ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEB INTELLIGENCE

Web έφαρμογή για ημιαυτόματο προσδιορισμό των παραμέτρων Eps και Minpts του αλγορίθμου συσταδοποίησης DBSCAN

(A web tool for semi-automatic Eps and MinPts parameters determination for DBSCAN Clustering)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΚΙΒΙΑΔΗ ΤΖΙΩΡΑ

Επιβλέπων: Στέφανος Ουγιάρογλου
Ε.ΔΙ.Π., ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Ιούλιος, 2021

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEB INTELLIGENCE

Web έφαρμογή για ημιαυτόματο προσδιορισμό των παραμέτρων Eps και Minpts του αλγορίθμου συσταδοποίησης DBSCAN

(A web tool for semi-automatic Eps and MinPts parameters determination for DBSCAN Clustering)

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΚΙΒΙΑΔΗ ΤΖΙΩΡΑ

Επιβλέπων: Στέφανος Ουγιάρογλου
Ε.ΔΙ.Π., ΔΙ.ΠΑ.Ε.

Εγκρίθηκε απο την τριμελή εξεταστική επιτροπή στις ... Ιουλίου 2021.

*(*Υπογραφή)	*(*Υπογραφή)	*(*Υπογραφή)
.....	.....	.....
Στέφανος Ουγιάρογλου Ε.ΔΙ.Π. ΔΙ.ΠΑ.Ε.	Λεωνίδας Καραμητόπουλος Διδάκτωρ	Δημήτριος Δέρβος Καθηγητής ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Ιούλιος, 2021

(Υπογραφή)

.....
Αλκιβιάδης Τζιώρας
Σχολή Μηχανικών - Τμήμα Μηχνικών Πληροφορικής και Ηλεκτρονικών Συστημάτων
Διεθνές Πανεπιστήμιο Θεσσαλονίκης

Πρόλογος

Το πρόγραμμα μεταπτυχιακών σπουδών ”Ευφυείς Τεχνολογίες Διαδικτύου” περιλαμβάνει μαθήματα που δίνουν την ευκαιρία στους φοιτητές να κατανοήσσουν ότι οι ραγδαίες εξελίξεις στην τεχνολογία έχουν οδηγήσει στην εκρηκτική αύξηση του όγκου των δεδομένων που παράγονται και διαδίδονται πλέον με μεγάλη ευκολία. Η εξαγωγή γνώσης για χρήση απο αυτά τα μεγάλα σύνολα δεδομένων γίνεται ολοένα και σημαντικότερη για τη λήψη αποφάσεων σχεδόν σε όλους τους τομείς της κοινωνίας. Ωστόσο, η ποσότητα (όγκος) των δεδομέων, η πολυπλοκότητα (ποικιλία) τους και ο ρυθμός με τον οποίο συλλέγονται και υφίστανται επεξεργασία έχουν μεγαλώσει πάρα πολύ καθιστώντας δύσκολη την ανάλυση τους. Επομένως, υπάρχει τεράστια ανάγκη για αυτοματοποιημένα εργαλεία εξαγωγής χρήσιμης πληροφορίας απο τα μεγάλα σύνολα δεδομένων, παρά τις δυσκολίες που θέτουν το τεράστιο μέγεθος και η διαφορετικότητα τους.

Οι άνθρωποι έχουν έμφυτη την ικανότητα να τοποθετούν τα πράγματα σε κατηγορίες, δηλαδή να εκτελούν απλές εργασίες κατηγοριοποίησης όπως το φιλτράρισμα των ανεπιθύμητων (spam) μηνυμάτων ηλεκτρονικού ταχυδρομείου ή πιο εξειδικευμένες, όπως η αναγνώριση ασθενειών μέσα από εικόνες που λαμβάνονται από απεικονιστικές συσκευές. Όλες αυτές οι ενέργειες προϋποθέτουν την γνώση των κατηγοριών που θα ενταχθούν τα σύνολα δεδομένων. Αντίθετα υπάρχουν σύνολα δεδομένων όπου δεν είναι εμφανής οι αντίστοιχες κατηγορίες που ανήκει κάθε στιγμιότυπο και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέρουσα δομή δεδομένων, όπου μπορεί να ενταχθεί. Η διαδικασία αυτή ονομάζεται συσταδοποίηση. Οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα στιγμιότυπα του συνόλου δεδομένων, ώστε να ανήκουν στην ίδια συστάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.

Οι αλγόριθμοι συσταδοποίησης χρειάζονται τον καθορισμό παραμέτρων για την σωστή εκτέλεση τους. Ωστόσο, είναι πιθανόν, μια εσφαλμένη παραμετροποίηση του αλγορίθμου, να οδηγήσει σε λάθος επιλογή συστάδων από τον αλγόριθμο. Χαρακτηριστικός αλγόριθμος που απαιτεί υπολογισμό παραμέτρων είναι ο Density-based spatial clustering of applications with noise (DBSCAN). Η εύρεση των τιμών αυτών των παραμέτρων μέσω μιας web εφαρμογής αποτελεί το αντικείμενο της παρούσας εργασίας.

Περίληψη

Οι διαδικασίες συσταδοποίησης συναντιούνται σε ένα ευρύ φάσμα των ανθρώπινων δραστηριοτήτων. Με τον όρο συσταδοποίηση εννοούμε την διαδικασία εκείνη κατά την οποία ένα σύνολο από στιγμιότυπα, διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση στιγμιότυπων στην ίδια ομάδα μεταφράζεται ως ομοιότητα των στιγμιότυπων αυτών και αντίστροφα (στιγμιότυπα που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια). Η ομοιότητα ή μη, μεταξύ των στιγμιότυπων, ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα που εξετάζεται και τη μορφή των στιγμιότυπων. Για την επίτευξη του στόχου της συσταδοποίησης έχουν αναπτυχθεί αρκετοί αλγόριθμοι και έχουν ενταχθεί σε κατηγορίες με βάση τον τρόπο λειτουργίας τους. Χαρακτηριστικό παράδειγμα αλγόριθμου που συσταδοποιεί βάσει πυκνότητας είναι ο DBSCAN [1]. Ο DBSCAN για να λειτουργήσει απαιτεί τον καθορισμό δύο σημαντικών παραμέτρων. Το καθορισμό των $MinPts$ και ε. Η δυσκολία στον εντοπισμό των τιμών αυτών των παραμέτρων αποτελούν το κίνητρο της παρούσας διπλωματικής εργασίας.

Στόχος της εργασίας ήταν η ανάπτυξη ενός εργαλείου που θα απλοποιεί και θα συντομεύει την διαδικασία επιλογής των παραμέτρων *MinPts* και ε που απαιτείται για την εκτέλση του αλγορίθμου DBSCAN. Η διαδικασία εύρεσης των παραμέτρων είναι μια ευρεστική διαδικασία που στηρίζεται στην παραγωγή των k-dist graphs. Παρόλου που τα γραφήματα αυτά βοηθάνε στον προσδιορισμό των παραμέτρων, οι τιμές ε και *MinPts* υπολογίζονται πάντα με σχετική ακρίβεια καθώς θα πρέπει να εντοπιστεί το σημείου του κατωφλίου από το παραγόμενο k-dist graph. Επιπρόσθετος στόχος ήταν η χρήση μιάς μεθόδου που να μπορεί να εντοπίσει το ακριβές σημείου του κατωφλίου.

Η παρούσα διπλωματική εργασία παρουσιάζει μια διαδικτυακή εφαρμογή που δίνει την δυνατότητα στον χρήστη να ανεβάσει το σύνολο δεδομένων, για το οποίο επιθυμεί να λάβει τις τιμές των παραμέτρων για την εκτέλεση του αλγορίθμου DBSCAN. Επιλέγει τιμή ή ένα εύρος τιμών *MinPts* για το οποίο επιθυμεί την δημιουργία k-dist graphs και παράλληλα υπολογίζεται με ακρίβεια το κατώφλι δηλαδή η τιμή ε.

Λέξεις Κλειδιά:Συσταδοποίηση, DBSCAN, k-dist graph, Υπολογισμός ε, Υπολογισμός minpts, Εφαρμογή διαδικτύου

Abstract

Clustering processes are encountered in a wide range of human activities. By clustering we mean the process by which a set of snapshots is separated into a set of logical groups. Entering snapshots in the same group translates as similarity of these snapshots and vice versa (snapshots belonging to different groups are dissimilar). The similarity or not between the snapshots essentially depends on the specific problem and the format of the snapshots. To achieve the goal of clustering, several algorithms have been developed and have been categorized based on how they work. A typical example of a density aggregation algorithm is DBSCAN. DBSCAN to operate requires the definition of two important parameters. The definition of $MinPts$ and e. These observations are the motivation for this dissertation.

The aim of this work was to develop a tool that will simplify and shorten the process of selecting the parameters *MinPts* and *e* required to execute the DBSCAN algorithm. The parameter finding process is an inventive process based on the production of k-dist graphs. Although these graphs help to determine the parameters, the value *e* always calculated with relative accuracy as the user would have to locate the threshold point from the generated k-dist graph. An additional goal was to use a method that could detect the exact threshold point of an exact value for *e*. This dissertation presented an online application that enables the user to upload the data set, for which he wishes to obtain the values of the parameters for the execution of DBSCAN. Selects a value or a range of values *MinPts* for which he wants to create k-dist graphs and at the same time accurately calculates the threshold, for the *e* value .

Keywords:«Clustering, k-dist graph, Calculation of ε, DBSCAN clustering parameters»

Ευχαριστίες

Θερμά ευχαριστώ τον επιβλέποντα της διπλωματικής εργασίας, μέλος ΕΔΙΠ του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, κύριο Ουγιάρογλου Στέφανο , για την συνδρομή του, στην εκπόνηση της συγκεκριμένης διπλωματικής εργασίας, την καθοριστική καθοδήγηση σε κρίσιμες περιόδους και τη διάθεση του χρόνου τους.Οφείλω να ομολογήσω οτι η συνδρομή του στην σύλληψη της ιδέας και ολοκλήρωση της διπλωματικής εργασίας ήταν καθοριστική.

Εξίσου θερμά ευχαριστώ τα μέλη της εξεταστικής επιτροπής, τον Καθηγητή του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, κύριο Δέρβο Δημήτριο και τον Διδάκτωρα, κύριο Καραμητόπουλο Λεωνίδα για τις υποδείξεις τους, οι οποίες συνέβαλαν στη βελτίωση της Διπλωματικής εργασίας.

Τέλος, θερμά ευχαριστώ την γυναίκα μου Παντελίτσα και την κόρη μου Αγγελική, για την αμέριστη συμπαράσταση και υπομονή που επέδειξαν κατά την διάρκεια της εκπόνησης της διπλωματικής εργασίας. Με την συμπαράστασή τους, έμεινα επικεντρωμένος στον στόχο μου και τελικά ολοκλήρωσα τη διπλωματική εργασία μου.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Συσταδοποίηση

Συσταδοποίηση είναι ο όρος που αναφέρεται στην ομαδοποίηση στιγμιότυπων ενός συνόλου δεδομένων σε σημαντικές υποκατηγορίες που ονομάζονται συστάδες. Η συσταδοποίηση δεδομένων είναι μια κλασική τεχνική εξόρυξης δεδομένων που βασίζεται στη μηχανική μάθηση και χωρίζει δεδομένα σε συστάδες παρόμοιων στιγμιότυπων. Σε γενικές γραμμές, η συσταδοποίηση μπορεί να οριστεί ως ένα πρόβλημα βελτιστοποίησης πολλαπλών στιγμιότυπων, καθώς μπορεί να επιτευχθεί με διάφορους αλγόριθμους, περισσότεροι από 100 αλγόριθμοι συσταδοποίησης έχουν προταθεί τις τελευταίες δεκαετίες, φυσικά έχουν αξιοσημείωτες διαφορές στην κατανόηση του τι συνιστά μια συστάδα. Αυτές οι διαφορές επηρεάζουν αναπόφευκτα τον τρόπο λειτουργίας κάθε αλγορίθμου ως προς την αποτελεσματική ανίχνευση συστάδων για ένα δεδομένο σύνολο δεδομένων. Στην πραγματικότητα, ο κατάλληλος αλγόριθμος ομαδοποίησης καθώς και οι ρυθμίσεις παραμέτρων που πιθανότατα θα απαιτήσει ενδέχεται να διαφέρουν ανάλογα με το πεδίο εφαρμογής, τον τύπο δεδομένων και την χρήση των αποτελεσμάτων. Ως επαναληπτική διαδικασία ανακάλυψης γνώσης, συνήθως απαιτεί εργασίες δοκιμής και σφάλματος (trial and error), ειδικά όταν εμπλέκονται παραμετρικοί αλγόριθμοι. Στην ορολογία της μηχανικής μάθησης, η συσταδοποίηση είναι μια μέθοδος μη εποπτευόμενης μάθησης, που χρησιμοποιείται συνήθως για διερευνητική εξόρυξη δεδομένων ή για ανάλυση στατιστικών δεδομένων. Άλλα πεδία στα οποία χρησιμοποιείται συχνά περιλαμβάνουν μηχανική μάθηση, αναγνώριση προτύπων, ανάλυση εικόνας, ανάκτηση πληροφοριών, βιο-πληροφορική, συμπίεση δεδομένων και γραφικά υπολογιστών.

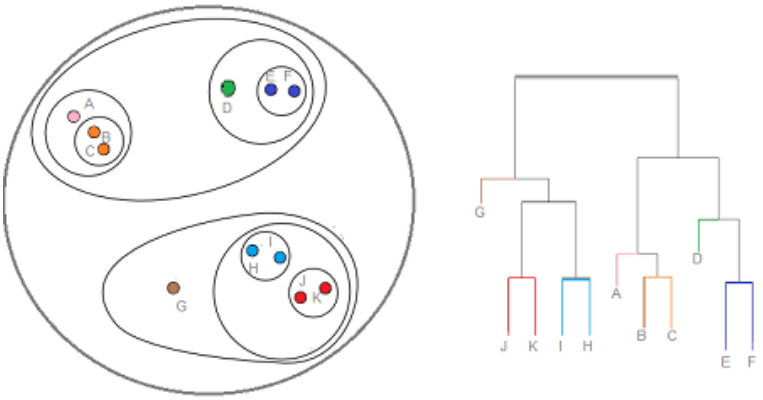
Όπως αναφέρθηκε, δεδομένου ότι η έννοια μιας συστάδας δεν είναι, ούτε και μπορεί να οριστεί ρητά, η μόνη γενίκευση στην οποία μπορούν να βασίζονται όλοι οι αλγόριθμοι συσταδοποίησης είναι ότι μια συστάδα είναι μια ομάδα στιγμιότυπων. Υπάρχουν φυσικά διαφορετικά μοντέλα συσταδοποίησης, ανάλογα με τις διαφορετικές προσεγγίσεις των ερευνητών, που στοχεύουν στον προσδιορισμό της συσταδοποίησης των στιγμιότυπων. Η κατανόηση αυτών των μοντέλων είναι ζωτικής σημασίας για την κατανόηση των διαφορετικών τύπων αλγορίθμων συσταδοποίησης και ενδεχομένως να προσδιοριστεί ποιοι είναι πιο κατάλληλοι για συγκεκριμένους τομείς και προβλήματα, αν και όπως σημειώνεται, στις περισσότερες περιπτώσεις, αυτή η απόφαση απαιτεί πειραματισμό.

1.1.1 Κατηγορίες αλγορίθμων συσταδοποίησης

Μερικά χαρακτηριστικά παραδείγματα κατηγοριών αλγορίθμων συσταδοποίησης παρατίθενται παρακάτω.

1. **Ιεραρχική συσταδοποίηση.** Η κύρια εικασία της συσταδοποίησης που βασίζεται στη συνδεσιμότητα είναι ότι τα γειτονικά στιγμιότυπα σχετίζονται περισσότερο μεταξύ τους παρά με στιγμιότυπα που βρίσκονται πιο μακριά στον χώρο δεδομένων. Οι αλγόριθμοι

αυτής της κατηγορίας [8] σχηματίζουν συστάδες συνδέοντας στιγμιότυπα λαμβάνοντας υπόψη την ενδιάμεση απόσταση τους. Σε αυτήν την περίπτωση, μια συστάδα μπορεί να περιγραφεί σε μεγάλο βαθμό από τη μέγιστη απόσταση που απαιτείται για τη σύνδεση των μερών της και διαφορετικές αποστάσεις θα οδηγήσουν σε διαφορετικές συστάδες. Αυτές οι συστάδες μπορούν αργότερα να αναπαρασταθούν σε ένα δενδρόγραμμα όπου ο άξονας y αντιπροσωπεύει τις διαφορετικές τιμές απόστασης και στον άξονα x τοποθετούνται τα στιγμιότυπα. Μέσω αυτής της αναπαράστασης ενός συνόλου δεδομένων, οι αλγόριθμοι δεν σχηματίζουν απλώς διακριτές συστάδες, αλλά παρέχουν επίσης μια λεπτομερή οπτική ιεραρχία, καθώς μπορούμε να δούμε ποιες από αυτές τις συστάδες μπορούν να συγχωνευτούν και για ποια τιμή απόστασης μπορεί να συμβαίνει αυτό.

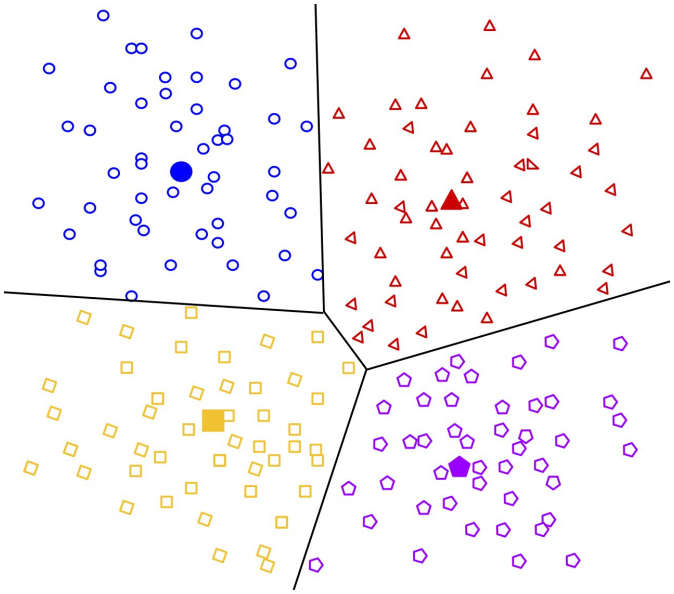


Σχήμα 1.1: Παράδειγμα ιεραρχικής συσταδοποίησης.

Οι αλγόριθμοι συσταδοποίησης με βάση τη συνδεσιμότητα μπορούν να ταξινομηθούν ανάλογα με τον τρόπο υπολογισμού των αποστάσεων, το κριτήριο σύνδεσης των στιγμιοτύπων (ελάχιστη, μέγιστη ή μέση απόσταση αποστάσεων) ή τον τρόπο με τον οποίο αρχίζουν να σχηματίζουν την ιεραρχία των συστάδων, είτε προσθέτοντας ένα-ένα τα στοιχεία ή διαιρώντας το σύνολο δεδομένων σε κατατμήσεις. Πρέπει να σημειωθεί ότι μπορούν εύκολα να παραπλανηθούν από ακραία στιγμιότυπα, δεδομένου ότι αυτά τα αντικείμενα ενδέχεται είτε να εμφανίζονται σε έναν ιεραρχικό αλγόριθμο ως πρόσθετες συστάδες είτε να προκαλούν τη συγχώνευση άλλων συστάδων, το οποίο είναι γνωστό ως φαινόμενο αλυσίδας.

2. **Διαμεριστική συσταδοποίηση.** Σε αυτήν την κατηγορία αλγορίθμων, υποθέτουμε ότι για έναν προκαθορισμένο αριθμό k συστάδων, κάθε στιγμιότυπο p του συνόλου δεδομένων αντιστοιχεί στη συστάδα του οποίου το μέσο ή κεντρικό στιγμιότυπο –ή μόνο κεντροειδές– βρίσκεται πιο κοντά στο p, με τρόπο που οι τετραγωνικές αποστάσεις από τη συστάδα ελαχιστοποιούνται. Με αυτόν τον τρόπο, κάθε συστάδα αντιπροσωπεύεται από το μέσο στιγμιότυπο, που χρησιμεύει ως πρωτότυπο χωρίς απαραίτητα να είναι ένα πραγματικό στιγμιότυπο του συνόλου δεδομένων, οπότε έχουμε όλες τις συστάδες με τη μορφή ενός κεντρικού διανύσματος. Χαρακτηριστικός εκπρόσωπος αυτής της κατηγορίας αλγορίθμων είναι ο k-means [9].

   Υπολογιστικά, αυτό το πρόβλημα βελτιστοποίησης θεωρείται δύσκολο και επομένως προτιμάται η αναζήτηση μόνο για κατά προσέγγιση λύσεις. Προκειμένου να βρεθεί ένα



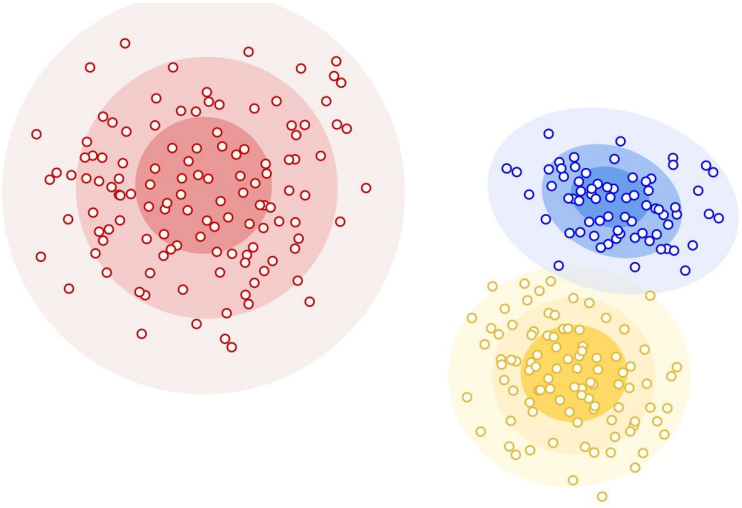
Σχήμα 1.2: Παράδειγμα διαμεριστικής συσταδοποίησης. [2]

ικανοποιητικό τοπικό βέλτιστο, απαιτούνται πολλές πειραματικές εκτελέσεις με τυχαίες αρχικές μεταβλητές. Ωστόσο, ορισμένες παραλλαγές του k-means είναι σε θέση να επιλέξουν τις καλύτερες αρχικές μεταβλητές μεταξύ πολλαπλών πειραματικών εκτελέσεων. Άλλες εναλλακτικές λύσεις περιορίζουν την επιλογή των κέντρων των συστάδων μεταξύ των υπαρχόντων στιγμιότυπων του συνόλου δεδομένων, ενώ άλλες επικεντρώνονται στην επιλογή των κέντρων κατά την πρώτη εκτέλεση με τυχαία επιλογή. Στις περισσότερες περιπτώσεις αυτών των αλγορίθμων, ο αριθμός των συστάδων πρέπει να καθοριστεί εκ των προτέρων, πράγμα που φυσικά αποτελεί μειονέκτημα. Επίσης, δεδομένου ότι εστιάζουν και εργάζονται γύρω από τα κέντρα των συστάδων, πιθανότατα ανιχνεύουν συστάδες παρόμοιου μεγέθους, αγνοώντας συχνά τα όρια των συστάδων.

Αυτή η μέθοδος είναι γενικά δημοφιλής στη μηχανική μάθηση λόγω των εννοιολογικών ομοιοτήτων που παρουσιάζει με τους αλγορίθμους ταξινόμησης εγγύτερων γειτόνων. Μια άλλη ενδιαφέρουσα χρήση του k-means, εκτός από την συσταδοποίηση, είναι ο τρόπος που μπορεί να χρησιμοποιηθεί ως τεχνική μείωσης δεδομένων με ανοχή στον θόρυβο, όπως αποδείχθηκε πειραματικά σε έρευνα [10].

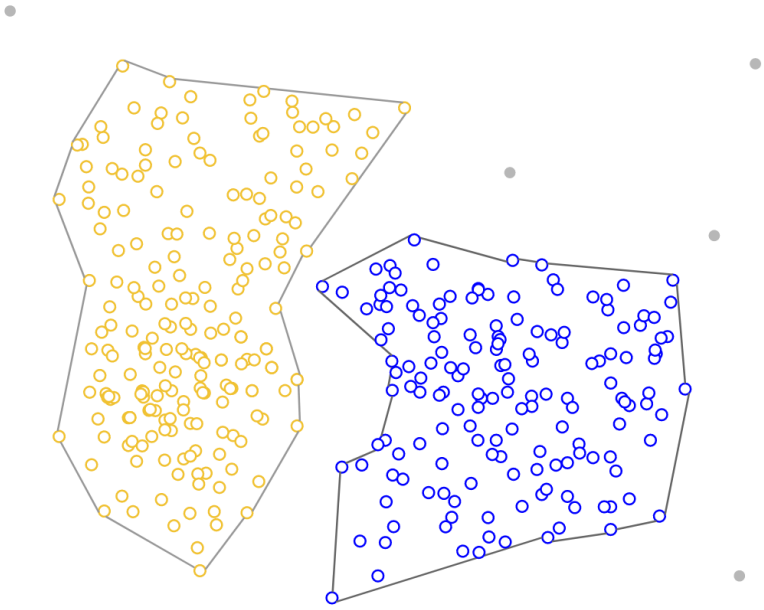
3. **Βάσει κατανομής ή μοντέλων.** Οι μέθοδοι χωρίζονται σε δύο κατηγορίες σε αυτές που προσεγγίζουν το πρόβλημα μέσω στατιστικής μοντελοποίησης [11] και σε αυτές που χρησιμοποιούν νευρωνικά δίκτυα [12]. Η ομαδοποίηση βάσει κατανομής σχετίζεται άμεσα με τη χρήση στατιστικών μοντέλων κατανομής (π.χ. Gaussian / Normal). Βρίσκουν το πλήθος των συστάδων και τον θόρυβο με βάση το μοντέλου που έχει επιλεχθεί. Αυτοί οι μέθοδοι είναι πολύ ισχυρές αφού μπορούν να εντοπίσουν τόσο το θόρυβο όσο και τις ακραίες τιμές με πολύ μεγάλη ακρίβεια. Η αδυναμία τους είναι ότι δύσκολα μπορεί να επιλεχθεί η πολυπλοκότητα του μοντέλου ή το σύνολο των περιορισμών που πρέπει να εφαρμοστούν. Σε περίπτωση χαμηλής ή μηδενικής πολυπλοκότητας, είναι εύκολο για αυτές τις μεθόδους να αποτύχουν λόγω υπερβολικής εφαρμογής. Όσο αφορά τα νευρωνικά δίκτυα διαθέτουν ιδιότητες ικανές στο να λύνουν προβλήματα συσταδοποίησης

και επειδή μπορεί να γίνεται η εκπαίδευση των νευρώνων με παράλληλους αλγορίθμους έχουν οδηγήσει στην επικράτηση των νευρωνικών δικτύων έναντι των στατιστικών μεθόδων.



Σχήμα 1.3: Παράδειγμα συσταδοποίησης βάσει κατανομής. [2]

4. **Βάσει πυκνότητας.** Η ομαδοποίηση που βασίζεται στη πυκνότητα δημιουργεί συστάδες με βάσει την υψηλή συγκέντρωση δεδομένων ή υψηλότερη από κάθε περιβάλλουσα περιοχή. Οι περιοχές που εμφανίζουν χαλαρή πυκνότητα στο σύνολο δεδομένων μπορεί να περιλαμβάνουν είτε θόρυβο είτε ακραίες τιμές, οπότε αντιμετωπίζονται αμφότερες με τον ίδιο τρόπο, αγνοούνται. Αντιπροσωπευτικό παράδειγμα αυτής της κατηγορίας είναι ο αλγόριθμος συσταδοποίησης DBSCAN [13], Ο οποίος είναι και το αντικείμενο μελέτης της παρούσας εργασίας. Περισσότερες λεπτομέρειες θα παρουσιαστούν παρακάτω καθώς ο καθορισμός παραμέτρων του DBSCAN για τον προσδιορισμό των συστάδων, πραγματοποιείται μετά από εργασίες δοκιμής και σφάλματος. Αυτή η δυσκολία προσδιορισμού των παραμέτρων, αποτέλεσε το κίνητρο για την εκπόνηση αυτής της διπλωματικής εργασίας.



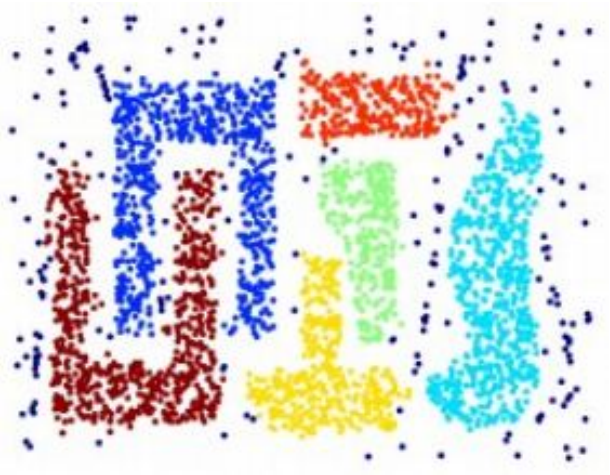
Σχήμα 1.4: Παράδειγμα συσταδοποίησης βάσει πυκνότηταςς. [2]

Συμπερασματικά, δεδομένου ενός συνόλου στοιχείων δεδομένων, χρησιμοποιούνται αλγόριθμοι συσταδοποίησης για την εκχώρηση καθενός από αυτά τα στοιχεία σε μια συγκεκριμένη συστάδα, ανάλογα με τις ιδιότητές τους. Θεωρητικά, στιγμιότυπα που έχουν παρόμοιες ιδιότητες και χαρακτηριστικά βρίσκονται φυσικά στην ίδια ή σε γειτονικές περιοχές στο χώρο δεδομένων και επομένως πιθανώς ανήκουν στην ίδια συστάδα, ενώ στοιχεία διαφορετικών ιδιοτήτων ανήκουν σε άλλες διακριτές περιοχές. Η απόσταση μεταξύ των διαφόρων ομάδων βασίζεται στη διαφορά μεταξύ των χαρακτηριστικών των στιγμιότυπων.

1.2 Αλγόριθμος συσταδοποίησης DBSCAN

Προκειμένου να προχωρήσουμε στην περιγραφή του αλγορίθμου συσταδοποίησης DBSCAN και του τρόπου με τον οποίο διαιρεί τα στοιχεία σύμφωνα με τη θέση τους κατά τη διαμόρφωση και την επέκταση των συστάδων, είναι σημαντικό να αναφερθούν ορισμένες σημαντικές έννοιας των συστάδων που βασίζονται στην πυκνότητα.

Στην περίπτωση ενός αλγορίθμου που λειτουργεί βάσει πυκνότητας, τα στοιχεία δεδομένων δεν θεωρούνται πλέον ως συγκεκριμένα στιγμιότυπα αλλά ως συστατικά που γεμίζουν το χώρο δεδομένων. Περιοχές μεγάλης ποσότητας στιγμιότυπων χαρακτηρίζονται ως περιοχές υψηλής πυκνότητας, ενώ σε αντίθετη περίπτωση αναφέρεται σε περιοχές χαμηλής πυκνότητας. Οι περιοχές υψηλής πυκνότητας, ή γενικά, οι περιοχές πυκνότητας που υπερβαίνουν την πυκνότητα του περιβάλλοντος χώρου τους, μπορούν εύκολα να αναγνωριστούν ως συστάδες. Στοιχεία περιοχών που βρέθηκαν εκτός των ορίων οποιασδήποτε συστάδας, τα οποία φυσικά έχουν πολύ χαμηλότερη πυκνότητα, μπορεί να είναι θόρυβος ή ακραίες τιμές. Προκειμένου να οριστεί ένα όριο στη διαφορά πυκνότητας που διαχωρίζει μια συστάδα από το περιβάλλον της, υποθέτουμε ότι η γειτονιά κάθε στιγμιότυπου μιας συστάδας για μια δεδομένη ακτίνα, περιέχει τουλάχιστον



Σχήμα 1.5: Παράδειγμα συνόλου δεδομένων με ακανόνιστες ή αλληλένδετες συστάδες - με βάση την πυκνότητά τους. [3]

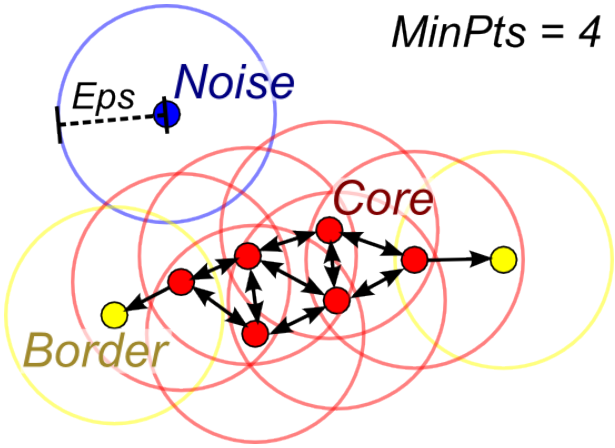
έναν ελάχιστο αριθμό στιγμιότυπων. Το σχήμα της γειτονιάς εξαρτάται σε μεγάλο βαθμό από τη μέτρηση που χρησιμοποιείται για τη συνάρτηση απόστασης μεταξύ δύο στιγμιότυπων p και q, που υποδηλώνεται με dist (p, q), αν και αυτός ο συγκεκριμένος αλγόριθμος δεν επιβάλλει τη χρήση μιας συγκεκριμένης μέτρησης. Για παράδειγμα, η χρήση της απόστασης Μανχάταν οδηγεί σε ορθογώνιες γειτονιές, ενώ η Ευκλείδεια απόσταση, η οποία χρησιμοποιείται συχνά για λόγους απλότητας, οδηγεί σε σφαιρικές γειτονιές. Σε γενικές γραμμές, η επιλογή της κατάλληλης μεθόδου υπολογισμού της απόστασης ποικίλλει μεταξύ διαφορετικών τομέων εφαρμογής και ανάλογα με την φύση των δεδομένων.

1.2.1 Βασικοί ορισμοί

Παρακάτω αναφέρονται βασικοί ορισμοί και έννοιες των συστάδων που βασίζονται στην πυκνότητα.

Ορισμός 1 (Eps-γειτονιά ενός στιγμιότυπου) *Η Eps-γειτονιά ενός στιγμιότυπου p, συμβολίζεται με* $N_{Eps}(p)$*, ορίζεται από* $N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$

Να σημειωθεί ότι δεν έχουν όλα τα στιγμιότυπα μιας συστάδας την ίδιο αριθμό γειτόνων. Στην πραγματικότητα, τα στιγμιότυπα που τοποθετούνται στις εσωτερικές περιοχές της συστάδας, γνωστά ως **βασικά στιγμιότυπα (core points)**, έχουν λογικά πολύ μεγαλύτερο αριθμό γειτόνων σε σύγκριση με τα στιγμιότυπα που βρίσκονται κοντά στα όρια της συστάδας, που αναφέρονται ως **οριακά στιγμιότυπα (border points)**. Για να προσδιορίσουμε πού βρίσκονται τα όρια μιας συστάδας, πρέπει να καθορίσουμε τον ελάχιστο αριθμό στιγμιότυπων που πρέπει να έχει το $N_{Eps}(p)$ έτσι ώστε το p να μπορεί να χαρακτηριστεί ως βασικό στιγμιότυπο (core). Πρόσθετα στιγμιότυπα που ανήκουν στη γειτονιά ενός βασικού στιγμιότυπου, χωρίς να είναι βασικά στιγμιότυπα, εξακολουθούν να αποτελούν μέρος της συστάδας, αλλά χαρακτηρίζονται ως οριακά στιγμιότυπα. Οι παραπάνω τύποι στιγμιότυπων απεικονίζονται στο σχήμα 1.6.



Σχήμα 1.6: Τύποι στιγμιότυπων [4]

Αυτό περιγράφεται καλύτερα στους ακόλουθους ορισμούς.

Ορισμός 2 (Directly Density Reachable) *Ένα στιγμιότυπο p είναι directly density reachable από ένα στιγμιότυπο q με Eps, MinPts εαν*

i $p \in N_{Eps}(q)$

ii $|N_{Eps}(q)| \geq MinPts$ *(core point condition)*

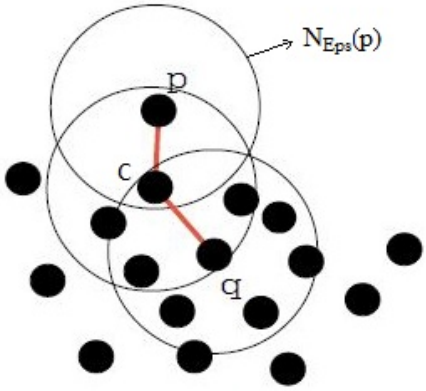
Directly density-reachable είναι συμμετρικό μόνο για ζεύγη βασικών στιγμιότυπων, το ίδιο δεν ισχύει για ένα βασικό στιγμιότυπο και ένα οριακό στιγμιότυπο όπως φαίνεται παρακάτω στο σχήμα 1.7.

Ορισμός 3 (Density-reachable) *Ένα στιγμιότυπο p είναι density-reachable από ένα στιγμιότυπο q με Eps και MinPts αν υπάρχει μία αλυσίδα από στιγμιότυπα* $p_1, ..., p_n, p_1 = q, p_n = p$ *όπου το* $p_i + 1$ *είναι directly density reachable από το pi.*

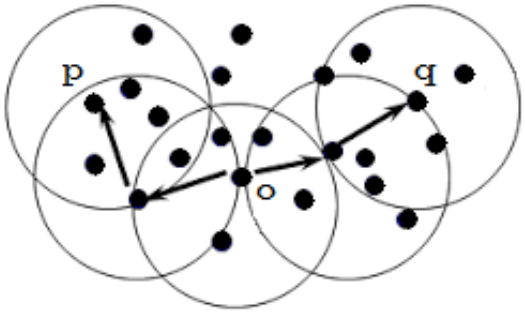
Η Density-reachable είναι μία επέκταση του directly density-reachable και όπου ισχύει η μεταβατική ιδιότητα αλλά όχι η συμμετρική. Γενικά αυτή η σχέση δεν ισχύει για όλα τα στιγμιότυπα αλλά ισχύει πάντα για τα βασικά στιγμιότυπα. Δύο οριακά στιγμιότυπα της ίδιας συστάδας μπορούν να μην είναι density reachable μεταξύ τους επειδή η συνθήκη του βασικού σημείου μπορεί να ισχύει και στα δύο. Όμως, πρέπει να υπάρχει και ένα βασικό στιγμιότυπο το οποίο να είναι density-reachable και για τα δύο.

Ορισμός 4 (Density-Connected) *Ένα στιγμιότυπο p είναι density connected με ένα στιγμιότυπο q με Eps, MinPts εαν υπάρχει ένα στιγμιότυπο ο τέτοιο ώστε τα δύο στιγμιότυπα p και q είναι density reachable από το στιγμιότυπο με Eps, MinPts .*

Η χρήση των παραπάνω ορισμών διευκολύνει επίσης τον ορισμό της συστάδας και του θορύβου.



Σχήμα 1.7: Για $MinPts = 6$, το στιγμιότυπο p είναι directly density reachable από το c και density reachable από το q. [5]



Σχήμα 1.8: Το στιγμιότυπο p είναι density connected με το στιγμιότυπο q μέσω του o

Ορισμός 5 (Συστάδα) *Έστω D ένα σύνολο από στιγμιότυπα. Μία συστάδα C με Eps και MinPts είναι ένα μη κενό υπο-σύνολο του D στο όποιο ισχύουν οι πιο κάτω συνθήκες:*

i $\forall p, q$: *αν* $p \in C$ *και q είναι density-reachable από το p με Eps και MinPts, τότε το* $q \in C$.

ii $\forall p, q \in C$: *p είναι density-connected με το q, με Eps και MinPts.*

Ορισμός 6 (Θόρυβος) *Έστω* $C_1, ..., C_k$ *είναι οι συστάδες του συνόλου D με παραμέτρους* Eps_i *and* $MinPts_i$, *i=1, ..., k. Τότε ως θόρυβος (σχήμα 1.6) ορίζεται το σύνολο των στιγμιότυπων τα οποία δεν ανήκουν σε καμία συστάδα* C_i, *δηλαδή* $= \{p \in D | \forall i : p \notin C_i\}$

1.2.2 Πλεονεκτήματα-μειονεκτήματα DBSCAN

Τα πλεονεκτήματα που προσφέρει η χρήση του αλγορίθμου DBSCAN είναι η δυνατότητα εντοπισμού θορύβου (outliers) σε ένα σύνολο δεδομένων. Μπορεί να εντοπίσει αυθείρετες συστάδες.

Είναι πάρα πολύ γρήγορος καθώς μπορεί να σαρώσει ολόκληρο το σύνολο δεδομένων σε μία επανάληψη και τέλος δεν χρειάζεται να προσδιοριστεί εκ των προτέρων ο αριθμός των συστάδων.

Στα μειονεκτήματα του αλγορίθμου ανήκει η αδυναμία του να δημιουργήσει καλές συστάδες, εάν τα δεδομένα έχουν διαφορετικές πυκνότητες. Δεν είναι κατάλληλος για πολλαπλών διαστάσεων δεδομένα. Έχει κάποιες δυσκολίες στη διάκριση των διαχωρισμένων συστάδων εάν βρίσκονται πολύ κοντά μεταξύ τους, παρόλο που έχουν διαφορετικές πυκνότητες. Τέλος απαιτεί από τον χρήστη τον καθορισμό της ακτίνας ε και τα ελάχιστα στιγμιότυπα *MinPts* της γειτονιάς.

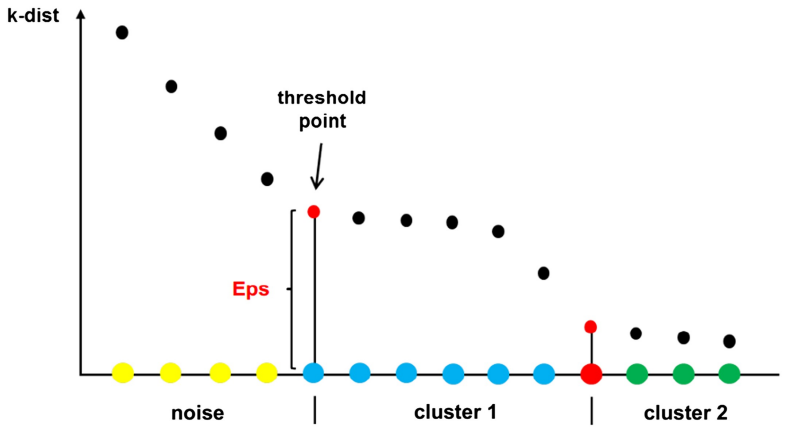
1.3 Καθορισμός παραμέτρων Eps και MinPts

Ο προσδιορισμός των παραμέτρων Eps και MinPts, γίνεται με μια ευρετική (heurestic) μέθοδο βασισμένη στην ακόλουθη παρατήρηση. Υποθέτοντας ότι έχουμε απόσταση d μεταξύ ενός στιγμιότυπου p και του εγγύτερου γείτονά του, η d-γειτονιά του p περιέχει ακριβώς k + 1 στιγμιότυπα για σχεδόν όλα τα στιγμιότυπα p. Η μόνη περίπτωση στην οποία η γειτονιά d συμπληρώνεται με περισσότερα από k + 1 στιγμιότυπα είναι εκείνη στην οποία αρκετά στιγμιότυπα της d-γειτονιάς έχουν ακριβώς την ίδια απόσταση d από p, η οποία είναι εξαιρετικά σπάνια. Επίσης, η αλλαγή k σε ένα στιγμιότυπο δεν επιφέρει αξιοσημείωτες αλλαγές στο d. Αυτό θα συνέβαινε μόνο εάν οι k εγγύτεροι γείτονες του p για k = 1, 2, 3 ... βρίσκονται περίπου σε ευθεία γραμμή, κάτι που γενικά είναι πολύ απίθανο για στιγμιότυπο σε μια συστάδα.

Για ένα δεδομένο k μπορούμε να ορίσουμε μια συνάρτηση $k - dist$ η οποία θα χρησιμοποιηθεί για τη χαρτογράφηση της απόστασης κάθε στιγμιότυπου από τον εγγύτερο γείτονά του. Η ταξινόμηση των στιγμιότυπων του συνόλου δεδομένων με φθίνουσα σειρά των k-dist τιμών τους, το γράφημα αυτής της συνάρτησης, γνωστό ως $k - distgraph$ μπορεί να αποκαλύψει κάποιες πληροφορίεσές σχετικά με την κατανομή πυκνότητας στο σύνολο δεδομένων. Εαν επιλέχθεί ένα αυθαίρετο σημείο p, οριστεί η παράμετρος Eps ίση με k-dist (p) και η παράμετρος MinPts ίση με k, όλα τα στιγμιότυπα με ίση ή μικρότερη τιμή k-dist θα είναι βασικά στιγμιότυπα [1]. Οι επιθυμητές τιμές παραμέτρων είναι εκείνες του κατωφλιού, του στιγμιότυπο με τη μέγιστη τιμή k-dist στη λιγότερο πυκνή συστάδα του συνόλου δεδομένων. Το στιγμιότυπο κατωφλίου (threshold point) είναι στην πραγματικότητα το πρώτο στιγμιότυπο της πρώτης καμπύλης στο ταξινομημένο γράφημα k-dist. Οι υψηλότερες τιμές k-dist θεωρούνται θόρυβος, ενώ όλα τα άλλα στιγμιότυπα, τοποθετημένα στα δεξιά του κατωφλίου, έχουν ήδη αντιστοιχιστεί σε κάποια συστάδα (σχήμα 1.9).

1.4 Αλγόριθμος συσταδοποίησης OPTICS

Ο OPTICS είναι ένας αλγόριθμος για την εύρεση συστάδων βάσει πυκνότητας σε ένα σύνολο δεδομένων. Η βασική ιδέα του αλγορίθμου είναι παρόμοια με του DBSCAN, αλλά έρχεται να αντιμετωπίσει μία από τις σημαντικότερες αδυναμίες του DBSCAN, το πρόβλημα της ανίχνευσης σημαντικών συστάδων σε δεδομένα διαφορετικής πυκνότητας [14]. Για να το επιτύχει αυτό, τα στιγμιότυπα του συνόλου δεδομένων ταξινομούνται (γραμμικά) έτσι ώστε τα πλησιέστερα



Σχήμα 1.9: Ταξινομημένο k-dist γράφημα.

χωρικά στιγμιότυπα γίνονται γείτονες στην σειρά. Επιπλέον, αποθηκεύεται μια ειδική απόσταση για κάθε στιγμιότυπο που αντιπροσωπεύει την πυκνότητα που πρέπει να γίνει αποδεκτή για μια συστάδα έτσι ώστε και τα δύο στιγμιότυπα να ανήκουν στην ίδια συστάδα.

Όπως ο DBSCAN, έτσι και ο OPTICS απαιτεί δύο παραμέτρους:

i ε, που περιγράφει τη μέγιστη απόσταση (ακτίνα) που πρέπει να ληφθεί υπόψη

ii Minpts, περιγράφει τον αριθμό των στιγμιότυπων που απαιτούνται για τη δημιουργία μιας συστάδας

Ο αλγόριθμος επικεντρώνεται στη σειρά με την οποία επεξεργάζονται τα στιγμιότυπα και αυτές οι πληροφορίες μπορούν αργότερα να χρησιμοποιηθούν από έναν εκτεταμένο αλγόριθμο DBSCAN για την εκχώρηση των στιγμιότυπων σε συστάδες. Για κάθε στιγμιότυπο, αυτές οι πληροφορίες μπορούν να περιγραφούν με δύο πρόσθετες τιμές, την core-distance και την reachability-distance.

Ορισμός 7 (Core distance ενός στιγμιότυπο p) *Έστω p είναι ένα στιγμιότυπο του συνόλου δεδομένων D, έστω ε είναι η τιμή της απόστασης, έστω $N_\varepsilon(p)$ είναι η ε-neighborhood του p, έστω το $MinPts$ είναι ένας φυσικός αριθμός και έστω $MinPts-distance(p)$ είναι η απόσταση του p από την MinPts γειτονιά. Τότε η $core-distance$ του p ορίζεται ως:*

$$core-distance_{\varepsilon,MinPts}(p) = \begin{cases} Εαν\ |N_\varepsilon(p)| < MinPts, & ΑΠΡΟΣΔΙΟΡΙΣΤΟ \\ αλλιώς, & MinPts\text{-}distance(p) \end{cases}$$

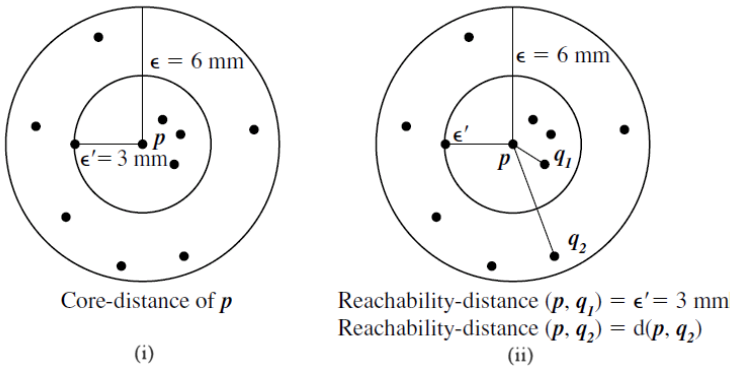
Με άλλα λόγια, η core distance του στιγμιότυπου p είναι η μικρότερη τιμή του ε' για την οποία το p είναι στιγμιότυπο πυρήνα, δοθείσας σταθερής τιμής MinPts, όπως φαίνεται στο σχήμα 1.10 (i).

Ορισμός 8 (Reachability distance στιγμιότυπου p σε σχέση με το στιγμιότυπο q) *Έστω στιγμιότυπα p και q του συνόλου δεδομένων D, έστω $N_{Eps}(q)$ είναι η ε γειτονιά του q, έστω $MinPts$*

είναι φυσικός αριθμός.Τότε, η $reachability - distance$ *του* p *σε σχέση με το* q *ορίζεται ως:*

$$reachability - distance_{e,MinPts}(p,q) = \begin{cases} \textit{Εαν } |N_{Eps}(q)| < MinPts, & \textit{ΑΠΡΟΣΔΙΟΡΙΣΤΟ} \\ \textit{Αλλιώς}, & \textit{max(core-distance(q), distance(q,p))} \end{cases}$$

Διαισθητικά, η reachability-distance μεταξύ ενός στιγμιότυπου p και q είναι το μέγιστο της core-distance p και της απόστασης μεταξύ p και q. Να σημειωθεί ότι η reachability-distance δεν καθορίζεται εάν το q δεν είναι βασικό στιγμιότυπο, όπως φαίνεται στο σχήμα 1.10 (ii).



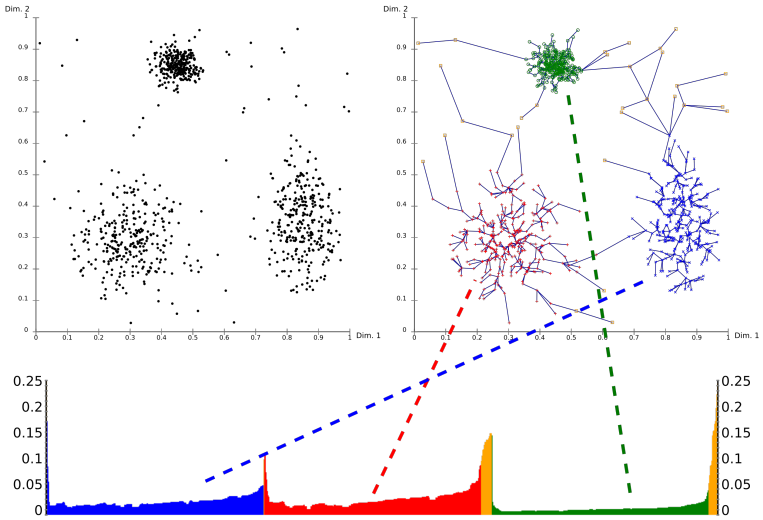
Σχήμα 1.10: Απεικόνιση core-distance και reachability-distances, για MinPts=5. [6]

1.4.1 Συσταδοποίηση με το reachability-plot

Χρησιμοποιώντας ένα reachability-plot (ένα ειδικό είδος δενδρογράμματος), η ιεραρχική δομή των συστάδων μπορεί να αποκτηθεί εύκολα. Πρόκειται για ένα διάγραμμα δύο διαστάσεων (2D), με τη σειρά των στιγμιότυπων όπως υποβάλλονται σε επεξεργασία από τον OPTICS στον άξονα x και την reachability-distance στον άξονα y. Δεδομένου ότι τα στιγμιότυπα που ανήκουν σε μια συστάδα έχουν μικρή reachability-distance από τον εγγύτερο γείτονά τους, οι συστάδες εμφανίζονται ως κοιλάδες σε ενα reachability-plot. Όσο πιο βαθιά είναι η κοιλάδα, τόσο πιο πυκνή είναι η συστάδα.

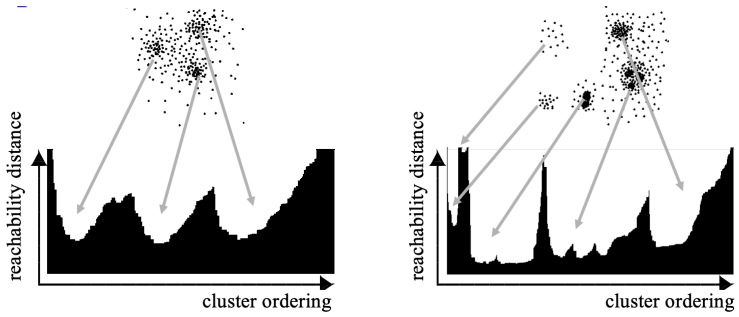
Παρατηρώντας το σχήμα 1.11 βλέπετε στην επάνω αριστερή περιοχή, εμφανίζεται ένα συνθετικό παράδειγμα συνόλου δεδομένων. Στο επάνω δεξί μέρος απεικονίζεται το δέντρο που δημιουργείτε από τον OPTICS και στο κάτω μέρος εμφανίζεται το reachability-plot όπως υπολογίζεται από τον OPTICS. Τα χρώματα σε αυτό το γράφημα είναι ετικέτες και δεν τα δημιουργεί ο αλγόριθμος, βοηθάνε στο να γίνουν οπτικά κατανοητές οι κοιλάδες που αντιστοιχούν στις συστάδες του παραπάνω σύνολο δεδομένων. Τα κίτρινα στιγμιότυπα σε αυτήν την εικόνα θεωρούνται θόρυβος και δεν υπάρχει κοιλάδα στο reachability-plot. Συνήθως δεν αντιστοιχίζονται σε συστάδες.

Η εξαγωγή συστάδων από αυτό το διάγραμμα μπορεί να γίνει χειροκίνητα επιλέγοντας ένα εύρος τιμών στον άξονα x μετά από οπτική επιθεώρηση, επιλέγοντας ένα κατώφλι στον άξονα y (το αποτέλεσμα είναι παρόμοιο με τα αποτέλεσματα του DBSCAN με τις ίδιες παραμέτρους ε και



Σχήμα 1.11: Reachability-plot και εξαγωγή συστάδων. [7]

minPts), ή με διαφορετικούς αλγόριθμους που προσπαθούν να εντοπίσουν κοιλάδες, ανίχνευση γόνατος ή τοπικά μέγιστα. Οι συστάδες που λαμβάνονται με αυτόν τον τρόπο συνήθως είναι ιεραρχικές και δεν μπορούν να επιτευχθούν με μία μόνο εκτέλεση του DBSCAN. Στο σχήμα 1.12 απεικονίζονται μερικά παραδείγματα της σχέσης μεταξύ συσταδοποίησης και reachability distance.



Σχήμα 1.12: Απεικόνιση σχέσης συστάδων και reachability-distance. [7]

1.4.2 Πλεονεκτήματα - μειονεκτήματα του OPTICS

Τα πλεονεκτήματα που προσφέρει η χρήση του αλγορίθμου OPTICS είναι ο αποτελεσματικός χειρισμός των συστάδων, ειδικά όταν τα δεδομένα έχουν ποικίλες πυκνότητες. Επίσης χάρη στον μηχανισμό ταξινόμησης που διαθέτει ανακτά στιγμιότυπα σε συγκεκριμένη σειρά. Τα μειονεκτήματα που χαρακτιρίζουν τον αλγόριθμο OPTICS είναι η μικρη ευαίσθησία του σε λανθασμένα δεδομένα καθώς και ότι απαιτεί μεγαλύτερο χρόνο εκτέλεσης.

1.5 Κίνητρο και Συνεισφορά

Οι αδυναμίες που αναφέρθηκαν παραπάνω τόσο στον αλγόριθμο DBSCAN, όσο και στον OPTICS εντοπίζονται στον προσδιορισμό από τον χρήστη των παραμέτρων eps και MinPts. Αν και στον αλγόριθμο OPTICS ο καθορισμός και η είσοδος των παραπάνω παραμέτρων συμβάλει στην μείωση του χρόνου εκτέλεσης και δεν επηρεάζει τα αποτελέσματα, στο DBSCAN είναι ζωτικής σημασίας ο προσδιορισμός και η σωστή επιλογή των δύο παραμέτρων, καθώς συμβάλει στο σωστό υπολογισμό των συστάδων και των στιγμιότυπων που θα ενταχθούν σε αυτές.

Αυτές οι διαπιστώσεις αποτελούν το κίνητρο για την εκπόνηση της παρούσας διπλωματικής εργασίας. Η παρούσα διπλωματική εργασία εκπονήθηκε με σκοπό να βοηθήσει στον προσδιορισμό των παραμέτρων eps και MinPts, που είναι ζωτικής σημασίας για την εκτέλεση του DBSCAN.Όπως αναφέρθηκε και στην ενότητα 1.3 ο καθορισμός των παραμέτρων γίνεται με μία ευριστική μέθοδο που παράγει ένα $k-distgraph$ και στην συνέχεια εντοπίζεται το κατώφλι από όπου και ξεκινά ο εντοπισμός του θορύβου στο σύνολο δεδομένων.

Στα πλαίσια αυτής της διπλωματικής εργασίας υλοποιήθηκε μια διαδικτυακή εφαρμογή, η οποία προσφέρει στον χρήστη μια φιλική προς όλες τις συσκευές διεπαφή από όπου μπορεί να ανεβάσει το σύνολο δεδομένων (data set), για το οποίο θέλει να προσδιορίσει τις παραμέτρους eps και MinPts. Η εφαρμογή ονομάζεται k-Dist-Graph WebApp. Ο χρήστης θέτει με σαφείνια τις προδιαγραφές του συνόλου δεδομένων. Η εφαρμογή παράγει ένα αριθμό από διαγράμματα, όπου ο χρήστης μπορεί να εντοπίσει την τιμή του κατωφλιού (threshold point). Η εφαρμογή κάνει μια εκτίμηση της τιμής του κατωφλίου και την εμφανίζει πάνω στα παραγώμενα $k-distgraph$. Η εφαρμογή δίνει την δυνατότητα να λάβει ο χρήστης τα αποτελέσματα είτε ασύγχρονα στο email του είτε σύγχρονα μέσω της εφαρμογής.

1.6 Οργάνωση Διπλωματικής Εργασίας

Τα υπόλοιπα τμήματα της παρούσας διπλωματικής εργασίας είναι οργανωμένα ως ακολούθως: Το κεφάλαιο 2 παρουσιάζει τις τεχνολογίες, τις γλώσσες προγραμματισμού και τις βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση της διαδικτυακής εφαρμογής. Το κεφάλαιο 3 παρουσιάζει την σχεδίαση και την υλοποίηση του k-Dist-Graph WebApp. Το κεφάλαιο 4 παρουσιάζει σενάρια χρήσης της εφαρμογής, πειράματα και περιγραφή των συνόλων δεδομένων που χρησιμοποιήθηκαν. Το κεφάλαιο 5 παρουσιάζει την δοκιμή εμπειρίας χρήστη (User experience testing). Τέλος, το κεφάλαιο 6 παρουσιάζει χρήσιμα συμπεράσματα που βασίζονται στα αποτελέσματα των πειραμάτων και δίνει κάποιες κατευθύνσεις για μελλοντική έρευνα στο συγκεκριμένο επιστημονικό πεδίο.

2 Τεχνολογίες

2.1 Τι είναι back-end και front-end

Οι όροι front end και back end χρησιμοποιούνται από προγραμματιστές και επαγγελματίες υπολογιστών για να περιγράψουν τα επίπεδα που απαρτίζουν το υλικό, ένα πρόγραμμα υπολογιστή ή έναν ιστότοπο και η διάκριση τους γίνεται με βάση το πόσο προσιτά είναι από ένα χρήστη είτε αυτός είναι φυσική οντότητα είτε ψηφιακή.

Το back-end αναφέρεται σε τμήματα μιας εφαρμογής υπολογιστή ή ενός κωδικού προγράμματος που του επιτρέπει να λειτουργεί και στα οποία δεν είναι δυνατή η πρόσβαση ενός χρήστη. Τα περισσότερα δεδομένα και οι εντολές λειτουργίας αποθηκεύονται και είναι προσβάσιμα από το back-end ενός συστήματος υπολογιστή. Συνήθως ο κώδικας αποτελείται από μία ή περισσότερες γλώσσες προγραμματισμού. Το back-end ονομάζεται επίσης επίπεδο πρόσβασης δεδομένων λογισμικού ή υλικού και περιλαμβάνει οποιαδήποτε λειτουργικότητα στην οποία απαιτείται πρόσβαση και πλοήγηση με ψηφιακά μέσα.

Το επίπεδο πάνω από το back-end είναι το front-end και περιλαμβάνει όλο το λογισμικό ή το υλικό που αποτελεί μέρος μιας διεπαφής χρήστη. Οι άνθρωποι ή οι ψηφιακοί χρήστες αλληλεπιδρούν άμεσα με διάφορες πτυχές του front-end ενός προγράμματος, συμπεριλαμβανομένων των δεδομένων που εισάγονται από τον χρήστη, των κουμπιών, των προγραμμάτων, των ιστότοπων και άλλων λειτουργιών. Τα περισσότερα από αυτά τα χαρακτηριστικά έχουν σχεδιαστεί από επαγγελματίες ειδικών στην εμπειρία χρήστη (user experience - UX) ώστε να είναι προσβάσιμα, με ευχάριστο και εύχρηστο τρόπο από τον τελικό χρήστη [15]. Οι λόγοι που μια εφαρμογή διαιρείται σε front-end και back-end είναι πάρα πολλοί.Ο κυριότερος λόγος είναι η απαίτηση διαφορετικών ικανοτήτων που χρειάζεται η ανάπτυξη τόσο του front-end, από web designers, όσο και του back-end από μηχανικούς και προγραμματιστές.

2.2 Τεχνολογίες back-end του K-Dist-Graph WebApp

Η ανάπτυξη του back-end τμήματος της εφαρμογής λαμβάνει το αρχείο δεδομένων από το χρήστη, τις παραμέτρους που έχει θέσει για τους υπολογισμούς, διενεργεί τις διεργασίες για την εξαγωγή των αποτελεσμάτων. Στηρίχθηκε στις παρακάτω γλώσσες προγραμματισμού και στις βιβλιοθήκες τους.

2.2.1 PHP

Η PHP [16] δημιουργήθηκε από τον φοιτητή Rasmus Lerdorf ως μια συλλογή από scripts γραμμένα στην γλώσσα προγραμματισμού perl [17] που τα χρησιμοποιούσε στην προσωπική του ιστοσελίδα. Η αρχική χρήση της PHP από τον Rasmus ήταν η παρακολούθηση στατιστικών στοιχείων που αφορούσαν την επισκεψιμότητα στο προσωπικό του βιογραφικό. Αργότερα

έγραψε ξανά τα scripts σε γλώσσα C [18], για λόγους καλύτερης απόδοσης, επεκτείνοντας ταυτόχρονα τις δυνατότητες της υποστηρίζοντας έτσι την χρήση διαδικτυακών forms και σύνδεση με βάσεις δεδομένων. Το πρώτο επίσημο όνομα της PHP ήταν PHP/FI από τα «Personal Home Page/Forms Interpreter» (Προσωπική Ιστοσελίδα / Διερμηνέας Φορμών). Μετά από αυτή την δημιουργία ο Rasmus διέθεσε τον κώδικα στην ιστοσελίδα του ώστε να επωφεληθούν κι άλλοι από αυτόν.

Η PHP είναι μια γλώσσα προγραμματισμού που σχεδιάστηκε για την δημιουργία δυναμικών ιστοσελίδων και είναι γνωστή ως Hypertext PreProcessor. Σε αντίθεση με άλλες γλώσσες scripting του διαδικτύου η γλώσσα PHP είναι μια server-side (εκτελείται στον εξυπηρετητή) γλώσσα που συνήθως γράφεται πλαισιωμένη από HTML για την εμφάνιση των αποτελεσμάτων. Έτσι η PHP δεν στέλνεται άμεσα σε έναν πελάτη (client), αντ' αυτού πρώτα αναλύεται και μετά αποστέλλεται το παραγόμενο αποτέλεσμα (έτσι προέκυψε και η ευρέως γνωστή ονομασία της σε HyperText PreProcessor). Η PHP είναι διαδεδομένη για την πληθώρα δυνατοτήτων που μπορεί να προσφέρει σε διαδικτυακές εφαρμογές, όπως να θέσει ερωτήματα σε βάσεις δεδομένων, να δημιουργήσει εικόνες, να διαβάσει και να γράψει αρχεία, να συνδεθεί σε απομακρυσμένους υπολογιστές, κ.α.

Για την ανάπτυξη των back-end scripts της εφαρμογής χρησιμοποιείται η έκδοση 7.4 της PHP. Για την λειτουργία της αποστολής email και την δημιουργία της αναφοράς των αποτελεσμάτων σε μορφή PDF, χρησιμοποιήθαν οι παρακάτω βιβλιοθήκη και functions της PHP.

- **PHP mail**. Η PHP mail() [19] είναι μία ενσωματωμένη function της PHP που χρησιμοποιείται για την αποστολή email μέσω script.Η mail function δέχεται σαν παραμέτρους την Email address, το Subject, το Message και το CC or BC email addresses. Η εφαρμογή χρησιμοποιεί αυτήν την function για την αποστολή των αποτελεσμάτων μέσω email.
- **PHP fpdf**. Το FPDF [20] είναι μια βιβλιοθήκη της PHP που επιτρέπει τη δημιουργία αρχείων PDF με καθαρή PHP, δηλαδή χωρίς τη χρήση της βιβλιοθήκης PDFlib. Το F από το FPDF σημαίνει δωρεάν (free). Το FPDF δεν απαιτεί κάποια επέκταση (εκτός από το Zlib για ενεργοποίηση συμπίεσης και GD για υποστήριξη GIF). Η εφαρμογή χρησιμοποιεί αυτήν την βιβλιοθήκη για την δημιουργία μιας αναφοράς με το σύνολο των παραγόμενων αποτελεσμάτων σε μορφή pdf.

2.2.2 Python

Η Python [21] είναι μια εύχρηστη, ισχυρή γλώσσα προγραμματισμού. Η ανάπτυξη της ξεκίνησε το 1990 στο Centrum Wiskunde Informatica (CWI) του Άμστερνταμ από τον Ολλανδό Guido van Rossum. Έχει αποδοτικές δομές δεδομένων υψηλού επιπέδου και μια απλή αλλά αποτελεσματική προσέγγιση στον αντικειμενοστρεφή προγραμματισμό. Η κομψή σύνταξη και η δυναμική πληκτρολόγηση της Python, μαζί με την ερμηνευμένη φύση της, την καθιστούν ιδανική γλώσσα για scripting και ταχεία ανάπτυξη εφαρμογών σε πολλές περιοχές στις περισσότερες πλατφόρμες. Η Python διαθέτει μεγάλη βιβλιοθήκη, που συνήθως αναφέρεται ως ένα από

τα μεγαλύτερα πλεονεκτήματά της, καθώς παρέχει εργαλεία κατάλληλα για πολλές εργασίες, όπως ανάλυση δεδομένων, επεξεργασία εικόνας, μηχανική μάθηση [22]. Για την επεξεργσία, τον υπολογισμό και την δημιουργία των $K - distgraphs$ χρησιμοποιήθαν οι παρακάτω βιβλιοθήκες.

- **NumPy** [23]. Είναι μια βιβλιοθήκη η οποία προσθέτει υποστήριξη για μεγάλες, πολυδιάστατους πίνακες, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου για λειτουργία σε αυτούς τους πίνακες.
- **sklearn.neighbors** [24]. Είναι μια βιβλιοθήκη που παρέχει λειτουργικότητα για μη εποπτευόμενες και εποπτευόμενες μεθόδους μάθησης με βάση τους γείτονες. Οι μη εποπτευόμενοι πλησιέστεροι γείτονες είναι το θεμέλιο πολλών άλλων μεθόδων μάθησης, ιδίως της πολλαπλής μάθησης και της συσταδοποίησης.
- **Matplotlib** [25]. Είναι μια βιβλιοθήκη σχεδίασης γραφημάτων για τη γλώσσα προγραμματισμού Python και την μαθηματική της βιβλιοθήκη NumPy. Παρέχει ένα object-oriented API για την ενσωμάτωση γραφημάτων σε εφαρμογές με χρήση εργαλείων γενικής χρήσης GUI.

2.2.3 Apache Cronjob

Το cron job [26] είναι μια εντολή Linux που χρησιμοποιείται για τον προγραμματισμό εργασιών που θα εκτελεστούν κάποια στιγμή στο μέλλον. Συνήθως χρησιμοποιείται για τον προγραμματισμό μιας εργασίας που εκτελείται περιοδικά. Για τα περισσότερα cron jobs, υπάρχουν τρία στοιχεία:

- Το script που πρόκειται να κληθεί ή να εκτελεστεί.
- Η εντολή που εκτελεί το script σε επαναλαμβανόμενη βάση.
- Η ενέργεια ή η έξοδος του script, η οποία εξαρτάται από το τι κάνει το σενάριο που καλείται.

Συχνά, τα scripts που καλούνται από cron jobs τροποποιούν αρχεία ή βάσεις δεδομένων. Ωστόσο, μπορούν να εκτελέσουν άλλες εργασίες που δεν τροποποιούν δεδομένα και στον διακομιστή, όπως η αποστολή ειδοποιήσεων μέσω email. Η εφαρμογή εκτελεί ένα script που διαγράφει τα αποτελέσματα και τα συνόλα δεδομένωνs 24 ώρες μετά την δημιοργία τους.

2.3 Τεχνολογίες front-end του K-Dist-Graph WebApp

Η ανάπτυξη του front-end τμήματος της εφαρμογής αναλαμβάνει την διεπαφή με τον χρήστη δύνοντας του την δυνατότητα να ανεβάσει στον server το αρχείο δεδομένων, να ει΄σαγει τις παραμέτρους για την επεξεργασία και να λάβει τα αποτελέσματα της. Στηρίχθηκε στις παρακάτω γλώσσες προγραμματισμού.

2.3.1 HTML

Το 1980, ο φυσικός Tim Berners-Lee, ανέπτυξε το ENQUIRE, ένα σύστημα διαμοιρασμού έγγραφων για τους ερευνητές του CERN. Το 1989, ο Berners-Lee προτείνει ένα σύστημα hyper-text που βασίζεται στο Διαδίκτυο. Ο Berners-Lee καθόρισε την HTML και έγραψε το πρόγραμμα περιήγησης και το διακομιστή στα τέλη του 1990. Αφού έληξαν τα προσχέδια HTML και HTML + στις αρχές του 1994, το Internet Engineering Task Force (IETF) δημιούργησε μια ομάδα εργασίας HTML, η οποία το 1995 ολοκλήρωσε το ”HTML 2.0” [27], την πρώτη προδιαγραφή HTML που προοριζόταν να αντιμετωπιστεί ως πρότυπο βάσει του οποίου θα έπρεπε να βασίζονται οι μελλοντικές υλοποιήσεις της. Τα αρχικά της προέρχονται από τα Hyper Text Markup Language δηλαδή σε ελεύθερη μετάφραση Γλώσσα Σήμανσης Υπερκειμένου και αποτελεί την κύρια γλώσσα σήμανσης για τις ιστοσελίδες. Η HTML γράφεται υπό μορφή στοιχείων HTML τα οποία αποτελούνται από ετικέτες (tags), οι οποίες ετικέτες περικλείονται μέσα σε σύμβολα < και >. Οι ετικέτες αυτές συνήθως λειτουργούν ανά ζεύγη, με την πρώτη να ονομάζεται ετικέτα έναρξης (ή ανοίγματος) και τη δεύτερη ετικέτα λήξης (ή κλεισίματος). Ανάμεσα στις ετικέτες μπορεί να τοποθετηθεί κείμενο, πίνακες, εικόνες κ.λ.π.

Η εφαρμογή K-Dist-Graph WebApp ακολουθεί το πρότυπο της HTML5, το οποίο είναι η εξέλιξη της γλώσσας HTML με πληθώρα καινούριων δυνατοτήτων που θα κάνουν τη ζωή πιο εύκολη στην κατασκευή ιστοσελίδων, καθώς προσφέρει και ένα πιο ευχάριστο περιβάλλον στον επισκέπτη της ιστοσελίδας.

2.3.2 CSS

Το CSS προτάθηκε για πρώτη φορά από τον Håkon Wium Lie στις 10 Οκτωβρίου 1994 [28]. Εκείνη την εποχή, ο Lie συνεργάστηκε με τον Tim Berners-Lee στο CERN. Περίπου την ίδια περίοδο προτάθηκαν αρκετες style sheets γλώσσες για τον Ιστό, σε δημόσιες λίστες αλληλογραφίας και εντός της Κοινοπραξίας World Wide Web που οδήγησαν στην κυκλοφορία της πρώτης σύστασης W3C CSS (CSS1) [29] το 1996. Συγκεκριμένα, μια πρόταση από τον Bert Bos ήταν η επιρροή. Έγινε συν-συγγραφέας του CSS1 και θεωρείται συν-δημιουργός του CSS [30]. Η ονομασία CSS προέρχεται από τις λέξεις Cascading Style Sheets δηλαδή, διαδοχικά φύλλα στυλ. Πρόκειται για μια τεχνολογία ορισμού προβολής δεδομένων σε μια ιστοσελίδα η οποία έρχεται και δένει μαζί με αυτή της HTML η οποία με την σειρά της καθορίζει την δομή μιας σελίδας.

Τα CSS καθορίζουν τον τρόπο μορφοποίησης με τον οποίο θα προβάλλεται το περιεχόμενο

μιας html σελίδας όπως είναι τα χρώματα και η θέση των στοιχείων της σελίδας. Ένα μεγάλο πλεονέκτημα των CSS είναι ότι μπορούν να αποθηκευτούν σε εξωτερικό αρχείο με κατάληξη .css και να χρησιμοποιηθούν από κοινού σε περισσότερες από μια .html σελίδες, εξοικονομώντας με τον τρόπο αυτό χρόνο και κόπο υλοποίησης. Αυτό πρακτικά σημαίνει πως μια αλλαγή σε αυτό το αρχείο θα ενημερώσει όλες τις html σελίδες με τις οποίες συνδέεται.

Η εφαρμογή K-Dist-Graph WebApp ακολουθεί το πρότυπο του CSS3.

2.3.3 Javascript

Τον Σεπτέμβριο του 1995, ένας προγραμματιστής της Netscape με το όνομα Brandan Eich ανέπτυξε μια νέα scripting γλώσσα. Αρχικά ονομάστηκε Mocha, αλλά γρήγορα έγινε γνωστό ως LiveScript και αργότερα ως JavaScript. Η javascript [31] είναι μια scripting γλώσσα η οποία έχει σχεδιαστεί και χρησιμοποιείται για να εισάγουμε την έννοια της διαδραστικότητας στις HTML σελίδες. Είναι μια ερμηνευτική γλώσσα, δηλαδή το script εκτελείται χωρίς να έχει περάσει από την διαδικασία της σύνταξης. Με την javascript μπορούμε να εκτελέσουμε κάποια πράγματα όταν συμβαίνει ένα γεγονός,για παράδειγμα όταν ο χρήστης κλικάρει σε ένα HTML στοιχείο, εκτελείται κάποιο script και λαμβάνουμε τα αντίστοιχα αποτελέσματα.Η javascript μπορεί να διαβάσει και να αλλάξει τα περιεχόμενα ενός HTML στοιχείου.

2.3.4 JQuery

Το jQuery δημιουργήθηκε τον Ιανουάριο του 2006 στο BarCamp NYC από τον John Resig, επηρεασμένο από την προηγούμενη βιβλιοθήκη cssQuery του Dean Edwards [32]. Αυτή τη στιγμή συντηρείται από μια ομάδα προγραμματιστών με επικεφαλείς τους Timmy Willison (jQuery selector engine) και τον Richard Gibson (Sizzle). Το jQuery είναι μια ελαφριά βιβλιοθήκη της JavaScript, και στηρίζεται στην φιλοσοφία ”γράψτε λιγότερα, κάντε περισσότερα”. Ο σκοπός του jQuery είναι να διευκολύνει τη χρήση της JavaScript στον ιστότοπό.Το jQuery αναλαμβάνει, με την χρήση μεθόδων, πολλές κοινές εργασίες της JavaScript που απαιτούν πολλές γραμμές κώδικα για να ολοκληρωθούν, να τις καλεί με μία μόνο γραμμή κώδικα. Το jQuery απλοποιεί επίσης πολλά από τα περίπλοκα πράγματα από το JavaScript, όπως είναι οικλήσεις AJAX και ο χειρισμός DOM.

2.3.5 AJAX

Ο όρος AJAX χρησιμοποιήθηκε δημόσια στις 18 Φεβρουαρίου 2005 από τον Jesse James Garrett [33]. Το AJAX είναι συντομογραφία του ”Asynchronous JavaScript and XML” και είναι ένα σύνολο τεχνικών ανάπτυξης ιστοσελίδων που χρησιμοποιούν πολλές τεχνολογίες ιστού στην πλευρά του πελάτη (client-side) για τη δημιουργία ασύγχρονων εφαρμογών ιστού. Με το Ajax, οι διαδικτυακές εφαρμογές μπορούν να στέλνουν και να ανακτούν δεδομένα από έναν διακομιστή ασύγχρονα (στο παρασκήνιο) χωρίς να παρεμβαίνουν στην εμφάνιση και τη συμπε-

ριφορά της υπάρχουσας σελίδας. Με την αποσύνδεση του επιπέδου ανταλλαγής δεδομένων από το επίπεδο παρουσίασης, το Ajax επιτρέπει στις ιστοσελίδες και, κατ ’επέκταση, στις εφαρμογές ιστού, να αλλάζουν περιεχόμενο δυναμικά χωρίς την ανάγκη επαναφόρτωσης ολόκληρης της σελίδας. Το Ajax δεν είναι μια νέα τεχνολογία ή μια διαφορετική γλώσσα, απλώς υπάρχουσες τεχνολογίες που χρησιμοποιούνται με νέους τρόπους.

2.3.6 Bootstrap

Το Bootstrap αναπτύχθηκε από τος Mark Otto και Jacob Thornton στα μέσα του 2010 για το Twitter με το όνομα Twitter Blueprint. Επίσημα ως Bootstrap ανακοινώθηκε από το Twitter τον Αύγουστο του 2011. Το Bootstrap είναι το πιο δημοφιλές δωρεάν HTML, CSS και JavaScript framework για την ανάπτυξη ενός ιστότοπου φιλικού προς φορητές συσκευές και κινητά τηλέφωνα. Είναι ένα framework για front-end που χρησιμοποιείται για ευκολότερη και ταχύτερη ανάπτυξη ιστοσελίδων και εφαρμογών. Περιλαμβάνει HTML και CSS πρότυπα σχεδίασης για τυπογραφία, φόρμες, κουμπιά, πίνακες, πλοήγηση, modals, καρουζέλ εικόνας και πολλά άλλα. Μπορεί επίσης να χρησιμοποιήσει JavaScript πρόσθετα (plugins).

3 Σχεδίαση και υλοποίηση του k-Dist-Graph WebApp

Σε αυτό το κεφάλαιο παρουσιάζονται τα βήματα που ακολουθήθηκαν για την ανάπτυξη της εφαρμογής k-Dist-Graph WebApp. η αρχιτεκτονική και η υλοποίηση.

Στην Ενότητα 3.1 δίνεται μια αναλυτική περιγραφή της καταγραφής των απαιτήσεων της εφαρμογής. Στην Ενότητα 3.2 δίνεται μία αναλυτική περιγραφή της σχεδίασης της αρχιτεκτονικής της εφαρμογης. Στην Ενότητα 3.3 δίνεται μία αναλυτική περιγραφή της υλοποίησης της εφαρμογης.

3.1 Καταγραφή απαιτήσεων εφαρμογής

Η εφαρμογή απευθύνεται σε ερευνητές, αναλυτές δεδομένων, πανεπιστημιακούς, σπουδαστές και αποτελεί ένα εργαλείο για τον προσδιορισμό των παραμέτρων ε και $MinPts$ του αλγορίθμου DBSCAN.

Με την είσοδο στον διαδικτυακό τόπο της εφαρμογής ο χρήστης θα βλέπει την αρχική οθόνη της εφαρμογής.
Η αρχική οθόνη θα περιλαμβάνει μία μπάρα πλοήγησης, το κυρίως σώμα της εφαρμογής και ένα υποσέλιδο.
Η αριστερή πλευρά της μπάρας θα περιέχει το όνομα της εφαρμογής.
Η δεξιά πλευρά της μπάρας θα έχει ένα σύνδεσμο με τίτλο «Rate Us», για αξιολόγηση της εφαρμογής.
Το κυρίως σώμα θα περιλαμβάνει ένα σύντομο κείμενο για το τι κάνει η εφαρμογή.
Θα ακολουθεί μια φόρμα εισόδου δεδομένων. Η φόρμα θα έχει τα παρακάτω πεδία:

- Input πεδίο για upload για το αρχείο δεδομένων. Υποχρεωτικό πεδίο.
- Radio πεδίο για την παράμετρο data set header με τιμές Yes | No . Υποχρεωτικό πεδίο.
- Select πεδίο για την παράμετρο data set delimiter με τιμές , | . | ; | space ή tab . Υποχρεωτικό πεδίο.
- Numeric input πεδίο για την παράμετρο data set class column. Υποχρεωτικό πεδίο.
- Numeric input πεδίο για την παράμετρο minPts range ”From” . Υποχρεωτικό πεδίο.
- Numeric input πεδίο για την παράμετρο minPts range ”To”. Υποχρεωτικό πεδίο.
- Input πεδίο για εισαγωγή email. Προαιρετικό πεδίο.
- Κουμπί «Submit» για αποστολή φόρμας.
- Κουμπί «Reset» για εκκαθάριση πεδίων φόρμας.

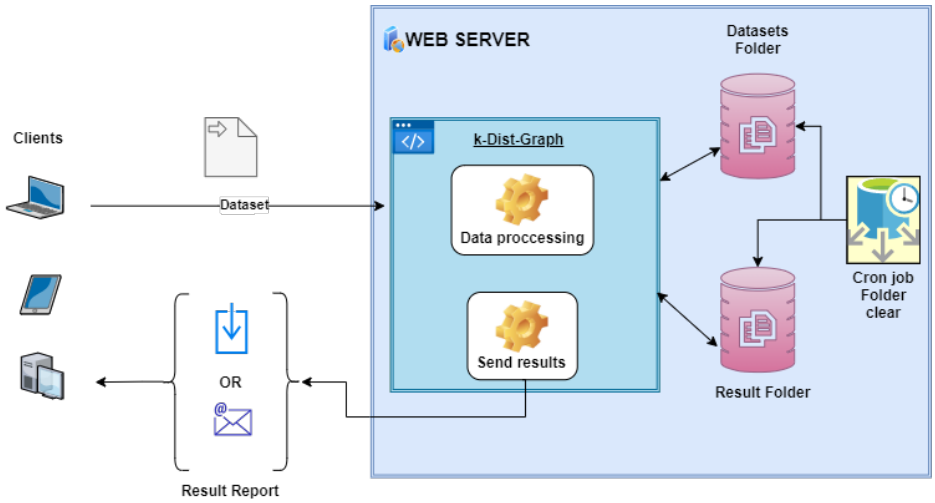
Σημείωση: Όταν το ποντίκι περνά πάνω από το εκάστοτε πεδίο της φόρμας θα εμφανίζει tooltip με πληροφορίες.

Τα αποτελέσματα της επεξεργσίας του συνόλο δεδομένων όταν δεν θα εισάγετε email από το χρήστη, θα εμφανίζεται πάνω από την φόρμα σε ένα πράσσινο box με σύνδεσμο για να εμφανιστεί η αναφορά σε μορφή PDF.

Το υποσέλιδο θα εμφανίζει πληροφορίες σχετικά με την εφαρμογή.

3.2 Αρχιτεκτονική εφαρμογής

Η λεπτομερής καταγραφή των απαιτήσεων του χρήστη οδήγησε στην σχεδίαση της εφαρμογής, όπως αυτή απεικονίζεται στο σχήμα 3.1.



Σχήμα 3.1: Αρχιτεκτονική εφαρμογής k-Dist-Graph.

Η διεπαφή του χρήστη με την εφαρμογή γίνεται μέσω της αρχικής οθόνης, σχεδιασμένη να αναπροσαρμόζεται σε όλες τις συσκευές φορητές και μη φορητές (responsive design). Ο χρήστης αποστέλει τα δεδομένα του από την διεπαφή. Αυτά λαμβάνονται και καλείται η διαδικασία upload. Γίνεται έλεγχος αν το αρχείο έχει μορφή csv. Αν είναι αυτής της μορφής προχωρά στο upload του αρχείου στο φάκελο «datasets» της εφαρμογής. Αν δεν είναι αυτής της μορφής καλείται python script το οιοίο αναλαμβάνει να προσπελάσει τα περιεχόμενα του αρχείου και να αφαιρέσει τις γραμμές είτε είναι κενές είτε αποτελούν σχόλια για το συνόλο δεδομένων. Γίνεται αποθήκευση, αφού προηγουμένς μετονομασθεί το αρχείο και γίνει τύπου csv στο φάκελο «datasets» της εφαρμογής.

Λαμβάνονται και οι υπόλοιπες παράμετροι από την διεπαφή και γίνεται έλεγχος, αν ο χρήστης ζητά να λάβει αποτελέσμα για μία τιμή MinPts ή για ένα εύρος τιμών. Καλείται το python script που θα αναλάβει την δημιουργία του $k - distgraph$ και τον υπολογισμό των ε και $MinPts$.

Μετά την επεξεργασία παράγεται ένα αρχείο PDF το οποίο είναι μια πλήρης αναφορά για τα εξαγώμενα αποτελέσματα της επεξεργασίας του συνόλου δεδομένων. Αν ο χρήστης έχει ζητήσει

αποστολή των αποτελεσμάτων στο email του θα κληθεί το αντίστοιχο PHP script, ειδάλως θα αποσταλεί πίσω στην διεπαφή ο σύνδεσμος για να ληφθεί η αναφορά. Τα αρχεία που ανεβάζει ο χρήστης από την διεπαφή και αυτά που παράγονται μετά την επεξεργασία από την εφαρμογή, αποθηκεύονται στους αντίστοιχους φακέλους της εφαρμογής, στο server, για χρονικό διάστημα 24 ωρών. Μετά το πέρας αυτού του χρονικού διαστήματος εκτελείται PHP script, που καλείται μέσω cron job του λειτουργικού συστήματος του web server, το οποίο και διαγράφει τα αρχεία αυτά.

3.3 Υλοποίηση εφαρμογής

Στο κεφάλαιο που ακολουθεί παρουσιάζεται η υλοποίηση των τμημάτων front-end και back-end της εφαρμογής.

3.3.1 Υλοποίηση front-end

Η καταγραφή των απαιτήσεων των χρηστών της, όπως περιγράφεται στο κεφάλαιο 3.1, οδήγησε στην ανάπτυξη μιας responsive διεπαφής, με λιτή και στοχευμένη προς τον χρήστη σχεδίαση.

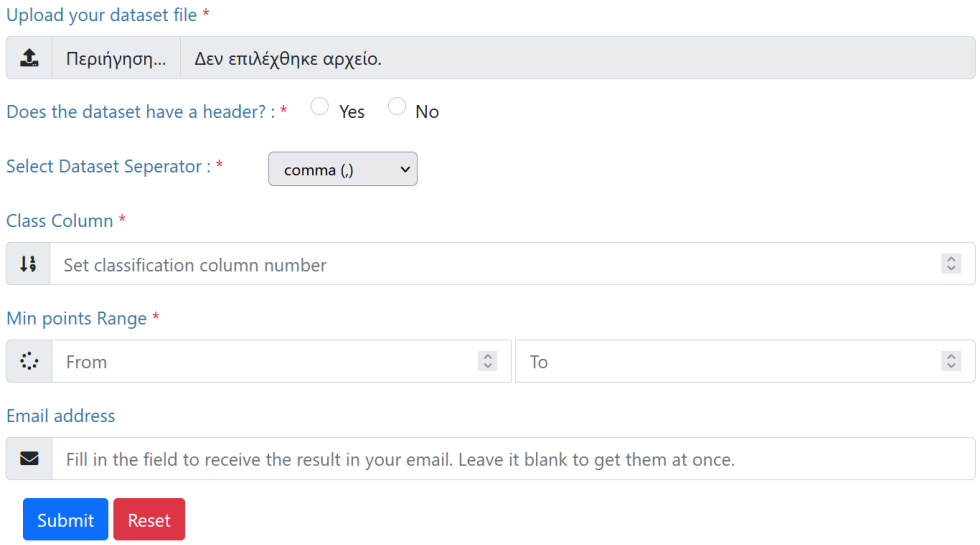
Αναπτύχθηκε ένα template σε HTML, που περιλαμβάνει πλήρη υποστήριξη στο framework του Bootstrap μέσω της ενσωμάτωσης του αντίστοιχου CSS στο header της HTML, την ενσωμάτωση του javascript πριν το τέλος του html αρχείου. Όπως έχει αναφερθεί παραπάνω στο 2.3.6 το Bootstrap εστιάζει πρώτα στις φορητές συσκευές, οπότε για να διασφαλιστεί η σωστή απόδοση και ζουμ αφής για όλες τις συσκευές προσθέτουμε το meta tag viewport.

```
<head>
 <!-- Bootstrap CSS -->
   <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.0.0/dist/css/
       bootstrap.min.css" rel="stylesheet"  crossorigin="anonymous">
   <meta name="viewport" content="width=device-width, initial-scale=1">
 <!-- JQuery plugin-->
   <script src="https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/
       jquery.min.js"></script>
</head>
...
 <!-- Bootstrap core JS-->
  <script src="https://cdn.jsdelivr.net/npm/bootstrap@4.6.0/dist/js/
      bootstrap.bundle.min.js"></script>
  <!-- Third party plugin JS-->
  <script src="https://cdnjs.cloudflare.com/ajax/libs/animejs/3.2.1/
      anime.min.js"></script>
 </body>
</html>
```

Απόσπασμα κώδικα 1: index.html - ενσωμάτωση Bootstrap framework

Σχεδιάσαμε την φόρμα εισόδου δεδομένων της διεπαφής αποτελούμενη απο 6 πεδία εισόδου

παραμέτρων, σύμφωνα με τις απαιτήσεις του χρήστη. Τα υποχρεωτικά πεδία σημάνθηκαν με ένα κόκκινο αστερίσκο, ώστε να δηλώνεται εμφανώς η υποχρέωση συμπλήρωσης του πεδίου (σχήμα 3.2).



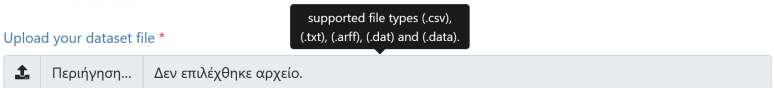
Σχήμα 3.2: Φόρμα εισόδου δεδομένων k-Dist-Graph.

Διασφαλίστηκε η ορθή εισαγωγή τιμών στα πεδία της φόρμας χρησιμοποιώντας το attribute < *type* > του html tag < *input* >.

```
1 <label for="input6" class="form-label">Email address</label>
2 <div class="input-group mb-3"><span class="input-group-text" id="
    basic-addon1"><i class="fa fa-envelope --bs-blue" ></i></span>
3  <input type="email" class="form-control" id="input6" name="input6"
     data-original-title="ToolTip Default Size" data-toggle="tooltip"
     title="if you fill the field, you can close the page." placeholder="
     Fill in the field to receive the result in your email. Leave it
     blank to get them at once." />
4 </div>
```

Απόσπασμα κώδικα 2: index.html - ορισμός ”email” στο attribute <type> του tag <input>

Έμφαση δόθηκε στην καλύτερη δυνατή πληροφόρηση των απαιτήσεων κάθε πεδίο εισαγωγής με την χρήση tooltips κάθε φορά που ο χρήστης περνά τον κέρσορα πάνω από κάποιο πεδίο (σχήμα 3.3). Υλοποιήθηκε με την δημιουργία JavaScript function και κλήση της στο αντίστοιχο HTML tag (Απόσπ. κώδικα 3).



Σχήμα 3.3: Εμφάνιση tooltip πάνω από πεδίο εισόδου.

```
<div class="input-group mb-3" data-toggle="tooltip" title="supported file
    types (.csv), (.txt), (.arff), (.dat) and (.data). ">
.
.
.
<!-- Tooltip JS function -->
 <script>
    $(document).ready(function(){
        $('[data-toggle="tooltip"]').tooltip();});
  </script>
```

Απόσπασμα κώδικα 3: index.html - ορισμός και εμφάνιση tooltip με χρήση JavaScript[1]

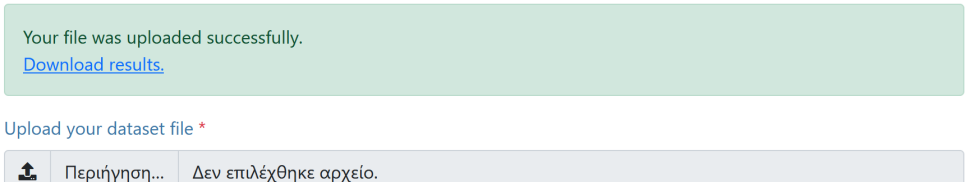
Με την συμπλήρωση όλων των υποχρεωτικών πεδίων ο χρήστης αποστέλει τα στοιχεία της φόρμας μέσω του κουμπιού *Submit*.

Καλείται Ajax call (Απόσπ. κώδικα 4) το οποίο αναλαμβάνει να αποστείλει ασύγχρονα τα δεδομένα της φόρμας στον webserver όπου και θα γίνει η επεξεργασία τους, ενώ παράλληλα αναμένει το τέλος της εκτέλεσης των απαραίτητων διεργασιών από την πλευρά του server, για να εμφανίσει τα αποτελέσματα στην διεπαφή του χρήστη (σχήμα 3.4) , αν δεν έχει επιλέξει την αποστολή αυτών με email.

```
<script>
            $(document).ready(function(e){
            // Submit form data via Ajax
            $("#fupForm").on('submit', function(e){
            e.preventDefault();
            var nameValue = document.getElementById("input6").value;
            var formData = new FormData(this);
            formData.append('session_id', document.cookie.match(/[^;]+/))
                ;
            $.ajax({
              type: 'POST',
              url: 'srun.php',
              data: formData,
              contentType: false,
              cache: false,
              processData:false,
            beforeSend: function(){
                $('.submitBtn').attr("disabled","disabled");
                $('#fupForm').css("opacity",".5");
            },
            success: function(response){
                 var response_f = $.trim(response);
                 //alert(response_f);
                if (response_f =='1') {
                     $('.statusMsg').html('');
                     $('#fupForm').css("opacity","");
                     $(".submitBtn").removeAttr("disabled");
                     $('#changeMe').html('<p class="alert alert-success">
                        You will receive results shortly in your email.
                        </p>');
```

```
28                      $('#fupForm')[0].reset();
29                  } else {
30                      $('.statusMsg').html('');
31                      $('#fupForm').css("opacity","");
32                      $(".submitBtn").removeAttr("disabled");
33                      $('#changeMe').html('<p class="alert alert-success">
                            ' + response_f + '</p>');
34                      $('#fupForm')[0].reset();
35                  }
36              }
37          });
38        });
39      });
40  </script>
```

Απόσπασμα κώδικα 4: index.html - ajax function για αποστολή δεδομένων φόρμας



Σχήμα 3.4: Λήψη αποτελεσμάτων από την διεπαφή μέσω συνδέσμου

3.3.2 Υλοποίηση back-end

Τα δεδομένα της φόρμας, λαμβάνονται από το ένα PHP script που είναι υπεύθυνο για την διαχελιρισή τους.

Η πρώτη εργασία που εκτελεί είναι η λήψη του *session id* από τον browser ή αν δεν υπάρχει δημιουργεί νέο *session id*. Το *session id* θα χρησιμοποιηθεί στην συνέχεια για την μοναδική σήμανση των αρχείων που θα παράξει η επεξεργασία των δεδομένων (Απόσπ. κώδικα 17).

```
1  $xid = ltrim($_POST['session_id'], "PHPSESSID =");
2  if ($xid == '' || $xid == 'null'){
3    session_start();
4    $xid =  session_id();
5  }
```

Απόσπασμα κώδικα 5: srun.php - Λήψη ή δημιουργία session id

Στο επόμενο βήμα εξετάζουμε τον τύπο αρχείου που έχει ανεβάσει ο χρήστης (Απόσπ. κώδικα 15). Αν είναι μορφής arff ή txt ή dat ή data, τότε καλούμε ένα python script που αναλαμβάνει να φιλτράρει το αρχείο από σχόλια, κενές γραμμές και ειδικούς χαρακτήρες (Απόσπ. κώδικα 7). Οι χαρακτήρες που χρησιμοποιούνται είτε σαν σχόλια στην πληθώρα των συνόλων δεδομένων είναι οι *%%*, *@* είτε περικλύουν τα δεδομένα όπως οι *{ και }*. Εφόσον το αρχείο επεξεργαστεί και εκκαθαριστεί, αποθηκεύεται σε μορφή csv στον φάκελο results.

```php
if(isset($_FILES["input2"]) && $_FILES["input2"]["error"] == 0){
  $target_dir = "datasets/";
  $file = $_FILES['input2']['name'];
  $path = pathinfo($file);
  $filename = $path['filename'];
  $ext = $path['extension'];
  $temp_name = $_FILES['input2']['tmp_name'];
  $path_filename_ext = $target_dir.$filename.".".$ext;
  move_uploaded_file($temp_name,$path_filename_ext);
  if ($ext == 'arff' || $ext == 'txt' || $ext == 'dat'
    || $ext == 'data'){
     $comm = escapeshellcmd('python file_conv.py ' .
     $path_filename_ext);
     exec($comm,$output_new, $ret_code);
     extract($output_new);
     $input2 = $output_new[0];
  }else{
        $input2 = $path_filename_ext; //Dataset name
      }
     echo "Your file was uploaded successfully.<br>";
}
```

Απόσπασμα κώδικα 6: srun.php - Έλεγχος τύπου αρχείου δεδομένων

```python
#import sys

file_import = str(sys.argv[1]) #file

split_string = file_import.split(".", 1)
substring = split_string[0]

fh = open(file_import, "r")
lines = fh.readlines()
fh.close()

keep = []

for line in lines:
    line = line.replace('{','')
    line = line.replace('}','')
    if not line.isspace() and (not line.startswith('@')) and (not line.
        startswith('%')):
         keep.append(line)

dir_s = substring + ".csv"
fh = open(dir_s, "w")
fh.write("".join(keep))
print (dir_s)
fh.close()
```

Απόσπασμα κώδικα 7: fileconv.py - Επεξεργασία και αλλαγή τύπου αρχείου δεδομένων

Τα δεδομένα έτοιμα προς επεξεργασία, αποστέλλονται στο python script που αναλαμβάνει την επεξεργασία και εξαγωγή των αποτελεσμάτων.

```
$command = escapeshellcmd('python eps.py ' . $input1fr . ' '
. $input2 . ' ' . $input3 . ' ' . $input4 . ' '
. $input5 . ' ' . $xid . ' ' . $input1to);
exec($command,$output, $ret_code);
```

Απόσπασμα κώδικα 8: srun.php - Κλήση Python script από το περιβάλλον της PHP

Παρακάτω παρουσιάζεται αναλυτικά η διαδικασία επεξεργασίας των δεδομένων, ο τρόπος δημιουργίας των $k - dist$ graphs και ο υπολογισμός της τιμής ε.

Στο script πρώτα εισάγονται οι απαραίτητες βιβλιοθήκες που θα χρησιμοποιηθούν για την ολοκλήρωση της διεργασίας (Απόσπ. κώδικα 9).

```
# -*- coding: utf-8 -*-
import sys
import numpy as np
from numpy import genfromtxt
from sklearn.neighbors import NearestNeighbors
from matplotlib import pyplot as plt
import time
import os
```

Απόσπασμα κώδικα 9: eps.py - Εισαγωγή libraries

Ανάγνωση των παραμέτρων από την PHP (Απόσπ. κώδικα 10, γραμμή 2 εως 8).

```
#Anagnosi parametron apo php
min_fr = int(sys.argv[1]) #k min
dataset = str(sys.argv[2]) #csv file
symb = str(sys.argv[3]) #delimiter
ds_class_col = int(sys.argv[4]) # class column
ds_titles = int(sys.argv[5]) # titles
save_dir = str(sys.argv[6]) # save_dir
min_to = int(sys.argv[7]) #k max
```

Απόσπασμα κώδικα 10: eps.py - Ανάγνωση παραμέτρων από την PHP

Θέτουμε το «result» ως φάκελο αποθήκευσης αποτελεσμάτων και γίνετε έλεγχος αν υπάρχει ο φάκελος, αλλιώς δημιουργείτε. Στην συνέχεια εκχωρείται το σύνολο δεδομένων στο πίνακα my_data, διαγράφεται η στήλη που αφορά την κατηγοριοποίηση των δεδομέων και διαγράφεται η πρώτη γραμμή αν στο σύνολο δεδομένων περιέχεται κεφαλίδα (header). Τέλος εκχωρείται το my_data στον πίνακα X.

```
#dir apothikeusis apotelesmaton
fp2 = "result\\"

degree_sign = u"\N{DEGREE SIGN}"

```

```
#an den yparxei to dir to dimiourgei
if not os.path.exists(fp2):
    os.makedirs(fp2)

#diavasma csv
my_data = genfromtxt(dataset, delimiter = symb)

#diagrafi stilis classification an yparxei
if not(ds_class_col == 0 or ds_class_col == '' or ds_class_col == ' '):
 my_data = np.delete(my_data, ds_class_col-1, 1)

#diagrafi headers an to dataset exei headers
if ds_titles == 1:
  my_data = np.delete(my_data, 0, 0)

X = my_data
```

Απόσπασμα κώδικα 11: eps.py - Προετοιμασία του συνόλου δεδομένων

Ορίζονται δύο functions (Απόσπ. κώδικα 12). Θεωρούμε πως η καμπύλη είναι αντιπροσωπευτική για όλες τις καμπύλες (π.χ. μονοτροπική και συνεχής), τότε μια γρήγορη και ”πονηρή” μέθοδος είναι να την περιστρέψετε σε κάποιο βαθμό και απλά να λάβετε την ελάχιστη τιμή. Η περιστροφή γίνεται με τον πολλαπλασιασμό με τον πίνακα περιστροφής [34] που ορίζεται ως

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

όπου το θ είναι η επιθυμητή γωνία σε ακτίνια (radians).

```
def find_elbow(gonia, theta):

    # make rotation matrix
    co = np.cos(theta)
    si = np.sin(theta)
    rotation_matrix = np.array(((co, -si), (si, co)))

    # rotate gonia vector
    rotated_vector = gonia.dot(rotation_matrix)

    # return index of elbow
    return np.where(rotated_vector == rotated_vector.min())[0][0]


def get_data_radiant(gonia):
  return np.arctan2(gonia[:, 1].max() - gonia[:, 1].min(),
                    gonia[:, 0].max() - gonia[:, 0].min())
```

Απόσπασμα κώδικα 12: eps.py - Ορισμός function υπολογισμού κατωφλιού

Εκκινείται βρόχος με αριθμό επαναλήψεων την τιμή ή το εύρος τιμών που έθεσε ο χρήστης.

Μέσα στο βρόχο αποθηκεύεται η χρονική στιγμή έναρξης της διεργασίας υπολογισμού (Απόσπ. κώδικα 13, γραμμή 2). Καλείται η βιβλιοθήκη NearestNeighboors [24] με παραμέτρους τον αριθμό των $MinPts$ και σαν αλγόριθμο υπολογισμού την τιμή $auto$, όπου και υπολογίζονται οι εγγύτεροι γείτονες. Καλούνται οι functions για τον υπολογισμό του ε (Απόσπ. κώδικα 13, γραμμή 8 εως 10).

Ορίζονται οι παράμετροι του γραφήματος και δημιουργούνται τα γραφήματα (Απόσπ. κώδικα 13, γραμμή 13 εως 22). Στις γραμμές 24 και 25 υπολογίζεται ο χρόνος εκτέλεσης. Στην γραμμή 26 αποθηκεύονται το γραφήμα ή τα γραφήματα k-dist στο φάκελο «result» σε μορφή εικόνας PNG. Στην γραμμή 27 ελέγχουμε αν η τιμή του ε είναι διάφορη του κενού και αν όχι αποστέλονται στο PHP σαν output η τιμή του ε, της γωνίας της καμπύλης, ο χρόνος εκτέλεσης, και το όνομα της παραχθείσας εικόνας.

```
1  while min_fr <= min_to:
2    start_time = time.time()
3    nbrs = NearestNeighbors(n_neighbors=min_fr, algorithm='auto').fit(X)
4    distances, indices = np.sort(nbrs.kneighbors(X))
5    new = np.sort(distances[:, min_fr-1])
6    f = np.arange(1, len(new)+1).astype(int)
7    x = np.array(new)
8    gonia = np.column_stack([f,x])
9    elbow_index = find_elbow(gonia, get_data_radiant(gonia))
10   epsillon = gonia[elbow_index,1].astype(float) #Timi tou epsillon
11   angle = np.rad2deg(get_data_radiant(gonia))
12 # Plotting
13   fig = plt.figure()
14   ax = plt.axes()
15   plt.grid(True)
16   ax.plot(new);
17   plt.title('K-dist Graph', fontsize=14)
18   plt.ylabel('MinPts =' + str(min_fr), fontsize=10);
19   plt.xlabel('Instances', fontsize=10);
20   plt.vlines(gonia[elbow_index, 0], ymin=gonia[:, 1].min(), ymax=gonia[:,
         1].max(), colors='red', linestyles='--')
21   plt.axhline(y = epsillon, color = 'orange', linestyle = '-.')
22   plt.savefig(fp2 + "\\" + save_dir + "_eps_mins_" + str(min_fr) +".png",
         bbox_inches='tight')
23 # end plotting
24   end_time= time.time()
25   data_time = end_time - start_time
26   img_png=fp2 + "\\" + save_dir +"_eps_mins_" + str(min_fr) +".png"
27   if str(round(epsillon,2)) == '':
28       print('e value is null.')
29       exit()
30   else:
31       print(str(round(epsillon,2)))
32       print('Elbow angle: ' + str(round(angle,2)) + degree_sign)
33       print(str(round(data_time,2)))
34       print(img_png)
35       min_fr += 1
```

Απόσπασμα κώδικα 13: eps.py - Βρόχος υπολογισμών

Μετά την ολοκλήρωση των υπολογισμών και την λήψη των σποτελεσμάτων, δημιουργείται η αναφορά σε μορφή PDF, με την χρήση της βιβλιοθήκης FPDF. Παρακάτω παρατίθεται ο κώδικας.

```
$pdf=new FPDF();

    $pdf->SetAuthor('Alkibiadis Tzioras');
    $pdf->SetTitle('k-dist graph WebApp');
    $pdf->SetFont('Arial','B',18);
    $pdf->AddPage('P');
    $pdf->SetDisplayMode(100,'default');
    $pdf->cell(100,14,"k-dist graph and eps for DBSCAN", 1,
    0, 'C');
    $pdf->SetFontSize(11);
    $pdf->SetXY(10,26);
    $pdf->Write(5, "Dataset filename: " . $file);
    $pdf->SetXY(10,31);
    $pdf->Write(5, "Min points: " . $input1fr);
    $pdf->SetXY(10,36);
    $pdf->Write(5, "---------------------------------");
    $pdf->SetFontSize(10);
    $pdf->SetXY(10,41);
    $pdf->Write(5,'Estimated epsilon value is ');
    $pdf->SetTextColor(255, 0, 0);
    $pdf->Write(5,$output[0]);
    $pdf->SetTextColor(0,0,0);
    $pdf->SetXY(10,45);
    $pdf->Write(5,$output[1]);
    $pdf->SetXY(10,49);
    $pdf->Write(5,"Process time: " . $output[2] . " sec");
    $pdf->Image($output[3],10,70,-100);

    //footer
    $pdf->SetFont('Arial','I',9);
    $pdf->SetXY(92,255);
    $pdf->Write(20,"Page : 1/1" );

    $pdf->output($pdf_name, 'F');
```

Απόσπασμα κώδικα 14: srun.php - Δημιουργία αρχείου αναφοράς σε PDF

Ελέγχεται αν έχει δημιουργηθεί το PDF και αν ο χρήστης επέλεξε την λήψη της αναφοράς μέσω email, καλείται το αρχείο *mail.php* και διεκπεραιώνει το αίτημα.

```

if (file_exists($pdf_name)){
```

```
3   if (!empty($input6)){
4     require_once "./php/mail.php";
5     clear_png($xid);
6     clear_data($input2);
7     exit;
8    }else{
9      $actual_link = (isset($_SERVER['HTTPS']) && $_SERVER['HTTPS']
10      === 'on' ? "https:" : "http:") . '//'. $_SERVER['HTTP_HOST']
11      . dirname($_SERVER['PHP_SELF']);
12     $filename = $actual_link . $pdf_name;
13     echo '<a href="' . $pdf_name . '" target="_blank">Download
14     results.</a>';
15     clear_png($xid);
16     clear_data($input2);
17     exit;
18    }
19  }
```

Απόσπασμα κώδικα 15: srun.php - Λήψη αποτελεσμάτων στην διεπαφή ή με email

Μετά την αποστολή της αναφοράς μέσω email ή την λήψη μέσω της διεπαφής, καλούνται (Απόσπ. κώδικα 15, γραμμες 5,6 και 15,16) οι functions *clear_png* και *clear_data* (Απόσπ. κώδικα 16), που προχωρούν στην διαγραφή των προσωρινών αρχείων που παρήχθησαν κατά την επεξεργασία του συνόλου δεδομένων .

```
1
2  function clear_png($xid){
3    $fullPath = __DIR__ . "/result/" ;
4    $del_file = $fullPath . $xid;
5    array_map('unlink', glob( "$del_file*.png"));
6  }
7
8  function clear_data($temp_csv){
9    $fullPathd = __DIR__ . "/";
10   $da_file = $fullPathd . $temp_csv;
11   array_map('unlink', glob( "$da_file*"));
12 }
```

Απόσπασμα κώδικα 16: srun.php - Functions εκκαθάρισης προσωρινών αρχείων

3.3.3 Υλοποίηση cron-job

Βασικός παράγοντας για την εύρυθμη λειτουργία του server που φιλοξενεί την εφαρμογή είναι η διαχείριση του αποθηκευτικού χώρου. Επιλέχθηκε η ανάπτυξη ενός PHP script (Απόσπ. κώδικα 20) που θα καλείται από το λειτουργικό σύστημα του server, ανά τακτα χρονικά διαστήματα και θα προχωρά στην διαγραφή των αποτελεσμάτων που αποθηκεύονται στον φάκελο $result$. Κάθε αρχείο με διάστημα δημιουργίας πάνω από 24 ώρες θα διαγράφεται οριστικά.

```php
<?php

//Delete results over 1 day.
$folderName = 'result';
if (file_exists($folderName)) {
    foreach (new DirectoryIterator($folderName) as $fileInfo) {
        if ($fileInfo->isDot()) {
        continue;
        }
        if ($fileInfo->isFile() && time() - $fileInfo->getCTime()
        >= 60 * 60 * 24) {
            unlink($fileInfo->getRealPath());
        }
    }
}

//Delete uploaded datasets over 1 day.
$folderName = 'datasets';
if (file_exists($folderName)) {
    foreach (new DirectoryIterator($folderName) as $fileInfo) {
        if ($fileInfo->isDot()) {
        continue;
        }
        if ($fileInfo->isFile() && time() - $fileInfo->getCTime()
        >= 60 * 60 * 24) {
            unlink($fileInfo->getRealPath());
        }
    }
}

?>
```

Απόσπασμα κώδικα 17: del.php - Script εκκαθάρισης αρχείων

4 Σενάρια Χρήσης

Στο κεφάλαιο που ακολουθεί θα παρουσιαστούν τέσσερα σενάρια χρήσης της εφαρμογής με απεικόνιση των παραμέτρων εισόδου στην διεπαφή και εμφάνιση των παραγόμενων αποτελεσμάτων.

4.1 Σύνολα δεδομένων

Τα σύνολα δεδομένων που θα χρησιμοποιηθούν περιγράφονται συνοπτικά στον παρακάτω πίνακα 4.1.

Dataset	Size	Attributes	Classes column
Letter Recognition	20000	16	17
Ecoli	336	7	8
Wine	178	13	14
Yeast	1484	8	9

Πίνακας 4.1: Περιγραφή συνόλων δεδομένων.

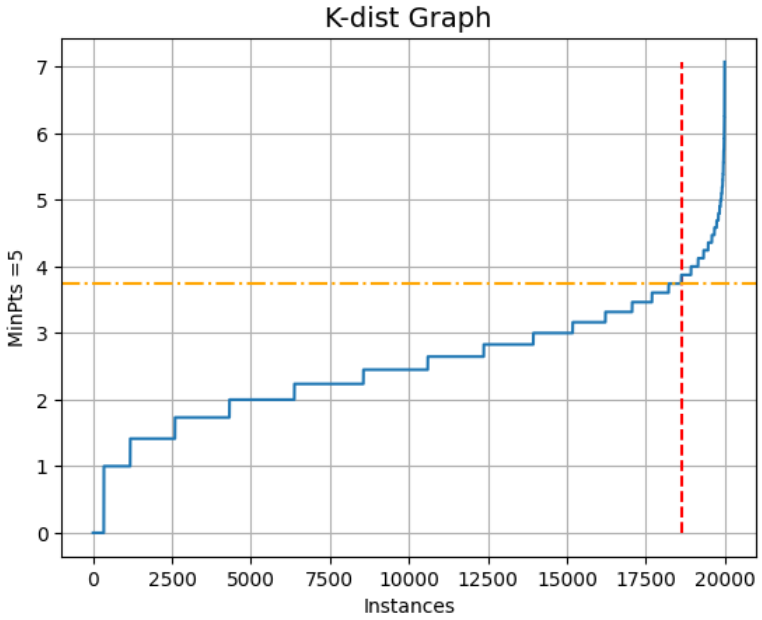
4.1.1 Letter Recognition (LR)

Ο σκοπός αυτού του συνόλου δεδομένων είναι να προσδιορίσει καθένα από τα 26 κεφαλαία γράμματα του αγγλικού αλφαβήτου μέσα από ένα μεγάλο αριθμό ασπρόμαυρων ορθογώνιων εικονοστοιχείων. Οι διαθέσιμες εικόνες χαρακτήρων βασίστηκαν σε 20 διαφορετικές γραμματοσειρές. Κάθε γράμμα σε αυτές τις 20 γραμματοσειρές παραμορφώθηκε τυχαία για να δημιουργήσει ένα αρχείο 20.000 μοναδικών στιγμιοτύπων. Τέλος, κάθε στιγμιότυπο μετατράπηκε σε 16 πρωτόγονα αριθμητικά χαρακτηριστικά (στατιστικές ροπές και μετρήσεις άκρων) τα οποία αργότερα κλι -μακώθηκαν ώστε να χωρέσουν σε μια σειρά ακέραιων τιμών από 0 έως 15. Τέτοια χαρακτηριστικά περιλαμβάνουν την οριζόντια και κατακόρυφη θέση του εμφανιζόμενου γράμματος, πλάτος και ύψος , το συνολικό αριθμς pixel, μέσος αριθμός pixel ανά άξονα, πλήθος άκρων από αριστερά προς τα δεξιά κ.λπ.

Η εφαρμογή παράγει τα αποτελέσματα του πίνακα 4.2, για εύρος τιμών MinPts 3 εως 5. Στιγμιότυπο του αποτελέσματος απεικονίζεται στο σχήμα 4.1. Μελετώντας τα αποτελέσματα διαπιστώνουμε πως το σημείο που υπάρχει η μέγιστη καμπυλότητα (κατώφλι σημειώνεται με την κόκκινη κάθετη γραμμή)) στο συγκεκριμένο σύνολο δεδομένων μοιάζει κοινό για όσες τιμές MinPts δοκιμάστηκαν, καθώς οι τιμές είναι στρογγυλοποιημένες στα δύο δακαδικά ψηφία. Η τιμή όμως για το ε (σημειώνεται με την πορτοκαλί οριζόντια γραμμή) διαφέρει ανάλογα με την τιμή MinPts που έχει επιλεχθεί, καθώς μεγαλώνει η ακτίνα της γειτονιάς των στιγμιότυπων. Παράλληλα διαπιστώνεται πως οι χρόνοι υπολογισμού είναι μικροί, αν και το σύνολο δεδομένων έχει 20000 στιγμιότυπα, και τείνουν να αυξάνονται καθώς αυξάνεται η τιμή του MinPts.

Τιμή *MinPts*	Τιμή ε	Γωνία καμπύλης	Χρόνος υπολογισμού (sec)
3	3.32	42.52°	2.01
4	3.46	42.52°	2.13
5	3.74	42.52°	2.35

Πίνακας 4.2: Αποτελέσματα για MinPts 3 εως 5 του συνόλου δεδομένων LR



Σχήμα 4.1: Στιγμιότυπο αποτελεσμάτων για σύνολο δεδομένων LR

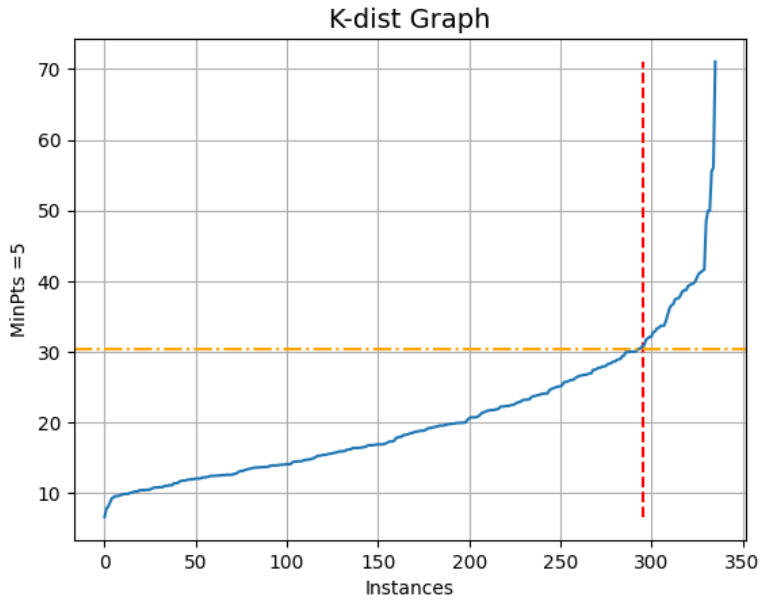
4.1.2 Ecoli (ECL)

Ο στόχος αυτού του συνόλου δεδομένων είναι η πρόβλεψη της θέσης των πρωτεϊνών χρησιμοποιώντας ορισμένες μετρήσεις σχετικά με το κύτταρο (κυτταρόπλασμα, εσωτερική μεμβράνη κ.α.).

Η εφαρμογή παράγει τα αποτελέσματα του πίνακα 4.3, για εύρος τιμών MinPts 3 εως 5. Στιγμιότυπο του αποτελέσματος απεικονίζεται στο σχήμα 4.2. Μελετώντας τα αποτελέσματα διαπιστώνουμε πως το σημείο που υπάρχει η μέγιστη καμπυλότητα (κατώφλι σημειώνεται με την κόκκινη κάθετη γραμμή)) στο συγκεκριμένο σύνολο δεδομένων η τιμή της γωνίας μειώνεται όσο η τιμή των MinPts αυξάνεται. Παρατηρώντας καλύτερα το σχήμα 4.2, στο k-dist graph η καμπύλη είναι πιο εμφανής σε σχέση με το σύνολο δεδομένων LR. Η τιμή για το ε (σημειώνεται με την πορτοκαλί οριζόντια γραμμή) διαφέρει ανάλογα με την τιμή MinPts που έχει επιλεχθεί, καθώς μεγαλώνει η ακτίνα της γειτονιάς των στιγμιότυπων. Παράλληλα διαπιστώνεται πως οι χρόνοι υπολογισμού είναι και σε αυτό το σύνολο δεδομένων πολύ μικροί και όσο αυξάνονται τα MinPts ο χρόνος υπολογισμού μειώνεται.

Τιμή *MinPts*	Τιμή ε	Γωνία καμπύλης	Χρόνος υπολογισμού (sec)
3	23.87	53.47°	0.24
4	28.88	53.43°	0.15
5	30.48	53.39°	0.15

Πίνακας 4.3: Αποτελέσματα για MinPts 3 εως 5 του συνόλου δεδομένων ECL



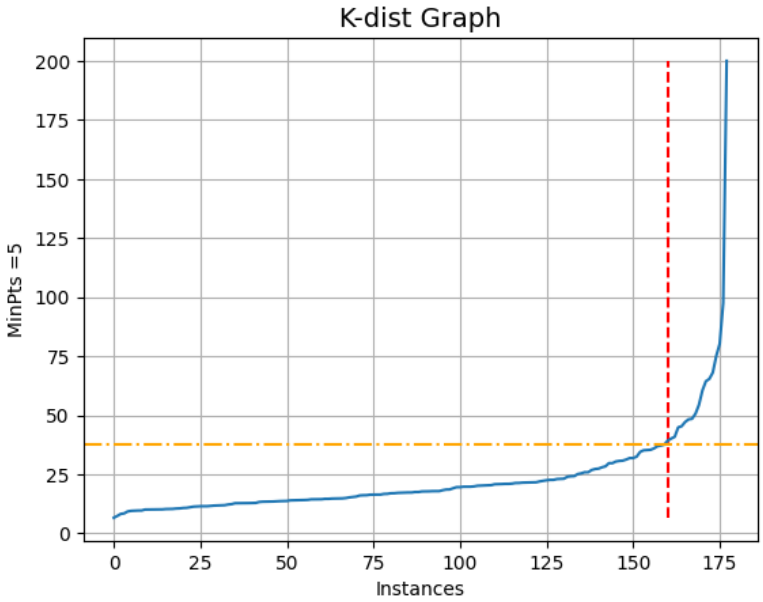
Σχήμα 4.2: Στιγμιότυπο αποτελεσμάτων για σύνολο δεδομένων ECL

4.1.3 Wine (WN)

Τα δεδομένα του συνόλου είναι τα αποτελέσματα μιας χημικής ανάλυσης των οίνων που καλλιεργούνται στην ίδια περιοχή στην Ιταλία, αλλά προέρχονται από τρεις διαφορετικές ποικιλίες. Η ανάλυση καθόρισε τις ποσότητες των 13 συστατικών που βρέθηκαν σε καθέναν από τους τρεις τύπους κρασιών. Η εφαρμογή παράγει τα αποτελέσματα του πίνακα 4.4, για εύρος τιμών MinPts 3 εως 5. Στιγμιότυπο του αποτελέσματος απεικονίζεται στο σχήμα 4.3. Μελετώντας τα αποτελέσματα διαπιστώνουμε πως το σημείο που υπάρχει η μέγιστη καμπυλότητα (κατώφλι σημειώνεται με την κόκκινη κάθετη γραμμή)) στο συγκεκριμένο σύνολο δεδομένων η τιμή της γωνίας διαφέρει όσο η τιμή των MinPts αυξάνεται. Παρατηρώντας καλύτερα το σχήμα 4.3, στο k-dist graph η καμπύλη είναι πιο εμφανής σε σχέση με το σύνολο δεδομένων LR. Η τιμή για το ε (σημειώνεται με την πορτοκαλί οριζόντια γραμμή) διαφέρει ανάλογα με την τιμή MinPts που έχει επιλεχθεί, καθώς μεγαλώνει η ακτίνα της γειτονιάς των στιγμιότυπων. Παράλληλα διαπιστώνεται πως οι χρόνοι υπολογισμού είναι και σε αυτό το σύνολο δεδομένων πολύ μικροί και παρόλο που αυξάνονται τα MinPts ο χρόνος υπολογισμού αυξομειώνεται.

Τιμή *MinPts*	Τιμή ε	Γωνία καμπύλης	Χρόνος υπολογισμού (sec)
3	34.54	84.07°	0.26
4	37.65	85.29°	0.15
5	37.58	5.04°	0.17

Πίνακας 4.4: Αποτελέσματα για MinPts 3 εως 5 του συνόλου δεδομένων WN



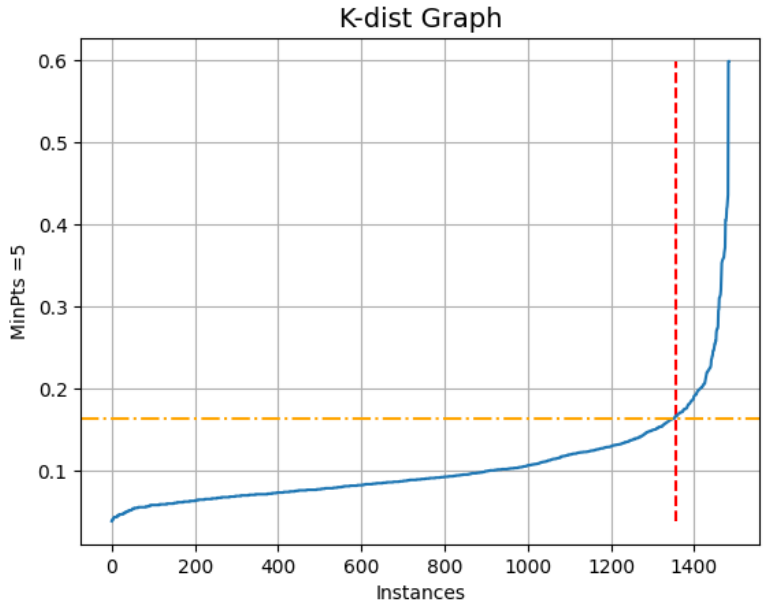
Σχήμα 4.3: Στιγμιότυπο αποτελεσμάτων για σύνολο δεδομένων WN

4.1.4 Yeast (YS)

Τα δεδομένα του συνόλου δεδομένων περιέχουν πληροφορίες σχετικά με ένα σύνολο κυττάρων ζύμης. Ο στόχος είναι να προσδιοριστεί η τοποθεσία κάθε κυττάρου ανάμεσα σε 10 πιθανές εναλλακτικές θέσεις, καθεμία από τις οποίες αντιπροσωπεύει μια κλάση, που καθορίζεται από ένα σύνολο 8 χαρακτηριστικών. Η εφαρμογή παράγει τα αποτελέσματα του πίνακα 4.5, για εύρος τιμών MinPts 3 εως 5. Στιγμιότυπο του αποτελέσματος απεικονίζεται στο σχήμα 4.4. Μελετώντας τα αποτελέσματα διαπιστώνουμε πως το σημείο που υπάρχει η μέγιστη καμπυλότητα (κατώφλι σημειώνεται με την κόκκινη κάθετη γραμμή) στο συγκεκριμένο σύνολο δεδομένων η τιμή της γωνίας παραμένει σχεδόν σταθερή όσο η τιμή των MinPts αυξάνεται. Η τιμή του ε (σημειώνεται με την πορτοκαλί οριζόντια γραμμή) παρατηρήθηκε ότι τείνει να παραμένει σταθερή παρόλου που αυξάνεται η τιμή των MinPts. Παράλληλα διαπιστώνεται πως οι χρόνοι υπολογισμού είναι και σε αυτό το σύνολο δεδομένων πολύ μικροί και παρόλο που αυξάνονται τα MinPts ο χρόνος υπολογισμού αυξομειώνεται.

Τιμή *MinPts*	Τιμή ε	Γωνία καμπύλης	Χρόνος υπολογισμού (sec)
3	0.13	42,51°	0.27
4	0.16	42,52°	0.15
5	0.16	42,52°	0.17

Πίνακας 4.5: Αποτελέσματα για MinPts 3 εως 5 του συνόλου δεδομένων YS



Σχήμα 4.4: Στιγμιότυπο αποτελεσμάτων για σύνολο δεδομένων YS

5 Δοκιμή εμπειρίας χρήστη

Η ”εμπειρία χρήστη” σαν όρος περιλαμβάνει όλες τις πτυχές της αλληλεπίδρασης του τελικού χρήστη με μια υπηρεσία, ένα προϊόν, μια ιστοσελίδα, μια εφαρμογή κ.α. Ο σχεδιασμός εμπειρίας χρήστη (UX) [35] είναι η διαδικασία σχεδιασμού που χρησιμοποιείται για τη δημιουργία προϊόντων που παρέχουν ουσιαστικές και σχετικές εμπειρίες στους χρήστες. Αυτό περιλαμβάνει το σχεδιασμό ολόκληρης της διαδικασίας απόκτησης και ολοκλήρωσης του προϊόντος, συμπεριλαμβανομένων των πτυχών της επωνυμίας, του σχεδιασμού, της χρηστικότητας και της λειτουργίας. Τέθηκε ως στόχο η εφαρμογή, που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής εργασίας, να προσφέρει μια ουσιαστική εμπειρία χρήσης στο τελικό χρήστη.

Η πιστοποίηση της επίτευξης του παραπάνω στόχου έγινε μέσω δοκιμής εμπειρίας χρήστη, με την χρήση ερωτηματολογίου. Το ερωτηματολόγιο περιέχει δέκα ερωτήσεις με πέντε επιλογές απόκρισης. Παρακάτω παρουσιάζεται το σύνολο των ερωτήσεων στην αγγλική γλώσσα.

1 I think that I would like to use this system frequently.

2 I found the system unnecessarily complex.

3 I thought the system was easy to use.

4 I think that I would need the support of a technical person to be able to use this system.

5 I found the various functions in this system were well integrated.

6 I thought there was too much inconsistency in this system.

7 I would imagine that most people would learn to use this system very quickly.

8 I found the system very cumbersome to use.

9 I felt very confident using the system.

10 I needed to learn a lot of things before I could get going with this system.

Σε κάθε απόκριση εκχωρείται μια τιμή για τον υπολογισμό της βαθμολογίας SUS. Η ανάλυση πόντων για τις απαντήσεις είναι:

- Strongly Disagree: 1 πόντος
- Disagree: 2 πόντοι
- Neutral: 3 πόντοι
- Agree: 4 πόντοι
- Strongly Agree: 5 πόντοι

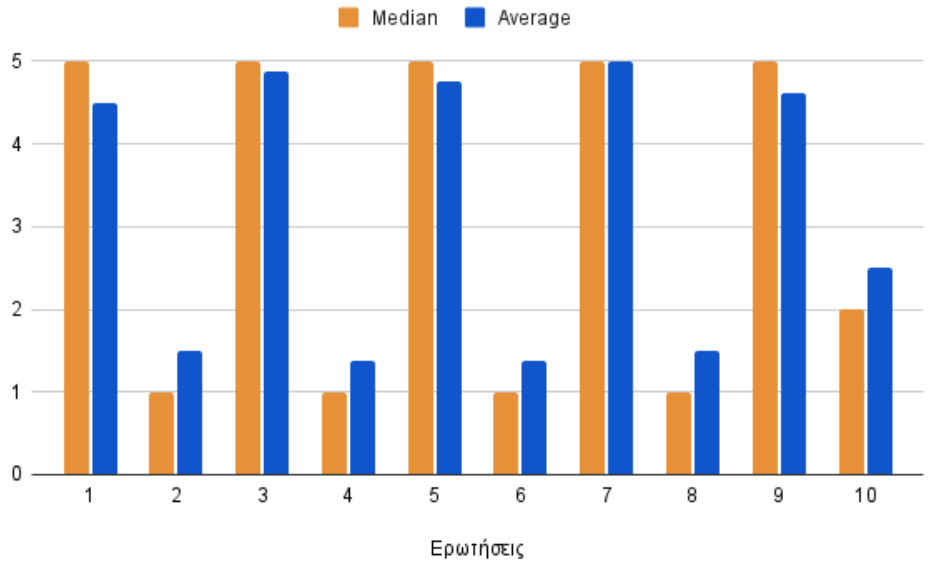
5.1 Αποτελέσμα έρευνας

Αρχικά υπήρξε η δυσκολία να βρεθούν εθελοντές για την έρευνα, καθώς προϋποθέτει την γνώση του αλγορίθμου DBSCAN. Τελικά επιλέχθηκε η αποστολή του ερωτηματολογίου σε 4 φοιτητές και 4 αποφοίτους πληροφορικής οι οποίοι γνωρίζουν τον αλγόριθμο DBSCAN και έχουν γνώση της δυσκολίας του προσδιορισμού των 2 παραμέτρων *MinPts* και *ε*. Συλλέγοντας τα αποτελέσματα του ερωτηματολογίου, υπήρχε η προσδοκία να επιτευχθεί μια βαθμολογία System Usability Scale πάνω από το 68 καθώς θεωρείτε ότι είναι πάνω από το μέσο όρο και οτιδήποτε κάτω από το 68 είναι κάτω από το μέσο όρο.

Ο μέσος όρος που έλαβε η εφαρμογή ανά ερώτηση παρουσίαζετια στο πίνακα 5.1. Αναλύοντας τα αποτελέσματα διαπιστώνεται ότι οι χρήστες στις ερωτήσεις με θετική χροιά (1,3,5,7,9) έδωσαν υψηλό σκορ με μέσο όρο 4,75. Στον αντίποδα οι ερωτήσεις με αρνητική χροιά (2,4,6,8,10) οι χρήστες έδωσαν χαμηλό σκορ και συσκεκριμένα μέσο όρο κάτω από 1,65. Στο σχήμα 5.1 παρουσάζεται και οπτικά ο παρακάτω πίνακας. Στο επόμενο βήμα υπολογίστηκε το SUS σκορ της εφαρμογής το οποίο είναι 88,75. το οποίο και παρουσιάζεται στο πίνακα 5.2.

Ερώτηση	**Μέσος όρος**	**Median**
Ερώτηση 1	4.5	5
Ερώτηση 2	1.5	1
Ερώτηση 3	4.875	5
Ερώτηση 4	1.375	1
Ερώτηση 5	4.75	5
Ερώτηση 6	1.375	1
Ερώτηση 7	5	5
Ερώτηση 8	1.5	1
Ερώτηση 9	4.625	5
Ερώτηση 10	2.5	2

Πίνακας 5.1: Μέσος όρος βαθμολογίας εφαρμογής ανα ερώτηση



Σχήμα 5.1: Γράφημα median και μέσου όρου ανα ερώτηση

User	SUS Raw Score	SUS Final Score
Χρήστης 1	31	77,5
Χρήστης 2	36	90
Χρήστης 3	36	90
Χρήστης 4	36	90
Χρήστης 5	40	100
Χρήστης 6	40	100
Χρήστης 7	32	80
Χρήστης 8	33	82,5
Μέσος Όρος	35,5	**88,75**

Πίνακας 5.2: Βαθμολογία εφαρμογής από το σύνολο των χρηστών

5.2 Συμεράσματα

Διαπιστώθηκε πως η πλειοψηφία των χρηστών έχει ευνοϊκή άποψη για την εφαρμογή, καθώς σίγουρα θα την χρησιμοποιούσε συχνά. Η πλειοψηφία των χρηστών ένιωσε σιγουριά και βρήκε πολύ εύκολη την χρήση της εφαρμογής. Διαμόρφωσαν θετική γνώμη για την λειτουργικότητά της.Επίσης οι περισσότεροι θεωρούν πως είναι πολύ εύκολο να μάθει κάποιος την χρήση της χωρίς να αντιμετωπίσει κάποια ιδιαίτερη δυσκολία. Το αποτέλεσμα που επετεύχθη στο SUS σκορ, κατατάσει την εφαρμογή, από πλευράς χρηστικότητας στο υψηλότερο επίπεδο, και δείχνει πως η εφαρμογή μπορεί να αποτελέσει ένα χρήσιμο και εύχρηστο εργαλείο για το χρήστη.

6 Συμπεράσματα και Μελλοντική Έρευνα

Υπάρχουν πολλές τεχνικές συσταδοποίησης συνόλων δεδομένων διαθέσιμες στη βιβλιογραφία και έχουν αναπτυχθεί οι αντίστοιχοι αλγόριθμοι. Ωστόσο, απαιτείται από την πλευρά του χρήστη η εισαγωγή παραμέτρων για την εξαγωγή ποιοτηκών αποτελεσμάτων από τους αλγορίθμους. Οι τιμές των παράμετρων δεν είναι τις περισσότερες φορές εύκολα εντοπίσημοι. Αρκετές φορές γίνονται διαισθητικά και άλλες φορές μετά από συνεχη πειραματισμό και έλεγχο. Ένας από αυτούς τους αλγορίθμους είναι και ο DBSCAN που ανήκει στην κατηγορία των αλγορίθμων που κάνουν συσταδοποίηση βάσει πυκνότητας. Αυτές οι παρατηρήσεις αποτέλεσαν το κίνητρο της παρούσας διπλωματικής εργασίας.

Στόχος της εργασίας ήταν η ανάπτυξη ενός εργαλείου που θα απλοποιεί και θα συντομεύει τη διαδικασία επιλογής των παραμέτρων *MinPts* και ε που απαιτείται για την εκτέλση του αλγορίθμου DBSCAN. Όπως ειπώθηκε στο κεφάλαιο 1.3 η διαδικασία εύρεσης των παραμέτρων είναι μία ευρεστική διαδικασία που στηρίζεται στην παραγωγή των k-dist graphs. Σε αυτά θα πρέπει να εντοπιστεί το σημείου του κατωφλίου, δηλαδή το σημείο που η γραφική παράσταση του k-dist graph δημιουργεί ένα «αγκώνα»,δηλαδή μια απότομη κλίση όπου το ζευγάρι MinPts και ε είναι η ιδανική επιλογή ως παράμετροι για τον DBSACAN. Παρόλου που τα γραφήματα αυτά βοηθάνε στον προσδιορισμό των παραμέτρων, η τιμή ε υπολογιζόταν πάντα με μία σχετική ακρίβεια. Επιπρόσθετος στόχος ήταν η χρήση μιάς μεθόδου που να μπορεί να εντοπίζει το ακριβές σημείου του κατωφλίου μία ακριβής τιμή για το ε.

Έτσι, η παρούσα εργασία παρουσίασε μια διαδυκτική εφαρμογή που δίνει την δυνατότητα στον χρήστη να ανεβάσει το σύνολο δεδομένων, για το οποίο επιθυμεί να λάβει τις τιμές των παραμέτρων για την εκτέλεση του DBSCAN. Επιλέγει τιμή ή ένα εύρος τιμών *MinPts* για το οποίο επιθυμεί την δημιουργία k-dist graphs και παράλληλα υπολογίζει με ακρίβεια το κατώφλι δηλαδή την τιμή ε.

Οι κατεύθύνσεις για μελλοντική επέκταση της διαδυκτιακής εφαρμογής περιλαμβάνουν την υποστήριξη περισσότερων μορφών συνόλων δεδομένων, που εξάγονται από άλλες εφαρμογές. Δημιουργία συστήματος διαχείρισης χρηστών. Ανάπτυξη βάσης δεδομένων που θα αποθηκεύει τα αποτελέσματα της επεξεργασίας, για άμεση ανάκληση από τους χρήστες. Τέλος η ανάπτυξη ενός API θα βοηθούσε στην έυκολη διασύνδεση του εργαλείου με άλλες εφαρμογές.

Βιβλιογραφία

[1] E. M., K. H-P., S. J., and X. X, "A density-based algorithm for discovering clusters in large spatial databases with noise," (Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 München, Germany), pp. 226–231, 1996.

[2] Google, "Machine Learning Crash Course - clustering algorithms." `https://developers.google.com/machine-learning/clustering/clustering-algorithms`. Accessed: 2020-06-01.

[3] G. Academic, "NCKU Data Mining: Clustering." `https://sejkai.gitbook.io/academic/ncku-data-mining/clustering`. Accessed: 2020-06-01.

[4] V. Agarwal, "Let's cluster data points using dbscan | by vibhor agarwal | medium," 2021.

[5] L. S. Marzena Kryszkiewicz, "Faster Clustering with DBSCAN," in *Intelligent Information Processing and Web Mining*, (Institute of Computer Science, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland), pp. 605 – 614, 2005.

[6] C. Sinclair, "Clustering Using OPTICS - a seemingly parameter-less algorithm." `https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7`, January 2019. Accessed: 2020-06-01.

[7] Z. Little, "Optics (ordering points to identify the clustering structure) | by z² little | medium," 2021.

[8] P. Berkhin, *A Survey of Clustering Data Mining Techniques*, pp. 25–71. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[9] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[10] S. Ougiaroglou and G. Evangelidis, "A simple noise-tolerant abstraction algorithm for fast k-nn classification," in *Hybrid Artificial Intelligent Systems* (E. Corchado, V. Snášel, A. Abraham, M. Woźniak, M. Graña, and S.-B. Cho, eds.), vol. 7209 of *Lecture Notes in Computer Science*, pp. 210–221, Springer Berlin Heidelberg, 2012.

[11] X. Xu, M. Ester, H. Kriegel, and J. Sander, "A distribution-based clustering algorithm for mining in large spatial databases," *Proceedings 14th International Conference on Data Engineering*, pp. 324–331, 1998.

[12] K.-L. Du, "Clustering: A neural network approach," *Neural networks : the official journal of the International Neural Network Society*, vol. 23, pp. 89–107, 08 2009.

[13] W.-K. Loh and Y.-H. Park, "A survey on density-based clustering algorithms," in *Ubiquitous Information Technologies and Applications* (Y.-S. Jeong, Y.-H. Park, C.-H. R. Hsu, and J. J. J. H. Park, eds.), (Berlin, Heidelberg), pp. 775–780, Springer Berlin Heidelberg, 2014.

[14] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, (New York, NY, USA), p. 49–60, Association for Computing Machinery, 1999.

[15] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: A survey approach," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, (New York, NY, USA), p. 719–728, Association for Computing Machinery, 2009.

[16] A. Mehdi, F. Betz, A. Dovgal, N. Lopes, H. Magnusson, G. Richter, D. Seguy, and J. Vrana, "PHP: PHP Manual - Manual."

[17] L. Wall, "The perl programming language," 01 2011.

[18] B. W. Kernighan and D. M. Ritchie, *The C programming language*. 2006.

[19] M. L. Hetland, "Extending python," *Beginning Python*, p. 321–336, 2017.

[20] Nov 2002.

[21] G. van Rossum, "Python tutorial," Tech. Rep. CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[22] R. Garreta and G. Moncecchi, *Learning Scikit-Learn: Machine Learning in Python*. Packt Publishing, 2013.

[23] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sept. 2020.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[25] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[26] L. Reznick, "Using cron and crontab," *Sys Admin*, vol. 2, no. 4, pp. 29–32, 1993.

[27] T. Berners-Lee and D. Connolly, "Hypertext markup language specification – 2.0 internet draft," tech. rep., November 1994.

[28] H. W. Lie, "Cascading html style sheets – a proposal," tech. rep., October 1994.

[29] H. W. Lie and B. Bos, "Cascading style sheets, level 1," tech. rep., 1996.

[30] B. Bos, "Simple style sheets for sgml $ html on the web," tech. rep., April 1995.

[31] D. Goodman, *JavaScript bible*. IDG Books Worldwide, 1998.

[32] R. York, *Beginning JavaScript and CSS Development with jQuery*. Wiley, 2009.

[33] J. J. Garrett, "Ajax: A new approach to web applications," 2007.

[34] G. Unterholzner, "Github - georg-un/kneebow: Knee or elbow detection for curves," August 2019.

[35] J. Brooke, *SUS – a quick and dirty usability scale*, pp. 189–194. 01 1996.

[36] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[37] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, p. 49–60, June 1999.

Αποσπάσματα Κώδικα και Αλγόριθμοι

Algorithm 18: Αλγόριθμος υπολογισμού k-dist graph και ε

```
# -*- coding: utf-8 -*-
import sys
import numpy as np
from numpy import genfromtxt
from sklearn.neighbors import NearestNeighbors
from matplotlib import pyplot as plt
import time
import os




#Anagnosi parametron apo php
min_fr = int(sys.argv[1]) #k min
dataset = str(sys.argv[2]) #csv file
symb = str(sys.argv[3]) #delimiter
ds_class_col = int(sys.argv[4]) # class column
ds_titles = int(sys.argv[5]) # titles
save_dir = str(sys.argv[6]) # save_dir
min_to = int(sys.argv[7]) #k max


#dir apothikeusis apotelesmaton
fp2 = "result\\"

degree_sign = u"\N{DEGREE SIGN}"

#an den yparxei to dir to dimiourgei
if not os.path.exists(fp2):
    os.makedirs(fp2)

#diavasma csv
my_data = genfromtxt(dataset, delimiter = symb)

#diagrafi stilis classification an yparxei
if not(ds_class_col == 0 or ds_class_col == '' or ds_class_col
    == ' '):
 my_data = np.delete(my_data, ds_class_col-1, 1)

#diagrafi headers an to dataset exei headers
if ds_titles == 1:
  my_data = np.delete(my_data, 0, 0)


X = my_data

def find_elbow(gonia, theta):
```

```
47
48      # make rotation matrix
49      co = np.cos(theta)
50      si = np.sin(theta)
51      rotation_matrix = np.array(((co, -si), (si, co)))
52
53      # rotate gonia vector
54      rotated_vector = gonia.dot(rotation_matrix)
55
56      # return index of elbow
57      return np.where(rotated_vector == rotated_vector.min())[0][0]
58
59
60  def get_data_radiant(gonia):
61    return np.arctan2(gonia[:, 1].max() - gonia[:, 1].min(),
62                      gonia[:, 0].max() - gonia[:, 0].min())
63
64
65  while min_fr <= min_to:
66    data_time = 0
67    start_time = time.time()
68    nbrs = NearestNeighbors(n_neighbors=min_fr,
          algorithm='auto').fit(X)
69    distances, indices = np.sort(nbrs.kneighbors(X))
70    new = np.sort(distances[:, min_fr-1])
71    f = np.arange(1, len(new)+1).astype(int)
72    x = np.array(new)
73    gonia = np.column_stack([f,x])
74    elbow_index = find_elbow(gonia, get_data_radiant(gonia))
75    epsillon = gonia[elbow_index,1].astype(float) #Timi tou
          epsillon
76    angle = np.rad2deg(get_data_radiant(gonia))
77  # Plotting
78    fig = plt.figure()
79    ax = plt.axes()
80    plt.grid(True)
81    ax.plot(new);
82    plt.title('K-dist Graph', fontsize=14)
83    plt.ylabel('MinPts =' + str(min_fr), fontsize=10);
84    plt.xlabel('Instances', fontsize=10);
85    plt.vlines(gonia[elbow_index, 0], ymin=gonia[:, 1].min(),
          ymax=gonia[:, 1].max(), colors='red', linestyles='--')
86    plt.axhline(y = epsillon, color = 'orange', linestyle = '-.')
87    plt.savefig(fp2 + "\\" + save_dir + "_eps_mins_" + str(min_fr)
          +".png", bbox_inches='tight')
88  # end plotting
89    end_time= time.time()
90    data_time = end_time - start_time
```

```
91  img_png=fp2 + "\\" + save_dir +"_eps_mins_" + str(min_fr)
        +".png"
92  if str(round(epsillon,2)) == '':
93      print('  value is null.')
94      exit()
95  else:
96      print(str(round(epsillon,2)))
97      print('Elbow angle: ' + str(round(angle,2)) + degree_sign)
98      print(str(round(data_time,2)))
99      print(img_png)
100     min_fr += 1
```

Algorithm 19: Αλγόριθμος επεξεργασίας και μετατροπής συνόλου δεδομένων σε csv

```
# -*- coding: utf-8 -*-
import sys

"""
Created on Sun Jun  6 13:54:34 2021

@author: info

"""

file_import = str(sys.argv[1]) #file

split_string = file_import.split(".", 1)
substring = split_string[0]

fh = open(file_import, "r")
lines = fh.readlines()
fh.close()

keep = []

for line in lines:
    line = line.replace('{','')
    line = line.replace('}','')
    if not line.isspace() and (not line.startswith('@')) and
        (not line.startswith('%%')):
          keep.append(line)

dir_s = substring + ".csv"
fh = open(dir_s, "w")
fh.write("".join(keep))
print (dir_s)
fh.close()
```

Algorithm 20: Αλγόριθμος διαγραφής αρχείων μετά από 24 ώρες

```php
<?php

//Delete results over 1 day.
$folderName = 'result';
if (file_exists($folderName)) {
    foreach (new DirectoryIterator($folderName) as $fileInfo) {
        if ($fileInfo->isDot()) {
        continue;
        }
        if ($fileInfo->isFile() && time() -
            $fileInfo->getCTime() >= 60 * 60 * 24) {
            unlink($fileInfo->getRealPath());
        }
    }
}

//Delete uploaded datasets over 1 day.
$folderName = 'datasets';
if (file_exists($folderName)) {
    foreach (new DirectoryIterator($folderName) as $fileInfo) {
        if ($fileInfo->isDot()) {
        continue;
        }
        if ($fileInfo->isFile() && time() -
            $fileInfo->getCTime() >= 60 * 60 * 24) {
            unlink($fileInfo->getRealPath());
        }
    }
}

?>
```

Algorithm 21: Αλγόριθμος διαχείρισης εφαρμογής

```php
<?php
require('fpdf/fpdf.php');

$input1fr = $input1to = $input2= $input3 = $input4 = $input5 =
    $input6  = $xid ="";

$xid = ltrim($_POST['session_id'], "PHPSESSID =");
if ($xid == '' || $xid == 'null'){
  session_start();
  $xid =  session_id();
}

if(isset($_FILES["input2"]) && $_FILES["input2"]["error"] == 0){

  $target_dir = "datasets/";
  $file = $_FILES['input2']['name'];
  $path = pathinfo($file);
  $filename = $xid . '_' . $path['filename'];
  $ext = $path['extension'];
  $temp_name = $_FILES['input2']['tmp_name'];
  $path_filename_ext = $target_dir.$filename.".".$ext;
  move_uploaded_file($temp_name,$path_filename_ext);

      if ($ext == 'arff' || $ext == 'txt' || $ext == 'dat' ||
          $ext == 'data'){
          $comm = escapeshellcmd('python fileconv.py ' .
              $path_filename_ext);
          exec($comm,$output_new, $ret_code);
          extract($output_new);
          $input2 = $output_new[0];
      }else{
          $input2 = $path_filename_ext; //Dataset name
       }

      echo "Your file was uploaded successfully.<br>";
}

function clear_png($xid){
  $fullPath = __DIR__ . "/result/" ;
  $del_file = $fullPath . $xid;
  array_map('unlink', glob( "$del_file*.png"));
}

function clear_data($temp_csv){
  $fullPathd = __DIR__ . "/";
  $da_file = $fullPathd . $temp_csv;
  array_map('unlink', glob( "$da_file*"));
```

```
45 }
46
47
48 $input1fr = $_POST['input1_from']; //kmax
49 $input1to = $_POST['input1_to']; //kmax
50 $input3 = $_POST['input3']; //delimiter of dataset
51 $input4 = $_POST['input4'];//Class column
52 $input5 = $_POST['input5'];//Dataset titles
53 $input6 = $_POST['input6'];// email
54
55 if ($input1fr > $input1to){
56     $swap = $input1fr;
57     $input1fr = $input1to;
58     $input1to = $swap;
59 }
60
61 $path = "result";
62
63 $pdf_name = $path . '/' . $xid . '_mins_' . $input1fr . '_to_' .
       $input1to . '.pdf';
64
65 $all=array();
66
67 if( $input1fr != $input1to ) {
68 $command = escapeshellcmd('python eps.py ' . $input1fr . ' ' .
       $input2 . ' ' . $input3 . ' ' . $input4 . ' ' . $input5 . ' '
       . $xid . ' ' . $input1to);
69     exec($command,$output, $ret_code);
70     array_push($all,$output);
71
72 $path = $output[4];
73 $pdf=new FPDF();
74 $pdf->SetAuthor('Alkibiadis Tzioras');
75     $pdf->SetTitle('k-dist graph WebApp');
76 $pdf->SetFont('Arial','B',18);
77
78 $loop = ($input1to-$input1fr);
79 $j = 0;
80 for ($x = 0; $x <= $loop; $x++) {
81   foreach ($all as $value) {
82
83     $eps = $value[$j];
84     $elbow = $value[++$j];
85     $time = $value[++$j];
86     $img = $value[++$j];
87
88     ++$j;
89   $count = $x + $input1fr;
```

```
90
91  //create a FPDF object
92      $pdf->AddPage();
93      $pdf->SetDisplayMode(100,'default');
94      $pdf->cell(100,14,"k-dist graph and eps for DBSCAN", 1, 0,
            'C');
95      $pdf->SetFontSize(11);
96      $pdf->SetXY(10,26);
97      $pdf->Write(5, "Dataset filename: " . $file);
98      $pdf->SetXY(10,31);
99      $pdf->Write(5, "Calculation for range " . $input1fr . " to "
            . $input1to . " min points.");
100     $pdf->SetXY(10,36);
101     $pdf->Write(5,
            "-------------------------------------------------------");
102     $pdf->SetFontSize(10);
103     $pdf->SetXY(10,41);
104     $pdf->Write(5,'Estimated epsilon value for ' . $count . " min
            points is ");
105     $pdf->SetTextColor(255, 0, 0);
106     $pdf->Write(5,$eps);
107     $pdf->SetTextColor(0,0,0);
108     $pdf->SetXY(10,46);
109     $pdf->Write(5,$elbow);
110     $pdf->SetXY(10,51);
111     $pdf->Write(5,"Process time : " . $time . " sec");
112     $pdf->Image($img,10,70,-100);
113     $pdf->SetFont('Arial','I',9);
114     $pdf->SetXY(92,250);
115     $pdf->Write(20,"Page : " . ($x + 1) . "/" . ($loop + 1));
116     $pdf->SetFont('Arial','B',18);
117          }
118     }
119
120   $pdf->Output($pdf_name, 'F');
121
122 if (file_exists($pdf_name)){
123      if (!empty($input6)){
124        require_once "./php/mail.php";
125        clear_png($xid);
126      //  clear_data($input2);
127        exit;
128      }else{
129        $actual_link = (isset($_SERVER['HTTPS']) &&
               $_SERVER['HTTPS'] === 'on' ? "https:" : "http:") .
               '//'.
               $_SERVER['HTTP_HOST'].dirname($_SERVER['PHP_SELF']);
130        $filename = $actual_link . $pdf_name;
```

```
131         echo '<a href="' . $pdf_name . '" target="_blank">Download
                results.</a>';
132         clear_png($xid);
133     //    clear_data($input2);
134         exit;
135         }
136     }
137   exit;
138 }else{
139     $command = escapeshellcmd('python eps.py ' . $input1fr . ' '
            . $input2 . ' ' . $input3 . ' ' . $input4 . ' ' . $input5
            . ' ' . $xid . ' ' . $input1to);
140
141     exec($command,$output, $ret_code);
142     extract($output);
143
144     $pdf=new FPDF();
145
146     $pdf->SetAuthor('Alkibiadis Tzioras');
147     $pdf->SetTitle('k-dist graph WebApp');
148     $pdf->SetFont('Arial','B',18);
149     $pdf->AddPage('P');
150     $pdf->SetDisplayMode(100,'default');
151     $pdf->cell(100,14,"k-dist graph and eps for DBSCAN", 1, 0,
            'C');
152     $pdf->SetFontSize(11);
153     $pdf->SetXY(10,26);
154     $pdf->Write(5, "Dataset filename: " . $file);
155     $pdf->SetXY(10,31);
156     $pdf->Write(5, "Min points: " . $input1fr);
157     $pdf->SetXY(10,36);
158     $pdf->Write(5,
            "-------------------------------------------------------");
159     $pdf->SetFontSize(10);
160     $pdf->SetXY(10,41);
161     $pdf->Write(5,'Estimated epsilon value is ');
162     $pdf->SetTextColor(255, 0, 0);
163     $pdf->Write(5,$output[0]);
164     $pdf->SetTextColor(0,0,0);
165     $pdf->SetXY(10,45);
166     $pdf->Write(5,$output[1]);
167     $pdf->SetXY(10,49);
168     $pdf->Write(5,"Process time: " . $output[2] . " sec");
169     $pdf->Image($output[3],10,70,-100);
170
171     //footer
172     $pdf->SetFont('Arial','I',9);
173     $pdf->SetXY(92,255);
```

```
174     $pdf->Write(20,"Page : 1/1" );
175
176
177 if (empty($input6)){
178     $actual_link = (isset($_SERVER['HTTPS']) &&
            $_SERVER['HTTPS'] === 'on' ? "https:" : "http:") . '//'.
            $_SERVER['HTTP_HOST'].dirname($_SERVER['PHP_SELF']);
179 $filename = $actual_link .'/' .$pdf_name;
180 }
181
182 $pdf->output($pdf_name, 'F');
183
184 if (file_exists($pdf_name)){
185     if (!empty($input6)){
186       require_once "./php/mail.php";
187       clear_png($xid);
188    //    clear_data($input2);
189       exit;
190     }else{
191
192 $actual_link = (isset($_SERVER['HTTPS']) && $_SERVER['HTTPS']
    === 'on' ? "https:" : "http:") . '//'.
    $_SERVER['HTTP_HOST'].dirname($_SERVER['PHP_SELF']);
193 $filename = $actual_link . $pdf_name;
194     echo '<a href="' . $pdf_name . '" target="_blank">Download
        results.</a>';
195     clear_png($xid);
196    // clear_data($input2);
197  exit;
198       }
199   }
200  exit;
201 }
202
203
204 ?>
```