ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ο ΑΛΓΟΡΙΘΜΟΣ K-MEANS ΣΕ PYTHON

Διπλωματική Εργασία

του

Κωνσταντίνου Τσολάκη

Θεσσαλονίκη, 6/2018

Ο ΑΛΓΟΡΙΘΜΟΣ K-MEANS ΣΕ PYTHON

Κωνσταντίνος Τσολάκης

Πτυχίο Ηλεκτρονικού Μηχανικού Τ.Ε., Α-ΤΕΙΘ, 2014

Διπλωματική Εργασία

υποβαλλόμενη για την πλήρη εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής
Νικόλαος Σαμαράς
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25/06/2018

Νικόλαος Σαμαράς Καθηγητής	Γεώργιος Ευαγγελίδης Καθηγητής	Γεωργία Κολωνιάρη Επίκουρη Καθηγήτρια
.....	.....	.....

Κωνσταντίνος Τσολάκης

.....

Περίληψη

Η Εξόρυξη Γνώσης από Δεδομένα είναι η εξεύρεση πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων. Ο αλγόριθμος K-Means, που μελετήθηκε εδώ, αποτελεί μία μέθοδο εξ αυτών για την επίτευξη του στόχου αυτού και συγκεκριμένα την μέθοδο της συσταδοποίησης σε *k* συστάδες. Η παρούσα διπλωματική εργασία, επιχειρεί την μελέτη και υλοποίηση αυτού του αλγορίθμου στην γλώσσα προγραμματισμού Python, μέσω του προγράμματος Spyder, ένα πρόγραμμα στα πρότυπα λογικής του Matlab. Παράλληλα έγινε μία προσπάθεια εφαρμογής αυτού του κώδικα σε πραγματικά δεδομένα και ανάλυση των αποτελεσμάτων των οποίων προέκυψαν.

Λέξεις Κλειδιά: *Εξόρυξη Γνώσης από Δεδομένα, Συσταδοποίηση, Μηχανική Μάθηση, Python, Spyder IDE.*

Abstract

Knowledge mining from Data is the method of extracting knowledge or finding patterns in big databases. The K-means algorithm, that was studied here, is one of these methods to accomplish that objective and more specifically to partition the dataset in *k* clusters. The present thesis, tries to study and implement the algorithm with programming language Python, through the Spyder program, a program that is very similar to Matlab. At the same time, an effort to implement the algorithm, test it on actual data and analyze the results that were found.

Keywords: *Data Mining, Knowledge mining, Clustering, Machine Learning, Python, Spyder IDE.*

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ 7
1.1 Πρόβλημα – Σημαντικότητα του θέματος.....7
1.2 Σκοπός της εργασίας.....9
1.3 Διάρθρωση της εργασίας.....10
2 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ.....11
2.1 Γενικά.....11
2.2 Μηχανική Μάθηση.....13
2.3 Συσταδοποίηση.....15
2.3.1 Ιεραρχική συσταδοποίηση.....18
2.3.2 Διαμεριστική διαμέριση.....21
2.4 Ο Αλγόριθμος k- means.....23
3 ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ…..…..……27
3.1 Υλοποίηση του αλγόριθμου k-means στην Python.....27
3.2 Αποτελέσματα.....29
3.2.1 Δεδομένα Wholesale.....29
3.2.2 Δεδομένα Sales.....35
3.2.3 Δεδομένα Google.....39
3.2.4 Δεδομένα Insurance.....42
3.3 Αξιολόγηση αλγόριθμου.....45
4 ΕΠΙΛΟΓΟΣ.....47
4.1 Σύνοψη και συμπεράσματα.....47
4.2 Μελλοντικές Επεκτάσεις.....47
5 ΒΙΒΛΙΟΓΡΑΦΙΑ.....48
5.1 Ελληνική.....48
5.2 Ξένη.....49
5.3 Ιστοσελίδες.....50
ΠΑΡΑΡΤΗΜΑΤΑ……..51
Παράρτημα 1: Κώδικας.....51
Παράρτημα 2: Δεδομένα Wholesale.....54
Παράρτημα 3: Δεδομένα Sales.....54
Παράρτημα 4: Δεδομένα Google.....54
Παράρτημα 5: Δεδομένα Insurance.....55

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

Εικόνα 1: Διάγραμμα ροής αλγόριθμου μηχανικής μάθησης (Γούλας, 2015).....15
Εικόνα 2: Λογική Ιεραρχικής και Διαιρετικής συσταδοποίησης (Χρυσός, 2006).....19
Εικόνα 3: Μέθοδοι Ιεραρχικής Συσταδοποίηση ς (Τσεργούλας, 2016).....20
Εικόνα 4: Σωστή συσταδοποίηση του k- means για εκτίμηση k= 3 (Zalik, 2008).....25
Εικόνα 5: Λάθος συσταδοποιήσεις του αλγόριθμου k-means για εκτιμήσεις k=1,2 και 4 (Zalik, 2008).....25
Εικόνα 6: Δεδομένα Wholesale, k=5.....30
Εικόνα 7: Δεδομένα Wholesale, k=5 (Μεγέθυνση).....30
Εικόνα 8: Δεδομένα Wholesale, k=10 (Μεγέθυνση).....31
Εικόνα 9: Δεδομένα Wholesale, k=10 (Μεγέθυνση).....31
Εικόνα 10: Δεδομένα Wholesale, k=20.....32
Εικόνα 11: Δεδομένα Wholesale, k=20 (Μεγέθυνση).....32
Εικόνα 12: Δεδομένα Wholesale, k=20 (Μεγέθυνση).....33
Εικόνα 13: Δεδομένα Wholesale, k=30.....33
Εικόνα 14: Δεδομένα Wholesale, k=30 (Μεγέθυνση).....34
Εικόνα 15: Δεδομένα Wholesale, k=50.....34
Εικόνα 16: Δεδομένα Wholesale, k=50 (Μεγέθυνση).....35
Εικόνα 17: Δεδομένα Sales , k=5.....36
Εικόνα 18: Δεδομένα Sales , k= 10.....36
Εικόνα 19: Δεδομένα Sales , k=10 (Μεγέθυνση).....37
Εικόνα 20: Δεδομένα Sales , k=25.....37
Εικόνα 21: Δεδομένα Sales , k=25 (Μεγέθυνση).....38
Εικόνα 22: Δεδομένα Sales , k=50.....38
Εικόνα 23: Δεδομένα Sales , k=50 (Μεγέθυνση).....39
Εικόνα 24: Δεδομένα Google , k=5.....40
Εικόνα 25: Δεδομένα Google , k=10.....40
Εικόνα 26: Δεδομένα Google , k=15.....41
Εικόνα 27: Δεδομένα Google , k=15 (Μεγέθυνση).....41
Εικόνα 28: Δεδομένα Insurance , k=5.....42
Εικόνα 29: Δεδομένα Insurance , k=10.....43
Εικόνα 30: Δεδομένα Insurance , k=25.....44
Εικόνα 31: Δεδομένα Insurance , k=25 (Μεγέθυνση).....44
Εικόνα 32: Δεδομένα Insurance , k=50.....45
Εικόνα 33: Δεδομένα Insurance , k=50 (Μεγέθυνση).....45

1 ΕΙΣΑΓΩΓΗ

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Στη σύγχρονη εποχή η πρόοδος της τεχνολογίας (βάσεις δεδομένων, πληροφοριακά συστήματα, διαδίκτυο) έχει οδηγήσει σε τεράστια αύξηση των διαθέσιμων δεδομένων, τα οποία μια επιχείρηση καλείται να αναλύσει προκειμένου να λάβει κατάλληλες αποφάσεις. Ωστόσο, ο όγκος, η ταχύτητα και η λεπτομέρεια συλλογής των δεδομένων αποτελεί τροχοπέδη στη λήψη αποφάσεων, αν αυτή δεν υποστηρίζεται από ένα κατάλληλο σύστημα ανάλυσης αυτών. Τέτοια συστήματα και τεχνικές (Συστήματα Υποστήριξης Αποφάσεων, Συστήματα Προγραμματισμού Επιχειρησιακών Πόρων, Αποθήκες Δεδομένων, Εξόρυξη Γνώσης από Δεδομένα) δίνουν τη δυνατότητα στις επιχειρήσεις που αναπτύσσουν συστήματα Επιχειρηματικής Ευφυΐας να δημιουργούν ολοένα και πιο αποτελεσματικά εργαλεία λήψης αποφάσεων. Στο πλαίσιο αυτό, σημαντικό ρόλο παίζει η διαδικασία της εξόρυξης γνώσης από δεδομένα, που συνίσταται σε αυτοματοποιημένη, μέσω μαθηματικών μοντέλων, αναζήτηση χρήσιμων πληροφοριών μέσα σε ένα μεγάλο πλήθος δεδομένων (Olson, 2008). Η τεχνική της εξόρυξης γνώσης απόδεδομένα δεν αφορά μόνο τις επιχειρήσεις, αλλά χρησιμοποιείται και σε κάθε τομέα όπου εμπλέκονται μεγάλες ποσότητες δεδομένων, όπως π.χ. στις εκλογικές διαδικασίες και στο χώρο της ιατρικής. Στον πίνακα 1 φαίνονται μερικές από τις περιοχές εφαρμογής της Εξόρυξης Γνώσης από Δεδομένα.

Πίνακας 1: Περιοχές εφαρμογής της Εξόρυξης Γώσης από Δεδομένα (Olson, 2008)

Περιοχή Εφαρμογής	**Εφαρμογές**	**Λεπτομέρειες**
Λιανικό Εμπόριο	Συσχέτιση Θέσης (Affinity Positioning)	Αποτελεσματική τοποθέτηση προϊόντων
	Διασταύρωση πωλήσεων (Cross- Selling)	Εύρεση περισσότερων προϊόντων για τους πελάτες
Τράπεζες	Διαχειριση πελατών	Αναγνώριση της αξίας του πελάτη
		Ανάπτυξη προγραμμάτων αύξησης εσόδων

Διαχείριση πιστωτικών καρτών	Ανάπτυξη (Lift)	Προσδιορισμός αποτελεσματικών περιοχών της αγοράς
	Απώλεια (Churn)	Προσδιορισμός πιθανών εσόδων των πελατών
Ασφάλειες	Ανίσχνευση απάτης	Προσδιορισμός αξιώσεων που χρήζουν διερεύνησης
Τηλεπικοινωνίες	Απώλεια (Churn)	Προσδιορισμός πιθανών εσόδων των πελατών
Τηλεπωλήσεις	On line πληροφορίες	Διευκόλυνση τηλεπωλητών με εύκολη πρόσβαση στα δεδομένα
Διαχείριση προσωπικού	Απώλεια (Churn)	Προσδιορισμός πιθανής δυνατότητας κινητικότητας και αντικατάστασης υπαλλήλων

Ο όρος *Εξόρυξη Γνώσης από Δεδομένα* είναι μία έννοια που συνήθως παραπέμπει σε κάθε είδος φόρμας με μεγάλη ποσότητα δεδομένων ή επεξεργασία δεδομένων (συλλογή, εξαγωγή δεδομένων, warehouse, ανάλυση δεδομένων και στατιστικής) αλλά επίσης γενικεύεται σε κάθε είδος συστήματος υποστήριξης αποφάσεων συμπεριλαμβανομένου της τεχνητής νοημοσύνης, της εκμάθησης μηχανής και της επιχειρηματικής ευφυΐας. Στην ορθή χρήση του όρου η λέξη κλειδί είναι η ανακάλυψη, που ορίζεται ως η ανίχνευση κάτι καινούριου.

Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (*Συσταδοποίηση*), ασυνήθιστες εγγραφές (*Anomaly Detection*) και εξαρτήσεις (Κ*ανόνες Συσχετίσεων*). Αυτό συνήθως συμπεριλαμβάνει τη χρήση βάσης δεδομένων όπως χωρικά ευρετήρια. Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων και η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα.

Στις διάφορες μεθόδους Εξόρυξης Γνώσης από Δεδομένα συγκαταλέγονται αυτές των Κανόνων Συσχέτισης (Association Rules), της Κατηγοριοποίησης (Classification), της Συσσταδοποίησης (Clustering), των Προβλέψεων (Predictions), των Διαδοχικών Προτύπων (Sequential Patterns) και των Όμοιων Χρονικών Ακολουθιών (Similar Time Sequences) (Olson, 2008). Στην παρούσα εργασία θα εστιάσουμε σε μια ειδική μεθοδολογία ανάλυσης που ανήκει στην τεχνική τη Συσταδοποίησης.

1.2 Σκοπός της εργασίας

Η παρούσα εργασία εντάσσεται στο πλαίσιο της τεχνικής Εξόρυξης γνώσης από Δεδομένα που ονομάζεται συσταδοποίηση (clustering). Η εν λόγω μέθοδος αφορά στη στατιστική ανάλυση και καταηγοριοποίηση όμοιων ή συσχετιζόμενων αντικειμένων σε υποσύνολα/ομάδες (clusters) έτσι ώστε αυτά να μοιράζονται κοινά χαρακτηριστικά (Δημητρακοπούλου, 2007). Η μέθοδος μπορεί να βρει εφαρμογή σε πλήθος πεδίων όπου απαιτείται ανάλυση/ συσταδοποίηση μεγάλου όγκου δεδομένων, όπως (Πιτουρά, 2011):

- γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία,
- weblog για εύρεση παρόμοιων προτύπων προσπέλασης,
- σχετιζόμενων αρχείων για browsing,
- κειμένων,
- πελατών με παρόμοια χαρακτηριστικά,
- εικόνων,
- χαρακτηριστικών ασθενειών,
- μετοχών με παρόμοια διακύμανση τιμών.

Ειδικότερα, η εργασία εστιάζει σε έναν από τους αλγόριθμους συσταδοποίησης, τον γνωστό αλγόριθμο k-means, ο οποίος παρότι παρουσιάστηκε πριν περίπου 50 χρόνια, παραμένει ο πιο διαδεδομένος λόγω της απλότητας, της ευκολίας στην εφαρμογή και της αποτελεσματικότητας που τον διακρίνουν (Jain, 2010). Ο εν λόγω αλγόριθμος υλοποιήθηκε με χρήση της γλώσσας προγραμματισμού Python (βλ. www.python.org).

1.3 Διάρθρωση της εργασίας

Στο δεύτερο κεφάλαιο παρουσιάζεται η τεχνική της εξόρυξης Γνώσης από δεδομένα, γίνεται μια σύντομη ανάλυση- παρουσίαση μεθόδων εξόρυξης όπως Κανόνες Συσχέτισης (Association Rules), του προβλήματος της Κατηγοριοποίησης *(Classification)* κ.α., και η επιμέρους μέθοδος της Συσταδοποίησης (*Clustering*), με την οποία ασχολείται η συγκεκριμένη εργασία, ενώ αναλύεται ο αλγόριθμος K-Means.

Στο τρίτο κεφάλαιο παρουσιάζεται ο αλγόριθμος, όπως υλοποιήθηκε στη γλώσσα προγραμματισμού Python. Επίσης γίνεται εφαρμογή του αλγορίθμου σε πραγματικά δεδομένα και αξιολογείται ο αλγόριθμος.

Στο τέταρτο κεφάλαιο αναλύονται τα προβλήματα που παρουσιάστηκαν κατά την υλοποίηση του υπό εξέταση αλγόριθμου και τα συμπεράσματα που προέκυψαν από τη χρήση του. Επιπλέον αναλύονται οι περιορισμοί της μεθόδου και προτείνονται τρόποι μελλοντικής επέκτασης της εργασίας.

2 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

2.1 Γενικά

Η ταχεία ανάπτυξη των ηλεκτρονικών υπολογιστών, του διαδικτύου και των εργαλείων μηχανοργάνωσης έχει οδηγήσει στην εκρηκτική ανάπτυξη του όγκου των διαθέσιμων δεδομένων. Σε παγκόσμιο επίπεδο, οι επιχειρήσεις και κάθε είδους επιστημονική και κοινωνική πρακτική παράγει, καθημερινά, τεράστιες ποσότητες δεδομένων από συναλλαγές πωλήσεων και μετοχών μέχρι αποτελέσματα επιστημονικών πειραμάτων και μετρήσεις των παραμέτρων υγείας των ασθενών. Στο πλαίσιο αυτό, έχει γίνει επιτακτική η ανάγκη για αναγνώριση και ανάκτηση του μέρους των δεδομένων που μπορεί να χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων. Έτσι, η συλλογή δεδομένων οδηγεί σε παραγωγή γνώσης.

Ο όρος 'Εξόρυξη Γνώσης από Δεδομένα' (Knowledge mining from Data) αναφέρεται στο σύνολο των τεχνικών που χρησιμοποιούνται για την ανακάλυψη προτύπων, μέσα στα δεδομένα, που έχουν ενδιαφέρουν και μπορούν να προσφέρουν χρήσιμη γνώση. Από τον ορισμό είναι προφανές ότι η ονομασία είναι ατελής και ορθότερα θα έπρεπε να μιλάει κανείς για 'Εξόρυξη Γνώσης'.

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτές της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα Νευρωνικά Δίκτυα, η Συσταδοποίηση, οι Γενετικοί Αλγόριθμοι(1950), τα Δέντρα Απόφασης (1960) και η Μηχανή Υποστήριξης Διανυσμάτων(1990). Η Εξόρυξη Γνώσης από Δεδομένα είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων. Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με την διαχείριση βάσεων δεδομένων κάνοντας

χρήση του τρόπου με τον οποίο αποθηκεύονται και κατατάσσονται στη βάση δεδομένων για να εκτελέσουν την θεωρία και τους διαθέσιμους αλγορίθμους περισσότερο αποτελεσματικά, επιτρέποντας σε τέτοιες μεθόδους να εφαρμόζονται σε μεγάλα σύνολα δεδομένων.

Κάποιες εκ των μεθοδων αυτών ειναι οι παρακάτω:

- **Κανόνες Συσχέτισης** (Association Rules). Οι Κανόνες Συσχέτισης είναι μια μέθοδος της Μηχανικής Μάθησης η οποία είναι βασισμένη σε κανόνες, και σκοπός της να ανακαλύπτει ενδιαφέρουσες συσχετίσεις μεταξύ μεταβλητών σε μεγάλες βάσεις δεδομένων. Στοχεύει να αναγνωρίσει δυνατούς κανόνες οι οποίοι ανακαλύπτονται μέσα στις βάσεις δεδομένων κρίνοντας κατά πόσο είναι ενδιαφέροντες.
- **Κατηγοριοποίηση** (Classification). Στη Μηχανική Μάθηση και την Στατιστική, Κατηριοποίηση είναι το πρόβλημα της αναγνώρισης σε ποιο σετ κατηγοριών ανήκει μια καινούρια παρατήρηση, βάσει ενός σετ δεδομένων εκπαίδευσης το οποίο περιέχει παρατηρήσεις οι οποίες ξέρουμε εκ των προτέρων σε ποια κατηγορία ανήκουν. Η Κατηγοριοποίση είναι ένα παράδειγμα *Αναγνώρισης Προτύπων*.
- **Δομημένης Πρόβλεψης** (Structured Prediction). Η Δομημένη Πρόβλεψη, ή απλά Πρόβλεψη, είναι ένας τίτλος "ομπρέλα" για τιε καθοδηγούμενες τεχνικές Μηχανικής Μάθησης, οι οποίες συμπεριλαμβάνουν προβλέψημα βαθμωτά αντικείμενα, αντί για βαθμωτούς ακεράιους ή φυσικούς αριθμούς.
- **Εξόρυξη Διαδοχικών Προτύπων** (Sequential Pattern Mining). Η Εξόρυξη Διαδοχικών Προτύπων ασχολείται με το να βρίσκει στατιστικά συσχετιζόμενα πρότυπα μεταξύ δεδομένων στα οποία οι τιμές παρουσίαζονται με μια διαδοχή (sequence) και συχνά οι τιμές αυτές θεωρούνται ακέραιες. Η Εξόρυξη Διαδοχικών Προτύπων έιναι μια ειδική περίπτωση της Δομημένης Εξόρυξης Γνώσης απο Δεδομένα.
- **Συσταδοποίησης** (Clustering). Η Συσταδοποίηση είναι διαδικάσια ομαδοποίησης των αντικειμένων με τέτοιων τρόπο έτσι ώστε τα αντικείμενα της ίδιας συστάδας να έχουνε περισσότερα κοινά χαρακτηρισρικά μεταξύ τους παρά με άλλες συστάδες.

Με την τελευταία θα ασχοληθεί αυτή η εργασία.

Έτσι, ο όρος ‘Εξόρυξη Γνώσης από Δεδομένα’ συχνά συμπεριλαμβάνεται στη διαδικασία ‘Ανακάλυψη Γνώσης από Δεδομένα’ (Knowledge discovery from data ή KDD) και η οποία αποτελείται από τα κάτωθι στάδια (Han, 2012):

1. Καθαρισμός δεδομένων για την εξάλειψη του θορύβου (μη συναφών δεδομένων),
2. Ενοποίηση δεδομένων, όπου δεδομένα συνενώνονται από διάφορες πηγές,
3. Επιλογή δεδομένων, ώστε να χρησιμοποιηθούν μόνο αυτά που έχουν σχέση με την ανάλυση,
4. Μετασχηματισμός δεδομένων, όπου τα δεδομένα μετατρέπονται σε, κατάλληλες για εξόρυξη Γνώσης, μορφές, διενεργώντας πράξεις περίληψης ή συγκέντρωσης,
5. Εξόρυξη δεδομένων, όπου εξάγονται πρότυπα δεδομένων με χρήση αυτοματοποιημένων μεθόδων,
6. Αξιολόγηση προτύπων, με βάση κάποιο κριτήριο, για να προσδιοριστούν αυτά που μπορεί να έχουν όφελος,
7. Παρουσίαση της γνώσης, με χρήση τεχνικών απεικόνισης.

Όπως φαίνεται παραπάνω ότι η Εξόρυξη Γνώσης απο Δεδομένα αποτελεί ένα μόνο βήμα στη διαδικασία ανακάλυψης γνώσης. Παρόλα αυτά, ο όρος συνηθίζεται να χρησιμοποιείται για το σύνολο της διαδικασίας.

Η Εξόρυξη Γνώσης απο Δεδομένα υιοθετεί ένα πλήθος τεχνικών και μεθόδων, από διάφορα πεδία, όπως Στατιστική (Statistics), Μηχανική Μάθηση (Machine Learning), Αναγνώριση Προτύπων (Pattern Recognition), Οπτικοποίηση (Visualization), Υπολογιστική Υψηλής Απόδοσης (High Performance Computing), Αποθήκες Δεδομένων (Data Warehouse) κ.α.

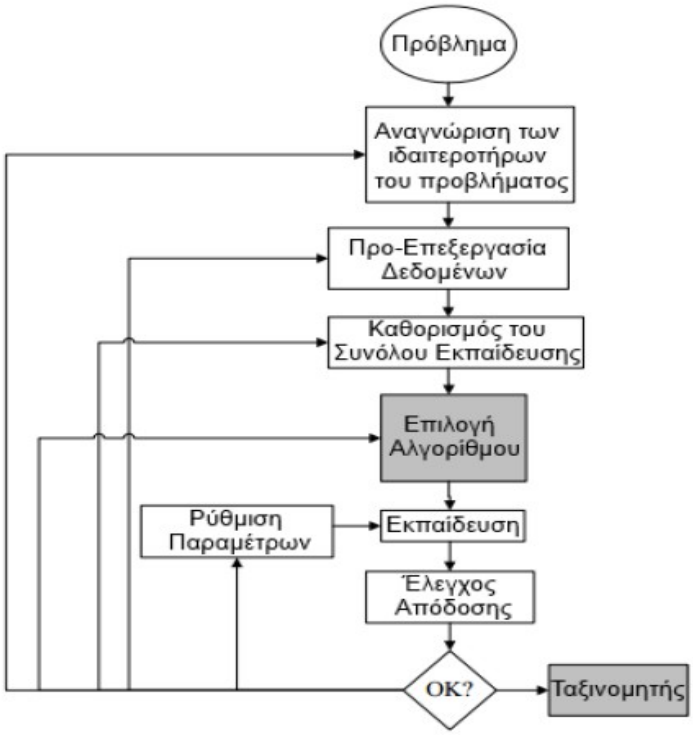
2.2 Μηχανική Μάθηση

Ο όρος ‘Μηχανική Μάθηση’ αναφέρεται στον τρόπο με τον οποίο, μέσω προγραμμάτων ηλεκτρονικών υπολογιστών, αναγνωρίζονται περίπλοκα πρότυπα μέσα στα δεδομένα και να λάβουν έξυπνες αποφάσεις. Το πεδίο περιλαμβάνει διάφορες εκδοχές, όπως (Han, 2012):

- Η καθοδηγούμενη μάθηση (Supervised learning), που στην ουσία είναι συνώνυμη της κατηγοριοποίησης. Στην περίπτωση αυτή, μηχανικά αναγνωρίσιμα πρότυπα χρησιμοποιούνται ως παραδείγματα εκπαίδευσης. Το μαθησιακό μοντέλο χρησιμοποιεί τα επισημασμένα αυτά πρότυπα και μαθαίνει να κατηγοριοποιεί τα δεδομένα που δέχεται στην είσοδο.Οι Αλγόριθμοι καθοδυγούμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει νέα παραδείγματα. Το βέλτιστο σενάριο επιτρέπει στον αλγόριθμο να καθορίσει σωστά την ετικέτα της κατηγορίας για άγνωστα μέχρι τώρα παραδείγματα. Για να επιτευχθεί αυτό, απαιτείται ο αλγόριθμος μάθησης να γενικεύει από τα δεδομένα εκπαίδευσης σε αθέατες καταστάσεις με ένα "λογικό" τρόπο. Παραρείγματα τέτοιων αλγορόθμων είναι: τα Νευρωνικά Δίκτυα, η Λογιστική Παλινδρόμηση, η Γραμμική Παλινδρόμηση, τα Δέντρα Αποφάσεων κ.α.
- Η μη καθοδηγούμενη μάθηση (Unsupervised learning). Είναι ουσιαστικά συνώνυμο της συσταδοποίησης. Στην εν λόγω διαδικασία μάθησης, τα δεδομένα εισόδου δεν φέρουν κάποιου είδους αναγνωριστικό. Έτσι, το μοντέλο δεν εποπτεύεται και ανακαλύπτει μόνο του πρότυπα (ομάδες) μέσα στα δεδομένα, χωρίς αυτές (σ.σ. οι ομάδες) να έχουν κάποια σημασιολογική σημασία. Καθώς τα δεδομένα προς διερεύνηση από τον αλγόριθμο μάθησης δεν είναι κατηγοριοποιημένα, δεν υπάρχει ξεκάθαρος τρόπος να αξιολογηθεί η ακρίβειά του στις δομές στις οποίες παράγει, ένα χαρακτηριστικό που διαχωρίζει τη μη καθοδηγούμενη μάθηση από την καθοδηγούμενη. Παραδείγματα τέτοιων αλγορίθμων είναι: η Ομαδοποίηση, η Ανίχνευση Ανωμαλιών, τα Νευρωνικά Δίκτυα κ.α.
- Η ημι-καθοδηγούμενη μάθηση (Semi-supervised learning). Αφορά σε τεχνικές μάθησης που κάνουν χρήση τόσο επισημασμένων όσο και μη προτύπων. Συνήθως χρησιμοποιείται σε προβλήματα δύο κατηγοριών προκειμένου να τελειοποιηθούν τα όρια μεταξύ τους. Αν, για παράδειγμα, η μία κατηγορία αφορά σε ‘θετικά δείγματα’ και η άλλη σε ‘αρνητικά δείγματα’, μπορούμε να χρησιμοποιήσουμε τα επισημασμένα πρότυπα (εποπτευόμενη μάθηση) για να οριστεί καλύτερα το όριο απόφασης για το αν ένα δεδομένο εισόδου ανήκει στη μία ή την άλλη κατηγορία.

- Η ενεργή μάθηση (Active learning) είναι ένα μαθησιακό μοντέλο, όπου οι χρήστες διαδραματίζουν ενεργό ρόλο στη διαδικασία. Για παράδειγμα, ένας εμπειρογνώμονας μπορεί να επισημάνει ένα πρότυπο παράδειγμα, που μπορεί να προέρχεται από μια διαδικασία εκαθοδηγούμενης ή μη μάθησης. Με τον τρόπο αυτό βελτιώνεται η ποιότητα του μαθησιακού μοντέλου.

Η διαδικασία της μηχανικής μάθησης μπορεί να καταγραφεί με το παρακάτω διάγραμμα ροής:



Εικόνα 1: Διάγραμμα ροής αλγόριθμου μηχανικής μάθησης (Γούλας, 2015)

Όπως έχει αναφερθεί, η παρούσα εργασία εντάσσεται στο πεδίο της μη καθοδηγούμενης μάθησης.

2.3 Συσταδοποίηση

Συνηθέστερα χρησιμοποιούμενη μέθοδος μη καθοδηγούμενης μάθησης αποτελεί αυτή της Συσταδοποίησης, η οποία αφορά στο διαχωρισμό ενός συνόλου αντικειμένων σε επιμέρους υποσύνολα (ομάδες) με βάση κάποιο κοινό χαρακτηριστικό ή σχέση. (Αφεντουλίδης, 2015). Η συσταδοποίηση είναι στην ουσία

μια διερευνητική τεχνική με σκοπό την ανεύρεση δομής σε ένα πλήθος δεδομένων. Διαφοροποιείται από την κατηγοριοποίηση καθώς δεν χρησιμοποιεί κάποιο προγενέστερο κριτήριο- αναγνωριστικό (Jain, 2010).

Προκειμένου να είναι δυνατή η μαθηματική μοντελοποίηση της μεθόδου, τα διάφορα αντικείμενα αναπαρίστανται ως διανύσματα ν διαστάσεων σε Ευκλείδειο ή μη χώρο. Κάθε διάσταση αφορά σε ένα συγκεκριμένο χαρακτηριστικό ή ιδιότητα του αντικειμένου. Η ομοιότητα, τότε, των αντικειμένων- διανυσμάτων εξετάζεται με τη χρήση κάποιου κριτηρίου. Στην περίπτωση που η αναπαράσταση αφορά τον Ευκλείδειο χώρο, παραδείγματα τέτοιων κριτηρίων είναι (Τσεργούλας, 2016):

- ✓ Η Ευκλείδεια απόσταση

$$L_2 - norm = \left(\sum_{k=1}^{D} \left| x_k - y_k \right|^2 \right)^{1/2}$$

- ✓ Η απόσταση Manhattan $L_1 - norm = \sum_{k=1}^{D} \left| x_k - y_k \right|$

- ✓ Η απόσταση Chebuchev $L\infty - norm = \max_{1 \le k \le D} \left| x_k - y_k \right|$

- ✓ Η απόσταση Minkowski $L_k = \left(\sum_{l=1}^{D} \left| x_l - y_l \right|^k \right)^{1/k}$

Η μέθοδοι συσταδοποίησης κατηγοριοποιούνται με διάφορους τρόπους, ανάλογα με τα χαρακτηριστικά των αντικειμένων ή/ και το σκοπό εφαρμογής της εκάστοτε μεθόδου. Έτσι, αυτές διαχωρίζονται ανάλογα με (Αφεντουλίδης, 2015):

- την εξάρτηση των ομάδων (επίπεδη έναντι ιεραρχικής)

Η επίπεδη συσταδοποίηση αφορά σε πλήρως ανεξάρτητες ομάδες και στόχο έχει την ανάδειξη ομοιοτήτων των αντικειμένων σε κάθε ομάδα ξεχωριστά. Αντίθετα, στην ιεραρχική, οι ομάδες συνδέονται σε διάταξη δενδρογράμματος, με αυτή της βάσης να περιγράφει πιο γενικά χαρακτηριστικά των αντικειμένων, ενώ τις ενδιάμεσες να είναι πιο εξειδικευμένες.

- την καθολικότητα της μεθόδου (ολική έναντι μερικής)

Σκοπός της ολικής συσταδοποίησης είναι η ανάθεση όλων των αντικειμένων σε κάποια ομάδα, σε αντίθεση με την μερική, όπου είναι πιθανό (ή αποτελεί εξαρχής επιλογή) κάποια από τα αντικείμενα να είναι απομονωμένα και εκτός κάποιας ομάδας.

- το είδος των ομάδων

Κριτήριο διαφοροποίησης είναι το πότε μια συγκέντρωση αντικειμένων αποτελεί ομάδα. Έτσι, έχουμε:

- ομάδες οι οποίες δημιουργούνται από τη συγκέντρωση των αντικειμένων γύρω από στοιχεία, που κατά μία έννοια αντιπροσωπεύουν κάθε ομάδα (prototype- based),
- ομάδες των οποίων τα αντικείμενα συγκεντρώνονται πιο κοντά σε ένα τουλάχιστον αντικείμενο της ίδιας ομάδας, παρά σε οποιαδήποτε άλλο κάποιας άλλης (continuity-based).
- ομάδες που χαρακτηρίζονται από υψηλή ή μη πυκνότητα αντικειμένων (density- based).

Τα πεδία εφαρμογής των μεθόδων συσταδοποίησης ποικίλλουν και συμπεριλαμβάνουν:

- Στο marketing, για την ανάλυση της αγοραστικής συμπεριφοράς των πελατών και την ανεύρεση πρότυπων συμπεριφοράς
- Στην ιατρική, για την κατηγοριοποίηση των ασθενών ανάλογα με την κατάσταση της υγείας τους, τον εντοπισμό ασθενειών ή μολύνσεων
- Στη βιολογία, για την συσταδοποίηση φυτών και ζώων με παρόμοια χαρακτηριστικά
- Στην αστρονομία, για την κατηγοριοποίηση των ουράνιων σωμάτων
- Στις ασφάλειες, για την ανεύρεση ασφαλιζόμενων με παρόμοια χαρακτηριστικά

- Στο διαδίκτυο για την ανάλυση της συμπεριφοράς των χρηστών, την κατηγοριοποίηση των εγγράφων κ.α.

Στα πλεονεκτήματα του αλγορίθμων περιλαμβάνονται (Ακακιάδου, 2007):

- Παράγουν λύσεις σε σχετικά μικρό χρόνο,
- Δεν απαιτούν την εκ των προτέρων εκτίμηση του αριθμού των ομάδων,
- Παρουσιάζουν ευελιξία, καθώς δίνουν τη δυνατότητα στο χρήστη να επιλέξει το επίπεδο τομής του δενδρογράμματος,
- Χειρίζονται με ευκολία κάθε κριτήριο ομοιότητα,
- Κατά τη διάρκεια της εκτέλεσης τους, είναι εύκολο να οπτικοποιηθούν τα αποτελέσματα που παράγουν,
- Μπορούν να χειριστεί πολλούς τύπους δεδομένων (π.χ. σε μορφή αλυσίδας, ομόκεντρα).

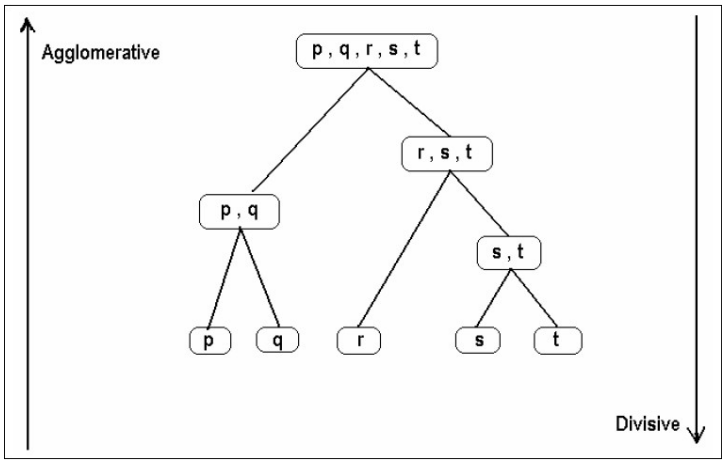
Ωστόσο, εμφανίζουν και αρκετά μειονεκτήματα, όπως (Ακακιάδου, 2007):

- Τα κριτήρια τερματισμού δεν ορίζονται με ακρίβεια,
- Δεν βελτιώνουν τις ενδιάμεσες ομάδες που δημιουργούν,
- Έχει μεγαλύτερες απαιτήσεις σε υπολογιστική ισχύ και μνήμη σε σχέση με τους αλγόριθμους διαμέρισης,
- Υπολογίζουν τοπικά βέλτιστες λύσεις και δεν εγγυώνται την εύρεση καθολικά βέλτιστης λύσης (Τσεργούλας, 2016)

2.3.1 Ιεραρχική συσταδοποίηση

Η ιεραρχική συσταδοποίηση (Hierarchical Clustering) δημιουργεί ένα δενδρόγραμμα αντικειμένων που υποδεικνύει τον τρόπο που συνδέονται οι ομάδες μεταξύ τους (Εικόνα 2). Οι ομάδες παράγονται, "κόβοντας" το δενδρόγραμμα στο επιθυμητό οριζόντιο επίπεδο. Το δενδρόγραμμα δημιουργείται με δύο τρόπους: συσσωρευτικά (agglomerative) ή διαιρετικά (divisive). Στην πρώτη περίπτωση ο αλγόριθμος ξεκινά θεωρώντας κάθε αντικείμενο ως μία ομάδα και πραγματοποιεί

συνδέσεις όμοιων αντικειμένων, ώστε να σχηματιστεί μια ιεραρχία (από κάτω προς τα πάνω). Την αντίθετη πορεία ακολουθεί η διαιρετική εκδοχή της μεθόδου, όπου γίνεται αρχή από όλα τα δεδομένα ενταγμένα σε μία ομάδα και η οποία διαιρείται σε μικρότερες ομάδες (Jain, 2010).



Εικόνα 2: Λογική Ιεραρχικής και Διαιρετικής συσταδοποίησης (Χρυσός, 2006)

Πιο αναλυτικά, οι δύο τρόποι ιεραρχικής συσταδοποίησης είναι (Χρυσός, 2006):

- **Συσσωρευτική Ιεραρχική Συσταδοποίηση (Agglomerative Hierarchical Clustering)**

Η μέθοδος θεωρεί κάθε δεδομένο ως ξεχωριστή ομάδα. Σε κάθε βήμα του αθροίζονται μικρότερες ομάδες με βάση κάποιο κριτήριο απόστασης ή ομοιότητας, μέχρι να δημιουργηθεί μία ενιαία ομάδα. Με βάση αυτό το κριτήριο διαχωρίζουμε τις μεθόδους:

1) Απλού συνδέσμου (Single linkage). Αρχικά, κάθε αντικείμενο βρίσκεται σε ξεχωριστή ομάδα και σε κάθε βήμα συνενώνονται αυτά που βρίσκονται πιο κοντά. Στην ομάδες που δημιουργούνται ως απόσταση θεωρείται αυτή μεταξύ των πιο κοντινών τους σημείων και η οποία, όταν έχει επιθυμητή τιμή, προκαλεί συνένωση των ομάδων σε μια μεγαλύτερη (Εικόνα 3).

2) Πλήρους συνδέσμου (Complete linkage). Η διαφορά της μεθόδου σε σχέση με την αντίστοιχη απλού συνδέσμου έγκειται στο ότι η μέτρηση της απόστασης

μεταξύ δύο ομάδων υπολογίζεται μεταξύ των πιο απομακρυσμένων σημείων τους (Εικόνα 3).

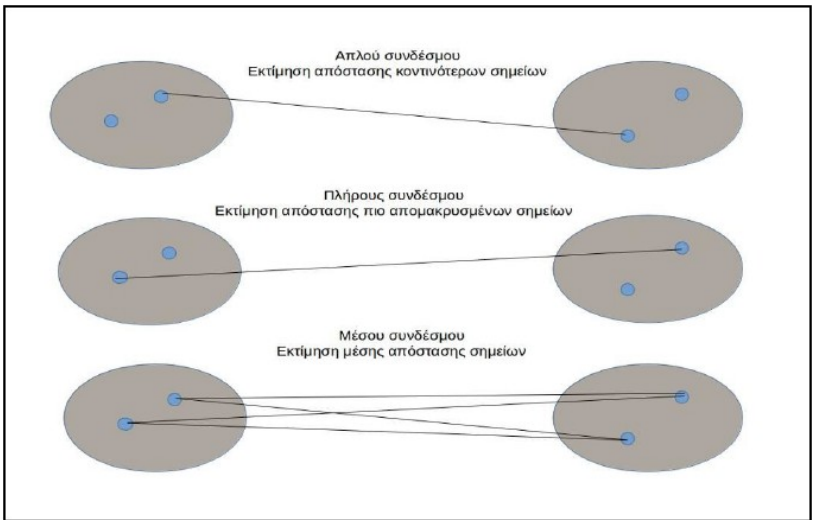
3) Μέσου συνδέσμου (Unweighted). Σε αυτή την περίπτωση η απόσταση μεταξύ δύο ομάδων υπολογίζεται ως σταθμισμένος μέσος όρος των αποστάσεων όλων των δυνατών ζευγαριών μεταξύ των αντικειμένων των ομάδων (Εικόνα 3).

4) Κεντροειδούς συνδέσμου (Centroid linkage). Σε αυτή τη μέθοδο η απόσταση των ομάδων είναι ίση με την ευκλείδεια απόσταση των κέντρων τους, δηλαδή του μέσου όρου των αντικειμένων που το αποτελούν.

5) Συνδέσμου Ward (Ward's linkage**).** Στην μέθοδο αυτή η απόσταση μεταξύ των ομάδων δίνεται από τύπο:

$$d^2(r,s) = n_r n_s \frac{\|\bar{x}_r - \bar{x}_s\|_2^2}{(n_r + n_s)}$$

όπου n_r και n_s ο αριθμός των αντικειμένων των ομάδων r και s και x_r και x_s ο μέσος όρος του (κέντρα των ομάδων). Η μέθοδος διαφοροποιείται από τις άλλες της ίδιας κατηγορίας, στο ότι η ένωση των ομάδων δεν προκαλείται όταν η τιμή της απόστασης d είναι μικρότερη από κάποιο επιθυμητό όριο. Αντίθετα, για κάθε πιθανή συνένωση ομάδων υπολογίζεται το άθροισμα των τετραγώνων των αποστάσεων όλων των στοιχείων από το κέντρο της ομάδας. Η συνένωση των ομάδων γίνεται αποδεκτή όταν το άθροισμά αυτό είναι το μικρότερο.



Εικόνα 3: Μέθοδοι Ιεραρχικής Συσταδοποίηση ς (Τσεργούλας, 2016)

- **Διαιρετική Ιεραρχική Συσταδοποίηση (Divisive Hierarchical Clustering)**

Η μέθοδος ακολουθεί την αντίθετη λογική της Αθροιστικής Συσταδοποίησης. Συγκεκριμένα, όλα τα δεδομένα βρίσκονται σε μια ενιαία ομάδα και στη συνέχεια διαιρούνται με βάση κάποιο κριτήριο, ώστε τελικά κάθε δεδομένο να αποτελεί ξεχωριστή ομάδα. Το εν λόγω κριτήριο διάσπασης βασίζεται στη διάμετρο μιας ομάδας (αλγόριθμος του Guènoche), η οποία ορίζεται ως η μέγιστη απόσταση μεταξύ δύο οποιοδήποτε αντικειμένων της. Η διάσπαση γίνεται όταν η μεγαλύτερη από τις παραγόμενες ομάδες έχει τη μικρότερη διάμετρο. Η διαδικασία τερματίζεται όταν κάθε αντικείμενο αποτελεί ξεχωριστή ομάδα.

2.3.2 Διαμεριστική διαμέριση

Σε αντίθεση με του αλγόριθμους ιεράρχησης, στη διαμεριστική διαμέριση (Partitional Clustering) ομαδοποιεί τα αντικείμενα, χωρίς να υπαγορεύει ιεραρχική δομή (Jain, 2010). Σε αυτού του τύπου συσταδοποίησης, ορίζονται εκ των προτέρων οι ομάδες και στη συνέχεια τα αντικείμενα μετακινούνται ανάμεσα στις ομάδες ανάλογα με τα κριτήρια που έχουν τεθεί.

- **Αλγόριθμος απλού περάσματος**

Οι αλγόριθμοι της κατηγορίας αυτής, ομαδοποιούν τα δεδομένα σε ένα στάδιο. Αρχικά, το πρώτο στοιχείο ορίζεται ως κέντρο της πρώτης ομάδας. Στη συνέχεια υπολογίζονται οι αποστάσεις των επόμενων στοιχείων από τα κέντρο μάζας των ήδη υπαρχόντων ομάδων. Γίνονται αποδεκτές οι μικρότερες αποστάσεις κάθε στοιχείου, ενώ απορρίπτονται οι άλλες. Αν η απόσταση που έχει γίνει αποδεκτή είναι μικρότερη από ένα επιθυμητό όριο, τότε το στοιχείο ανατίθεται στην ομάδα από την οποία απέχει λιγότερο. Σε αντίθεση περίπτωση, το εν λόγω στοιχείο γίνεται κέντρο μιας νέας ομάδας. Ο αλγόριθμος τερματίζεται όταν ομαδοποιηθούν όλα τα αντικείμενα.

- **Αλγόριθμοι συσταδοποίηση τετραγωνικού λάθους**

Οι εν λόγω αλγόριθμοι χρησιμοποιούν ως κριτήριο ομοιότητας την παράμετρο του τετραγωνικού λάθους, η οποία δίνεται από τη σχέση:

$$e^2(K,L) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \left\| x_i^{(j)} - c_j \right\|^2$$

Όπου $x_i^{(j)}$ είναι το i στοιχείο της j ομάδας, c_j το κέντρο μάζας της j ομάδας, k ο αριθμός των ομάδων, N_j *ο* αριθμός των στοιχείων της j ομάδας. Συνήθως, οι αλγόριθμοι αυτής της κατηγορίας υλοποιούνται ως εξής:

Βήμα 1: Επιλογή αντικειμένων-εκπροσώπων
Βήμα 2: Ανάθεση των υπόλοιπων ομάδων στου κοντινότερους εκπροσώπους
Βήμα 3: Εκ νέου υπολογισμός των εκπροσώπων
Βήμα 4: Επανάληψη των βημάτων 2 και 3 μέχρι να μην παρατηρούνται αλλαγές

Ο πιο διαδεδομένος αλγόριθμος αυτή της κατηγορίας είναι ο k-means, ο οποίος αποτελεί αντικείμενο της παρούσας εργασίας και αναλύεται παρακάτω.

- **Ασαφής συσταδοποίηση (Fuzzy clustering)**

Η ασαφής συσταδοποίηση θεωρεί ότι κάθε αντικείμενο μπορεί να ανήκει ταυτόχρονα, σε κάποιο βαθμό, δύο ή περισσότερες ομάδες. Ο βαθμός με τον οποίο συμμετέχει ο αντικείμενο στην ομάδα εκφράζεται με την συνάρτηση συμμετοχής, η οποία μπορεί να πάρει τιμές από 0 (= το αντικείμενο σίγουρα δεν ανήκει στην ομάδα) έως 1 (= το αντικείμενο σίγουρα ανήκει στην ομάδα). Σημειώνεται ότι, συνάρτηση συμμετοχής μπορεί να οριστεί και στην περίπτωση των κλασικών αλγόριθμων συσταδοποίησης, ωστόσο στις περιπτώσει αυτές οι τιμές της θα είναι διακριτές (0 και 1). Αντίθετα, στην ασαφή συσταδοποίηση η συνάρτηση συμμετοχής είναι συνεχής (από 0 έως1).

2.4 Ο Αλγόριθμος k- means

Αναπτύχθηκε το 1957 από τον Stuart Llloyd, γι' αυτό αναφέρεται και ως αλγόριθμος Lloyd. Ομαδοποιεί δεδομένα στο ν- διάστατο ευκλείδειο χώρο, συνήθως με χρήση του Ευκλείδειας απόστασης. Ο αντιπρόσωπος κάθε ομάδας ονομάζεται κεντροειδές (centroid) και υπολογίζεται συνήθως ως ο μέσος όρος των αντικειμένων κάθε ομάδας. Η μέθοδος υλοποιείται σε δύο φάσεις (Nidheesh, 2017):

> **1η Φάση**: ορίζουμε, συνήθως τυχαία, κάποια αντικείμενα ως κεντροειδή. Στην συνέχεια ο αλγόριθμος υπολογίζει την απόσταση κάθε αντικειμένου από κάθε κεντροειδές (με χρήση της Ευκλείδειας απόστασης) και το αντιστοιχεί στην ομάδα του κεντροειδούς με το οποίο η απόσταση αυτή είναι η μικρότερη.
>
> **2η Φάση**: επαναϋπολογίζουμε τα κεντροειδή και επαναλαμβάνουμε τη διαδικασία. Τα νέα κεντροειδή υπολογίζονται ως μέσοι όροι των αντικειμένων κάθε ομάδας που έχει δημιουργηθεί από τη 1η φάση.

Συνήθως, επιλέγουμε η διαδικασία να ολοκληρώνεται μετά από κάποιο αριθμό επαναλήψεων, που θεωρούμε ότι θα αποδώσει ικανοποιητική συσταδοποίηση. Εναλλακτικά, μπορούμε να ορίσουμε μια απόσταση- όριο, η οποία είναι η μέγιστη τιμή της απόστασης που επιτρέπουμε στα κεντροειδή να μετακινηθούν. Έτσι, σε περίπτωση που τα νέα κεντροειδή απέχουν από τα προηγούμενα λιγότερο από την τιμή- κατώφλι, ο αλγόριθμος τερματίζεται. Επιπλέον, ένα σύνηθες κριτήριο τερματισμού που επιλέγεται είναι η παράμετρος RSS. Αυτή ισούται με το άθροισμα των τετραγώνων των αποστάσεων των διανυσμάτων- αντικειμένων από το κεντροειδές. Ονομάζεται άθροισμα τετραγώνων σφαλμάτων ή υπολειμματικό άθροισμα τετραγώνων (Residual Sum of Squares- RSS) και χαρακτηρίζει κάθε ομάδα Μπορούμε έτσι, να θέσουμε κάποιο ανώτατο όριο για την τιμή της RSS, πέραν του οποίου η διαδικασία συσταδοποίηση ς τερματίζεται.

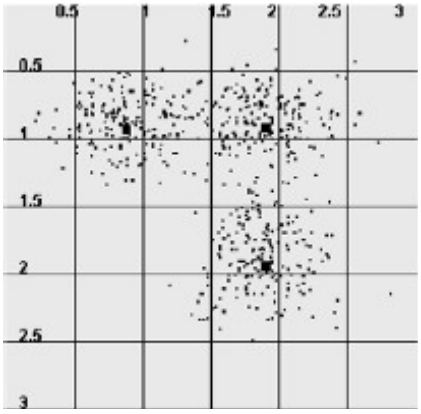
Η χρονική, T(n) και χωρική, S(n) πολυπλοκότητα του αλγόριθμου k- means, εξαρτάται από τον αριθμό των διανυσμάτων- αντικειμένων (n), το πλήθος των διαστάσεων τους (d), τον αριθμό των ομάδων που αναζητούμε (k) και τον αριθμό των

επαναλήψεων (i). Με βάση τα μεγέθη αυτά, έχουμε: T(n) = n*d*k*i και S(n) = (n+k)*d. Σημειώνεται ότι, συνήθως τα k,d και i είναι αρκετά μικρότερα από το n συνεπώς ο αλγόριθμος είναι κατά κύριο λόγο γραμμικός ως προς τα αντικείμενα που ομαδοποιεί. Επιπρόσθετα, είναι σαφές από τον τύπου του S(n) ότι ο αλγόριθμος απαιτεί την παρουσία όλων των δεδομένων και επιπρόσθετα των κεντροειδών κάθε επανάληψης να είναι διαθέσιμα στη μνήμη του υπολογιστή (Αφεντουλίδης, 2015).

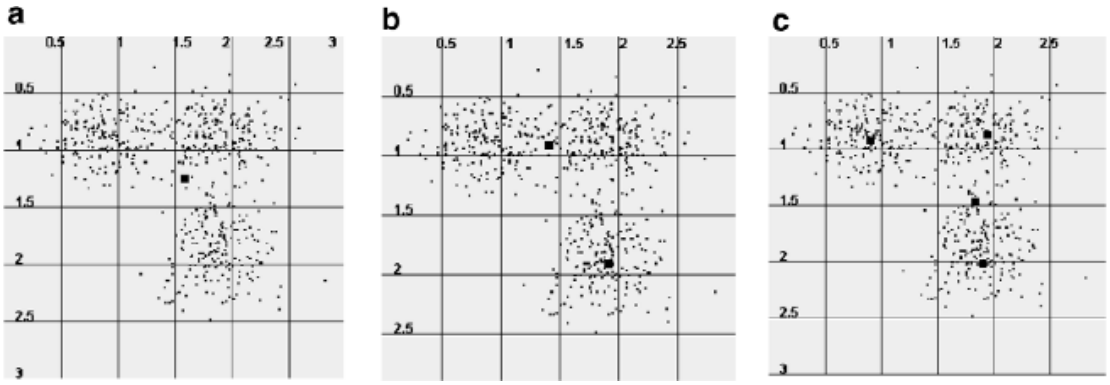
Ο αλγόριθμος k-means αποτελεί μια ελκυστική λύση συσταδοποίησης, λόγω της ευκολίας στη χρήση του και της αξιοπιστίας των αποτελεσμάτων του. Επιπλέον, έχει το σημαντικό πλεονέκτημα να μπορεί εύκολα να παραλληλοποιηθεί. Συγκεκριμένα, τα δεδομένα διαιρούνται σε κομμάτια και την συσταδοποίηση για κάθε κομμάτι αναλαμβάνει διαφορετικό υπολογιστικό σύστημα. Ακολουθεί ενημέρωση των συστημάτων με τα νέα κεντροειδή σε κάθε επανάληψη. Ωστόσο, εμφανίζει κι σημαντικές αδυναμίες, όπως (Ακακιάδου, 2007):

- Ο αλγόριθμος συγκλίνει σε τοπικό βέλτιστο και όχι σε καθολικό βέλτιστο,
- Ο ορισμός των αρχικών κεντροειδών δεν είναι σαφώς καθορισμένος,
- Τα αποτελέσματα εξαρτώνται και από το μέτρο απόστασης που χρησιμοποιείται,
- Αδυνατεί να αναγνωρίσει ανόμοιες, ως προς το σχήμα και το μέγεθος, ομάδες, κυρίως σε μεγάλους όγκους δεδομένων. Συνήθως, το πρόβλημα σχετίζεται με την ποικιλία στην πυκνότητα των στοιχείων.

Το κύριο μειονέκτημα του αλγόριθμου k- means είναι ότι απαιτεί τον εκ των προτέρων προσδιορισμό του πλήθους των ομάδων. Το γεγονός αυτό είναι κρίσιμο για την επιτυχία της μεθόδου και απαιτεί πρότερη γνώση γύρων από τα δεδομένα εισόδου. Έτσι, μόνο αν το πλήθος των ομάδων εκτιμηθεί σωστά και συμπίπτει με τον πραγματικό, ο αλγόριθμος δίνει σωστά αποτελέσματα (Zalik, 2008). Στις παρακάτω εικόνες 4 και 5 φαίνονται τα αποτελέσματα του αλγορίθμου για την περίπτωση συσταδοποίησης με επιλογή k= 1,2 και 4, όταν οι πραγματικές ομάδες είναι τρείς.



Εικόνα 4: Σωστή συσταδοποίηση του k- means για εκτίμηση k= 3 (Zalik, 2008).



Εικόνα 5: Λάθος συσταδοποιήσεις του αλγόριθμου k-means για εκτιμήσεις k=1,2 και 4 (Zalik, 2008).

Είναι προφανές ότι ενδεχόμενη κακή επιλογή των κεντροειδών οδηγεί σε λύσεις χαμηλής ποιότητας. Επιπλέον, είναι γνωστό ότι προκαλεί εκθετική αύξηση του χρόνου υλοποίησης (Capóa, 2017). Η εκτίμηση του σωστού αριθμού ομάδων, γίνεται συνήθως με πολλές εκτελέσεις του αλγόριθμου, ενώ έχουν γίνει πολλές ερευνητικές προσπάθειες για την ανάπτυξη μεθοδολογίας εκτίμησης των αρχικών κεντροειδών.

Μία από τις πρώτες και πιο δημοφιλείς στρατηγικές αρχικού προσδιορισμού κεντροειδών, συνίσταται σε τυχαία επιλογή τους στο κέντρο συμπλεγμάτων αντικειμένων που εμφανίζουν υψηλή πυκνότητα. Η λογική είναι ότι είναι πιο πιθανό να επιλεγούν τα σωστά κεντροειδή, καθώς συνήθως αυτά βρίσκονται σε περιοχές υψηλής συγκέντρωσης δεδομένων. Η προσέγγιση αυτή, όμως, μειονεκτεί, καθώς δεν εγγυάται ότι δεν υπάρχουν δύο ή περισσότερα κεντροειδή κοντά στο κέντρο του συμπλέγματος. Επιπλέον, υπάρχουν διαδικασίες αρχικοποίησης κεντροειδών, που βασίζονται στη θεωρία πιθανοτήτων.

Ωστόσο, ο αλγόριθμος k-means έχει δύο κύρια θεωρητικές ελλείψεις:

- Πρώτον, έχει αποδειχτεί ότι στην χειρότερη περίπτωση ο χρόνος που διαρκεί για να εκτελεστεί ο αλγόριθμος πολυονιμικός σε σχέση με το μέγεθπς εισόδου (Arthur D., Vassilvitskii S., 2006)
- Δεύτερον, η προσέγγιση που βρέθηκε μπορεί να είναι αυθαίρετα κακή σε σχέση με την αντικειμενική λειτουργία σε σύγκριση με τη βέλτιστη συστοδοποίηση.

Ο αλγόριθμος k-means ++ αντιμετοπίζει το δεύτερο από αυτά τα εμπόδια καθορίζοντας μια διαδικασία αρχικοποίησης των κέντρων των συστάδων πριν προχωρήσει με τις τυπικές επαναλήψεις βελτιστοποίησης του k-means. Αρχικά επιλέγεται (τυχαία) μόνο το πρώτο κεντροειδές. Κάθε επόμενο προσδιορίζεται με μια πιθανότητα ανάλογη προς την απόσταση σε σχέση με την προηγουμένως επιλεγμένη ομάδα κεντροειδών. Η μέθοδος πλεονεκτεί ως προς ό,τι διατηρεί την ποικιλομορφία των κεντροειδών και είναι ανθεκτική στις αποκλίσεις, ωστόσο είναι διαδοχική στη φύση της και δεν επιτρέπει την παραλληλοποίηση της (Capóa, 2017). Με την αρχικοποίηση k-means++, ο αλγόριθμος είναι σίγουρο ότι θα βρει μια λύση η οποία είναι O(log$k$) ανταγωνιστική με την βέλτιστη k-means λύση.

Ο k-means++ είναι δηλαδή ένας αλγόριθμος για να επιλέγει τις αρχικές τιμ΄ς συσταδοποίησης του k-means αλγορίθμου. Παρουσιάστηκε πρώτη φορά από τους David Arthur kai Sergei Vassilvitskii, ως ένας προσεγγιστικός τρόπος για το NP-hard πρόβλημα του k-means – έανσ τρόπος να αποφεύγεται τη κάποιες φορές φτωχή συσταδοποίηση που γίνεται με τον κανονικό k-means αλγόριθμο.

3 ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ

3.1 Υλοποίηση του αλγόριθμου k-means στην Python

Για την υλοποίηση του αλγόριθμου χρησιμοποιήθηκε η διανομή της Python από την Anaconda (https://www.anaconda.com/) και πιο συγκεκριμένα το Spyder IDE. Ο αλγόριθμος εφαρμόστηκε σε τέσσερα αρχεία δεδομένων, που παρουσιάζονται παρακάτω. Το πρόγραμμα διαβάζει το εκάστοτε αρχείο (συγκεκριμένες στήλες που ορίζει ο χρήστης). Ο χρήστης ορίζει τον αριθμό των κεντροειδών και το πρόγραμμα τα ορίζει. Στη συνέχεια ο αλγόριθμος μετράει τις αποστάσεις των δεδομένων από τα κεντροειδή με χρήση της Ευκλείδειας απόστασης και επιστρέφει τα νέα κεντροειδή. Η διαδικασία οπτικοποιείται. Παρακάτω παρουσιάζεται ο κώδικας, όπως υλοποιήθηκε:

1. Εισαγωγή απαιτούμενων βιβλιοθηκών για την ανάλυση των δεδομένων και την οπτικοποίηση των αποτελεσμάτων

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import animation
```

2. Δημιουργία τυχαίων σημείων

```
points = np.vstack(((np.random.randn(150, 2) * 0.75 + np.array([1, 0])),
                    (np.random.randn(50, 2) * 0.25 + np.array([-0.5, 0.5])),
                    (np.random.randn(50, 2) * 0.5 + np.array([-0.5, -0.5]))))
```

3. Επιλογή του αριθμού των κεντροειδών από τον χρήστη

```
noCentroids = int(input('Give the number of the Centroids: '))
```

4. Ορισμός αρχικών κεντροειδών με βάση τον αριθμό που δόθηκε

```
def initialize_centroids(points, k):
    centroids = points.copy()
    np.random.shuffle(centroids)
    return centroids[:k]
```

5. Σχηματισμός πίνακα για το κοντινότερο κεντροειδές για κάθε σημείο

```
def closest_centroid(points, centroids):
    distances = np.sqrt(((points - centroids[:,np.newaxis])**2).sum(axis=2))
```

```
    return np.argmin(distances, axis=0)
```

6. Επιστροφή νέων κεντροειδών που ανατέθηκαν από τα σημεία που ήταν κοντινότερα σε αυτά

```
def move_centroids(points, closest, centroids):
    return np.array([points[closest==k].mean(axis=0) for k in
range(centroids.shape[0])])
```

7. Οπτικοποίηση αρχικών κεντροειδών

```
plt.subplot(121)
plt.scatter(points[:, 0], points[:, 1])
centroids = initialize_centroids(points, noCentroids)
plt.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
plt.grid(True)
plt.title('Αρχικά k')
```

8. Οπτικοποίηση τελικών κεντροειδών

```
plt.subplot(122)
plt.scatter(points[:, 0], points[:, 1])
closest = closest_centroid(points, centroids)
centroids = move_centroids(points, closest, centroids)
plt.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
plt.grid(True)
plt.title('Τελικά k')
```

9. Animation

    # Ορισμός σχήματος, αξόνων και στοιχείων

```
plt.show()

fig = plt.figure()
ax = plt.axes(xlim=(-4, 4), ylim=(-4, 4))
centroids = initialize_centroids(points, noCentroids)
centroids, = ax.plot([],[])
```

    # Συνάρτηση αρχικοποίησης: σχεδιασμός του φόντου (background) κάθε πλαισίου

```
def init():
    centroids.set_data([],[])
    return centroids,
```

Συνάρτηση animation

```
def animate(i):
    global centroids

    closest = closest_centroid(points, centroids)
    centroids = move_centroids(points, closest, centroids)
    ax.cla()
    ax.scatter(points[:, 0], points[:, 1], c=closest)
    ax.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
    return centroids,
```

Κλήση του animator. Η σχέση blit=True, σημαίνει ότι επανασχεδιάζονται μόνο τα μέρη που αλλάζουν θα Ορισμός σχήματος, αξόνων και στοιχείων

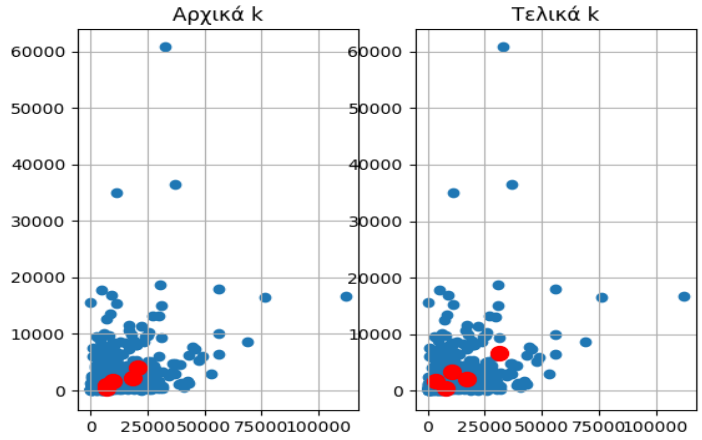
```
animation.FuncAnimation(fig, animate, init_func=init,
                        frames=30, interval=200, blit=True)

plt.show()
```

3.2 Αποτελέσματα

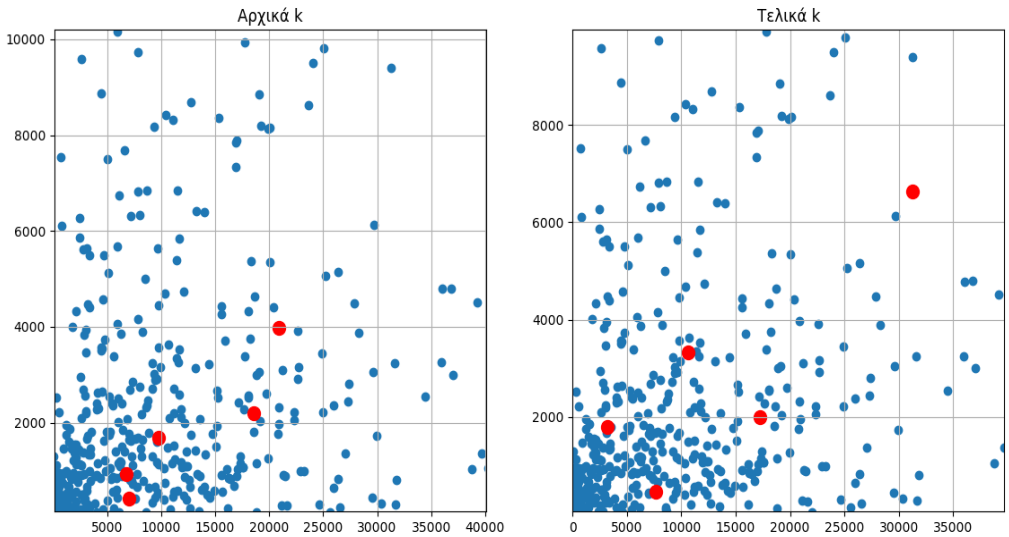
Ο αλγόριθμος (βλ. παράρτημα 1) εφαρμόστηκε σε τέσσερα αρχεία δεδομένων, όπως φαίνεται παρακάτω.

3.2.1 Δεδομένα Wholesale

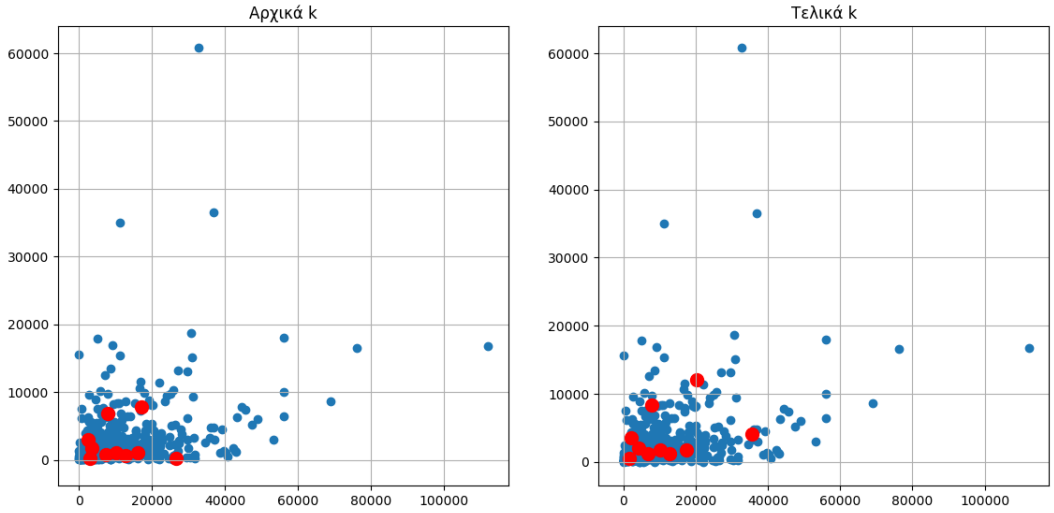
Το αρχείο περιέχει 440 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 2) με δεδομένα πωλήσεων. Επιλέγονται μόνο οι στήλες Φρέσκα (Fresh) και Κατεψυγμένα (Frozen). Παρακάτω παρουσιάζονται τα αποτελέσματα της συσταδοποίησης για k= 5,10,20,30 και 50.



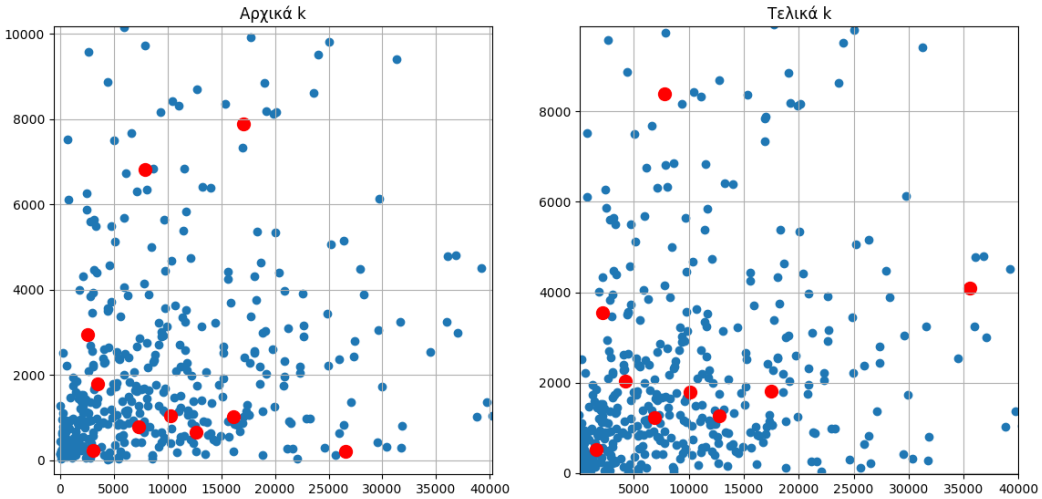
Εικόνα 6: Δεδομένα Wholesale, k=5



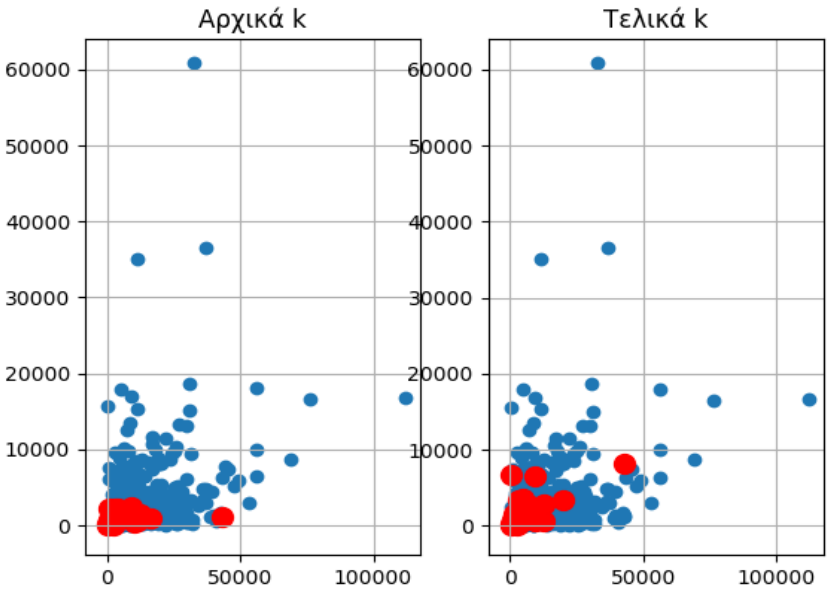
Εικόνα 7: Δεδομένα Wholesale, k=5 (Μεγέθυνση)



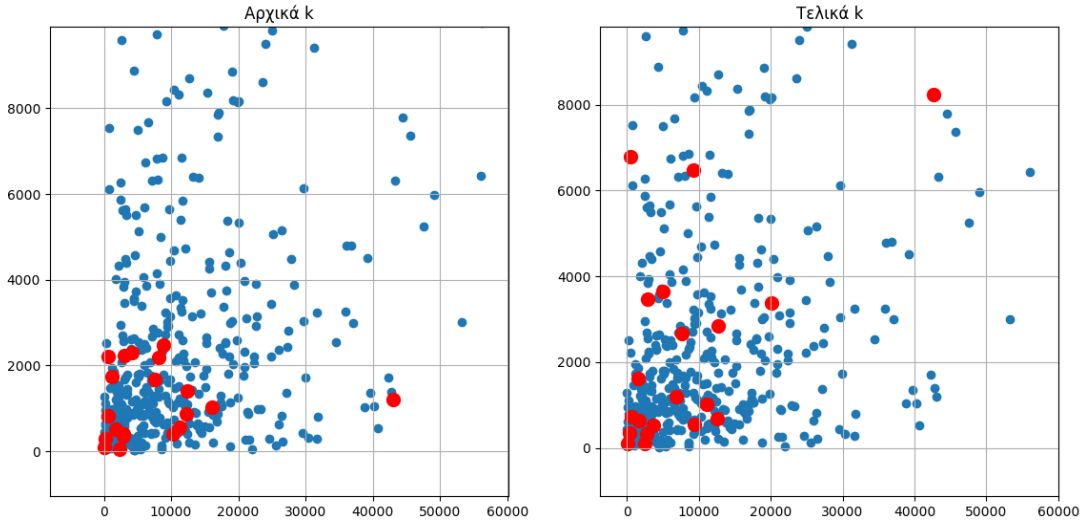
Εικόνα 8: Δεδομένα Wholesale, k=10 (Μεγέθυνση)



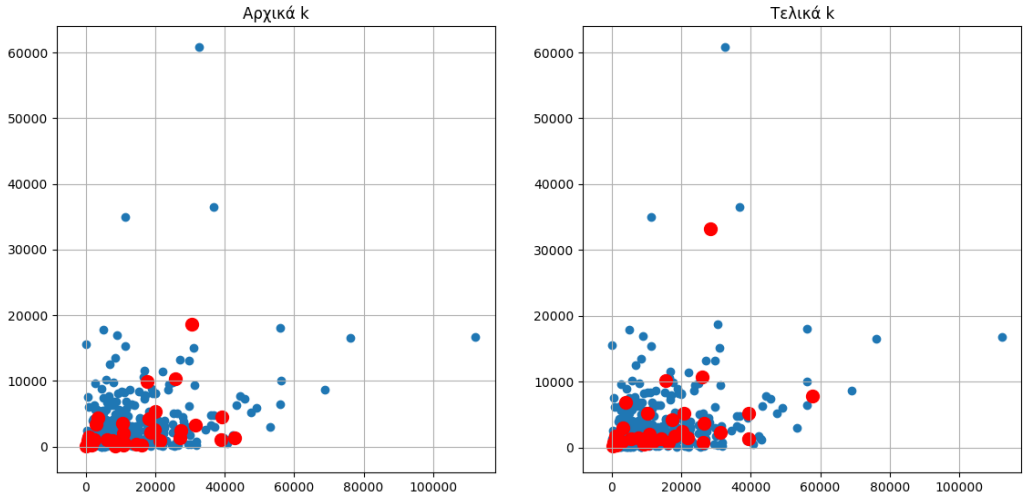
Εικόνα 9: Δεδομένα Wholesale, k=10 (Μεγέθυνση)



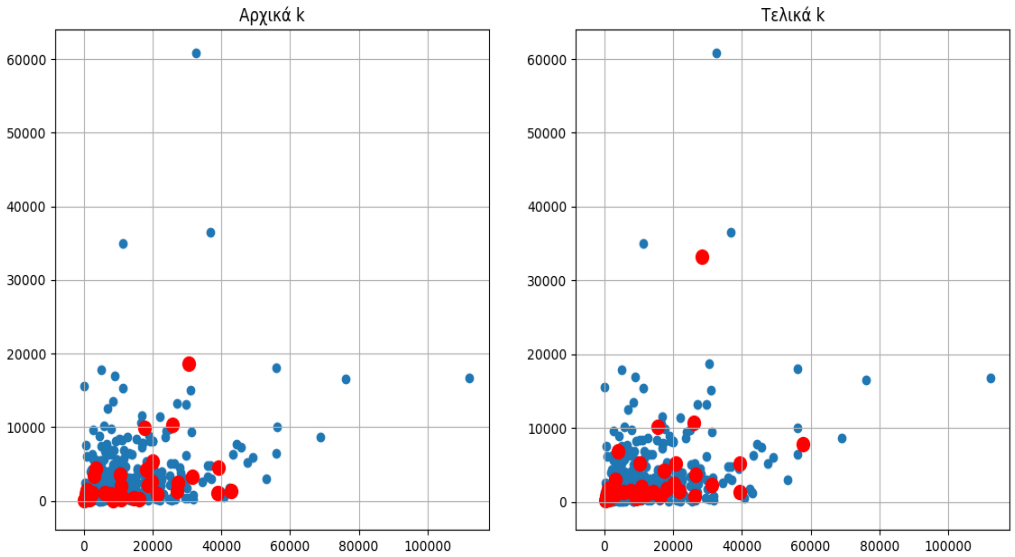
Εικόνα 10: Δεδομένα Wholesale, k=20



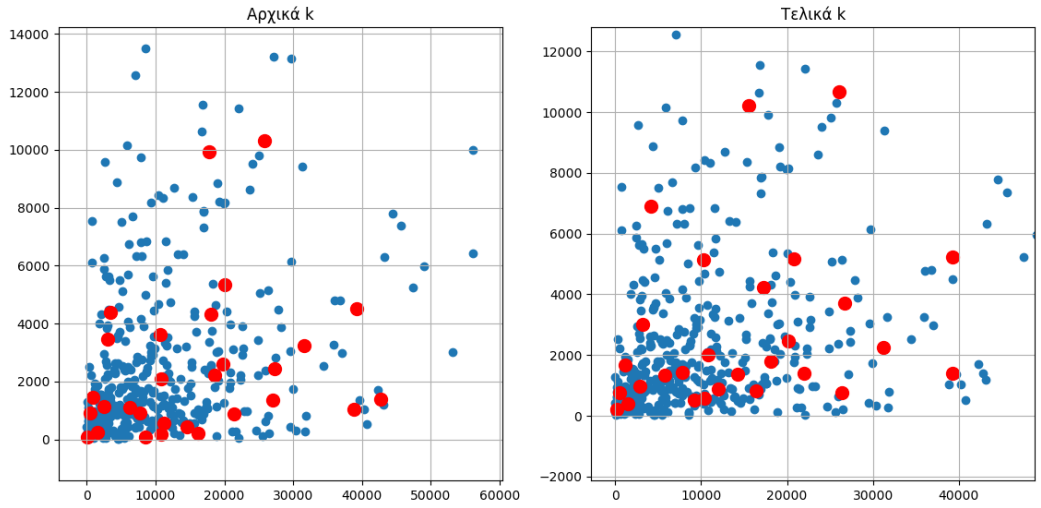
Εικόνα 11: Δεδομένα Wholesale, k=20 (Μεγέθυνση)



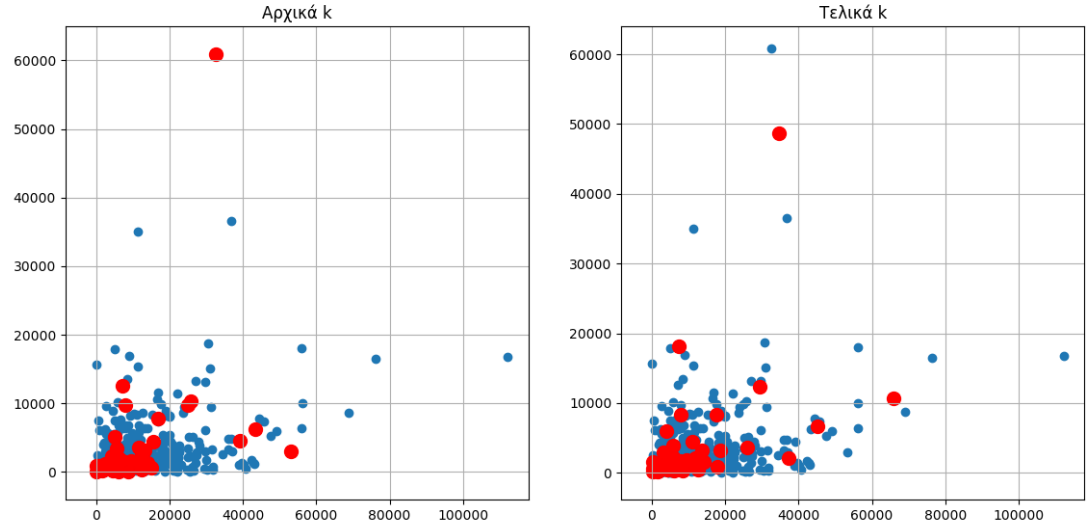
Εικόνα 12: Δεδομένα Wholesale, k=20 (Μεγέθυνση)



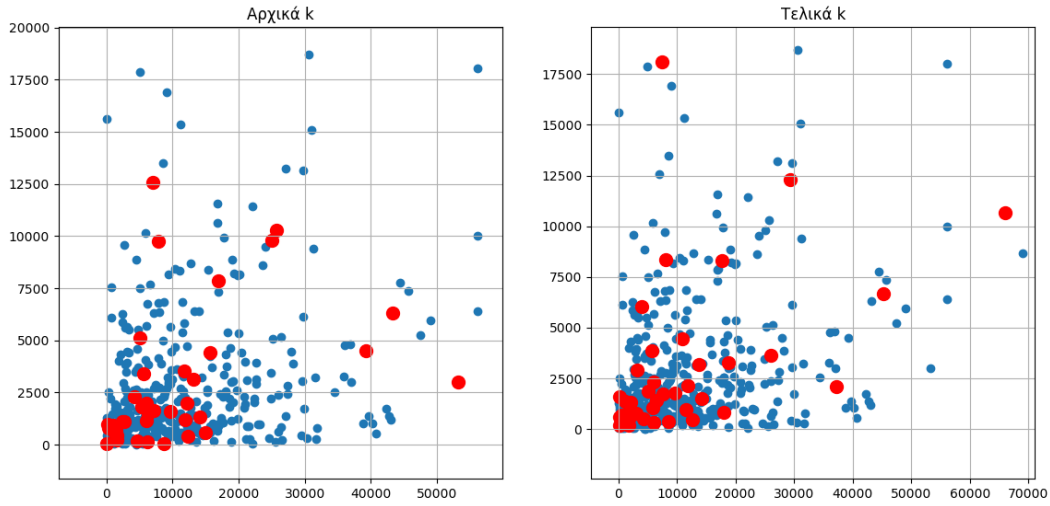
Εικόνα 13: Δεδομένα Wholesale, k=30



Εικόνα 14: Δεδομένα Wholesale, k=30 (Μεγέθυνση)



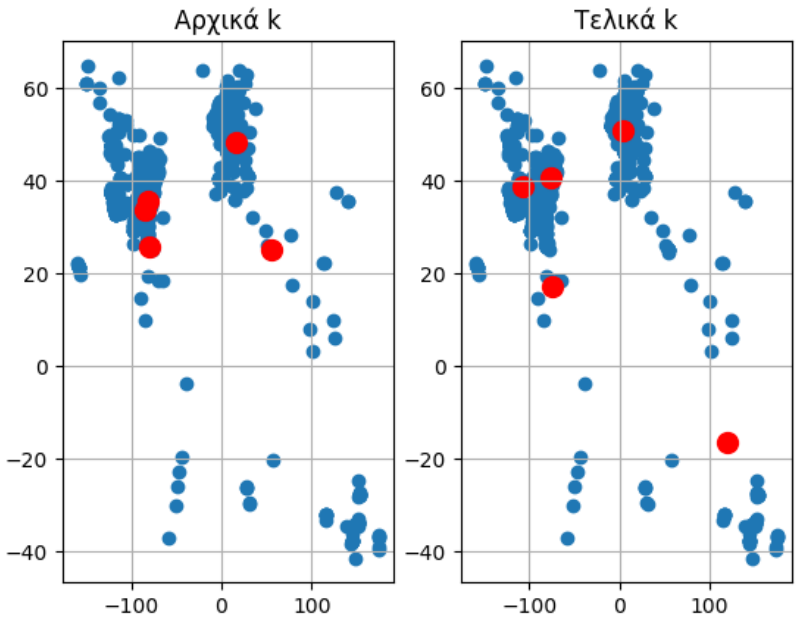
Εικόνα 15: Δεδομένα Wholesale, k=50



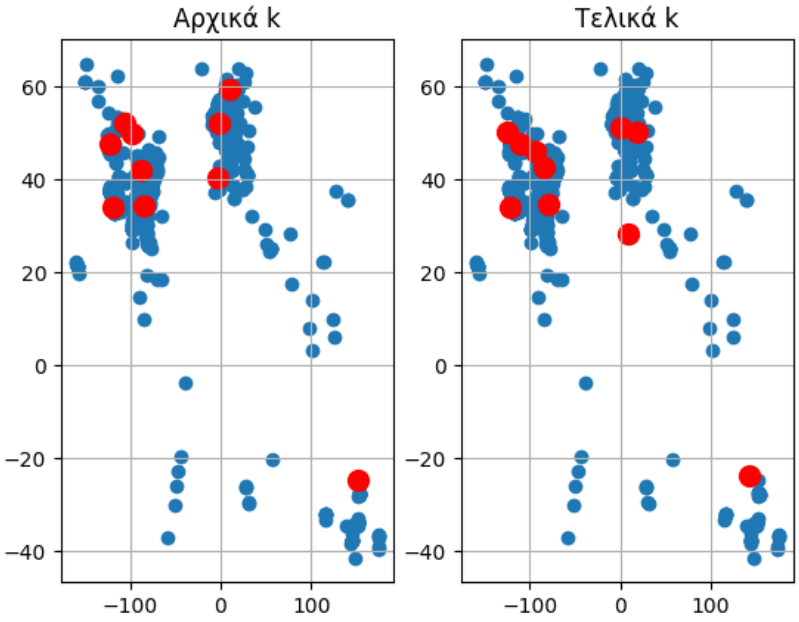
Εικόνα 16: Δεδομένα Wholesale, k=50 (Μεγέθυνση)

3.2.2 Δεδομένα Sales

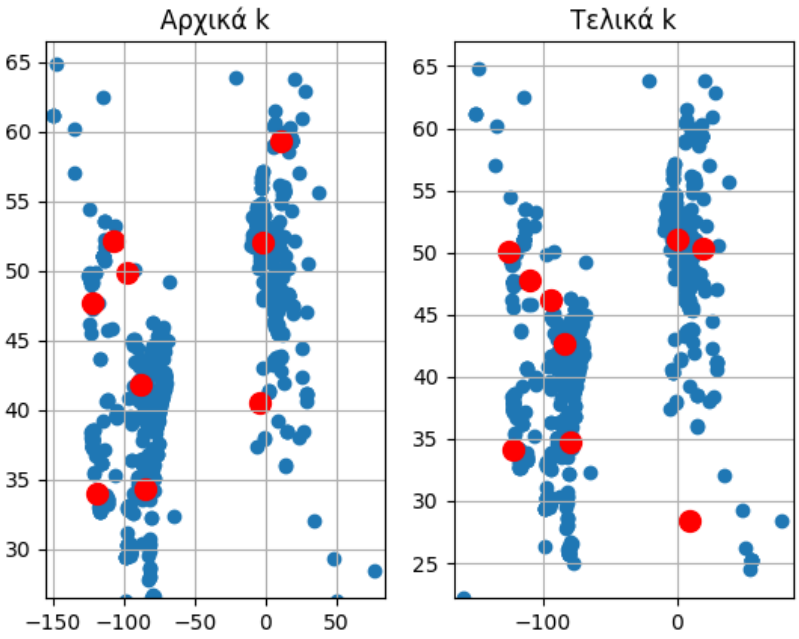
Το αρχείο περιέχει 998 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 3) με δεδομένα πωλήσεων. Επιλέγονται μόνο οι στήλες Γεωγραφικό πλάτος (Latitude) και Γεωγραφικό μήκος (Longtitude). Παρακάτω παρουσιάζονται τα αποτελέσματα της συσταδοποίησης για k= 5,10,25 και 50.



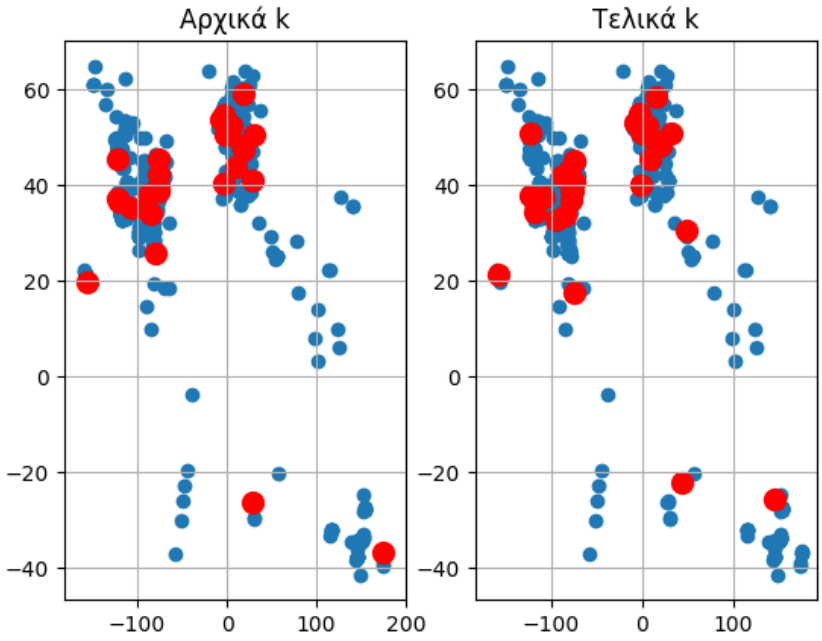
Εικόνα 17: Δεδομένα Sales , k=5



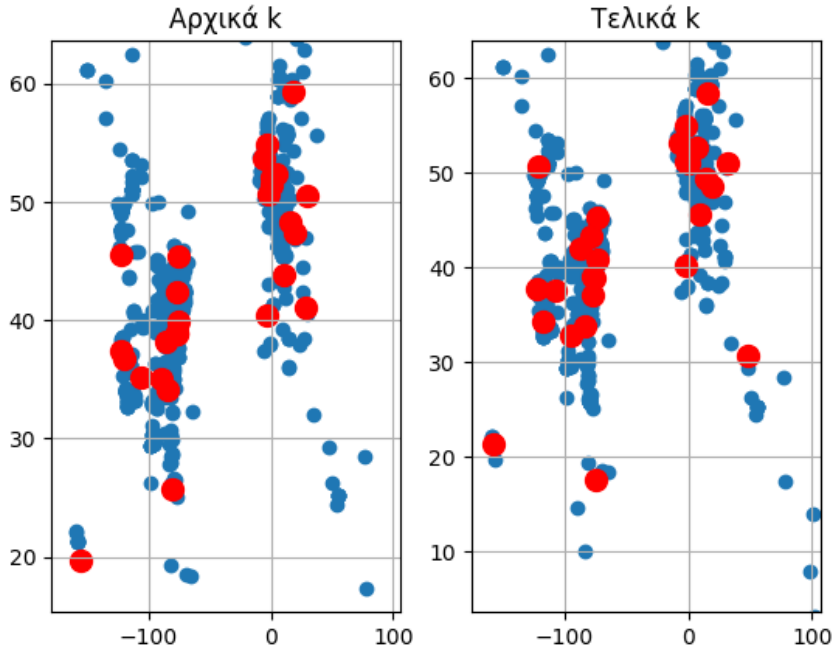
Εικόνα 18: Δεδομένα Sales , k= 10



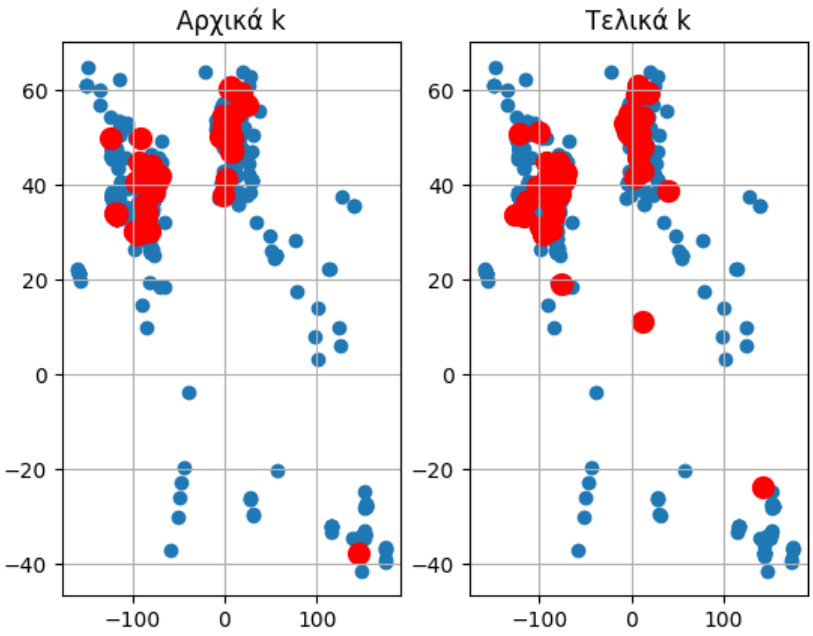
Εικόνα 19: Δεδομένα Sales , k=10 (Μεγέθυνση)



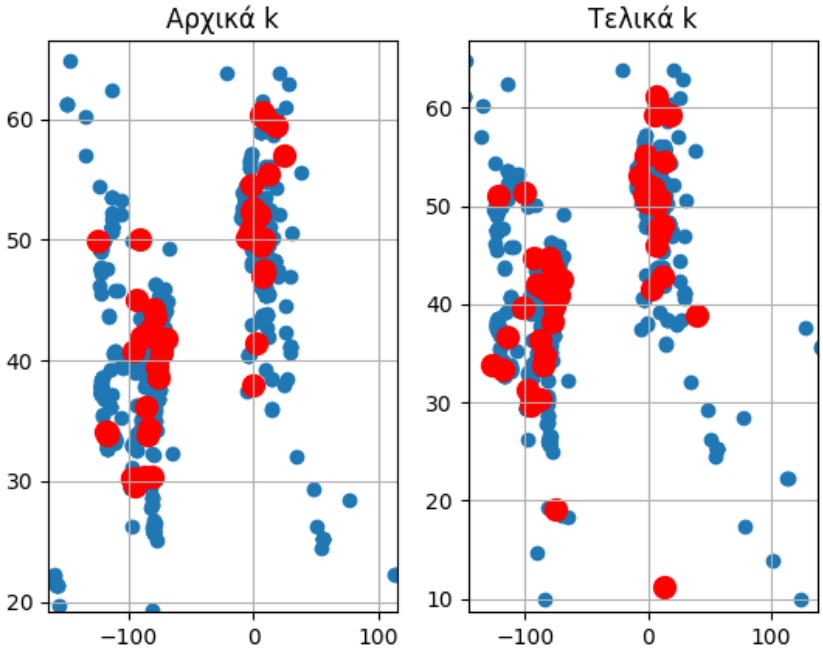
Εικόνα 20: Δεδομένα Sales , k=25



Εικόνα 21: Δεδομένα Sales , k=25 (Μεγέθυνση)



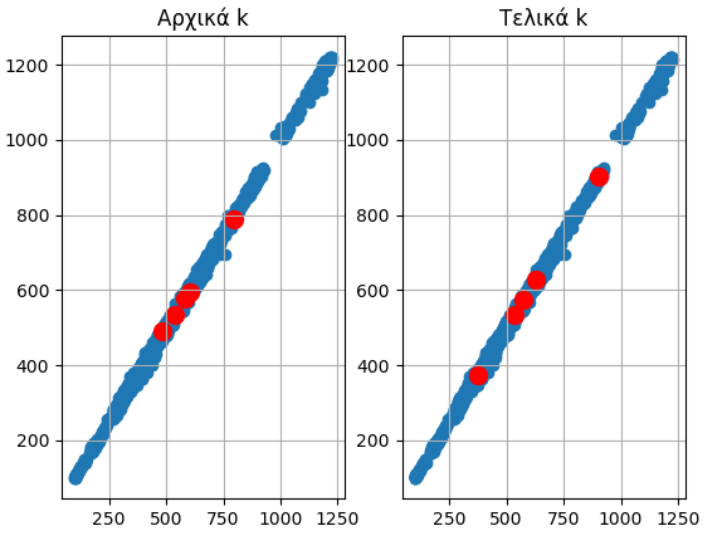
Εικόνα 22: Δεδομένα Sales , k=50



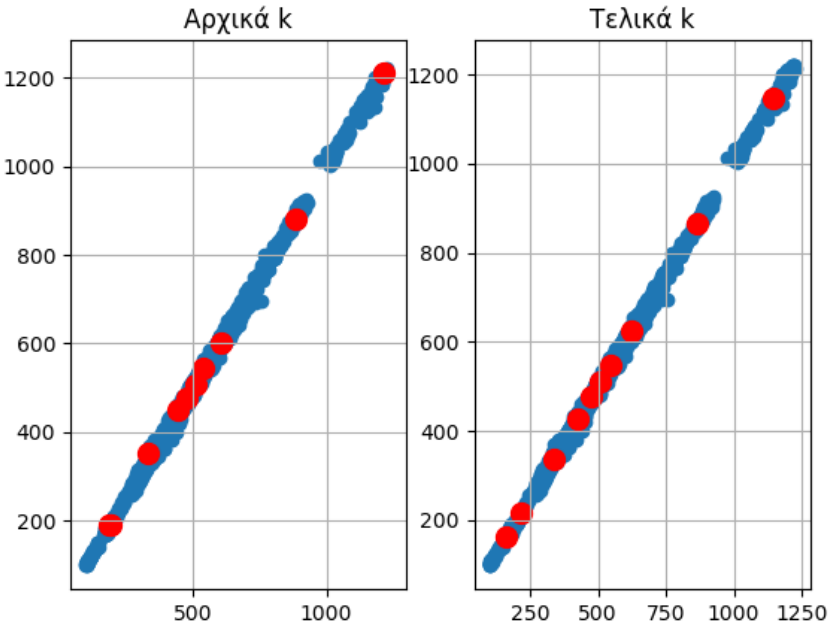
Εικόνα 23: Δεδομένα Sales , k=50 (Μεγέθυνση)

3.2.3 Δεδομένα Google

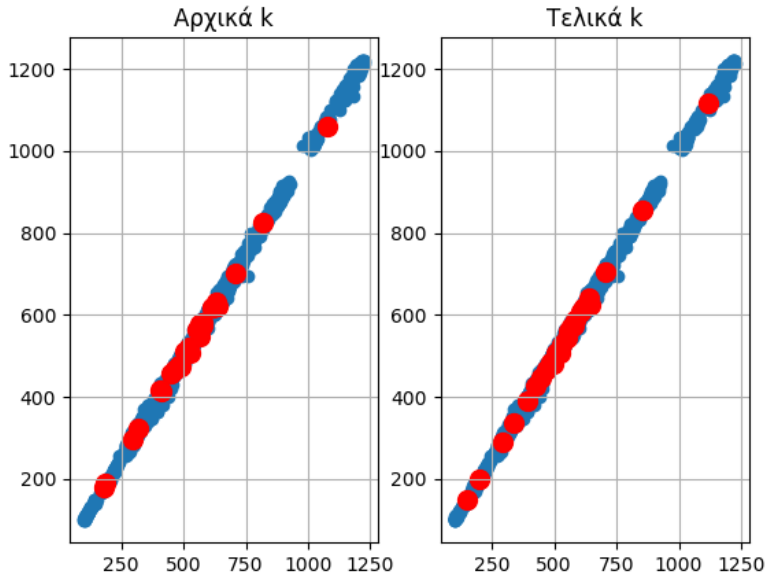
Το αρχείο περιέχει 2.518 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 4) με δεδομένα τιμών της μετοχής της Google. Επιλέγονται μόνο η τιμή ανοίγματος (Open) και η τιμή κλεισίματος (Close). Παρακάτω παρουσιάζονται τα αποτελέσματα της συσταδοποίησης για k= 5,10 και 15.



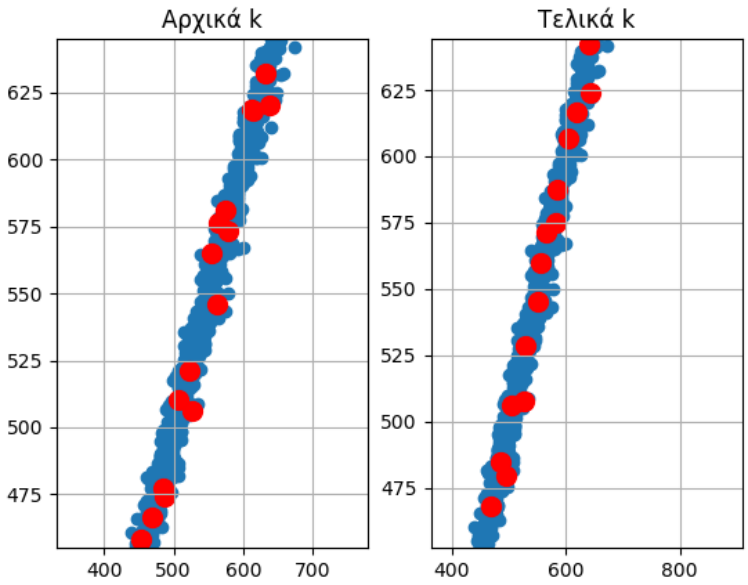
Εικόνα 24: Δεδομένα Google , k=5



Εικόνα 25: Δεδομένα Google , k=10



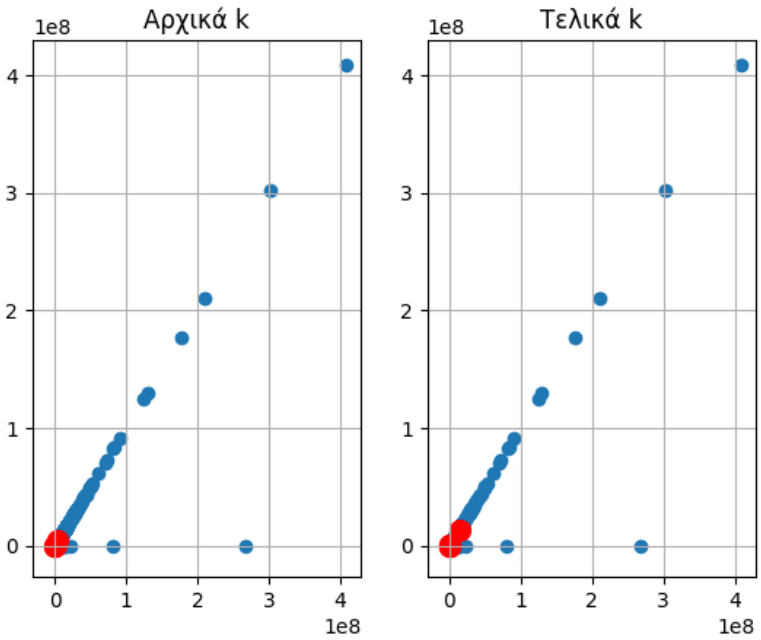
Εικόνα 26: Δεδομένα Google , k=15



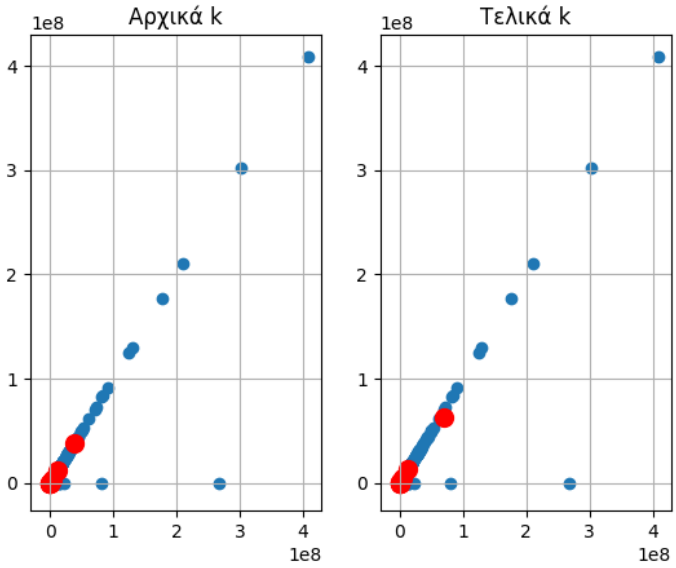
Εικόνα 27: Δεδομένα Google , k=15 (Μεγέθυνση)

3.2.4 Δεδομένα Insurance

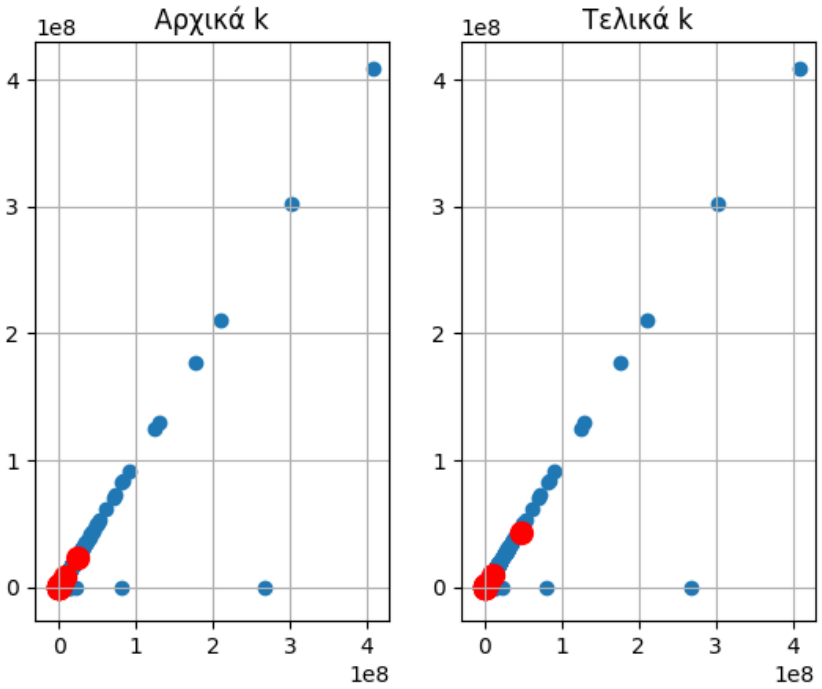
Το αρχείο περιέχει 36.634 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 5) με δεδομένα ασφαλιστικών συμβολαίων. Επιλέγονται μόνο οι στήλες E (hu_site_limit) και F(fl_site_limit). Παρακάτω παρουσιάζονται τα αποτελέσματα της συσταδοποίησης για k= 5,10, 25 και 50.



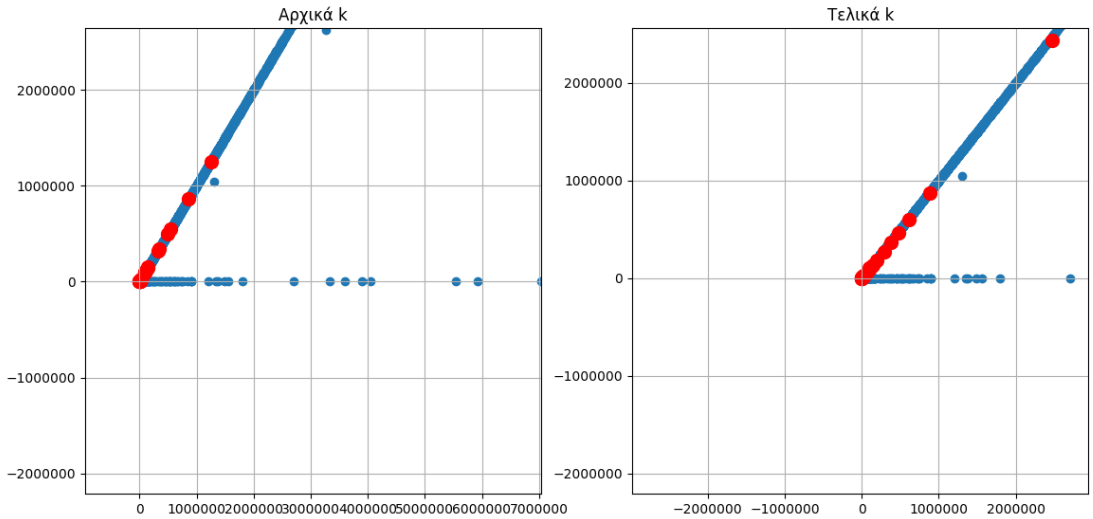
Εικόνα 28: Δεδομένα Insurance , k=5



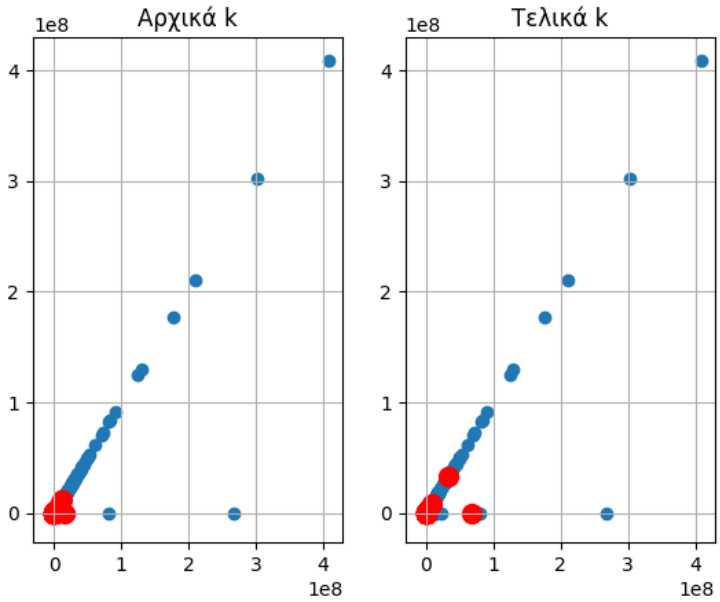
Εικόνα 29: Δεδομένα Insurance , k=10



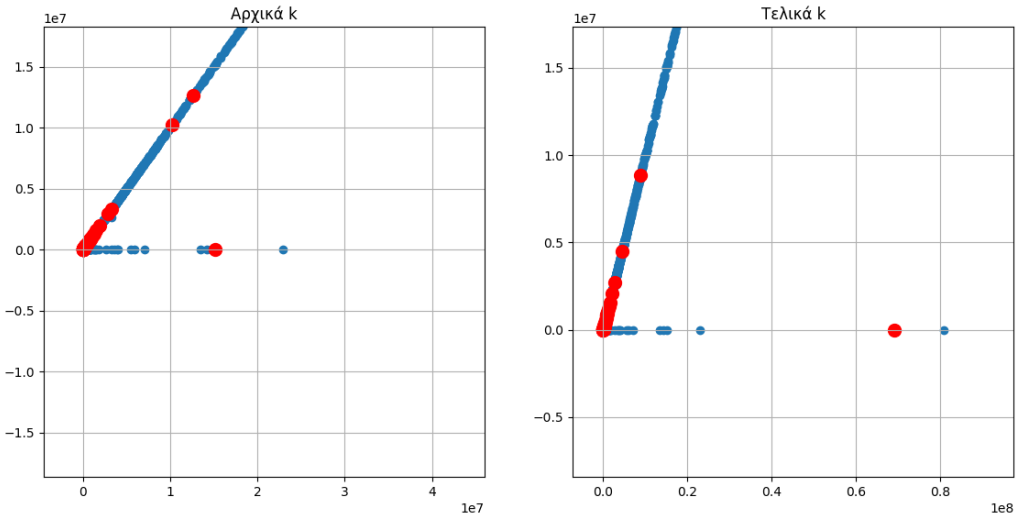
Εικόνα 30: Δεδομένα Insurance , k=25



Εικόνα 31: Δεδομένα Insurance , k=25 (Μεγέθυνση)



Εικόνα 32: Δεδομένα Insurance , k=50



Εικόνα 33: Δεδομένα Insurance , k=50 (Μεγέθυνση)

3.3 Αξιολόγηση αλγόριθμου

Γενικά, ο αλγόριθμος ολοκληρώνεται σε μικρό χρονικό διάστημα (μερικά δευτερόλεπτα). Πρόβλημα παρουσιάστηκε με τα δεδομένα INSURANCE, καθώς ο όγκος τους είναι μεγάλος και το πρόγραμμα απαιτεί μεγάλο χρόνο να τα διαβάσει.

Τα δεδομένα WHOLESALES δεν παρουσιάζουν γραμμικότητα. Η μεγαλύτερη μάζα των δεδομένων συγκεντρώνεται στην περιοχή μηδέν, ενώ υπάρχουν και πολλά απομονωμένα στοιχεία. Ο αλγόριθμος διασπείρει τα κεντροειδή, συμπεριλαμβάνοντας και τις απομονωμένες περιοχές για k= 30 και πάνω.
Τα δεδομένα SALES παρουσιάζουν διασπορά με δύο μεγάλες μάζες και αρκετά απομονωμένα στοιχεία. Οι απομονωμένες περιοχές συμπεριλαμβάνονται στον προσδιορισμός των κεντροειδών για k= 25 και πάνω.

Τα δεδομένα GOOGLE παρουσιάζουν γραμμική συσχέτιση με μία ασυνέχεια, ώστε να δημιουργούνται δύο διακριτά τμήματα. Για k= 5 ο αλγόριθμος δεν δημιουργεί κεντροειδές στο ένα τμήμα, ούτε υπολογίζει νέο κεντροειδές σε εκείνη την περιοχή. Για μεγαλύτερες τιμές k, ο αλγόριθμος συμπεριλαμβάνει και τη δεύτερη περιοχή. Ακόμα και για k= 15, όμως, ορίζει μόνο ένα κεντροειδές, το οποίο φαίνεται να προσαρμόζεται στο κέντρο των δεδομένων της περιοχής. Σε κάθε περίπτωση, ο αλγόριθμος ‘απλώνει’ τα κεντροειδή μέχρι και στις μικρότερες τιμές των δεδομένων.

Τα δεδομένα INSURANCE παρουσιάζουν επίσης γραμμικότητα, αλλά και ασυνέχεια, με ένα κύριο σώμα και μερικά απομονωμένα στοιχεία τόσο στο διαγώνιο άξονα όσο και στον οριζόντιο. Το κύριο μέρος των δεδομένων συγκεντρώνονται στην περιοχή κοντά στο μηδέν. Ο αλγόριθμος δεν υπολογίζει κεντροειδή για τα απομονωμένα στοιχεία στον οριζόντιο άξονα ούτε για k = 25.

4 ΕΠΙΛΟΓΟΣ

4.1 Σύνοψη και συμπεράσματα

Η παρούσα εργασία αφορά σε μία από τις πιο διαδομένες μεθόδους συσταδοποίησης και συγκεκριμένα ο αλγόριθμος k- means. Εφαρμόστηκε σε δεδομένα δύο διαστάσεων σε τέσσερα διαφορετικά σύνολα δεδομένων. Η εφαρμογή της μεθόδους αφορούσε σε ορισμό του αριθμού των κεντροειδών και υπολογισμού των νέων με χρήση της Ευκλείδειας απόστασης.

Από την υλοποίηση προκύπτει ότι ο αλγόριθμος είναι εύκολος στην υλοποίηση και εύχρηστος. Υλοποιείται ταχύτατα ακόμα και σε μεγάλους όγκους δεδομένων. Τα κεντροειδή που παράγει μετακινούνται από τα αρχικώς ορισμένα και εμφανίζουν καλή διασπορά στα δεδομένα. Όμως, φαίνεται να αντιμετωπίζει πρόβλημα στην περίπτωση που τα δεδομένα εμφανίζουν ασυνέχειες, καθώς μάλλον ‘αγνοεί’ αυτά που είναι απομονωμένα. Λύση στο φαινόμενο, στην περίπτωση που τα απομονωμένα δεδομένα έχουν ενδιαφέρον, δίνει ο ορισμός μεγάλου αριθμού κεντροειδών.

4.2 Μελλοντικές Επεκτάσεις

Στην παρούσα εργασία εξετάστηκαν δεδομένα δύο διαστάσεων. Θα ήταν χρήσιμο να επεκταθεί η εφαρμογή της μεθόδους και στην περίπτωση περισσότερων διαστάσεων, αλλά και να συγκριθεί με άλλου αλγόριθμους συσταδοποίησης. Επιπλέον, ο αλγόριθμος ίσως παρουσιάσει βελτίωση με χρήση άλλων αποστάσεων, πέραν της Ευκλείδειας. Προτείνεται, επίσης, η εκτίμηση του αριθμού των κεντροειδών, μέσω μιας από τις μεθόδους που προτείνονται στη βιβλιογραφία και όχι ο τυχαίος ορισμός τους, κάτι που αναμένεται να βελτιώσει τα αποτελέσματα. Ενδιάφερον, τέλος, θα είχε η παραλληλοποίηση του αλγόριθμου και η χρήση του στην ανάλυση δεδομένων μεγάλου όγκου.

5 ΒΙΒΛΙΟΓΡΑΦΙΑ

5.1 Ελληνική

Αφεντουλίδης Α. Γ. (2015). *Ανάλυση επίδοσης και μοντελοποίηση του αλγορίθμου συσταδοποίησης k-means σε κεντρικό και κατανεμημένο περιβάλλον*, Διπλωματική Εργασία, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστώ, Εθνικό Μετσόβιο Πολυτεχνείο

Ακακιάδου Γ. (2007). *Μελέτη του αλγορίθμου συσταδοποίησης k-means σε δεδομένα του παγκόσμιου ιστού*, Διπλωματική Εργασία, Τμήμα Πληροφορικής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Γούλας Χ. (2015). Σχεδιασμός και Ανάπτυξη Αλγορίθμου Συσταδοποίησης Μεγάλης Κλίμακας Δεδομένων- Εφαρμογή σε Χρονολογικά και Μη Δεδομένα,
Μεταπτυχιακή Διπλωματική Εργασία, Μ.Π.Σ. «Μεταπτυχιακό Πρόγραμμα Επιστήμη και Τεχνολογία των Υπολογιστών», Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Πατρών

Δημητρακοπούλου Κ. (2007). *Αναγνώριση λειτουργικών υπο-δομών στο πρωτεϊνικό δίκτυο του Saccharomyces Cerevisiae συνδυάζοντας δεδομένα έκφρασης γονιδίων και αλληλεπίδρασης πρωτεϊνών*, Διπλωματική Εργασία, Δ.Π.Μ.Σ. «Βιοιατρική Τεχνολογία», Πάτρα

Ιωάννου Ζ.Μ. (2014). *Αποτελεσματικές Τεχνικές διαχείρισης Δεδομένων στον Παγκόσμιο Ιστό*, Μεταπτυχιακή Εργασία, Μ.Π.Σ. «Επιστήμη και Τεχνολογία Υπολογιστών», Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Πατρών

Τσεργούλας Ηλίας (2016). Προσδιορισμός Βέλτιστων Επιχειρηματικών Συστάδων (Business Clusters) με τη Χρήση Αλγόριθμων Συσταδοποίησης και Γεωγραφικών Συστημάτων Πληροφοριών, Διπλωματική Εργασία, Μεταπτυχιακή Εξειδίκευση στα Πληροφορικά Συστήματα, Σχολή Θετικών Επιστημών και Τεχνολογίας, Ελληνικό Ανοικτό Πανεπιστήμιο

Χρυσός Ε.Γ. (2006). *Μετάδοση πληροφορίας σε ασύρματο δίκτυο αισθητήρων με ομαδοποιημένους κόμβους και με χρήση διευθύνσεων από κανόνες Golomb,* Διπλωματική Εργασία, Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Πολυτεχνείο Κρήτης

5.2 Ξένη

Capóa M., Pérez A., Lozano A.J. (2017). *An efficient approximation to the K -means clustering for massive data*, Knowledge-Based Systems 117, pp. 56–69

Han J., Kamber M., Pei J. (2012). *Data mining- Concepts and Techniques*, 3rd Editiion, Morgan Kaufmann Publishers, Elsevier, USA

Olson L.D., Delen D. (2008). *Advanced Data Mining Techniques*, Springer-Verlag Berlin

Jain K. A. (2010). *Data clustering: 50 years beyond K-means*, Pattern Recognition Letters 31, pp. 651–666

Nidheesh N., Abdul Nazeer K.A., Ameer P.M. (2017). *An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data*, Computers in Biology and Medicine 91, pp. 213–221

Zalik K.R. (2008). *An efficient k' -means clustering algorithm,* Pattern Recognition Letters 29, pp. 1385–1391

Arthur D., Vassilvitskii S. (2006), "How slow is the *k*-means method?", *ACM New York, NY, USA*, pp. 144–153

5.3 Ιστοσελίδες

Πιτουρά Ε. (2011). Συσταδοποίηση Ι, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf. Ανακτήθηκε στις 5/2/2018

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα 1: Κώδικας

```
# -*- coding: utf-8 -*-
"""
Created on Sun May  7 20:07:20 2017

@author: Konstantinos
"""

import matplotlib.pyplot as plt
import numpy as np
from matplotlib import animation

"""δημιουργία τυχαίων σημείων"""
points = np.vstack(((np.random.randn(150, 2) * 0.75 + np.array([1, 0])),
                    (np.random.randn(50, 2) * 0.25 + np.array([-0.5, 0.5])),
                    (np.random.randn(50, 2) * 0.5 + np.array([-0.5, -0.5]))))

"""επιλογή του χρήστη του αριθμού των κεντροειδών"""
noCentroids = int(input('Give the number of the Centroids: '))

"""επιστρέφει τα κεντροειδή από τα αρχικά σημεία βάσει του αριθμού που δώσαμε πριν"""
def initialize_centroids(points, k):
    centroids = points.copy()
    np.random.shuffle(centroids)
    return centroids[:k]

"""επιστρέφει έναν πίνακα που περιέχει την θέση για το κοντινότερο κεντροειδές για το κάθε σημείο"""
def closest_centroid(points, centroids):
```

```
            distances = np.sqrt(((points - centroids[:,np.newaxis])**2).sum(axis=2))
            return np.argmin(distances, axis=0)

        """επιστρέφει τα καινούρια κεντροειδή που ανατέθηκαν από τα σημεία που ήταν κοντινότερα σε αυτά"""
        def move_centroids(points, closest, centroids):
                    return np.array([points[closest==k].mean(axis=0) for k in range(centroids.shape[0])])



        plt.subplot(121)
        plt.scatter(points[:, 0], points[:, 1])
        centroids = initialize_centroids(points, noCentroids)
        plt.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
        plt.grid(True)
        plt.title('Αρχικά k')



        plt.subplot(122)
        plt.scatter(points[:, 0], points[:, 1])
        closest = closest_centroid(points, centroids)
        centroids = move_centroids(points, closest, centroids)
        plt.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
        plt.grid(True)
        plt.title('Τελικά k')

        plt.show()

        fig = plt.figure()
        ax = plt.axes(xlim=(-4, 4), ylim=(-4, 4))
        centroids = initialize_centroids(points, noCentroids)
        centroids, = ax.plot([],[])
```

```
def init():
    centroids.set_data([],[])
    return centroids,

def animate(i):
    global centroids

    closest = closest_centroid(points, centroids)
    centroids = move_centroids(points, closest, centroids)
    ax.cla()
    ax.scatter(points[:, 0], points[:, 1], c=closest)
    ax.scatter(centroids[:, 0], centroids[:, 1], c='r', s=100)
    return centroids,

animation.FuncAnimation(fig, animate, init_func=init,
                        frames=30, interval=200, blit=True)
plt.show()
```

Παράρτημα 2: Δεδομένα Wholesale

Channel,Region,Fresh,Milk,Grocery,Frozen,Detergents_Paper,Delicassen
2,3,12669,9656,7561,214,2674,1338
2,3,7057,9810,9568,1762,3293,1776
2,3,6353,8808,7684,2405,3516,7844
1,3,13265,1196,4221,6404,507,1788
2,3,22615,5410,7198,3915,1777,5185
2,3,9413,8259,5126,666,1795,1451
2,3,12126,3199,6975,480,3140,545
2,3,7579,4956,9426,1669,3321,2566
1,3,5963,3648,6192,425,1716,750
2,3,6006,11093,18881,1159,7425,2098

Παράρτημα 3: Δεδομένα Sales

Transaction_date,Product,Price,Payment_Type,Name,City,State,Country,Account_Created,Last_Login,Latitude,Longitude
01-02-09 6:17,Product1,1200,Mastercard,carolina,Basildon,England,United Kingdom,01-02-09 6:00,01-02-09 6:08,51.5,-1.1166667
01-02-09 4:53,Product1,1200,Visa,Betina,Parkville ,MO,United States,01-02-09 4:42,01-02-09 7:49,39.195,-94.68194
01-02-09 13:08,Product1,1200,Mastercard,Federica e Andrea,Astoria ,OR,United States,01-01-09 16:21,01-03-09 12:32,46.18806,-123.83
01-03-09 14:44,Product1,1200,Visa,Gouya,Echuca,Victoria,Australia,9/25/05 21:13,01-03-09 14:22,-36.1333333,144.75
01-04-09 12:56,Product2,3600,Visa,Gerd W ,Cahaba Heights ,AL,United States,11/15/08 15:47,01-04-09 12:45,33.52056,-86.8025
01-04-09 13:19,Product1,1200,Visa,LAURENCE,Mickleton ,NJ,United States,9/24/08 15:19,01-04-09 13:04,39.79,-75.23806
01-04-09 20:11,Product1,1200,Mastercard,Fleur,Peoria ,IL,United States,01-03-09 9:38,01-04-09 19:45,40.69361,-89.58889
01-02-09 20:09,Product1,1200,Mastercard,adam,Martin ,TN,United States,01-02-09 17:43,01-04-09 20:01,36.34333,-88.85028
01-04-09 13:17,Product1,1200,Mastercard,Renee Elisabeth,Tel Aviv,Tel Aviv,Israel,01-04-09 13:03,01-04-09 22:10,32.0666667,34.7666667
01-04-09 14:11,Product1,1200,Visa,Aidan,Chatou,Ile-de-France,France,06-03-08 4:22,01-05-09 1:17,48.8833333,2.15

Παράρτημα 4: Δεδομένα Google

Date, Open, High, Low, Close, Volume, Adj Close
8/19/2014,585.002622,587.342658,584.002627,586.862643,978600,586.862643
8/18/2014,576.11258,584.512631,576.002598,582.162619,1284100,582.162619
8/15/2014,577.862619,579.382595,570.522603,573.482626,1519100,573.482626

8/14/2014,576.182596,577.902645,570.882599,574.652582,985400,574.652582
8/13/2014,567.312567,575.002602,565.752564,574.782577,1439200,574.782577
8/12/2014,564.522567,565.902572,560.882518,562.732562,1542000,562.732562
8/11/2014,569.992585,570.492553,566.002578,567.882551,1214700,567.882551
8/8/2014,563.562536,570.252576,560.352561,568.772565,1494700,568.772565
8/7/2014,568.00257,569.89258,561.102543,563.362525,1110900,563.362525
8/6/2014,561.782569,570.702601,560.002541,566.376589,1334300,566.376589

Παράρτημα 5: Δεδομένα Insurance

policyID,statecode,county,eq_site_limit,hu_site_limit,fl_site_limit,fr_site_limit,tiv_2011, tiv_2012,eq_site_deductible,hu_site_deductible,fl_site_deductible, fr_site_deductible, point_latitude,point_longitude,line,construction,point_granularity
119736,FL,CLAY COUNTY,498960,498960,498960,498960,498960,792148.9,0,9979.2,0,0,30.102261,-81.711777,Residential,Masonry,1
448094,FL,CLAY COUNTY,1322376.3,1322376.3,1322376.3,1322376.3,1322376.3,1438163.57,0,0,0,0,30.063936,-81.707664,Residential,Masonry,3
206893,FL,CLAY COUNTY,190724.4,190724.4,190724.4,190724.4,190724.4,192476.78,0,0,0,0,30.089579,-81.700455,Residential,Wood,1
333743,FL,CLAY COUNTY,0,79520.76,0,0,79520.76,86854.48,0,0,0,0,30.063236,-81.707703,Residential,Wood,3
172534,FL,CLAY COUNTY,0,254281.5,0,254281.5,254281.5,246144.49,0,0,0,0,30.060614,-81.702675,Residential,Wood,1
785275,FL,CLAY COUNTY,0,515035.62,0,0,515035.62,884419.17,0,0,0,0,30.063236,-81.707703,Residential,Masonry,3
995932,FL,CLAY COUNTY,0,19260000,0,0,19260000,20610000,0,0,0,0,30.102226,-81.713882,Commercial,Reinforced Concrete,1
223488,FL,CLAY COUNTY,328500,328500,328500,328500,328500,348374.25,0,16425,0,0,30.102217,-81.707146,Residential,Wood,1
433512,FL,CLAY COUNTY,315000,315000,315000,315000,315000,265821.57,0,15750,0,0,30.118774,-81.704613,Residential,Wood,1
142071,FL,CLAY COUNTY,705600,705600,705600,705600,705600,1010842.56,14112,35280,0,0,30.100628,-81.703751,Residential,Masonry,1