure των αλγορίθμων PADC και UOE-DBSCAN στο Arbitrary Shaped dataset ως προς τη παράμετρο προστασίας ε.

Στις Εικόνες 65, 66, 67 παρατηρούνται τα αποτελέσματα των μετρήσεων της αποδοτικότητας των αλγορίθμων PADC και UOE-DBSCAN στο Arbitrary Shaped dataset. Πιο συγκεκριμένα, η Εικόνα 65 παρουσιάζει τα αποτελέσματα του F-measure ως προς τη παράμετρο προστασίας ε. Παρατηρώντας τα αποτελέσματα γίνεται αντιληπτό ότι ο νέος αλγόριθμος που αναπτύχθηκε υπερέχει σταθερά εναντίον του PADC κατά 9% (99.8%/90.8% αντίστοιχα) για όλες τις τιμές του ε από [0.1 – 1]. Άρα προκύπτει το συμπέρασμα ότι ο νέος αλγόριθμος είναι περισσότερο αποδοτικός και αποτελεσματικός στη ποιότητα των αποτελεσμάτων που εξάγει. Επιπλέον, η Εικόνα 66 δείχνει ότι ο νέος αλγόριθμος είναι αποδοτικότερος κατά 9.7% (99.8%/90.1% αντίστοιχα) για όλες τις τιμές της παραμέτρου προστασίας ε από [0.1 – 1]. Επίσης, η Εικόνα 67 παρουσιάζει την απόδοση του Precision ως προς τη παράμετρο των outliers r και παρατηρείται ότι ο νέος αλγόριθμος συνεχίζει να είναι αποδοτικότερος σε ποσοστό από 3.8% (για r=1) έως 90% (για r=0.1) ανεξάρτητα από τις διαφορετικές τιμές r. Είναι φανερό ότι ο νέος αλγόριθμος UOE-DBSCAN είναι πολύ ποιοτικός καθώς, στο συγκεκριμένο dataset, τα αποτελέσματα που αποδίδει βρίσκονται στο βέλτιστο επίπεδο.



Εικόνα 66 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Arbitrary Shaped dataset ως προς τη παράμετρο προστασίας ε.



Εικόνα 67 - Συγκριτικά αποτελέσματα του μετρικού μεγέθους Precision των αλγορίθμων PADC και UOE-DBSCAN στο Arbitrary Shaped dataset ως προς τη παράμετρο ακραίων τιμών r.

6. Αποτίμηση της αποδοτικότητας του νέου αλγορίθμου

Στο προηγούμενο κεφάλαιο μελετήθηκε η σύγκριση των αλγορίθμων συσταδοποίησης που εφαρμόζονται στο Διαδίκτυο των Πραγμάτων με στόχο την επίτευξη της ανωνυμίας των δεδομένων των χρηστών. Σε αυτό το κεφάλαιο θα πραγματοποιηθεί η αποτίμηση του νέου αλγορίθμου με βάση τα αποτελέσματα που παρατηρήθηκαν στο προηγούμενο κεφάλαιο.

Το νέο αλγοριθμικό μοντέλο UOE-DBSCAN που αναπτύχθηκε στη παρούσα διπλωματική, εμφανίζει πολύ ποιοτικά αποτελέσματα όταν χειρίζεται δεδομένα που δημιουργούν σχηματικές δομές. Ο πυρήνας του νέου αλγορίθμου είναι ο απλός DBSCAN, με αποτέλεσμα να κληρονομούνται όλες οι ιδιότητες, τα θετικά χαρακτηριστικά και οι ιδιαιτερότητες του βασικού αυτού αλγορίθμου. Βέβαια, ο νέος αλγόριθμος αποτελεί βελτιστοποιημένη εκδοχή του απλού DBSCAN με αποτέλεσμα στα περισσότερα γραφήματα που παρουσιάστηκαν στο κεφάλαιο 5, να παρατηρείται ότι ο νέος αλγόριθμος παράγει καλύτερα και πιο αξιόπιστα αποτελέσματα.

Από τα datasets που χρησιμοποιήθηκαν και από τα αποτελέσματα που παρατηρήθηκαν προκύπτουν διάφορα και ενδιαφέροντα συμπεράσματα για τη συμπεριφορά του νέου αλγορίθμου UOE-DBSCAN τα οποία διαφέρουν από dataset σε dataset. Πιο συγκεκριμένα, το νέο αλγοριθμικό σχήμα παρουσίασε καλύτερη συμπεριφορά στο Iris dataset, στο Breast Cancer Coimbra dataset και στο Arbitrary Shaped dataset εξάγοντας ποιοτικότερα αποτελέσματα συγκριτικά με τους υπόλοιπους αλγορίθμους. Τα παραπάνω τρία datasets διαθέτουν σχετικά μικρό αριθμό διαστάσεων (4, 9 και 2 αντίστοιχα), γεγονός που οδηγεί στο συμπέρασμα ότι ο UOE-DBSCAN είναι αποδοτικότερος όταν διαχειρίζεται δεδομένα μικρών διαστάσεων. Βέβαια, τα πειράματα πραγματοποιήθηκαν σε δεδομένα που λαμβάνουν πραγματικές και ακέραιες τιμές. Θα μπορούσε να αποτελέσει αντικείμενο μελλοντικής έρευνας η μελέτη του αλγορίθμου σε δεδομένα που λαμβάνουν αλφαριθμητικές τιμές.

Το νέο αλγοριθμικό μοντέλο που αναπτύχθηκε, βασίστηκε στη φιλοσοφία ότι πρέπει να βελτιστοποιηθεί η ποιότητα των αποτελεσμάτων. Κατά συνέπεια, ο αλγόριθμος υλοποιήθηκε με τρόπο τέτοιο ώστε να ελαχιστοποιεί τις ακραίες τιμές και να τις μετατρέπει σε αξιοποιήσιμα και ασφαλή δεδομένα. Το πόρισμα αυτό γίνεται αντιληπτό όταν παρατηρήσουμε τα αποτελέσματα των αλγορίθμων DBSCAN και UOE-DBSCAN, κυρίως στο Iris dataset, όπως φαίνεται στις Εικόνες 36, 37. Στα αποτελέσματα των δύο αλγορίθμων είναι φανερή η αύξηση των σημείων που συμμετέχουν στο σχηματισμό των συστάδων όσον αφορά το νέο αλγοριθμικό μοντέλο.

Ένα ακόμη ενδιαφέρον στοιχείο του νέου αλγορίθμου είναι ότι, παρά το γεγονός ότι αναπτύχθηκε με στόχο την εξάλειψη όσο το δυνατόν περισσότερων ακραίων τιμών, στα αποτελέσματα του κεφαλαίου 5 παρατηρήθηκε ότι ο νέος αλγόριθμος δεν επηρεάζεται καθόλου από τις ακραίες τιμές. Το συμπέρασμα αυτό επιβεβαιώνεται από τις Εικόνες 52, 55, 58, 61 στις οποίες το νέο μοντέλο παρατηρείται να έχει σταθερή συμπεριφορά ανεξαρτήτως των τιμών της παραμέτρου r σε σχέση με τις διακυμάνσεις που εμφανίζει ο αλγόριθμος PADC.

Τέλος, η ιδέα για την υλοποίηση του συγκεκριμένου αλγοριθμικού μοντέλου σχηματίστηκε παρατηρώντας τα αποτελέσματα της Εικόνας 33. Πιο συγκεκριμένα, εκτελώντας τον νέο αλγόριθμο αρκετές φορές επαναληπτικά, παρατηρήθηκε ότι αυξάνεται το ποσοστό της απόδοσης της μετρικής F-measure για το Iris dataset. Αυτό το πόρισμα, οδήγησε στο να γίνει αντιληπτό ότι κάποιες φορές, τα αποτελέσματα των αλγοριθμικών μοντέλων δεν είναι πάντα τα βέλτιστα δυνατά. Πάντα υπάρχει περιθώριο για περαιτέρω βελτίωση κάθε αλγορίθμου, με γνώμονα πάντα τη διασφάλιση της ιδιωτικότητας των χρηστών, έχοντας σαν φυσικό συμπέρασμα τη εξέλιξη της κοινωνίας σε μια πιο εκσυγχρονισμένη και ασφαλή κοινωνία.

7. Συμπεράσματα

Σε αυτή τη διπλωματική εργασία, πραγματοποιήθηκε η μελέτη της έννοιας του Διαδικτύου των Πραγμάτων (IoT) και αναλύθηκαν τα χαρακτηριστικά και οι ιδιότητές του.

Στη συνέχεια, έγινε μια λεπτομερής αναφορά των αλγορίθμων ομαδοποίησης K-means, DBSCAN, ιεραρχικής συσταδοποίησης και PADC, οι οποίοι έχουν ως κύριο στόχο την επίτευξη ανωνυμίας και τη διατήρηση της ιδιωτικότητας των δεδομένων των χρηστών που είναι συνδεδεμένοι στο Διαδίκτυο των Πραγμάτων μέσω των συσκευών που διαθέτουν.

Επίσης, αναπτύχθηκε ένα νέο αλγοριθμικό μοντέλο, με τίτλο UOE-DBSCAN, το οποίο αποτελεί βελτιστοποιημένη εκδοχή του βασικού αλγορίθμου DBSCAN και συμβάλλει στη μείωση των ανωμαλιών που δημιουργούν στα τελικά αποτελέσματα οι ακραίες τιμές. Από την παρατήρηση, την ανάλυση και την σύγκριση των αποτελεσμάτων του νέου αλγορίθμου σε σχέση με τους προαναφερθέντες αλγόριθμους συσταδοποίησης, προκύπτει το συμπέρασμα ότι ο UOE-DBSCAN είναι αρκετά αποδοτικότερος από τους υπόλοιπους αλγόριθμους που αναλύθηκαν παραπάνω όταν τα δεδομένα είναι μικρών διαστάσεων ή όταν αυτά αναπαριστούν σχηματικές δομές.

Αναφορικά, το νέο αλγοριθμικό σχήμα λειτουργεί αρκετά αποδοτικά στο Iris dataset εμφανίζοντας ποιοτικότερα αποτελέσματα στην μετρική του F-measure σε σχέση με τον PADC κατά [0.7% – 3.2%] όταν η παράμετρος προστασίας παίρνει τιμές στο διάστημα [0.1 - 0.87]. Βέβαια, στο διάστημα [0.88 – 1] ο PADC παράγει καλύτερα αποτελέσματα σε ποσοστό της τάξης του 1%. Ο αλγόριθμος PADC υπερτερεί σε σύγκριση με τον νέο αλγόριθμο στη μετρική του Precision για όλες τις δυνατές τιμές της παραμέτρου προστασίας, αλλά η διαφορά είναι της τάξεως του 1.9%. Επίσης, ο UOE-DBSCAN είναι πιο αποδοτικός κατά [50%-60%] στο διάστημα [0.1 – 0.4] και κατά 12% στο διάστημα [0.6 – 0.7] ενώ στο υπόλοιπο διάστημα ο PADC υπερτερεί σχεδόν κατά 1% - 10% στην μετρική του Precision για διαφορετικές τιμές του r. Επίσης, συγκρίνοντας όλους τους αλγορίθμους στο Iris dataset σε σχέση με τις τρεις μετρικές F-measure, Precision και Recall προκύπτει το συμπέρασμα ότι ο αλγόριθμος K-means είναι ο πιο αποδοτικός ως προς το F-measure φτάνοντας σε ποσοστό 99% με τον PADC και το νέο μοντέλο UOE-DBSCAN να έχουν αρκετά απόδοση της τάξεως του 95% και 92% αντίστοιχα. Ως προς το Precision προκύπτει το συμπέρασμα ότι ο αλγόριθμος K-means είναι ο αποδοτικότερος φτάνοντας σε ποσοστό 93%, ενώ ο νέος αλγόριθμος UOE-DBSCAN πλησιάζει αρκετά την απόδοση του PADC με διαφορά 3% ενώ

είναι καλύτερος σε Precision από τον απλό DBSCAN με διαφορά 2%. Αντιθέτως, το Recall του K-means είναι πολύ χαμηλό σε ποσοστό 56% ενώ οι αλγόριθμοι PADC και UOE-DBSCAN είναι αρκετά αποδοτικοί σε ποσοστό 95% και 92% αντίστοιχα.

Ακόμη, ο UOE-DBSCAN αποδίδει καλύτερα στο Breast Cancer Coimbra dataset και υπερέχει σταθερά εναντίον του PADC στην μετρική του F-measure κατά 5% για τιμές του ε από [0.1 – 0.4] και κατά 4.5% για τιμές του ε από [0.45 – 1] συμπεραίνοντας ότι ο νέος αλγόριθμος που αναπτύχθηκε είναι πιο αποδοτικός και αποτελεσματικός έναντι των υπόλοιπων αλγορίθμων. Επίσης, παρατηρούμε ότι το Precision του UOE-DBSCAN σε σχέση με τη παράμετρο προστασίας ε είναι αποδοτικότερο κατά 8,8% για τιμές από [0.1 – 0.4] και για τιμές από [0.45 – 1] αποδίδει κατά 8.35% καλύτερα ενώ σε σχέση με τη παράμετρο των ακραίων τιμών r το νέο αλγοριθμικό σχήμα είναι αποδοτικότερο σε ποσοστό της τάξεως από 1% έως 88,7% ανεξάρτητα από τις διαφορετικές τιμές r. Συγκρίνοντας όλους τους αλγορίθμους στο Breast Cancer Coimbra dataset εξάγεται το συμπέρασμα ότι ο αλγόριθμος UOE-DBSCAN παράγει συνολικά καλύτερα αποτελέσματα σε σύγκριση με τον PADC με διαφορά της τάξεως του 2% στη μετρική του F-measure. Επίσης, παρατηρείται η μεγαλύτερη απόδοση σε Precision του νέου αλγοριθμικού μοντέλου UOE-DBSCAN σε σύγκριση με τον PADC με τιμές 93% και 88% αντίστοιχα. Ο απλός DBSCAN αποδίδει Precision σε ποσοστό 93% ενώ ο αλγόριθμος K-means αποδίδει Precision σε ποσοστό 81%. Επιπλέον, προκύπτει το συμπέρασμα ότι το Recall των αλγορίθμων DBSCAN και UOE-DBSCAN κυμαίνεται σε ποσοστό 93%, για τον PADC παρατηρείται μεγάλη απόδοση στο Recall σε ποσοστό 97% ενώ η τιμή του Recall για τον K-means είναι 59%.

Επίσης, συμπεραίνεται ότι ο UOE-DBSCAN αποδίδει καλύτερα στο Arbitrary Shaped dataset παρατηρώντας ότι ο νέος αλγόριθμος υπερέχει σταθερά εναντίον του PADC κατά 9% για όλες τις τιμές του ε στην μετρική του F-measure. Το παραπάνω συμπέρασμα επιβεβαιώνεται, παρατηρώντας το Precision σε σχέση με τη παράμετρο προστασίας ε και τη παράμετρο r. Πιο συγκεκριμένα, ο νέος αλγόριθμος είναι αποδοτικότερος κατά 9.7% για όλες τις τιμές της παραμέτρου προστασίας ε ενώ η απόδοση του Precision ως προς τη παράμετρο των outliers r είναι ποιοτικότερη σε ποσοστό από 3.8% έως 90% ανεξάρτητα από τις διαφορετικές τιμές r. Είναι φανερό ότι, στο συγκεκριμένο dataset, ο νέος αλγόριθμος είναι περισσότερο αποδοτικός και αποτελεσματικός στη ποιότητα των αποτελεσμάτων που εξάγει. Συγκρίνοντας όλους τους αλγορίθμους στο Arbitrary Shaped dataset στη μετρική F-measure, οι αλγόριθμοι K-means και PADC υστερούν σε αυτό το κομμάτι. Πιο συγκεκριμένα, οι αλγόριθμοι K-means και PADC αποδίδουν χαμηλότερο ποσοστό F-measure σε ποσοστό 90% αμφότεροι. Εν αντιθέσει, τα

αποτελέσματα του F-measure των δύο αλγορίθμων βασισμένων στη πυκνότητα αγγίζουν το ποσοστό του 99%. Σχετικά με τη μετρική του Precision, τα αποτελέσματα όλων των αλγορίθμων πέραν του PADC αγγίζουν το ποσοστό του 99% ενώ ο PADC αποδίδει σε ποσοστό 90%. Επίσης, από τα αποτελέσματα προκύπτει το συμπέρασμα ότι το Recall των αλγορίθμων DBSCAN και UOE-DBSCAN κυμαίνεται στις ίδιες τιμές με το Precision τους, για τον PADC παρατηρείται απόδοση στο Recall σε ποσοστό 91% ενώ ο K-means φτάνει σε ποσοστό 82%. Τα αποτελέσματα του νέου αλγοριθμικού μοντέλου είναι ποιοτικότερα έναντι άλλων αλγορίθμων εξαιτίας του γεγονότος ότι τα δεδομένα αντικατοπτρίζονται ως σχηματικές δομές.

Αντιθέτως, ο UOE-DBSCAN, από τις αναλύσεις των αποτελεσμάτων, παρατηρείται ότι δεν αποδίδει εξίσου καλά στο Wine dataset. Από τη σύγκριση των αλγορίθμων PADC και UOE-DBSCAN στο Wine dataset εξάγεται το συμπέρασμα ότι ο αλγόριθμος PADC αποδίδει καλύτερα από τον UOE-DBSCAN καθώς παρατηρείται ότι για οποιαδήποτε τιμή της παραμέτρου προστασίας ε, ο PADC είναι αποδοτικότερος σε ποσοστό από 26.5% έως 31%. Στο συγκεκριμένο dataset, παρατηρείται ότι τα αποτελέσματα εμφανίζουν παρόμοιες τιμές στη μετρική του F-measure και στη μετρική του Precision. Ακόμη, ο αλγόριθμος PADC υπερέχει έναντι του νέου μοντέλου UOE-DBSCAN στην μετρική του Precision για διαφορετικές τιμές του r. Πιο συγκεκριμένα, ο PADC είναι πιο αποδοτικός κατά [19%-38%] σε όλο το διάστημα των τιμών του r. Στο συγκεκριμένο dataset, παρατηρείται ότι ο PADC είναι σε όλες τις μετρικές ανώτερος σε απόδοση από τον UOE-DBSCAN. Συγκρίνοντας όλους τους αλγορίθμους στο Wine dataset ως προς το F-measure, Precision και το Recall, γίνεται αντιληπτό η υπεροχή του αλγορίθμου K-means έναντι του νέου μοντέλου κατά 45% στη μετρική F-measure. Επίσης, παρατηρείται η μεγαλύτερη απόδοση του K-means σε σύγκριση με τον UOE-DBSCAN (92% και 52% αντίστοιχα) στη μετρική του Precision. Ο PADC αποδίδει εξίσου καλό Precision σε ποσοστό 84% ενώ ο αλγόριθμος ο DBSCAN αποδίδει Precision σε ποσοστό 52%.

Επιπλέον, το νέο αλγοριθμικό σχήμα εμφανίζει ανθεκτικότητα απέναντι στις παρεμβολές των ακραίων τιμών, παρά το γεγονός ότι έχει ως βασική του λειτουργία την εξάλειψη όσο το δυνατόν περισσότερων εξ' αυτών.

Επιπρόσθετα, το νέο αλγοριθμικό μοντέλο προσφέρει δυνατότητες για μελλοντική εργασία και έρευνα, ως προς τον σχεδιασμό και τη γενικότερη βελτιστοποίηση της απόδοσής του. Υπάρχει η δυνατότητα μελλοντικής εργασίας πάνω στην παραμετροποίηση του νέου αλγοριθμικού μοντέλου ως προς τη παράμετρο των ακραίων τιμών r. Η παράμετρος r είναι αρκετά σημαντική

καθώς βοηθάει στη μείωση των παρεμβολών που δημιουργούν οι ακραίες τιμές και η εφαρμογή του ως μέρος του αλγορίθμου θα ενίσχυε σημαντικά την απόδοση του συγκεκριμένου μοντέλου. Οπότε, υπάρχει η δυνατότητα υλοποίησης και εφαρμογής ενός αλγορίθμου ο οποίος θα επεκτείνει τον αλγόριθμο UOE-DBSCAN και θα χρησιμοποιεί την παράμετρο r με παρόμοιο τρόπο όπως χρησιμοποιείται στον αλγόριθμο PADC. Ταυτόχρονα, θα μπορούσε να αποτελέσει αντικείμενο μελλοντικής έρευνας η επέκταση του νέου αλγοριθμικού μοντέλου χρησιμοποιώντας διαφορετικό τρόπο υπολογισμού της απόστασης. Θα μπορούσε να χρησιμοποιηθεί η Σχετική απόσταση, όπως εφαρμόστηκε στον αλγόριθμο PADC, ή κάποιο άλλου είδους απόστασης αντί της Ευκλείδειας που χρησιμοποιείται στον νέο αλγόριθμο. Επίσης, ο νέος αλγόριθμος UOE-DBSCAN αναλύθηκε και εξετάστηκε σε δεδομένα που περιέχουν πραγματικές και ακέραιες τιμές. Σημαντική μελλοντική εργασία μπορεί να αποτελέσει η μελέτη της συμπεριφοράς του νέου αλγορίθμου σε δεδομένα με αλφαριθμητικές τιμές. Τέλος, όπως προαναφέρθηκε, υπάρχει ένα μειονέκτημα του DBSCAN, το οποίο, βέβαια, αποτελεί πρόβλημα και άλλων αλγορίθμων, και ονομάζεται ‘κατάρα των διαστάσεων’ και αφορά τα προβλήματα που δημιουργούνται όταν τα δεδομένα που διαχειρίζεται ο εκάστοτε αλγόριθμος είναι μεγάλων διαστάσεων. Έτσι, το νέο αλγοριθμικό σχήμα θα μπορούσε να βελτιστοποιηθεί ως προς αυτήν την πτυχή, με συνέπεια τα αποτελέσματα που θα δημιουργούνται να είναι ακριβείς, ασφαλείς και ποιοτικά και να μην επηρεάζονται από τις διαστάσεις των δεδομένων.

8. Ακρωνύμια

AOL	America Online
BLE	BlueTooth Low-Energy
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DDoS	Distributed Denial of Service
GPS	Global Positioning System
IMDb	Internet Movie Database
IL	Information Loss
IoT	Internet of Things
IP	Internet Protocol
LoRaWAN	Long Range Wide-Area Network
M2M	Machine to Machine
PADC	Privacy and Availability Data Clustering
RFID	Radio Frequency Identification
UCI	University of California, Irvine
UNB	Ultra Narrow Band
UOE-DBSCAN	Utilized Outlier Effect- Density Based Spatial Clustering of Applications with Noise
USB	Universal Serial Bus
WANs	Wide Area Networks
T.K.	Ταχυδρομικός Κώδικας

9. Ευρετήριο εικόνων

10. Ευρετήριο πινάκων

11. Βιβλιογραφία

1: Penerio, The History of the Internet of Things, 2019, [online] Available:https://perenio.com/blog/the-history-of-the-internet-of-things

2: Sachin A. Goswami , Bhargav P. Padhya , Ketan D. Patel, Internet of Things: Applications, Challenges and Research Issues, 2019

3: A. M. Ortiz, D. Hussein, S. Park, The cluster between internet of things and social networks: Review and research challenges, 2014

4: J. Lin, W. Yu, N. Zhang, A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications, 2017

5: Zhihong Yang ; Yingzhao Yue ; Yu Yang ; Yufeng Peng ; Xiaobo Wang ; Wenji Liu, Study and Application on the Architecture and KeyTechnologies for IOT , 2011

6: Slubne-Suknie, IOT ARCHITECTURE LAYERS, , [online] Available:https://slubne-suknie.info/?n=iot+architecture+layers

7: AVSystem, What Technologies are Used in IoT – Technology Behind Internet of Things, May 04, 2020, [online] Available:www.avsystem.com/blog/iot-technology/

8: Steven Lerner, 12 IoT Security ChallengesAnd How to Address Them in the Enterprise, 2019, [online] Available:https://www.enterprisedigi.com/iot/articles/iot-security-challenges

9: Panagiotis I.Radoglou Grammatikis, Panagiotis G.Sarigiannidis, Ioannis D.Moscholios, Securing the Internet of Things: Challenges, threats and solutions, March 2019

10: Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou , Jeffrey Voas, DDoS in the IoT: Mirai and Other Botnets, 2017

11: Avast, What is Malware?, 2020, [online] Available:https://www.avast.com/c-malware

12: Yavuz Canbay, Yilmaz Vural , Seref Sagiroglu, Privacy Preserving Big Data Publishing, 2018

13: Michael Arrington, AOL Proudly Releases Massive Amounts of Private Data, 2006, [online] Available:https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data

14: SABRINA DE CAPITANI DI VIMERCATI, SARA FORESTI, GIOVANNI LIVRAGA and PIERANGELA SAMARATI, DATA PRIVACY: DEFINITIONS AND TECHNIQUES, 2012

15: Rashad Saeed , Azhar Rauf, Anatomization through Generalization (AG): AHybrid Privacy-Preserving Approach to PreventMembership, Identity and Semantic SimilarityDisclosure Attacks, 2018

16: B. Sreevidya, M. Rajesh,T. Sasikala, Performance Analysis of VariousAnonymization Techniques for PrivacyPreservation of Sensitive Data, 2018

17: Sang Ni, Mengbo Xie, Quan Qian, Clustering Based K-anonymity Algorithm for Privacy Preservation, Revised 2017

18: Y. Xu, T. Ma, M. Tang, and W. Tian, A survey of privacy preservingdata publishing using generalization and suppression, 2014

19: L. Sweeney, Achieving k-anonymity privacy protection usinggeneralization and suppression, 2002

20: Sung-Bong Jang, A Study of Performance Enhancement in Big DataAnonymization , 2017

21: A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, 2007

22: N. Li, T. Li, and S. Venkatasubramanian, t-closeness: Privacy beyondk-anonymity and l-diversity, 2007

23: T. Dalenius, Towards a methodology for statistical disclosure control, 1977

24: Qianqian Meng, Lijuan Zhou, Research on Differential Privacy Preserving Clustering AlgorithmBased on Spark Platform , 2017

25: Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pol Mac Aonghusa, (k,)-Anonymity :k-Anonymity with -Differential Privacy, 2017

26: C. Dwork, Differential privacy, 2006

27: C. Dwork, Differential privacy: A survey of results, 2008

28: Y. Li, Z. Hao, W. Wen, Research on differential privacy preserving K-means clustering, Mar.2013

29: LINA NI, CHAO LI, XIAO WANG, HONGLU JIANG AND JIGUO YU , DP-MCDBSCAN: Differential PrivacyPreserving Multi-core DBSCANClustering for Network User Data, 2018

30: C. Dwork, F. McSherry, K. Nissim, Calibrating noise to sensitivity in private data analysis, 2006

31: C.C. Aggarwal, P.S. Yu, (Eds.), Privacy-Preserving Data Mining: Models and Algorithms, 2015

32: SAURAV KAUSHIK, An Introduction to Clustering and different methods of clustering, 2016, [online] Available:https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

33: A. K. Jain and R. C. Dubes, Algorithms for clustering data, 1988

34: Sanatan Mishra, Unsupervised Learning and Data Clustering, 2017, [online] Available:https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

35: J. Macqueen, Some methods for classification and analysis of multivariate observations, 1966

36: Lloyd, Least squares quantization in pcm, 1982

37: Dong Su,Jianneng Cao,Ninghui Li,Elisa Bertino,Hongxia Jin, Differentially Private K-Means Clustering, 2016

38: UCI Machine Learning Repository , UCI Datasets, [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html.

39: B. Angelin , A. Geetha, Outlier Detection using Clustering Techniques – K-means and K-median, 2020

40: A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, 2014

41: Ester, Martin;Hans-Peter Kriegel; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.), A density-based algorithm for discovering clusters in large spatial databases with noise, 1996

42: Jinfei Liu, Jun Luo, Joshua Zhexue Huang, Li Xiong, Privacy Preserving Distributed DBSCAN Clustering, 2012

43: Li Meng'Ao , Meng Dongxue , Gu Songyuan , Liu Shufen, Research and Improvement of DBSCAN Cluster Algorithm, 2015

44: Liping Zhang, Song Deng, Shiyue Li, Analysis of Power Consumer Behavior Based on the Complementation of K-means and DBSCAN,

45: Data Novia, DBSCAN: Density-Based Clustering Essentials, 2020, [online] Available:https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/

46: Vijaya, A. S., & Bateja, R., A Review on Hierarchical Clustering Algorithms, 2017

47: R. Yogita, Dr.R. Harish, A Study of Hierarchical Clustering Algorithm, Nov 2013

48: Gowda, K. C., & Krishna, G., Agglomerative clustering using the concept of mutual nearest neighbourhood, 1978

49: Gower, J. C., & Ross, G. J., Minimum spanning trees and single linkage cluster analysis, 1969

50: Dawyndt, P., De Meyer, H., & De Baets, B., The complete linkage clustering algorithm revisited, 2005

51: Defays, D., An efficient algorithm for a complete link method, 1977

52: Murtagh, F., & Legendre, P. , Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?, 2014

53: Roux, M., A comparative study of divisive hierarchical clustering algorithms, 2015

54: Xiong, Jun Ren, Lei Chen, Zhiqiang Yao, Mingwei Lin, Dapeng Wu and Ben Niu, Enhancing privacy and availability for data clustering in intelligent electrical service of IoT, 2018