ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΡΟΗΓΜΕΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

«Έρευνα και επισκόπηση αλγορίθμων μηχανικής μάθησης για συσταδοποίηση και χρήση κανόνων σε δεδομένα εκπαίδευσης με χρήση του εργαλείου Weka»

Συγγραφέας: Μπέρκ Αναστασία – Μαρία

ΑΜ: mscacs21017

Επιβλέπων:

Δρ. Φοίβος Μυλωνάς, Αναπληρωτής καθηγητής

Αθήνα

Φεβρουάριος 2023

UNIVERSITY OF WEST ATTICA SCHOOL
SCHOOL OF ENGINEERING
DEPARTMENT OF INFORMATION AND COMPUTER ENGINEERING
POSTGRADUATE PROGRAM MSC-ACS
ADVANCED COMPUTER SYSTEMS TECHNOLOGIES

Diploma thesis:

«Research and overview of machine learning algorithms for clustering and usage rules on educational data using the Weka tool»

Postgraduate student: Berk Anastasia - Maria
RN: mscacs21017

Supervisor:
Dr. Phivos Mylonas, Associate Professor

Athens
February 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΠΡΟΗΓΜΕΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

«Έρευνα και επισκόπηση αλγορίθμων μηχανικής μάθησης για συσταδοποίηση και χρήση κανόνων σε δεδομένα εκπαίδευσης με χρήση του εργαλείου Weka»

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η μεταπτυχιακή διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	Φοίβος Μυλωνάς	Αναπληρωτής Καθηγητής Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής / Εισηγητής – Επιβλέπων	Φοίβος Απόστολος Μυλωνάς - Ψηφιακά υπογεγραμμένο 2023.02.24 18:23:26 +02'00'
2	Περικλής Ανδρίτσος	Αναπληρωτής Καθηγητής Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής / Μέλος Εξεταστικής Επιτροπής	Periklis Andritsos Digitally signed by Periklis Andritsos Date: 2023.03.03 10:33:21 +02'00'
3	Ιωάννης Βογιατζής	Καθηγητής - Πρόεδρος Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής / Μέλος Εξεταστικής Επιτροπής	Ioannis Voyiatzis Digitally signed by Ioannis Voyiatzis Date: 2023.03.08 15:52:47 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Η κάτωθι υπογεγραμμένη **Μπέρκ Αναστασία – Μαρία** του **Νικολάου - Ερόλ**, με αριθμό μητρώου **mscacs21017** φοιτήτρια του Προγράμματος Μεταπτυχιακών Σπουδών **Προηγμένες Τεχνολογίες Υπολογιστικών Συστημάτων** του **Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών** της Σχολής **Μηχανικών** του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Η Δηλούσα

Μπέρκ Αναστασία – Μαρία

(Υπογραφή)

Ψηφιακή Υπογραφή Επιβλέποντα

Φοίβος-Απόστολος Μυλωνάς - Ψηφιακά υπογεγραμμένο 2023.02.24 18:23:48 +02'00'

ANASTASIA - MARIA BERK

Digitally signed by ANASTASIA - MARIA BERK
Date: 2023.02.21 19:44:27 +02'00'

Ευχαριστίες

Αρχικά θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα αναπληρωτή καθηγητή μου κ. Μυλωνά Φοίβο, για την εμπιστοσύνη και την επιστημονική του καθοδήγηση στη διάρκεια εκπόνησης της εργασίας αυτής.

Επίσης, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για όλη τη στήριξη, τη συμπαράσταση και την κατανόησή τους και σε αυτόν τον αγώνα που κληθήκαμε να αντιμετωπίσουμε όλοι μαζί ώστε να επιτύχω την εκπλήρωση των σπουδών μου.

Τέλος, ιδιαίτερες ευχαριστίες θέλω να απευθύνω στην επιστήθια φίλη μου Πηνελόπη για την συνεχή υποστήριξη, επιμονή, υπομονή και τις εύστοχες παρατηρήσεις της καθόλη τη διάρκεια εκπόνησης της εργασίας αυτής με σκοπό ένα αρτιότερο αποτέλεσμα.

Περίληψη - Λέξεις κλειδιά

Η συγκεκριμένη εργασία αποτελεί επισκόπηση που αφορά τα προβλήματα μηχανικής μάθησης με δεδομένα από τον εκπαιδευτικό χώρο και εξαρτώνται σε μεγάλο βαθμό από τους αλγόριθμους που χρησιμοποιούνται για την εκπαίδευση του εκάστοτε μοντέλου. Υπάρχουν διάφορες προσεγγίσεις και αλγόριθμοι για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Η εποπτευόμενη και η μη εποπτευόμενη μάθηση είναι οι δύο πιο σημαντικές από αυτές τις προσεγγίσεις. Προβλήματα που έχουν αναδυθεί και αποτελούν σημαντικό πεδίο έρευνας του εκπαιδευτικού τομέα σχετίζονται με τις εκπαιδευτικές πρακτικές που εφαρμόζονται στην τηλεκπαίδευση και θα μπορούσαν εύκολα να ερευνηθούν και να επιλυθούν με τη βοήθεια μιας μορφής μάθησης χωρίς επίβλεψη, γνωστή και ως Clustering ή στα ελληνικά συσταδοποίηση (ή και ομαδοποίηση).

Όπως προκύπτει και από την ονομασία της, η συσταδοποίηση περιλαμβάνει το χωρισμό σημείων δεδομένων σε πολλές συστάδες με παρεμφερείς τιμές. Δηλαδή, η συσταδοποίηση στοχεύει να διαχωρίσει ομάδες στοιχείων με παρόμοια χαρακτηριστικά όπου στη συνέχεια θα ομαδοποιηθούν σε διαφορετικές μεταξύ τους συστάδες. Αν και για τους ανθρώπους είναι εύκολο να εκπαιδευτούν στα να διαχωρίζουν ένα μήλο από ένα πορτοκάλι το ίδιο δεν ισχύει για ένα μηχάνημα παρά μόνο εάν εκπαιδευτεί αποτελεσματικά σε ένα σχετικά μεγάλο σύνολο δεδομένων. Αυτή η εκπαίδευση επιτυγχάνεται με αλγόριθμους μάθησης χωρίς επίβλεψη (ή μη εποπτευόμενη μάθηση) και συγκεκριμένα με συσταδοποίηση.

Σε απλή γλώσσα, οι συστάδες είναι η συγκέντρωση σημείων δεδομένων που έχουν παρεμφερείς τιμές ή χαρακτηριστικά και οι αλγόριθμοι συσταδοποίησης είναι οι μέθοδοι για την συσταδοποίηση παρεμφερών σημείων δεδομένων σε ετερογενή ενιαία σύνολα ετερόκλητων στοιχείων συνδεμένων μεταξύ τους με διαφορετικούς κανόνες με βάση τις τιμές ή τα χαρακτηριστικά που τα διακρίνουν.

Σκοπός αυτής της διπλωματικής είναι η χρήση αλγόριθμων μηχανικής μάθησης για συσταδοποίηση και χρήση κανόνων σε ένα σύνολο δεδομένων χρησιμοποιώντας το εργαλείο Weka, ιδανικό για τη διεξαγωγή συμπερασμάτων σε δεδομένα εκπαίδευσης.

Συσταδοποίηση, Weka, Μηχανική μάθηση, Εξόρυξη εκπαιδευτικών δεδομένων, Εξόρυξη δεδομένων

Abstract – Keywords

This paper is an overview of machine learning problems with data from the educational field, which are highly dependent on the algorithms used to train each model. There are various approaches and algorithms for training a machine learning model. Supervised and unsupervised learning are the two most important of these approaches. Problems that have emerged and are an important field of research in the educational sector are related to the educational practices applied in distance education and could easily be investigated and solved with the help of a form of unsupervised learning, also known as Clustering or in Greek grouping).

As its name suggests, clustering involves dividing data points into several clusters with similar values. That is, clustering aims to separate groups of elements with similar characteristics where they will then be grouped into different clusters. Although it is easy for humans to be trained to tell an apple from an orange, the same is not true for a machine unless it is effectively trained on a relatively large data set. This training is achieved with unsupervised learning (or unsupervised learning) algorithms, namely clustering. In plain language, clusters are the aggregation of data points that have similar values or characteristics, and clustering algorithms are the methods for clustering similar data points into heterogeneous single sets of disparate elements connected together by different rules based on the values or characteristics that distinguish them.

The purpose of this diploma is to use machine learning algorithms for clustering and applying rules to a dataset using the Weka tool, ideal for conducting inference on training data..

Clustering, Weka, Machine learning, Educational Data Mining, Data mining

Παράρτημα εικόνων

Παράρτημα Πινάκων

Εισαγωγή

Σε μία εποχή όπου υπάρχει πρόσβαση σε μεγάλους όγκους δεδομένων (big data) και ειδικότερα σε δεδομένα που προέρχονται από εκπαιδευτικά ιδρύματα είναι επιβεβλημένη η ανάγκη για την συστηματική οργάνωση και αξιοποίηση αυτών, με σκοπό την εξόρυξη και τελικά την ανακάλυψη νέας γνώσης, η οποία μπορεί να αποτελέσει σημαντικό ερευνητικό εύρημα και να βελτιώσει την ποιότητα της παρεχόμενης εκπαίδευσης. Ο τομέας αυτός αφορά την εξόρυξη εκπαιδευτικών δεδομένων και ασχολείται με τη βελτίωση της απόδοσης και της επιτυχίας των μαθητών, τη βελτίωση της απόδοσης και υποστήριξης των εκπαιδευτικών και την αποτελεσματική ενίσχυση της εκπαιδευτικής διαδικασίας γενικότερα.

Με τον όρο εξόρυξη δεδομένων (data mining) αναφερόμαστε κυρίως στις μεθόδους και τεχνικές που χρησιμοποιούνται με σκοπό την ανακάλυψη γνώσης από βάσεις δεδομένων με σκοπό την υποστήριξη της λήψης αποφάσεων. Πρόκειται δηλαδή για μία διαδικασία ανάδειξης άγνωστων προηγουμένως μοτίβων στα δεδομένα που είναι δυνητικά χρήσιμα και κατανοητά. Επίσης, η διαδικασία αυτή περιλαμβάνει ανάλυση των δεδομένων για εξεύρεση μη αναμενόμενων μεταξύ τους συσχετίσεων με σκοπό την εξαγωγή κανόνων για πρόβλεψη μελλοντικής συμπεριφοράς -πληροφορίες, δηλαδή οι οποίες δεν μπορούν να εξαχθούν από τον χρήστη με «γυμνό» μάτι.

Η εξόρυξη δεδομένων χρησιμοποιεί διάφορους τύπους μεθόδων και τεχνικών εξόρυξης, αυτό εξαρτάται από το είδος της βάσης δεδομένων που θα χρησιμοποιηθεί, το είδος της γνώσης το οποίο απαιτείται για να εξαχθούν και τέλος το είδος των τεχνικών που θα χρησιμοποιηθούν για τη διαδικασία της εξόρυξης. Εμείς, στην παρούσα εργασία θα ασχοληθούμε με τη μέθοδο της Συσταδοποίησης ή αλλιώς Ομαδοποίησης (clustering) την οποία και θα αναλύσουμε περαιτέρω στην συνέχεια.

Πολύ συχνά παρατηρείται το φαινόμενο, να συγχέεται η έννοια της εξόρυξης δεδομένων με την έννοια της μηχανικής μάθησης (machine learning). Αν και οι δύο τομείς μπορεί να αλληλεπικαλύπτονται, ειδοποιός μεταξύ τους διαφορά αποτελεί αφενός η πρόβλεψη στην οποία αποσκοπεί η μηχανική μάθηση βασιζόμενη σε γνωστές ιδιότητες ενός συνόλου δεδομένων αφετέρου και η ανακάλυψη μη γνωστών ιδιοτήτων ενός συνόλου δεδομένων στην οποία αποσκοπεί η εξόρυξη δεδομένων.

Με τον όρο μηχανική μάθηση αναφερόμαστε στη δημιουργία προτύπων ή μοντέλων από ένα σύνολο δεδομένων από ένα υπολογιστικό σύστημα. Τα κύρια είδη μηχανικής μάθησης είναι δύο, η μάθηση με επίβλεψη (supervised learning) και η μάθηση χωρίς

επίβλεψη (unsupervised learning) και ανάλογα με τη φύση του υπό διερεύνηση προβλήματος χρησιμοποιούνται διάφορες τεχνικές που εμπίπτουν στο ένα ή το άλλο είδος. Στη περίπτωση μας όπου θα γίνει εφαρμογή της διαδικασίας της Συσταδοποίησης έχουμε μάθηση χωρίς επίβλεψη όπου στόχος του συστήματος είναι η ανακάλυψη συσχετίσεων και ομάδων από τα δεδομένα βάση των ιδιοτήτων τους[1], [2].

Υπό αυτό το πρίσμα σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση των αλγορίθμων συσταδοποίησης, η ανάλυση και η κατανόηση του τρόπου υλοποίησης τους, η παρουσίαση παραδειγμάτων εφαρμογής τους, η παρουσίαση των πλεονεκτημάτων και των μειονεκτημάτων τους και τέλος η εφαρμογή τους μέσω του περιβάλλοντος Weka και η πραγματοποίηση σύγκρισης μεταξύ τους. Επίσης, θα επιχειρηθεί από τη διαδικασία συσταδοποίησης η ανακάλυψη σημαντικών πληροφοριών που αφορούν την στάση αλλά και την ποιότητα της παρεχόμενης εκπαίδευσης εν μέσω της πανδημίας του κορονοϊού σε μαθητές ανά τον κόσμο, ο εντοπισμός κοινών χαρακτηριστικών και η ανακάλυψη προτύπων που θα μπορέσουν να εισφέρουν στην καλύτερη οργάνωση και διαχείριση της μαθησιακής διαδικασίας που μπορεί να παρασχεθεί γενικότερα αλλά και ειδικότερα υπό συνθήκες τηλεκπαίδευσης.

1. Συσταδοποίηση

Μία από τις πιο δημοφιλείς μεθόδους εξόρυξης δεδομένων αποτελεί η Συσταδοποίηση ή όπως πολλές φορές απαντάται Ομαδοποίηση (clustering), η οποία αποτελεί μέθοδο μάθησης χωρίς επίβλεψη (unsupervised learning). Στη συγκεκριμένη μέθοδο επιχειρείται ο διαχωρισμός ενός συνόλου δεδομένων σε επιμέρους ομάδες με στόχο τη σύσταση ομάδων με όσο το δυνατόν πιο ομοιογενή μεταξύ τους χαρακτηριστικά. Δεν υπάρχουν προκαθορισμένες κατηγορίες ή συγκεκριμένες κλάσεις στις οποίες είναι κατανεμημένα τα δεδομένα αλλά ομαδοποιούνται βάση της κοινής ομοιότητας η οποία δύναται να τα χαρακτηρίζει. Βέλτιστη επιδίωξη αποτελεί η σύσταση ομάδων ή συστάδων τα δεδομένων των οποίων εμφανίζουν μεταξύ τους τις μεγαλύτερες ομοιότητες σε σχέση με τις λοιπές διαφορετικές συστάδες. Τελικός στόχος της συγκεκριμένης διαδικασίας είναι ο εντοπισμός προτύπων σε «ιδανικές» συστάδες που θα μας οδηγήσουν σε εξαγωγή συμπερασμάτων για αυτές αλλά και η ανακάλυψη των μεταξύ τους ομοιοτήτων και διαφορών[2]–[5].

Υποκατηγορία της τεχνητής νοημοσύνης και της μηχανικής μάθησης αποτελεί η μάθηση με επίβλεψη. Προσδιορίζεται από τη χρήση προκαθορισμένων συνόλων δεδομένων για την εκπαίδευση αλγορίθμων που κατατάσσουν δεδομένα ή εκτιμούν αποτελέσματα με ακρίβεια. Όσο το μοντέλο ενισχύεται με δεδομένα εισόδου, τροποποιεί τα βάρη του έως ότου το μοντέλο εναρμονιστεί κατάλληλα, κατάσταση που είναι μέρος της διεργασίας διασταυρούμενης επικύρωσης (cross validation). Η μάθηση με επίβλεψη είναι εφικτό υπό προϋποθέσεις να βοηθήσει τους οργανισμούς να λύσουν μια ποικιλία προβλημάτων της καθημερινότητας σε μία πραγματική κλίμακα.[6], [7]

Στον αντίποδα, η μάθηση χωρίς επίβλεψη είναι ένα είδος μηχανικής μάθησης όπου ένα μοντέλο πρέπει να αναζητήσει μοτίβα σε ένα μη επισημασμένο σύνολο δεδομένων και με ελάχιστη ανθρώπινη επίβλεψη. Στη συγκεκριμένη μέθοδο προσφέρονται μόνο οι είσοδοι και ένα μοντέλο πρέπει να εντοπίσει ενδιαφέροντα μοτίβα στα δεδομένα [8]. Η μάθηση χωρίς επίβλεψη περιγράφει μια κατηγορία προβλημάτων όπου απαιτείται ένα μοντέλο για την περιγραφή ή την εξαγωγή σχέσεων σε δεδομένα, είναι επωφελής στην διερευνητική ανάλυση δεδομένων, ανιχνεύοντας αυτόματα τη δομή και τις σχέσεις στα δεδομένα και παρέχει αρχικές γνώσεις που χρησιμοποιούνται για τον έλεγχο μεμονωμένων υποθέσεων[9].

Οι τεχνικές μάθησης με επίβλεψη είναι καταλληλότερες όταν επιδιώκεται η πρόβλεψη μιας συγκεκριμένης τιμής στόχου για μια ομάδα δεδομένων ενώ στη μάθηση

χωρίς επίβλεψη επιδιώκεται η ανάδειξη μιας πιθανής κρυφής δομής και σχέσης μεταξύ των δεδομένων[10]. Τα αντικείμενα που επιχειρεί να μελετήσει η εξόρυξη δεδομένων εκπαίδευσης ή Educational Data Mining (EDM) είναι η μοντελοποίηση μαθητών, τα συστήματα υποστήριξης αποφάσεων, τα προσαρμοστικά συστήματα και ζητήματα που αφορούν την αξιολόγηση και την έρευνα. Η μοντελοποίηση μαθητών στοχεύει στην πρόβλεψη της απόδοσης και στην πρόβλεψη χαρακτηριστικών που υποδηλώνουν ανεπιθύμητες συμπεριφορές. Πέρα από την ανακάλυψη του προφίλ των αντικειμένων μιας δομής και την ομαδοποίηση τους, μελετάτε και η ανάλυση κοινωνικών δικτύων. . Στα συστήματα υποστήριξης αποφάσεων ο στόχος είναι η παροχή αναφορών, η δημιουργία ειδοποιήσεων για τους ενδιαφερόμενους, ο σχεδιασμός και ο προγραμματισμός, η δημιουργία διδακτικού υλικού, η ανάπτυξη εννοιολογικών χαρτών, η δημιουργία συστάσεων[11]. Στην εξόρυξη δεδομένων εκπαίδευσης (EDM), όπως θα ήταν αναμενόμενο, χρησιμοποιούνται διάφορες προγνωστικές ή περιγραφικές δραστηριότητες και μέθοδοι εξόρυξης δεδομένων, η ταξινόμηση και η ομαδοποίηση είναι οι πιο συνήθης από αυτές [12]. Επιπλέον, οι περιγραφικές μέθοδοι που χρησιμοποιούνται σε εργασίες ομαδοποίησης, δημιουργούν ομάδες που έχουν παρόμοια δομή, σχέσεις και διασυνδέσεις των δεδομένων που έχουν εξορυχθεί με τη χρήση συναρτήσεων μάθησης χωρίς επίβλεψη [13].

1.1 Μέθοδοι και Αλγόριθμοι συσταδοποίησης

1.1.1 Κατηγοριοποίηση αλγορίθμων με βάση τη μέθοδο συσταδοποίησης

Στη συγκεκριμένη μέθοδο ανάλογα με τον τρόπο με τον οποίο γίνεται ο καθορισμός των συστάδων έχουμε τους εξής τύπους αλγορίθμων: Διαιρετικής συσταδοποίησης, Ασαφούς συσταδοποίησης, Μη ασαφούς συσταδοποίησης, Συσταδοποίηση βασισμένη σε δίκτυα Kohonen, Ιεραρχική συσταδοποίηση, Συσταδοποίηση βάση πυκνότητας, Συσταδοποίηση βασισμένη σε πλέγμα και τέλος Συσταδοποίηση υποχώρων [2]. Ενδεικτικοί αλγόριθμοι είναι οι K – Means, Fuzzy C – Means, PAM (Partitioning Around Medoids), Cure, Birch, DBSCAN, Sting, Clique, Rock κ.α. [14]. Συγκεκριμένα, σε εκπαιδευτικά περιβάλλοντα βρίσκουν εφαρμογή όπως έχει αποδειχθεί όλες οι τεχνικές και έχουν αναπτυχθεί διαφορετικές περιπτωσιολογικές μελέτες για την αξιολόγηση της απόδοσης διαφορετικών τεχνικών και την επίτευξη των

κύριων στόχων της εκπαιδευτικής εξόρυξης δεδομένων, που είναι ο εντοπισμός προτύπων συμπεριφοράς των μαθητών στο ακαδημαϊκό τους περιβάλλον,

1. Ταξινόμηση των μαθητών με βάση τις καταγεγραμμένες επιδόσεις,
2. Κατηγοριοποίηση των εκπαιδευτικών με βάση τις δραστηριότητές τους και τη χρήση των πλατφορμών,
3. Προσδιορισμός πρότυπων επιτυχίας στη χρήση εικονικών περιβαλλόντων μάθησης κτλ. [15], [16].

Η επιλογή της κατάλληλης μεθόδου συσταδοποίησης εξαρτάται από πολλές και διαφορετικές συνιστώσες όπως είναι οι απαιτήσεις των διαφορετικών υπό έρευνα εφαρμογών, ο τύπος του συνόλου δεδομένων που εξετάζουμε αλλά και τα ζητούμενα και οι ειδικές παράμετροι που θέτονται από τον ίδιο τον ερευνητή ώστε να ικανοποιούνται με τον καλύτερο δυνατό τρόπο οι στόχοι που έχουν τεθεί για την εξόρυξη των δεδομένων με σκοπό να προσδιορίσουν και να περιγράψουν τα πρότυπα γνώσης που μπορούν δυνητικά να εξαχθούν από ένα σύνολο δεδομένων άρα να επιτευχθεί και η ανακάλυψη νέας γνώσης.

1.1.1.1 Διαιρετική συσταδοποίηση (Partitional clustering)

Γενική περιγραφή: Βασίζεται στη διαίρεση των δεδομένων του συνόλου σε ένα σύνολο συστάδων που δεν αλληλοσχετίζονται. Στόχος είναι η μείωση της ανομοιότητας μεταξύ των δεδομένων της ίδιας συστάδας και η αύξηση της ανομοιότητας μεταξύ των διαφορετικών συστάδων [2]. Γενικά υπάρχουν δύο τύποι διαιρετικής συσταδοποίησης hard και soft ανάλογα αφενός με το αν ένα σημείο δεδομένων ανήκει σε μία συστάδα ή όχι (hard clustering), και αφετέρου εάν ένα σημείο δεδομένων μπορεί να ανήκει σε μία ή περισσότερες συστάδες (soft clustering). Οι κυριότεροι αλγόριθμοι αυτής της κατηγορίας είναι ο K – Means, PAM και CLARA [17]. Αρχικά, πρέπει να καθοριστεί ο αριθμός των συστάδων που πρέπει να δημιουργηθούν για τις μεθόδους συσταδοποίησης. Στη διαιρετική συσταδοποίηση, όταν μία βάση δεδομένων περιέχει πολλαπλά αντικείμενα, τότε η διαιρετική μέθοδος κατασκευάζει κατατμήσεις των δεδομένων που καθορίζονται από τον χρήστη, στις οποίες κάθε διαχωρισμός αντιπροσωπεύει μία συστάδα και μια συγκεκριμένη περιοχή [18], [19]. Δημοφιλή αλγόριθμο της συγκεκριμένης μεθόδου αποτελεί ο K – Means.

Ο γενικός τύπος του αλγορίθμου K - Means έχει την παρακάτω μορφή: [20]

$$E = \sum_{i=1}^{c} \sum_{x \in Ci} d(x, m_i)$$

Όπου m_i το κέντρο της συστάδας C_i, ενώ d (x, m_i) η Ευκλείδεια απόσταση μεταξύ ενός στοιχείου x και του κέντρου m_i.

Παράδειγμα εφαρμογής: Ευρέως διαδεδομένη εφαρμογή του K – means απαντάται στον τομέα του μάρκετινγκ όπως για παράδειγμα η μελέτη διάφορων παραμέτρων τις οποίες καλείται να εξετάσει μία αλυσίδα πίτσας ώστε να ανοίξει νέα υποκαταστήματα - κέντρα διανομής σε μία πόλη. Εφόσον οριστεί η γεωγραφική περιοχή δημιουργήθηκαν περιοχές γειτονιάς και καθορίστηκαν τα γεωγραφικά κεντροειδή. Στη συνέχεια, καθορίστηκαν οι καλύτερες συστάδες γειτονιών όπου αναλογικά υπήρχαν περισσότερα εστιατόρια πιτσαρίας. Τέλος, εντοπίστηκαν από αυτές όσες θεωρητικά ήταν οι «καλύτερες» λόγω του ότι υπήρχαν σχετικά λίγα εστιατόρια πίτσας άρα αυτές μπορούν δυνητικά να αποτελέσουν καλές υποψήφιες περιοχές για το άνοιγμα μίας νέας πιτσαρίας [21].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Ο αλγόριθμος K – Means βρίσκει εφαρμογές σε διάφορους τομείς της πραγματικής ζωής μεταξύ των οποίων και ο τομέας της εξόρυξης δεδομένων σχετικών με την εκπαίδευση. Όπως για παράδειγμα:

1. Αξιολόγηση του τρόπου διδασκαλίας μαθημάτων βασικών δεξιοτήτων υπολογιστή σε μαθητές με αγροτική ή αστική καταγωγή.
2. Μελέτη περίπτωσης σύγκρισης ερωτηματολογίων με σκοπό την αξιολόγηση της συναισθηματικής νοημοσύνης μαθητών.
3. Περίπτωση όπου για να βοηθηθούν οι προπτυχιακοί φοιτητές να αποδώσουν καλύτερα, υπάρχουν ορισμένοι σημαντικοί παράγοντες που πρέπει να ληφθούν υπόψη. Μεταξύ αυτών είναι το υπόβαθρο, οι εμπειρίες, οι στόχοι των μαθητών κ.α. [22]. Ανάλογα λοιπόν, με τα επιμέρους χαρακτηριστικά, δηλαδή, τον τόπο καταγωγής (1), την κατάταξη των απαντήσεων από το ερωτηματολόγιο σε ομάδες (2) και παραμέτρους που μπορούν να συντελέσουν στην επίδοση ενός φοιτητή (3) πραγματοποιήθηκε σε κάθε επιμέρους περίπτωση δημιουργία συστάδων και ανάθεση στοιχείων σε αυτές βάση των ειδικών χαρακτηριστικών που τίθενται κάθε φορά κατά περίπτωση.

Συμπεράσματα: Σε κάθε περίπτωση εφαρμογής διαιρετικής συσταδοποίησης σε ένα σύνολο δεδομένων έχουμε τη διαίρεση του συνόλου σε επιμέρους συστάδες με σκοπό

τη δημιουργία ομάδων στοιχείων με κοινά χαρακτηριστικά όσο το δυνατών πιο ομοιογενών μεταξύ τους και όσο το δυνατόν περισσότερο ανομοιογενών μεταξύ των γειτονικών τους συστάδων, ενώ ανάλογα τον τύπο ένα στοιχείο μπορεί να ανήκει αποκλειστικά σε μία συστάδα ή όχι. Από αυτή τη διαδικασία μπορούμε να εξάγουμε σημαντικές πληροφορίες για τις ομαδοποιήσεις που προκύπτουν και τα μοτίβα που ανακαλύπτονται μέσα από αυτό για τα υπό έρευνα δεδομένα.

1.1.1.2 Ασαφής συσταδοποίηση (Fuzzy clustering)

Γενική περιγραφή: Χρησιμοποιεί ασαφή λογική ώστε να ομαδοποιήσει τα δεδομένα σε συστάδες και θεωρεί ότι κάθε αντικείμενο μπορεί να ταξινομηθεί σε πολλές συστάδες για αυτό και προσομοιάζει σε μεγάλο βαθμό την αβεβαιότητα που χαρακτηρίζει δεδομένα της πραγματικής ζωής. Μερικοί από τους πιο ευρέως εφαρμοζόμενους αλγόριθμους ασαφούς συσταδοποίησης είναι οι αλγόριθμοι Fuzzy C-means (FCM), ο EM (Expectation Maximization) και ο Apriori [2]. Η λειτουργία του Fuzzy C-means είναι παρόμοια με του αλγόριθμου K-means και αποτελείται από τα ακόλουθα βήματα::

1. Επιλογή ενός πλήθους συστάδων.
2. Εκχώρηση τυχαίων συντελεστών σε κάθε σημείο δεδομένων στο σύμπλεγμα συστάδων.
3. Επανάληψη των βημάτων II, III έως ότου υπάρξει σύγκλιση του αλγορίθμου δηλαδή, η μεταβολή των συντελεστών μεταξύ δύο επαναλήψεων δεν υπερβαίνει τη μεταβλητή ε που αποτελεί το δεδομένο όριο ευαισθησίας:
4. Υπολογισμός του κέντρου κάθε συστάδας.
5. Υπολογισμός των συντελεστών κάθε σημείου δεδομένων που βρίσκεται στις συστάδες [23]–[27].

Ο γενικός τύπος του αλγόριθμου Fuzzy C - Means έχει την παρακάτω μορφή:

$$J_m\,(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} U_{ik}^{m}\; d^2\left(X_{k,}\; V_i\right)$$

Όπου U ∈ M_{fcn} ο πίνακας των συστάδων, $V = \left[v_{1,} \ldots, v_c\right] \in R^{sxc}$ τα κέντρα των ομάδων, c ο αριθμός των εξεταζόμενων ομάδων, n ο αριθμός στοιχείων και $m \geq 1$ η παράμετρος ασάφειας ή ασαφοποιητής.

- Εάν $m \rightarrow 1$ τότε οι συστάδες τείνουν να είναι μη ασαφής δηλαδή $U_{ik} \rightarrow 1$ ή $U_{ik} \rightarrow 0$.

- Εάν $m \rightarrow \infty$ τότε $U_{ik} \rightarrow 1/c$ τα στοιχεία ανήκουν σε όλες τις συστάδες.

Παράδειγμα εφαρμογής: Αποτελεί αλγόριθμο ο οποίος γενικά προτιμάτε για την ανάλυση μεγάλων συνόλων δεδομένων (big data) με ποικίλες εφαρμογές όπως για παράδειγμα βϊοιατρικές εφαρμογές όπως η φυσιολογική παρακολούθηση της υγείας, η απεικόνιση και ο προσδιορισμός της αλληλουχίας ορισμένων βιολογικών μετρήσεων. Χρησιμοποιήθηκαν διαφορετικά μέτρα για τον προσδιορισμό χαρακτηριστικών ομάδων σε ένα σύνολο δεδομένων όπου οι μετρήσεις από 45 διαφορετικούς αισθητήρες είχαν μια σειρά τιμών και διέφεραν ως προς τις δραστηριότητες που μπορούσαν να εκτελεστούν. Αυτό περιλάμβανε τη χρήση τριών διαφορετικών μεθόδων κανονικοποίησης (χωρίς κανονικοποίηση, χρησιμοποιώντας το μέγιστο για κάθε χαρακτηριστικό και στατιστική τυποποίηση με τον μέσο όρο και την τυπική απόκλιση κάθε χαρακτηριστικού) και τεσσάρων μέτρων απόστασης και διαφοροποίησης της μέτρησης της απόστασης δραστηριότητας (Ευκλείδεια απόσταση, κλιμακούμενη Ευκλείδειος, συνημίτονο και συσχέτιση) [28].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Άλλο χαρακτηριστικό παράδειγμα εφαρμογής σε ότι αφορά την εξόρυξη δεδομένων εκπαίδευσης του συγκεκριμένου αλγορίθμου αποτελεί η συσταδοποίηση των προτύπων πρόσβασης των μαθητών ή της συμπεριφοράς πλοήγησης σε περιβάλλοντα ηλεκτρονικής μάθησης με σκοπό τη βελτίωση των συστημάτων ηλεκτρονικής μάθησης. Οι εκπαιδευτές αυτών των συστημάτων μπορούν να χρησιμοποιήσουν τεχνικές οπτικοποίησης για να αποκτήσουν μια γενική εικόνα των δεδομένων χρήσης του μαθητή ή να εφαρμόσουν τεχνικές ασαφούς ομαδοποίησης προκειμένου να λάβουν τις ακριβείς ομάδες στις οποίες μπορούν να χωριστούν οι μαθητές. Ενώ επίσης, μπορούν να εφαρμόσουν την εξόρυξη κανόνων συσχέτισης με εφαρμογή του αλγορίθμου Apriori για να ανακαλύψουν εάν υπάρχει κάποια σχέση μεταξύ αυτών των χαρακτηριστικών και άλλων ιδιοτήτων αλλά και στον εντοπισμό των πηγών οποιωνδήποτε ανακόλουθων τιμών που αποκτήθηκαν από τους μαθητές [22]

Συμπεράσματα: Ανεξαρτήτως του πεδίου εφαρμογής της ασαφούς συσταδοποίησης αποτελεί μέθοδο η οποία δημιουργεί κατατμήσεις δεδομένων που περιέχουν παρόμοια θέματα. Είναι ένας αυτόνομος τύπος μάθησης χωρίς επίβλεψη για την ταξινόμηση των προτύπων των συνόλων δεδομένων μέσω της διερεύνησης τους. Η τάση της υιοθέτησης

της μηχανικής μάθησης, της επιστήμης μεγάλων δεδομένων, του υπολογιστικού νέφους σε διάφορους κλάδους εξαρτάται από τη μη εποπτευόμενη μάθηση στις δομές δεδομένων για να «αφηγηθεί» την ιστορία για τη συμπεριφορά για παράδειγμα των καταναλωτών, τον εντοπισμό απάτης και την τμηματοποίηση της αγοράς και πολλών άλλων ζητημάτων που ερευνώνται στην εξόρυξη δεδομένων [29].

1.1.1.3 Μη ασαφής συσταδοποίηση (Crisp clustering)

Γενική περιγραφή: Θεωρεί ότι κάθε δεδομένο ενός συνόλου ανήκει ή όχι σε μία συγκεκριμένη κατηγορία [2]. Ένα χαρακτηριστικό της μεθόδου μη ασαφούς συσταδοποίησης είναι ότι το όριο μεταξύ των συστάδων ορίζεται πλήρως [30]. Η συνήθης επικύρωση στον μη ασαφή διαμερισμό συνοψίζεται στις εξής δύο υποθέσεις: Έστω ότι Α και Β δύο τμήματα ενός δεδομένου συνόλου δεδομένων. Το τμήμα Α είναι καλύτερο από το τμήμα Β εάν και μόνο εάν η μέση κανονικότητα των συστάδων στο Α είναι υψηλότερη από τη μέση κανονικότητα των συστάδων στο Β ή σε διαφορετική περίπτωση το τμήμα Α είναι καλύτερο από το τμήμα Β εάν και μόνο εάν η μέση αρνητική εντροπία όλων των συστάδων στο Α είναι χαμηλότερη από τη μέση αρνητική εντροπία όλων των συστάδων στο Β [31]. Σε αυτή την κατηγορία εντάσσεται και ο γενετικός αλγόριθμος (Genetic Algorithm) με εφαρμογές στην επιστήμη και τη μηχανική αλλά και σε άλλους τομείς της πραγματικής ζωής. Τα κυριότερα στοιχεία που είναι κοινά σε όλους σχεδόν τους γενετικούς αλγόριθμους είναι:

1. Η συνάρτηση καταλληλόλητας (fitness function) για βελτιστοποίηση.
2. Πληθυσμός χρωμοσωμάτων.
3. Επιλογή των προς αναπαραγωγή χρωμοσωμάτων.
4. Διασταύρωση (crossover) για την παραγωγή χρωμοσωμάτων νέας γενιάς.
5. Τυχαία μετάλλαξη χρωμοσωμάτων στη νέα γενιά [32].

Παράδειγμα εφαρμογής: Χαρακτηριστικό παράδειγμα αποτελεί η βελτιστοποίηση της ταξινόμησης μιας αποθήκης με σκοπό την καλύτερη διαχείριση της διάρκειας των χρόνων επεξεργασίας των αποστολών και του αριθμού των εργαζομένων που απασχολούνται για αυτό. Η ποσότητα του όγκου διαλογής σε μια αποθήκη δεν είναι σταθερή και τείνει να κυμαίνεται ανάλογα με τον όγκο των αποστολών κάθε μέρα επομένως δεν είναι γνωστό εκ των προτέρων πόσοι ακριβώς υπάλληλοι θα χρειαστούν κάθε μέρα, σε καθημερινή βάση απαιτείται να διασφαλίζεται ότι οι ταξινομημένες

αποστολές φορτώνονται σε ρυμουλκούμενα μέσα μεταφοράς το συντομότερο δυνατό προς όλες τις κατευθύνσεις. Υπάρχει μια διαδικασία για τον προσδιορισμό του αριθμού των εργαζομένων που χρειάζονται για την αποθήκη διαλογής για έναν ορισμένο αριθμό αποστολών σε διάφορα χρονικά διαστήματα χρησιμοποιώντας γενετικούς αλγόριθμους. Αυτό απαιτείται προκειμένου να δημιουργηθούν προσομοιώσεις ώστε να μπορεί να προσδιοριστεί ο βέλτιστος αριθμός εργαζομένων [33].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Επίσης, παράδειγμα εφαρμογής μεθόδου Μη – ασαφούς συσταδοποίησης σε εκπαιδευτικά δεδομένα αποτελεί η μελέτη και πρόβλεψη της ανάρμοστης συμπεριφοράς φοιτητών Ανώτατων Εκπαιδευτικών Ιδρυμάτων μέσω της δημιουργίας μοντέλων ομαδοποίησης και κανόνων για την ομαδοποίηση και την πρόβλεψη της συμπεριφοράς των φοιτητών και τον εντοπισμό της ανάρμοστης συμπεριφοράς στις πανεπιστημιουπόλεις. Στόχος είναι η μείωση των ανάρμοστων συμπεριφορών εντοπίζοντας τους παράγοντες που τις προκαλούν στις πανεπιστημιουπόλεις για αυτό το λόγο, λαμβάνονται υπόψιν και ερευνώνται οι εξής παράγοντες: Φοιτητική κατεύθυνση, Επίπεδο Φοιτητή, Φύλο, Σωρευτικός μέσος όρος βαθμολογίας, Τοπική Διεύθυνση, Εθνικότητα και χρόνο παραπτώματος ανά μήνα. Για το σκοπό αυτό έχει εφαρμοστεί η μέθοδος της μη ασαφούς συσταδοποίησης για την εξόρυξη των δεδομένων των φοιτητών και τον εντοπισμό των παραγόντων που προκαλούν παραπτώματα στην πανεπιστημιούπολη με εφαρμογή των τεχνικών εξόρυξης δεδομένων των Νευρωνικών Δικτύων (ANN) και των Δέντρων απόφασης (DT) [34].

Συμπεράσματα: Έτσι, διαπιστώνεται ότι ανεξαρτήτως πεδίου εφαρμογής η συγκεκριμένη μέθοδος επιδιώκει τη δημιουργία σαφών συστάδων, αρκετοί λοιπόν αλγόριθμοι οδηγούν σε δημιουργία σαφών συστάδων όπου ένα στοιχείο του συνόλου δεδομένων είτε ανήκει σε μία κατηγορία, είτε όχι.

1.1.1.4 Συσταδοποίηση βασισμένη σε δίκτυα Kohonen (Kohonen net clustering)

Γενική περιγραφή: Βασίζεται στη λογική των Νευρωνικών δικτύων έτσι αλλάζοντας τα βάρη των συνδέσεων καθορίζεται η θέση των κόμβων εξόδου οι οποίοι εντέλει σχηματίζουν συστάδες [2]. Ο συγκεκριμένος αλγόριθμος θεωρείται αρκετά περίπλοκος.

Όλα τα συμπλέγματα διασυνδέονται σύμφωνα με τον τοπολογικό χάρτη. Όσον αφορά τον χάρτη τοπολογίας, όταν το σύμπλεγμα που βρίσκεται πιο κοντά στο σημείο δεδομένων (το πρωτεύον σύμπλεγμα) ενημερώνεται, ενημερώνεται και το σύμπλεγμα αμέσως μετά από αυτό (το πλησιέστερο σύμπλεγμα). Από λειτουργική άποψη, αρχικά οι γείτονες του πρωτεύοντος συμπλέγματος μπορεί να μην είναι οι τοπολογικοί του γείτονες. Λόγω της φύσης του κανόνα ενημέρωσης, οι γείτονες του πρωτεύοντος συμπλέγματος στον χώρο τοπολογίας γίνονται γείτονες στο χώρο χαρακτηριστικών μετά από κάποιο χρονικό διάστημα [35]–[37]. Τα δίκτυα Kohonen βρίσκουν εφαρμογές σε τομείς όπως η ομαδοποίηση ροών δεδομένων σε ερευνητικό επίπεδο μέσω νευρο – ασαφών δικτύων Kohonen (neuro-fuzzy Kohonen network). Επίσης, σε συνδυασμό με την υποστήριξη διανυσματικών μηχανών μπορούν να χρησιμοποιηθούν με σκοπό την ανίχνευση εισβολών, την αναγνώριση εικόνας, την επεξεργασία φωνής ενώ μπορεί επίσης να χρησιμοποιηθεί και για την ανάλυση δεδομένων.

Τα κυριότερα βήματα του αλγορίθμου Kohonen είναι:

1. Για κάθε νευρώνα λαμβάνεται αντίγραφο προτύπου εισόδου.
2. Βρίσκουμε το νευρώνα με το μικρότερο επίπεδο ενεργοποίησης, το «νικητή».

$$AL_j = \sqrt{\sum_{i=1}^{n} (W_{ij} - X_i)^2}$$

3. Για κάθε νευρώνα «νικητή» και τους φυσικούς γειτονικούς του κόμβους εφαρμόζεται ο παρακάτω κανόνας εκπαίδευσης για την τροποποίηση των βαρών:

$$W_{IJ}(t+1) = W_{ij}(t) + \alpha(t) * gamma(t) * \left[X_i - W_{ij}(t)\right] gamma(t)$$
$$= exp\left\{-0.5 * \left[\frac{r_{ij}}{sigma(t)}\right]^2\right\}$$

Όπου α ο ρυθμός μάθησης, r_{ij} είναι η απόσταση μεταξύ του «νικητή» και του κόμβου που θα ενημερωθεί και sigma η ακτίνα γειτονίας.

4. Επανάληψη βημάτων 1 – 3 για κάθε νέο πρότυπο εισόδου.
5. Επανάληψη βήματος 4 έως ότου όλα τα πρότυπα εισόδου εξεταστούν.
6. Επανάληψη του βήματος 5 για έναν καθορισμένο αριθμό φορών [38].

Παράδειγμα εφαρμογής: Χαρακτηριστικό παράδειγμα ανάλυσης δεδομένων μέσω δικτύων Kohonen αποτελεί η αναγνώριση λημμάτων της διαλέκτου Bankga της φυλής των Malay. Αυτού του τύπου η έρευνα σκοπεύει να ελαχιστοποιήσει τον φόρτο αναζήτησης πολλών δεδομένων λημμάτων χρησιμοποιώντας ένα δίκτυο Kohonen. Έτσι,

αρκεί να γίνει εκπαίδευση του δείγματος των παρεχόμενων δεδομένων που είναι πολύ λιγότερα από τα πραγματικά δεδομένα. Το αποτέλεσμα αυτής της εκπαίδευσης χρησιμοποιείται στη συνέχεια για να ληφθεί ο αριθμός των λημμάτων που αναγνωρίζονται με επιτυχία από το δίκτυο Kohonen επίσης, μπορεί να προσδιοριστεί η κατάσταση αποδοχής του προτεινόμενου λήμματος [39].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Σε ότι αφορά την εξόρυξη και μηχανική μάθηση δεδομένων εκπαίδευσης και εκεί φυσικά βρίσκουμε εφαρμογές των Νευρωνικών δικτύων και ειδικότερα δικτύων Kohonen. Παράδειγμα αποτελεί η ανάπτυξη ενός έξυπνου μοντέλου συστήματος διδασκαλίας (Intelligent Tutoring System) που είναι ικανό να οδηγήσει στη διδακτική αντιμετάθεση του περιεχομένου. Αρχικά, οι αντιδράσεις του συστήματος διδασκαλίας βασίζουν τη συμπεριφορά του σε κανόνες που ορίζονται από έναν ειδικό δάσκαλο. Μετά από αυτό, ένα νευρωνικό δίκτυο που μαθαίνει από τη συμπεριφορά του μαθητή όταν μελετάει προσαρμόζει αυτούς τους κανόνες. Με αυτόν τον τρόπο, το νευρωνικό δίκτυο βελτιώνει τους κανόνες του δασκάλου και, κατά συνέπεια, ορίζει μια στρατηγική μάθησης που είναι πιο προσαρμοστική και αλληλεπιδραστική στο προφίλ του μαθητή [40].

Συμπεράσματα: Η συγκεκριμένη μέθοδος προσπαθεί να προσομοιάσει τη χωρική οργάνωση των λειτουργιών ενός ανθρώπινου εγκεφάλου, ανεξάρτητα του πεδίου εφαρμογής μπορούμε να πούμε ότι μπορεί να εφαρμοστεί σε συστήματα που απαιτούν να αυτορυθμίζονται και να αυτοοργανώνοται. Εισαγάγει ένα μοντέλο συστήματος που αποτελείται από τουλάχιστον δύο αλληλεπιδρώντα υποσυστήματα διαφορετικής φύσης, ένα ανταγωνιστικό νευρωνικό δίκτυο που υλοποιεί τη συνάρτηση «ο νικητής τα παίρνει όλα» (winner – take - all) και ένα άλλο υποσύστημα που ελέγχεται από το νευρωνικό δίκτυο και το οποίο τροποποιεί την τοπική συναπτική πλαστικότητα των νευρώνων στη μάθηση. Η μάθηση περιορίζεται χωρικά στην τοπική γειτονιά των πιο ενεργών νευρώνων και εντέλει καθίσταται εφικτή η εφαρμογή ενός αποτελεσματικού και ισχυρού συστήματος αυτοργάνωσης [41].

1.1.1.5 Ιεραρχική συσταδοποίηση (Hierarchical clustering)

Γενική περιγραφή: Βασίζεται σε σύνδεση μικρότερων συστάδων σε μεγαλύτερες ή στο διαχωρισμό μεγαλύτερων συστάδων σε μικρότερες. Αποτέλεσμα αυτής της διαδικασίας αποτελεί το δεντρογράφημα στο οποίο αποτυπώνεται η συσχέτιση των συστάδων [2]. Οι μέθοδοι ιεραρχικής συσταδοποίησης διαφοροποιούνται ως προς τον τρόπο υπολογισμού των αποστάσεων. Εκτός από την επιλογή της συνάρτησης απόστασης, ο χρήστης πρέπει επίσης να αποφασίσει για τα κριτήρια σύνδεσης (καθώς ένα σύμπλεγμα αποτελείται από πολλά αντικείμενα, υπάρχουν πολλοί υποψήφιοι για τον υπολογισμό της απόστασης) που θα χρησιμοποιήσει. Οι δημοφιλείς επιλογές είναι γνωστές ως συσταδοποίηση μονής σύνδεσης (single – linkage) -η ελάχιστη απόσταση αντικειμένων-, συσταδοποίηση πλήρους σύνδεσης (complete linkage) -η μέγιστη απόσταση αντικειμένων- και η μέθοδος ομάδας μη σταθμισμένου ή σταθμισμένου ζεύγους με αριθμητικό μέσο όρο UPGMA ή WPGMA (Unweighted or Weighted Pair Group Method with Arithmetic Mean) γνωστή ως συσταδοποίηση σύνδεσης μέσου όρου. Επίσης, η ιεραρχική ομαδοποίηση μπορεί να είναι αθροιστική, ξεκινώντας με μεμονωμένα στοιχεία και συγκεντρώνοντάς τα σε συστάδες ή διαιρετική, ξεκινώντας από το πλήρες σύνολο δεδομένων και χωρίζοντάς το σε χωριστές συστάδες. Η παραπάνω προσέγγιση δεν παράγει ένα μοναδικό διαμέρισμα του συνόλου δεδομένων, αλλά μάλλον μια ιεραρχία από την οποία ο χρήστης πρέπει να επιλέξει το κατάλληλο σύμπλεγμα. Τέλος, έχουν μικρή ανοχή ως προς τα ακραία σημεία, τα οποία είτε θα εμφανιστούν ως πρόσθετες συστάδες είτε ακόμη και θα προκαλέσουν τη συγχώνευση άλλων συστάδων (γνωστό ως "φαινόμενο αλυσίδων", ιδίως στη συσταδοποίηση μονής σύνδεσης) [42]–[44].

Ο γενικός τύπος για τον αλγόριθμο UPGMA είναι ο παρακάτω:

$$D_{X.Y} = \frac{\Sigma Dxy}{n_x \cdot n_y}$$

Όπου X και Y είναι οι δύο συστάδες, n_x και n_y είναι ο αριθμός των αντικειμένων στις συστάδες X και Y, αντίστοιχα, x και y είναι αντικείμενα στις συστάδες X και Y, και D_{xy} είναι η απόσταση μεταξύ αντικειμένων x και y, και D_{XY} είναι η απόσταση μεταξύ των συστάδων X και Y [45]

Παράδειγμα εφαρμογής: Χαρακτηριστική μελέτη περίπτωσης εφαρμογής συσσωρευτικής ιεραρχικής συσταδοποίησης σχετίζεται με την εξόρυξη ενός αρκετά μεγάλου όγκου δεδομένων που αφορά τις αξιολογήσεις από περίπου ένα εκατομμύριο

αξιολογήσεις για 5253900 ταινίες από 6040 χρήστες και στόχο έχει να κατηγοριοποιήσει τις ταινίες βάσει των καταχωρημένων αξιολογήσεων. Όπου το σύνολο των δεδομένων αποτελείται από μικτούς τύπους χαρακτηριστικών αριθμητικά και κατηγορικά. Η ιεραρχική μέθοδος που χρησιμοποιήθηκε βασίζεται στον K-means όπου ορίστηκε ο αριθμός των συστάδων προς τον συνολικό αριθμό των αρχικών αντικειμένων δεδομένων και με βάση τα αποτελέσματα του K – Means δημιουργήθηκαν ιεραρχίες από τον αλγόριθμο Unweighted Pair-Group Method of Average (UPGMA) και τον αλγόριθμο Single Linkage (SLINK) στα κύρια πειράματα και από τον αλγόριθμο Unweighted Pair-Group Method of Centroids (UPGMC) στα εκτεταμένα πειράματα. Οι ιεραρχίες δημιουργήθηκαν απευθείας στα αρχικά αντικείμενα δεδομένων για σκοπούς σύγκρισης. Τέλος, ο συντελεστής συσχέτισης Pearson εκτελέστηκε για να καλύψει κάθε ζεύγος ιεραρχιών σε ένα ίδιο σύνολο δεδομένων, ένα από κεντροειδή που δημιουργήθηκε με K - Means και ένα από μεμονωμένα αντικείμενα [45].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Άλλη μελέτη περίπτωσης με εφαρμογή ιεραρχικής συσταδοποίησης σε εκπαιδευτικά δεδομένα αποτελεί η χαρτογράφηση των προσεγγίσεων στα προφίλ διδασκαλίας των εκπαιδευτικών στην τριτοβάθμια εκπαίδευση με βάση τα αποτελέσματα τους στον κατάλογο απογραφής διδακτικών προσεγγίσεων (Approaches to Teaching Inventory). Αυτό απαιτεί την κατηγοριοποίηση των ατόμων σε ομάδες που μοιράζονται παρόμοιες προσεγγίσεις στα στυλ διδασκαλίας. Διεξήχθη ιεραρχική ομαδική ανάλυση στα δεδομένα του ερωτηματολογίου χρησιμοποιώντας τις μεταβλητές «εννοιολογική αλλαγή», «συζήτηση: δάσκαλος-μαθητής», «συζήτηση: μαθητές», «μεταφορά πληροφοριών» και «εστίαση τεστ» για να κατηγοριοποιηθούν παρόμοιες περιπτώσεις στα δεδομένα. Χρησιμοποιήθηκε η μέση σύνδεση εντός των ομάδων και η τετραγωνισμένη Ευκλείδεια απόσταση χρησιμοποιήθηκε για τον υπολογισμό της μέσης απόστασης μεταξύ όλων των πιθανών ζευγών συστάδων. Η μέση συνδεσιμότητα ομάδων ασχολείται περισσότερο με την ομοιογένεια εντός των ομάδων, αυτό θεωρήθηκε ως το πιο σημαντικό χαρακτηριστικό της ομαδοποίησης. επομένως, τονίζεται η ομοιογένεια. Το τετράγωνο της Ευκλείδειας απόστασης είναι η προεπιλογή για τα αριθμητικά δεδομένα που έχουν τιμή διαστήματος. Ο αριθμός των ομάδων υπολογίστηκε με βάση τις ακόλουθες απαιτήσεις: κάθε περίπτωση έπρεπε να είναι μέρος μιας ομάδας, η μεγαλύτερη από τις ομάδες έπρεπε να έχει έναν λογικό αριθμό περιπτώσεων και έπρεπε να υπάρχει μια σημαντική αύξηση στην απόσταση [22].

Συμπεράσματα: Η ιεραρχική ομαδοποίηση είναι ιδιαίτερα σημαίνουσα στην ανάλυση δεδομένων, ειδικά λόγω της αυξητικής τάσης των δεδομένων του πραγματικού κόσμου. Μια μέθοδος συσταδοποίησης που απαιτεί λιγότερο υπολογιστικό κόστος μπορεί να είναι επωφελής στη γενική εξόρυξη δεδομένων και στην ανακάλυψη γνώσης, καθώς και σε συγκεκριμένους τομείς π.χ. στη βιοπληροφορική, στην παρακολούθηση χρήσης ιστού και στην ανάλυση κοινωνικών δικτύων αλλά και σε άλλα πεδία εφαρμογής του πραγματικού κόσμου. Λόγω της ευρείας διάδοσης των διαδικτυακών εφαρμογών, των φορητών συσκευών και του δικτύου αισθητήρων, ο όγκος των προς ανάλυση δεδομένων αυξάνεται πολύ πιο γρήγορα από την υπολογιστική ισχύ, ειδικά τα τελευταία χρόνια.

1.1.1.6 Συσταδοποίηση βάση πυκνότητας (Density – based clustering)

Γενική περιγραφή: Βασίζεται στο σχηματισμό συστάδων με κριτήριο την πυκνότητα που εντοπίζεται μεταξύ γειτονικών στοιχείων [2]. Η θεμελιώδης έννοια της ομαδοποίησης με βάση την πυκνότητα είναι ότι για κάθε περίπτωση ενός συμπλέγματος, η γειτονιά μιας δεδομένης ακτίνας (Eps) πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό περιπτώσεων (MinPts). Ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους για ομαδοποίηση δεδομένων με βάση την πυκνότητα είναι ο DBSCAN [11]. Η ομαδοποίηση με βάση την πυκνότητα είναι μια μη παραμετρική μέθοδος που θεωρεί τις συστάδες ως περιοχές υψηλής πυκνότητας, με πυκνότητα p(x). Οι μέθοδοι συσταδοποίησης που βασίζονται στην πυκνότητα δεν απαιτούν ως παράμετρο εισόδου των αριθμό των συστάδων, ούτε κάνουν υποθέσεις σχετικά με την υποκείμενη πυκνότητα p(x) ή τη διακύμανση εντός των συστάδων που μπορεί να υπάρχει στο σύνολο δεδομένων. Ως αποτέλεσμα, οι ομάδες με βάση την πυκνότητα δεν είναι απαραίτητα ομάδες σημείων με χαμηλή ανομοιότητα μεταξύ συστάδων, όπως μετράται από μια συνάρτηση ανομοιότητας, και επομένως δεν έχουν απαραίτητα κυκλικό σχήμα, αλλά μπορούν να διαμορφωθούν αυθαίρετα στο χώρο δεδομένων [46], [47]. Η συγκεκριμένη μέθοδος βρίσκει εφαρμογή σε διάφορους τομείς του πραγματικού κόσμου και θεωρείται αρκετά επιτυχημένη στη χωρική ομαδοποίηση εφαρμογών με θόρυβο βάσει πυκνότητας [19].

Η διαδικασία του DBSCAN περιγράφεται ως εξής, ο αλγόριθμος αποτελείται από τα ακόλουθα βήματα:

1. Βρίσκει τα σημεία κοντά σε κάθε σημείο που βρίσκονται εντός eps κάθε σημείου και προσδιορίζει τα σημεία πυρήνα με τον μεγαλύτερο αριθμό γειτόνων minPts.
2. Προσδιορίζει τα συνδεδεμένα στοιχεία των σημείων πυρήνα στο γράφημα γειτονικών, αγνοώντας όλα τα σημεία που δεν είναι σημεία πυρήνα.
3. Αντιστοιχίζει κάθε μη πρωτεύον σημείο με ένα κοντινό σύμπλεγμα εάν το σύμπλεγμα βρίσκεται εντός ε (eps) από ένα πρωτεύον σημείο, διαφορετικά το αντιστοιχίζει με θόρυβο [48].

Παράδειγμα εφαρμογής: Μια τυπική μελέτη περίπτωσης είναι η ομαδοποίηση του φασματικού χώρου (5-διάστατα σημεία) που δημιουργείται από δορυφορικές εικόνες διαφορετικών φασματικών καναλιών, η οποία είναι μια συνήθης εργασία στην ανάλυση εικόνων τηλεπισκόπησης.. Η εύρεση συστάδων σε τέτοιους χαρακτηριστικούς χώρους είναι συνηθισμένη εργασία στην ανάλυση ψηφιακής εικόνας με τηλεπισκόπηση για τη δημιουργία θεματικών χαρτών σε γεωγραφικά συστήματα πληροφοριών. Η υπόθεση είναι ότι τα ιδιοδιανύσματα ομοιογενών σημείων στο υπέδαφος της Γης σχηματίζουν συστάδες σε έναν ιδιοχώρο υψηλών διαστάσεων. Για το σκοπό αυτό έγινε εφαρμογή του Generalized DBSCAN (GDBSCAN) [49].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Άλλη μελέτη περίπτωσης εφαρμογής σε εκπαιδευτικά δεδομένα της συσταδοποίησης βάση πυκνότητας αποτελεί η σύσταση - πρόταση εκπαιδευτικών πόρων σε εκπαιδευόμενους. Με την ταχεία ανάπτυξη των εκπαιδευτικών πλατφορμών στο διαδίκτυο, οι εκπαιδευόμενοι χρειάζονται έναν αποτελεσματικό τρόπο για να επιτύχουν κατάλληλους πόρους μάθησης. Είναι επομένως χρήσιμο να ομαδοποιήσουμε παρόμοια εκπαιδευτικά αντικείμενα και να προτείνουμε εξατομικευμένα δεδομένα στους εκπαιδευόμενους. Ωστόσο, λόγω της ποικιλίας των πόρων μάθησης και των αντικειμένων θορύβου, είναι δύσκολο για τους εκπαιδευόμενους να αποκτήσουν κατάλληλους πόρους. Οι αλγόριθμοι πυκνότητας μπορούν να εκτελεστούν με βάση την πυκνότητα των δεδομένων για να προτείνουν ένα συμπαγές και εξατομικευμένο σύνολο αντικειμένων στον χρήστη [50].

Συμπεράσματα: Η συσταδοποίηση βάση πυκνότητας μπορεί να εφαρμοστεί σε μία πληθώρα ερευνητικών πεδίων και τομέων της πραγματικής ζωής στην περίπτωση όπου η επιδίωξη δεν είναι η πρόβλεψη μιας συγκεκριμένης μεταβλητής αλλά η εύρεση μοτίβων και συσχετίσεων βάσει κάποιων χαρακτηριστικών σε ένα σύνολο δεδομένων. Επίσης όταν ο όγκος των δεδομένων είναι τόσο μεγάλος όπου δεν μπορούν να

εντοπιστούν χειροκίνητα συσχετίσεις και μοτίβα αλλά και όταν υπάρχει πολύ θόρυβος στα δεδομένα.

1.1.1.7 Συσταδοποίηση βασισμένη σε πλέγμα (Grid – based clustering)

Γενική περιγραφή: Βασίζεται στο χωρισμό του χώρου σε συγκεκριμένο αριθμό κελιών μέσα στον οποίο επιτελείται η διαδικασία της συσταδοποίησης [2]. Αποτελεί γρήγορη τεχνική και έχει χαμηλή υπολογιστική πολυπλοκότητα. Υπάρχει αφενός η μέθοδος συσταδοποίησης που βασίζεται σε πλέγμα: η STING (Statistical Information Grid – based method) όπου εφαρμόζει και τον ομώνυμο αλγόριθμο και αφετέρου ο αλγόριθμος WaveCluster.

Τα βήματα που εκτελούνται στον αλγόριθμο συσταδοποίησης που βασίζεται σε πλέγμα είναι:

1. Διαίρεση του χώρου δεδομένων σε έναν πεπερασμένο αριθμό κελιών.
2. Επιλογή ενός τυχαίου κελιού «c», όπου το c δεν πρέπει να διασχιστεί εκ των προτέρων.
3. Υπολογισμός πυκνότητας του «c».
4. Εάν η συγκέντρωση του 'c' είναι μεγαλύτερη από την συγκέντρωση κατωφλίου, τότε το κελί 'c' επισημαίνεται ως νέο σύμπλεγμα, υπολογίζεται η πυκνότητα όλων των γειτόνων του κελιού, εάν η συγκέντρωση ενός γείτονα είναι μεγαλύτερη από το όριο , στη συνέχεια το κελί προστίθεται στο σύμπλεγμα, τα προηγούμενα βήματα επαναλαμβάνονται έως ότου δεν υπάρχει γείτονας με μεγαλύτερη συγκέντρωση από την συγκέντρωση κατωφλίου..
5. Τα βήματα II, III και IV επαναλαμβάνονται έως ότου διασχιστούν όλα τα κελιά.
6. Τέλος αλγορίθμου [51]–[53].

Ο αλγόριθμος υπολογίζει τους αριθμούς των μοτίβων και τον χωρικό όγκο του μπλοκ βάσει του παρακάτω τύπου:

$$v_B = \prod_i e_B i \qquad i = 1 \dots d$$

Όπου ο χωρικός όγκος V_B ενός μπλόκ B είναι το καρτεσιανό γινόμενο των εκτάσεων του e του μπλόκ B για κάθε διάσταση και i υποδηλώνει μία μετάθεση του δείκτη που αντικατοπτρίζει την ταξινομημένη σειρά [54].

Παράδειγμα εφαρμογής: Μια αντιπροσωπευτική μελέτη περίπτωσης αποτελεί η ομαδοποίηση αρχείων καταγραφής δικτύου μεγάλης κλίμακας με σκοπό την παροχή σύντομων πηγών δεδομένων για τη μετέπειτα ανάλυση καταγραφής των συνδέσεων του δικτύου. Ως αποτέλεσμα η μέθοδος αυτή μπορεί να συμπιέσει αποτελεσματικά την αποθήκευση αρχείων καταγραφής, να μειώσει τη χρονική πολυπλοκότητα, να διαχειριστεί πραγματικά δυναμικά δεδομένα και να πραγματοποιήσει σταδιακά αυξανόμενη ομαδοποίηση. Η διαδικασία συσταδοποίησης έχει ως εξής γίνεται καταγραφή και αναλύση του μήνυματος πρωτοκόλλου που φτάνει και μέσω των συσκευών σύνδεσης δικτύου (NIC), αναλύονται οι σχετικές τιμές ιδιοτήτων, μέσω της επιλογής χαρακτηριστικών και της εξαγωγής χαρακτηριστικών, επιλέγεται το πεδίο χαρακτηριστικών που είναι χρήσιμο για την συσταδοποίηση αρχείων καταγραφής , γίνονται οι υπολογισμοί ομοιότητας των επιλεγμένων τιμών χαρακτηριστικών και τέλος επιτελείται η διαδικασία συσταδοποίησης αρχείων καταγραφής. Αρχικά, χωρίζεται σε πλέγμα ανάλογα με τη διάσταση του χρόνου όπου έχει ληφθεί το μήνυμα, υλοποιείται η αρχική συσταδοποίηση, και δημιουργείται η πρώτη συστάδα της συσταδοποίησης. Δεύτερον, κρίνεται η ομοιότητα με την πρώτη σύστάδα της συσταδοποίησης, εκ νέου συσταδοποίηση και τελικά δημιουργείται η τελική εγγραφή καταγραφής. Τέλος, εξάγετε το αρχείο καταγραφής που δημιουργείται από τις δύο διαδικασίες συσταδοποίησης καθώς και μερικά αραιά δεδομένα και δεδομένα ακραίων τιμών [55].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Άλλη μελέτη περίπτωσης σχετική με την εκπαίδευση που αξιοποιεί μεθόδους συσταδοποίησης βασισμένες σε πλέγμα αποτελεί και ο διαμοιρασμός εκπαιδευτικών πόρων βάση μοντέλου ομάδας κοινότητας πλέγματος. Ως κοινότητα νοείται μια λογική περιορισμένη διαίρεση στον άπειρο χώρο του πλέγματος, και κάθε μία αντιπροσωπεύει μια διαίρεση αυτού. Οι εκπαιδευτικοί πόροι που βασίζονται στην κοινότητα πλέγματος (Grid Community) είναι μερικοί ακέραιοι που σχηματίστηκαν αφού ο εκπαιδευτικός πόρος ταξινόμησε αυτούς τους κόμβους σύμφωνα με κάποιο είδος εκπαιδευτικού προτύπου μεταδεδομένων. Η κοινότητα είναι ένας χώρος διαχείρισης που ελέγχεται ανεξάρτητα και έχει στρατηγικές για την κοινή χρήση πόρων προκειμένου να επιτευχθεί η κοινή χρήση πόρων και η διαχείριση των κοινοτικών πόρων. Σύμφωνα με τις αρχές ομαδοποίησης των

εκπαιδευτικών πληροφοριών στο Learning Objects Metadata (LOM - Μεταδεδομένα Μαθησιακών Αντικειμένων), και τα χαρακτηριστικά της κοινότητας πλέγματος και της κατανομής πόρων στο πλέγμα, στη βάση ενός προτεινόμενου μοντέλου κοινότητας πλέγματος βασισμένου σε εκπαιδευτικούς πόρους, προτείνεται η κατασκευή της έννοιας της ομάδας κοινότητας ισοδύναμου πλέγματος, η οικοδόμηση του μηχανισμού της ανταλλαγής πόρων και της διάδοσης πληροφοριών μεταξύ των κοινοτήτων, που συμβάλλουν στην επίτευξη αμοιβαίας επικοινωνίας και ανταλλαγής πληροφοριών μεταξύ των εκπαιδευτικών κοινοτήτων. Υπάρχουν super - peers στην κοινότητα πλέγματος, στην οποία αποθηκεύονται όλες οι διευθύνσεις IP των κόμβων και άλλες πληροφορίες αυτής της κοινότητας. Για να δημιουργηθεί μια κοινοτική ομάδα, ο super - peer μιας κοινότητας θα πρέπει να υπολογίσει την αξία της συνάρτησης μέλους αυτής της κοινότητας, να βρει τον εκπαιδευτικό πόρο που βασίζεται στην κοινότητα με τον ίδιο τύπο στο επίπεδο ευρετηρίου μέσω των ταξινομημένων κόμβων και στη συνέχεια να ενταχθεί στο κοινοτικό ευρετήριο. Τέλος, με βάση το κάθε αντίστοιχο σύστημα εικονικής κοινότητας ακολουθείται η διαδικασία σχεδιασμού του συστήματος πλέγματος [56].

Συμπεράσματα: Η συσταδοποίηση βασισμένη στο πλέγμα αποτελεί γενικά μέθοδο που βρίσκει εφαρμογή σε διάφορα πεδία της πραγματικής ζωής και ειδικότερα σε πολύ μεγάλους όγκους δεδομένων πολλαπλής ανάλυσης στο υπολογιστικό πλέγμα όπου θέλουμε να δημιουργήσουμε κάποιο περιορισμένο ανεξάρτητο χώρο.

1.1.1.8 Συσταδοποίηση υποχώρων (Subspace clustering)

Γενική περιγραφή: Βασίζεται στον εντοπισμό των καλύτερων αποτελεσμάτων συσταδοποίησης στα υποσύνολα του αρχικού χώρου. Γνωστοί αλγόριθμοι της συγκεκριμένης μεθόδου είναι ο CLIQUE με κριτήριο συσταδοποίησης την εύρεση πυκνών περιοχών σε υποχώρους και ο PROCLUS με κριτήριο συσταδοποίησης την εύρεση ομοίων στοιχείων σε υποχώρους [2]. Αυτού του είδους οι αλγόριθμοι επιχειρούν να αντιμετωπίσουν το πρόβλημα που εντοπίζονται συχνά σε πολυδιάστατα δεδομένα, πολλές διαστάσεις είναι άσχετες και μπορούν να καλύψουν υπάρχουσες συστάδες με θορυβώδη δεδομένα. Η επιλογή χαρακτηριστικών καταργεί άσχετες και περιττές διαστάσεις αναλύοντας ολόκληρο το σύνολο δεδομένων. Οι αλγόριθμοι ομαδοποίησης

υποχώρων εντοπίζουν αναζητήσεις σχετικών διαστάσεων, επιτρέποντάς τους να εντοπίζουν συστάδες που υπάρχουν σε πολλαπλούς, πιθανώς αλληλοεπικαλυπτόμενους, υποχώρους. Υπάρχουν δύο κύριοι κλάδοι της συσταδοποίησης υποχώρων με βάση τη στρατηγική που χρησιμοποιούν για αναζήτηση. Ένας αλγόριθμος από πάνω προς τα κάτω βρίσκει ένα αρχικό σύμπλεγμα στο πλήρες σύνολο διαστάσεων και αξιολογεί τον υποχώρο κάθε συμπλέγματος, βελτιώνοντας συνεχώς το αποτέλεσμα. Οι μέθοδοι από κάτω προς τα πάνω βρίσκουν πυκνές περιοχές σε χώρους μικρών διαστάσεων και τις συνδυάζουν για να σχηματίσουν συμπλέγματα [57].

Ο αλγόριθμος για να μοντελοποιηθεί μια συλλογή σημείων δεδομένων με μια ένωση υποχώρων διαμορφώνεται ανάλογα με τις απαιτήσεις που καλείται να καλύψει στους υποχώρους. Έστω $\{x_j \in R^D\}_{j=1}^N$ ένα δεδομένο σύνολο σημείων που προέρχονται από μια άγνωστη ένωση $n \geq 1$ γραμμικών ή ομοπαράλληλων υποχώρων $\{S_i\}_{i=1}^n$ αγνώστων διαστάσεων $d_i = \dim(S_i)$, $0 < d_i < D$, $i = 1, \ldots, n$. Οι υποχώροι μπορούν να περιγραφούν ως:

$$S_i = \{x \in R^D : x = \mu_i + U_i y\}, \; i = 1, \ldots, n$$

όπου $\mu_i \in R^D$ είναι ένα τυχαίο σημείο στον υποχώρο S_i ($\mu_i = 0$ για γραμμικούς υποχώρους), $U_i \in R^{D \times d_i}$ είναι μια βάση για τον υποχώρο S_i και $y \in R^{d_i}$ είναι μια αναπαράσταση χαμηλών διαστάσεων για το σημείο x. Στόχος της συσταδοποίησης υποχώρων είναι να βρεθεί ο αριθμός των υποχώρων n, οι διαστάσεις τους $\{d_i\}_{i=1}^n$, οι βάσεις υποχώρων $\{U_i\}_{i=1}^n$, τα σημεία $\{\mu_i\}_{i=1}^n$ (στην περίπτωση των ομοπαράλληλων υποχώρων), και τον διαχωρισμό των σημείων σύμφωνα με τους υποχώρους [58].

Παράδειγμα εφαρμογής: Μια τυπική μελέτη περίπτωσης περιλαμβάνει προβλήματα όρασης που σχετίζονται με υπολογιστή που απαιτούν την αποτελεσματική και αποτελεσματική οργάνωση μεγάλων ποσοτήτων δεδομένων για την ανάκτηση πληροφοριών. Στην συσταδοποίηση υποχώρων, τα δείγματα δεδομένων θεωρείται ότι διανέμονται σε μια συλλογή υποχώρων. Αυτό το πρόβλημα εμφανίζεται σε διάφορες εφαρμογές όρασης, όπως τμηματοποίηση κίνησης, ομαδοποίηση προσώπων με μεταβαλλόμενο φωτισμό, χρονική τμηματοποίηση βίντεο κ.λπ. Η συσταδοποίηση υποχώρων μπορεί επίσης να χρησιμοποιηθεί για τη λήψη μιας πολυσχιδούς τμηματικής γραμμικής προσέγγισης όπως αποδεικνύεται από πειράματά με πραγματικά δεδομένα [59].

Παράδειγμα εφαρμογής σε εκπαιδευτικά δεδομένα: Στο πεδίο της εξόρυξης δεδομένων σε εκπαιδευτικά δεδομένα έχουμε την εφαρμογή συσταδοποίησης υποχώρων με σκοπό την μοντελοποίηση των ακαδημαϊκών επιδόσεων των φοιτητών. Πιο συγκεκριμένα, οι μεταρρυθμίσεις στο εκπαιδευτικό σύστημα εστιάζουν περισσότερο στη συνεχή αξιολόγηση. Η μελέτη των επιδόσεων των μαθητών είναι η κύρια πρόκληση για κάθε μάθημα που έχει συνεχή αξιολόγηση για να επικυρώσει εάν οι στόχοι του μαθήματος επιτυγχάνονται και επίσης να εντοπίσει τους τομείς της δομής του μαθήματος που χρήζουν βελτίωσης. Μέσω του αλγορίθμου συσταδοποίησης υποχώρων PROCLUS αναλύεται η επίδοση των φοιτητών σε διάφορα βασικά θέματα του μαθήματος με διάφορους τύπους συνιστωσών ικανοτήτων όπως παρουσίαση, εργασία, κουίζ, μελέτη περίπτωσης κ.λπ. μαζί με γραπτή εξέταση προκειμένου να ελεγχθεί η γνώση των φοιτητών καθώς και το ενδιαφέρον τους για ο θέμα. Ο αλγόριθμος PROCLUS έχει χρησιμοποιηθεί στον πειραματισμό καθώς ο αλγόριθμος προσδιορίζει ομοιότητες μεταξύ συνόλων δεδομένων και ομαδοποιεί παρόμοιους υποχώρους. Ο αλγόριθμος όχι μόνο λαμβάνει υπόψη τυχαία σημεία στα δεδομένα, αλλά σαρώνει επίσης με επιτυχία ολόκληρο το σύνολο δεδομένων για να εντοπίσει τις απαραίτητες διαστάσεις που οδηγούν σε πραγματικές ομάδες. Τα πειραματικά αποτελέσματα αποδεικνύουν την αποτελεσματικότητα του αλγορίθμου PROCLUS στην κατηγοριοποίηση των μαθητών σύμφωνα με τις ικανότητές τους καθώς και στην πρόβλεψη των μελλοντικών τους επιδόσεων. Η ποιότητα των συμπλεγμάτων συστάδων που σχηματίζονται με τον PROCLUS έχει τεκμηριωθεί ότι είναι υψηλής διάστασης που ανακαλύπτει μοτίβα στις επιδόσεις των μαθητών [60].

Συμπεράσματα: Πολλά προβλήματα του πραγματικού κόσμου αφορούν συλλογές δεδομένων υψηλής κλίμακας διαστάσεων, όπως εικόνες, βίντεο, κείμενο και έγγραφα Ιστού, δεδομένα μικροσυστοιχίας DNA κ.α.[61] Ανεξάρτητα από το πεδίο εφαρμογής η συγκεκριμένη μέθοδος αποτελεί μία καλή επιλογή όταν πρόκειται να ερευνήσουμε πολυδιάστατα δεδομένα με πλήθος χαρακτηριστικών για αυτό το λόγο και η συσταδοποίηση σε δεδομένα υψηλών διαστάσεων χαρακτηρίζεται από αρκετή πολυπλοκότητα.

1.1.2 Κατηγοριοποίηση με βάση τη θεωρία ορισμού συστάδας

Στη συγκεκριμένη μέθοδο οι αλγόριθμοι διαφοροποιούνται ως προς τον τρόπο χειρισμού της αβεβαιότητας που χαρακτηρίζει επικαλυπτόμενες συστάδες [2]. Ο ορθός αλγόριθμος για την ομαδοποίηση συστάδων και οι παράμετροι, συμπεριλαμβανομένου του ορίου πυκνότητας ή του αριθμού των αναμενόμενων συστάδων που προορίζονται να προκύψουν από τα δεδομένα, εξαρτώνται από το συγκεκριμένο σύνολο δεδομένων και τον επιδιωκόμενο σκοπό των αποτελεσμάτων [19], [62]. Στην συγκεκριμένη περίπτωση αναφερόμαστε στην πραγματικότητα σε οποιοδήποτε αλγόριθμο συσταδοποίησης μπορούμε δυνητικά να εφαρμόσουμε σε ένα σύνολο δεδομένων λαμβάνοντας υπόψιν τις παραμέτρους που προαναφέρθηκαν.

Μια λύση στο πρόβλημα της οργάνωσης συμπλέγματος είναι να βρεθεί ένα διαμέρισμα που να πληροί ένα κριτήριο βελτιστοποίησης. Αυτό το κριτήριο βελτιστοποίησης προσανατολισμένο στο στόχο μπορεί να εκφραστεί ως συνάρτηση f που αντικατοπτρίζει την τιμή διαφορετικών συνδυασμών ή ομάδων. Αυτός είναι ο στόχος που επιδιώκεται. Στην οικονομία, την κοινωνιολογία και την επιστήμη και την τεχνολογία, προκύπτουν πολλά ερωτήματα σχετικά με την ομαδοποίηση των πεπερασμένων συνόλων.

Αποτελεί η ανάλυση συστάδων σε τραπεζικά δάνεια. Τα δανειακά συμβόλαια (λογαριασμοί) με απόλυτα γνωστή «φάση αποταμίευσης» πρέπει να κατανεμηθούν σε συστάδες για σκοπούς σχεδιασμού ρευστότητας της τράπεζας. Υποθέτοντας ότι υπάρχουν n λογαριασμοί (αντικείμενα) και m χαρακτηριστικά για κάθε λογαριασμό, επιδιώκουμε να κατηγοριοποιήσουμε αυτούς τους n λογαριασμούς σε m διαστασιακά κενά σε σημαντικές συστάδες, K σε αριθμό. Η συσταδοποίηση επιτυγχάνεται με την ελαχιστοποίηση της ομοιότητας εντός συστάδων και τη μεγιστοποίηση της ανομοιότητας εκτός συστάδων [63].

Επίσης, στο πεδίο της εκπαίδευσης υπάρχουν πολλά παραδείγματα συσταδοποίησης στον τομέα των μαθησιακών αναλύσεων (Learning Analytics) που απεικονίζουν το εύρος των αντιθετικών και περιστασιακών κριτηρίων που επιλέγονται για τη δημιουργία συστάδων: συχνότητα πρόσβασης στο υλικό μαθημάτων, επιλογή σύγχρονης έναντι ασύγχρονης επικοινωνίας κατά τη διάρκεια διαδικτυακών στρατηγικών συλλογικής εργασίας που χρησιμοποιούνται από τους μαθητές κατά τη διάρκεια μίας ατομικής διαδικτυακής καθοδήγησης. Πιο συγκεκριμένη μελέτη περίπτωσης αποτελεί η ανάλυση συστάδων για να εντοπίσει τέσσερις πρωτότυπες πορείες εμπλοκής σε τρία MOOC που προσφέρονται από το Πανεπιστήμιο του

Στάνφορντ στην πλατφόρμα Coursera: Ολοκλήρωση, Έλεγχος, Αποσύνδεση και Δειγματοληψία. Ιδιαίτερη έμφαση δίνεται στην αντικειμενική και «φυσική» ποιότητα των συστάδων που προκύπτουν λόγω του ότι «έχουν νόημα από εκπαιδευτική άποψη». Ως εκ τούτου, οι συστάδες ερμηνεύονται ως υποπληθυσμοί μαθητών που θα μπορούσαμε ρεαλιστικά να περιμένουμε να «ανακαλύψουμε» σε μια σειρά από διαφορετικά περιβάλλοντα διαδικτυακής μάθησης [64].

1.1.3 Κατηγοριοποίηση με βάση τον τύπο δεδομένων

Στη συγκεκριμένη μέθοδο οι αλγόριθμοι κατηγοριοποιούνται ανάλογα με το είδος των δεδομένων τα οποία σκοπεύουμε να μελετήσουμε, αριθμητικά ή κατηγορικά και κειμενικά [2]. Μια καλή μέθοδος συσταδοποίησης παράγει συστάδες υψηλής ποιότητας με ελάχιστη απόσταση εντός συστάδας (υψηλή ομοιότητα) και μέγιστη απόσταση μεταξύ των υπερκλάσεων (χαμηλή ομοιότητα) [65]. Οι κατηγορικές μεταβλητές χαρακτηρίζονται από ιδιότητες που αποτελούν κατηγορίες. Μπορούν να διακριθούν δύο κύριοι τύποι αυτών των μεταβλητών: διχοτομικές, για τις οποίες υπάρχουν μόνο δύο κατηγορίες, και πολύ – κατηγορικές [66]. Ανάλογα με τον τύπο του συνόλου δεδομένων σκληρά, ασαφή και αδρά ορισμένα (rough sets) μπορούν να εφαρμοστούν διάφοροι κατηγορικοί αλγόριθμοι συσταδοποίησης που αποτελούν συνήθως παραλλαγές γνωστών διαιρετικών και ιεραρχικών αλγορίθμων [67]. Στον αντίποδα, για τα αριθμητικά δεδομένα υπάρχουν πολλές περισσότερες μέθοδοι συσταδοποίησης που μπορούν να εφαρμοστούν είτε διαιρετικές, είτε ιεραρχικές, είτε άλλου τύπου με δημοφιλή τον αλγόριθμο K- means [68].

Όπως αναφέρθηκε προηγουμένως, η ομαδοποίηση έχει χρησιμοποιηθεί σε πολλούς τομείς, όπως η επιχειρηματική ευφυΐα, η ανάλυση εικόνων, η αναζήτηση ιστού, η βιολογία, η ασφάλεια και άλλες εφαρμογές του πραγματικού κόσμου. Οι διαφορετικοί τύποι εφαρμογών οδηγούν στην ανάπτυξη πολυάριθμων αλγορίθμων ομαδοποίησης, κάτι που είναι σημαντικό για τη διάκρισή τους. Κάθε αλγόριθμος για την ανάλυση συστάδων είναι συγκεκριμένος για έναν συγκεκριμένο τύπο δεδομένων, όπως αριθμητικά, κατηγορικά, πολυμεταβλητά ή χωρικά, αλλά δεν υπάρχει συνδυασμός αριθμητικών και γλωσσικών δεδομένων που να είναι ασαφής. Η ομαδοποίηση είναι η διαδικασία κατηγοριοποίησης ενός συνόλου δεδομένων με βάση τα διάφορα χαρακτηριστικά τους σε ομάδες. Ο βαθμός στον οποίο δύο αντικείμενα είναι παρόμοια

υπολογίζεται χρησιμοποιώντας μια προκαθορισμένη μέτρηση. Υπάρχουν διαφορετικές συναρτήσεις απόστασης που εξαρτώνται από τη φύση του τύπου χαρακτηριστικού του αντικειμένου δεδομένων. Παράδειγμα συνάρτησης απόστασης που χρησιμοποιείται συνήθως για αριθμητικά δεδομένα είναι η Ευκλείδεια απόσταση. Τα δεδομένα από διαφορετικές περιοχές έχουν συχνά διαφορετικές ιδιότητες, ως αποτέλεσμα, η συνάρτηση που χρησιμοποιείται για τον υπολογισμό της ομοιότητας είναι διαφορετική και ο αλγόριθμος που χρησιμοποιείται για την ομαδοποίηση θα πρέπει να είναι διαφορετικός ή να οδηγεί με άλλο τρόπο στη δημιουργία ενός νέου αλγορίθμου για ομαδοποίηση [69].

Σε ότι αφορά την εξόρυξη εκπαιδευτικών δεδομένων (EDM) συνήθως έχουμε να διαπραγματευτούμε με κατηγορικά δεδομένα, μερικοί αλγόριθμοι συσταδοποίησης που χρησιμοποιούνται συνήθως είναι οι K-modes, Robust Clustering algorithm για Κατηγορικές ιδιότητες (ROCK) και Chameleon.

Ο αλγόριθμος ROCK βασίζεται στην έννοια του γείτονα (neighbor) και των δεσμών (links) για να εκτιμήσει την ομοιότητα μεταξύ των στοιχείων ενός συνόλου δεδομένων. Ως γείτονες ενός σημείου θεωρούνται τα σημεία τα οποία εμφανίζουν σημαντική ομοιότητα με αυτό. Ως σύνδεση link θεωρείται ο αριθμός των κοινών γειτόνων μεταξύ των στοιχείων p_i, p_j.

Η συνάρτηση κριτήριο που χρησιμοποιεί ο αλγόριθμος ROCK για την εύρεση των καλύτερων συστάδων είναι η παρακάτω:

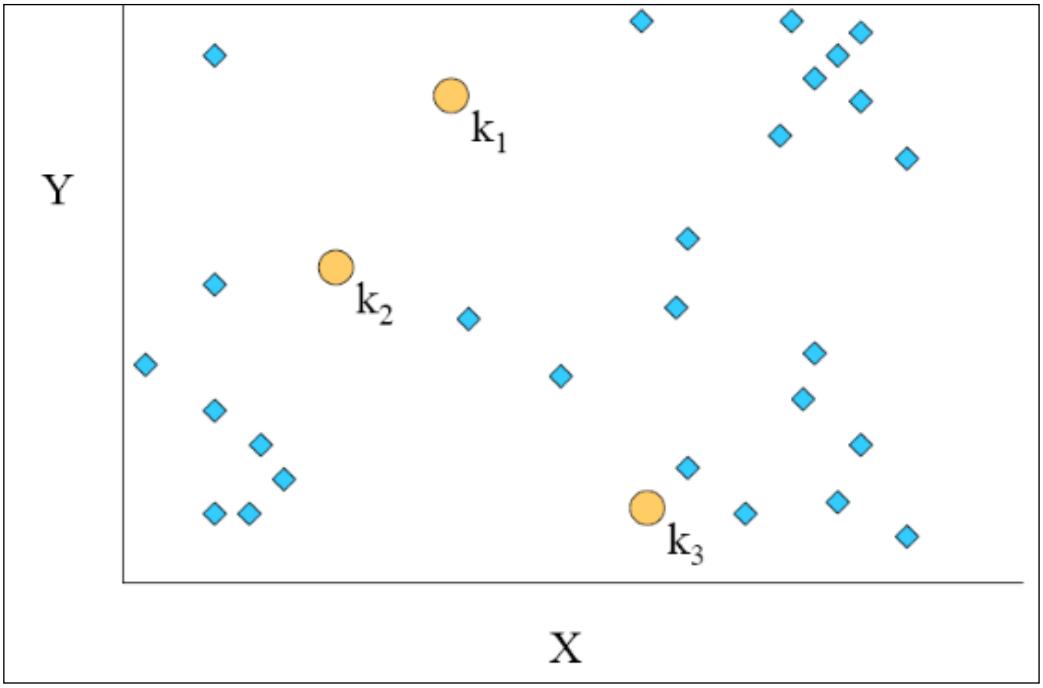
$$E_1 = \sum_{i=1}^{k} n_i \sum_{p_{k,pr \in Ci}} \frac{link\left(p_{q,p_r}\right)}{n_i^{1+2f(\theta)}}$$

Όπου Ci δηλώνει τη συστάδα i μεγέθους n_i. p_q και p_r είναι τα ζεύγη τιμών, $\sum_{p_{k,pr \in Ci}} link\left(p_{q,p_r}\right)$ το άθροισμα των δεσμών μεταξύ των σημείων στην ίδια συστάδα και $n_i^{1+2f(\theta)}$ ο αριθμός δεσμών ανάμεσα στα σημεία της συστάδας Ci [70].

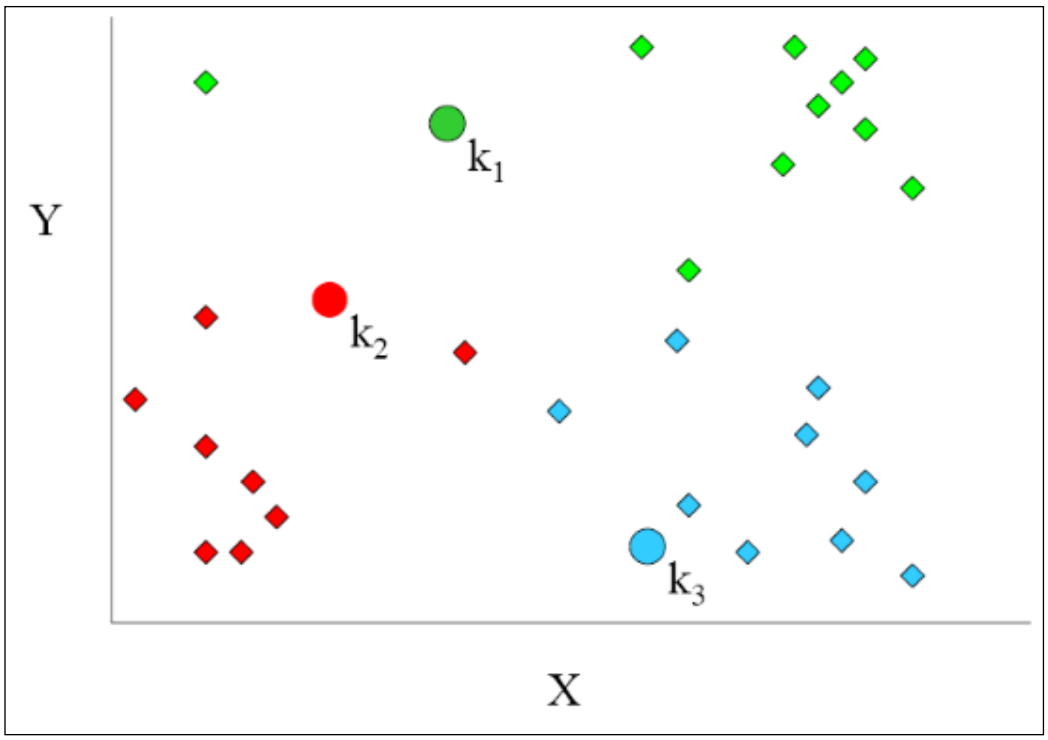
2 Αλγόριθμοι Συσταδοποίησης WEKA

2.1 Αλγόριθμος K – Means (Simple K – Means)

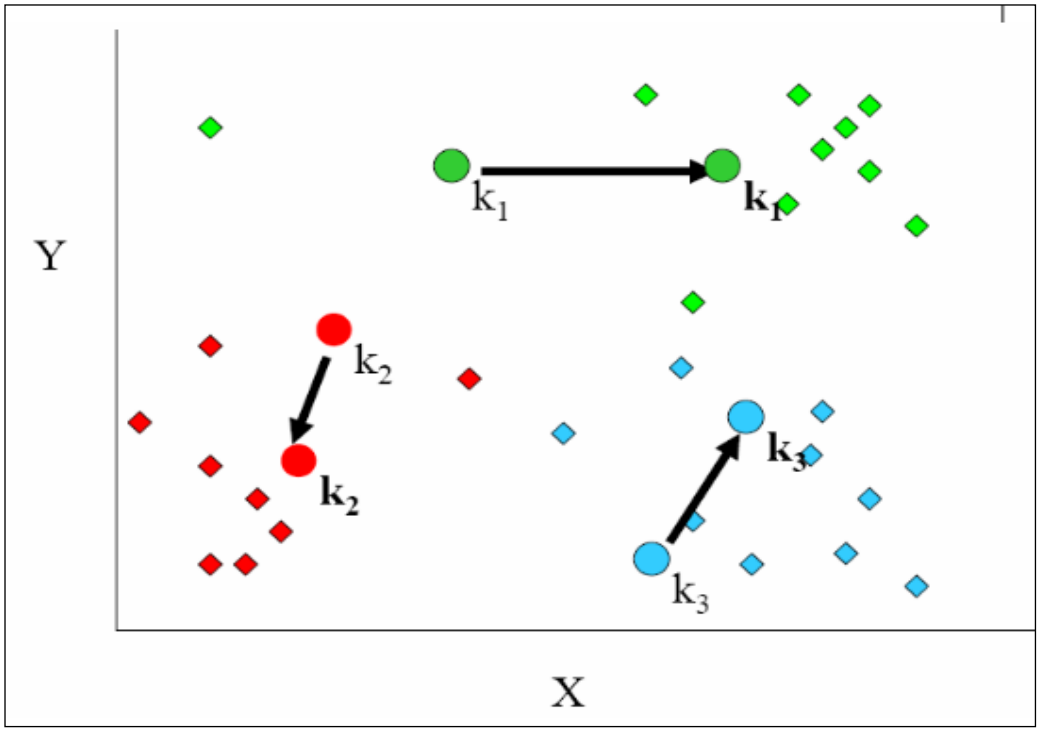
Έναν από τους πλέον δημοφιλείς και με ευρεία εφαρμογή αλγορίθμους διαιρετικής συσταδοποίησης αποτελεί ο αλγόριθμος K – μέσων (K – means) ο οποίος επιχειρεί να κατατμήσει ένα σύνολο δεδομένων σε K διακριτές και μη επικαλυπτόμενες συστάδες. Αρχικά ο αλγόριθμος εκκινεί αφού καθορίσουμε τον επιθυμητό αριθμό συστάδων K. Στη συνέχεια, εκτελεί τη διαίρεση των αντικειμένων σε συστάδες που μοιράζονται ομοιότητες αλλά είναι ανόμοια με τα αντικείμενα που ανήκουν σε διαφορετικές συστάδες [71]. Έπειτα, υπολογίζει για κάθε συστάδα το μέσο όρο όλων των σημείων της και επανακαθορίζει νέο κέντρο για τη συστάδα. Η εν λόγω διαδικασία συνεχίζεται έως ότου σταματήσουν να αλλάζουν τα κέντρα των συστάδων[2]–[4], [6].



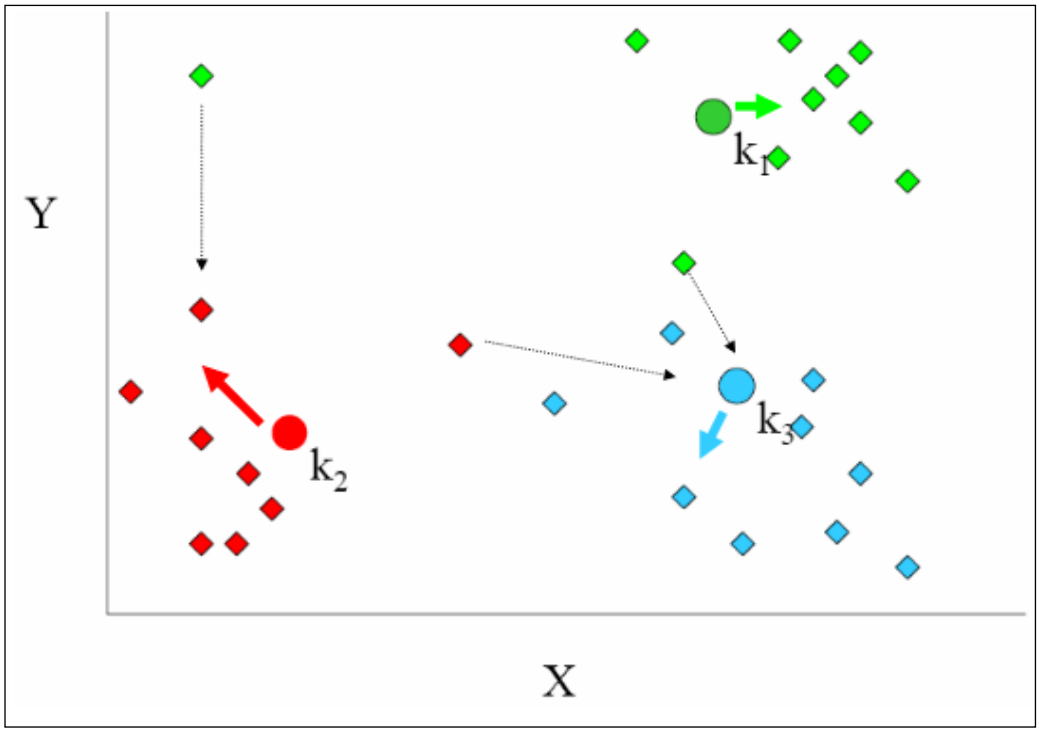
Εικόνα 1: Παράδειγμα δεδομένων στην αρχική κατάσταση με k=3 αρχικά σημεία και 3 συστάδες [5].



Εικόνα 2: Ανάθεση σημείων στις κοντινότερες τους συστάδες [5].



Εικόνα 3: Επαναϋπολογισμός των κέντρων των σημείων [5].



Εικόνα 4: Νέα ανάθεση σημείων στα νέα κέντρα [5]**.**

Έστω ότι αρχικά όπως φαίνεται στην εικόνα 1 υπάρχουν 30 σημεία, ο αλγόριθμος εκτελείται με k = 3. Αρχικά επιλέγονται τρία τυχαία σημεία ως κέντρα k1, k2, k3 για τρεις αντίστοιχες συστάδες. Στη συνέχεια κάθε σημείο ανατίθεται στην ομάδα της οποίας το κέντρο είναι πιο κοντά οπότε το σημείο k1 ανατίθεται στην πράσινη ομάδα, το σημείο k2 στην κόκκινη ομάδα και το σημείο k3 στην γαλάζια ομάδα (εικόνα 2). Έπειτα επανϋπολογίζονται τα κέντρα κάθε ομάδας (εικόνα 3) και γίνεται εκ νέου ανάθεση των σημείων στις κοντινότερες τους ομάδες (εικόνα 4) και με αυτό τον τρόπο ολοκληρώνεται ο πρώτος κύκλος εκτέλεσης του αλγορίθμου. Η διαδικασία θα επαναληφθεί για προκαθορισμένο αριθμό βημάτων ή όσες φορές είναι απαραίτητο έως ότου να μην προκύπτουν αλλαγές στο διαχωρισμό των σημείων σε ομάδες.

Πλεονεκτήματα

1. Για ένα μεγάλο αριθμό μεταβλητών, ο K - Means μπορεί να είναι υπολογιστικά ταχύτερος σε σχέση με την ιεραρχική συσταδοποίηση όταν έχουμε μικρό αριθμό συστάδων.
2. Εύκολος στην υλοποίηση.
3. Μπορεί να παράγει αυστηρότερες συστάδες σε σχέση με την ιεραρχική συσταδοποίηση.

4. Όταν επανυπολογιστούν τα κεντροειδή ένα στοιχείο μπορεί να αλλάξει συστάδα [72].

Μειονεκτήματα

1. Δύσκολο να προβλεφθεί ο αριθμός των συστάδων (K – Value).
2. Η σειρά των δεδομένων έχει αντίκτυπο στα τελικά αποτελέσματα.
3. Εμφανίζει μικρή ανοχή στην κλιμάκωση: Η αναπροσαρμογή της κλίμακας των συνόλων δεδομένων λόγω κανονικοποίησης ή τυπικής προσαρμογής θα αλλάξει εντελώς τα αποτελέσματα.
4. Οι αρχικές παράμετροι έχουν ισχυρή επίδραση στα τελικά αποτελέσματα [73].

2.2 Αλγόριθμος Farthest First

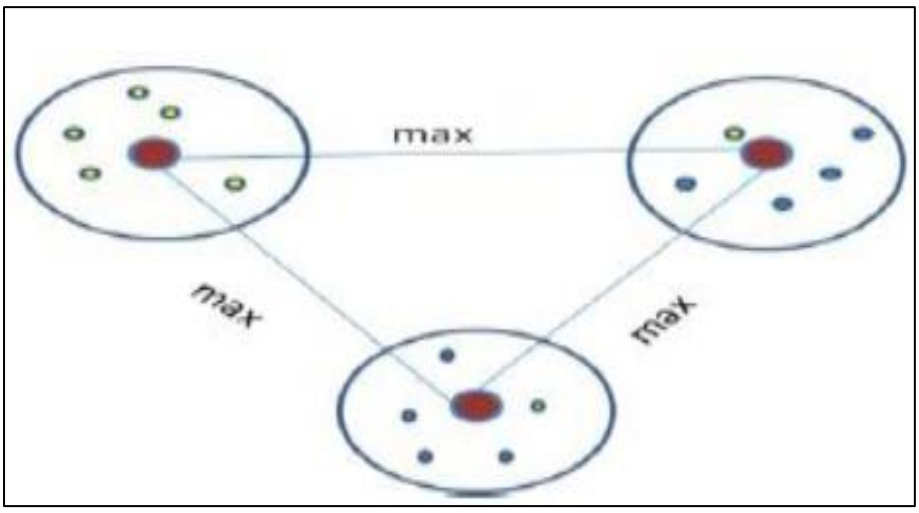
Ο αλγόριθμος Farthest First Traversal (FTT) των k- center αποτελεί ένα γρήγορο και άπληστο αλγόριθμο. Αποτελεί τροποποιημένο αλγόριθμο της μεθόδου K - Means. Με βάση τον συγκεκριμένο αλγόριθμο πρώτα επιλέγονται τα k – σημεία ως κέντρα συστάδων. Πρώτα το αρχικό κέντρο επιλέγεται βάση τυχαιότητας, το δεύτερο επιλέγεται άπληστα ως το πιο απομακρυσμένο σημείο σε συνάρτηση με το πρώτο και κάθε κέντρο που απομένει επιλέγει το πιο απομακρυσμένο σημείο από το ήδη επιλεγμένο σύνολο κέντρων, καθορίζεται άπληστα και προστίθεται στα υπόλοιπα σημεία στο σύμπλεγμα με το κέντρο του να είναι το πιο κοντινότερο του [74].

Βήματα αλγορίθμου:

1. Διάσχιση Farthest first (D: σύνολο δεδομένων, k: ακέραιος) {
2. Τυχαία επιλογή πρώτου κέντρου;
3. //επιλογή κέντρων
4. για (I= 2,...,k) {
5. για (κάθε σημείο που απομένει) { υπολόγισε την απόσταση για το τρέχον σύνολο κέντρου; }
6. επέλεξε το σημείο με τη μεγαλύτερη απόσταση ως νέο κέντρο; }
7. //ανάθεση υπόλοιπων σημείων
8. για (κάθε σημείο που απομένει) {
9. υπολόγισε την απόσταση κάθε κέντρου συστάδας;
10. Τοποθέτησε το στη συστάδα με την μικρότερη απόσταση; }}

Ντετερμινιστική μέθοδος για την επιλογή του πρώτου σημείου:

Για κάθε $Xi = [xi, 1, xi, 2, \dots\dots\dots, xi, m]$ στο D που περιγράφεται από m κατηγορικές ιδιότητες, χρησιμοποιούμε τη συνάρτηση $(f(x_{i,j}|\mathrm{D})$ για να δηλώσουμε τον αριθμό συχνοτήτων της τιμής του χαρακτηριστικού $x_{i,j}$ στο σύνολο δεδομένων και στη συνέχεια, μια συνάρτηση βαθμολόγησης για την αξιολογεί κάθε σημείο, η οποία ορίζεται ως εξής: $Score\ (X_i) = \sum_{j=1}^{m} f(x_{i,j}|D)$ [75]



Εικόνα 5: Ανάθεση σημείων σε συστάδες σύμφωνα με τον αλγόριθμο FFT [76]

Πλεονεκτήματα

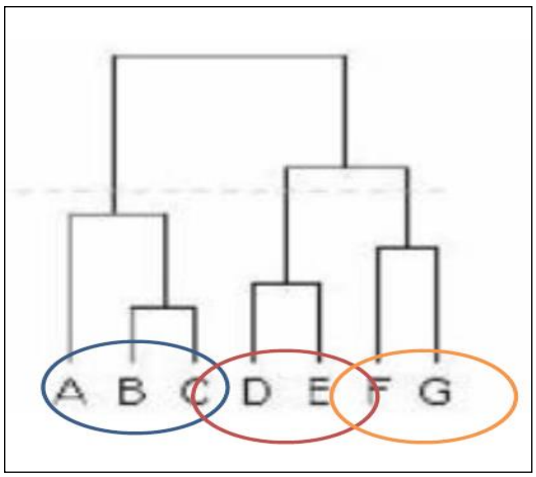
1. Αυτή η μέθοδος είναι αποτελεσματική σε εφαρμογές όπου ο χρόνος είναι σημαντικός γιατί σε αρκετές περιπτώσεις αυξάνει την ταχύτητα αφού κάνει λιγότερες αλλαγές θέσης.
2. Γρήγορη και κατάλληλη μέθοδος για εφαρμογές εξόρυξης δεδομένων μεγάλης κλίμακας [77].

Μειονεκτήματα

1. Σε σχέση με άλλους αλγορίθμους δημιουργεί ανομοιόμορφες συστάδες.
2. Εξαιτίας του ότι δεν υπάρχει περιορισμός τιμής κατωφλίου για το μέγεθος των συστάδων τα αντικείμενα ομαδοποιούνται ακατάστατα σε μία μόνο συστάδα [78].

2.3 Αλγόριθμος Ιεραρχικής Συσταδοποίησης (Hierarchical Clusterer)

Οι ιεραρχικοί αλγόριθμοι συσταδοποίησης όπως υποδηλώνει και το όνομα τους επιδιώκουν να οικοδομήσουν μια ιεραρχία συστάδων. Η ιεραρχική ομαδοποίηση είναι μια μέθοδος για την ομαδοποίηση αντικειμένων δεδομένων σε ένα δέντρο συμπλέγματος. Οι αλγόριθμοι ιεραρχικής ομαδοποίησης μπορούν περαιτέρω να ταξινομηθούν σε αλγόριθμους αθροιστικούς (agglomerative) και στους διαιρετικούς (divisive) αλγορίθμους ανάλογα με το αν η ιεραρχική αποσύνθεση μοντελοποιείται με τρόπο από κάτω προς τα πάνω ή από πάνω προς τα κάτω. Οι ιεραρχικές τεχνικές παράγουν μια πρόσθετη ακολουθία διχοτομήσεων, με ένα συνολικό σύμπλεγμα που τα περικλείει όλα στην κορυφή και μεμονωμένα συμπλέγματα διαφορετικών αντικειμένων στο κάτω μέρος. Κάθε ενδιάμεσο επίπεδο δύναται να θεωρηθεί ότι ενώνει δύο συστάδες από το επόμενο χαμηλότερο επίπεδο ή διαχωρίζει μία συστάδα από το επόμενο υψηλότερο επίπεδο. Το αποτέλεσμα ενός αλγορίθμου ιεραρχικής ομαδοποίησης μπορεί να εμφανιστεί σε γραφική μορφή ως δέντρο, που ονομάζεται δενδρικό διάγραμμα. Αυτό το δέντρο εμφανίζει σε γραφική αναπαράσταση τη διαδικασία συγχώνευσης και τις ενδιάμεσες συστάδες. Αυτή η γραφική δομή δείχνει πώς τα σημεία μπορούν να συγχωνευθούν σε μία ενιαία συστάδα [79]



Εικόνα 6: Δενδρόγραμμα Ιεραρχικής συσταδοποίησης [4].

Αρχικά ο αλγόριθμος θεωρεί κάθε μεμονωμένο σημείο ως μια ομάδα και μετρά τις μεταξύ τους αποστάσεις. Έπειτα, εντοπίζεται το πιο κοντινό ζευγάρι ομάδων και συγχωνεύεται σε μία ομάδα άρα προκύπτουν λιγότερες ομάδες. Στη συνέχεια υπολογίζονται εκ νέου οι αποστάσεις μεταξύ των ομάδων. Η διαδικασία

επαναλαμβάνεται έως ότου κάθε σημείο τοποθετηθεί σε μία και μοναδική ομάδα. Στο τελευταίο βήμα σχεδιάζεται το δενδροδιάγραμμα το οποίο μπορεί να κατατμηθεί στο σημείο το οποίο ανταποκρίνεται στον ζητούμενο αριθμό συστάδων[2], [6], [80]. Στην παραπάνω εικόνα εντοπίζονται 3 διαφορετικές συστάδες, η ABC, η DE και η FG.

Πλεονεκτήματα

1. Κάποιοι είναι πιο αποδοτικοί στη διαχείριση του θορύβου σε σύγκριση με διαιρετικούς αλγόριθμους.
2. Εύκολοι στην υλοποίηση.
3. Δημιουργεί ένα πιο δομημένο σύνολο συστάδων καθώς εξάγει μία ιεραρχία από την οποία μπορεί να αποφασιστεί ο αριθμός των συστάδων [81].

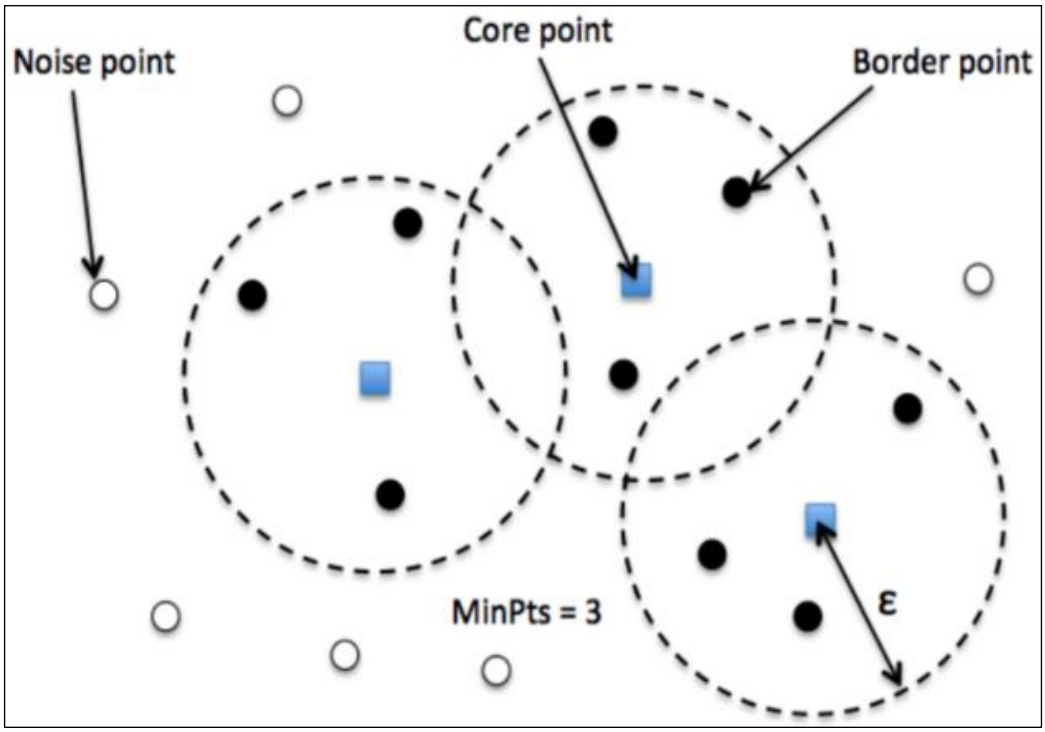
Μειονεκτήματα

1. Όταν ένα στιγμιότυπο αντιστοιχηθεί σε μία συστάδα τότε δεν μπορεί να μετακινηθεί σε διαφορετική συστάδα.
2. Η μη γραμμική πολυπλοκότητα του τον καθιστά ακατάλληλο για μεγάλο όγκο δεδομένων.
3. Μεγάλη ευαισθησία σε ακραία σημεία.
4. Η σειρά των δεδομένων έχει αντίκτυπο στα τελικά αποτελέσματα.
5. Οι αρχικές παράμετροι έχουν ισχυρή επίδραση στα τελικά αποτελέσματα [82].

2.4 Αλγόριθμος DBSCAN (Make Density Based Clusterer)

Ο αλγόριθμος Density - Based Spatial Clustering of Applications with Noise (DBSCAN) αποτελεί αλγόριθμο που βασίζεται στην πυκνότητα και χρησιμοποιείται για τον προσδιορισμό συστάδων διαφορετικών σχημάτων και μεγεθών σε ένα σύνολο δεδομένων. Ο DBSCAN έχει δύο συνιστώσες, η πρώτη αφορά την ακτίνα ε (esp) η οποία ορίζει τη μεγαλύτερη επιτρεπτή απόσταση μεταξύ δύο σημείων μέσα στην ίδια συστάδα και η δεύτερη είναι τα ελάχιστα σημεία (MinPts), τα οποία θέτουν τον ελάχιστο αριθμό σημείων δεδομένων που προαπαιτούνται για να σχηματιστεί μία μεμονωμένη συστάδα. Έπομένως, MinPts είναι ο ελάχιστος αριθμός γειτονικών σημείων που περιλαμβάνονται σε μία συστάδα με ακτίνα ή μέγιστο μήκος esp. Η ακτίνα eps καθορίζει την εγγύτητα των σημείων μεταξύ τους ώστε να θεωρούνται μέρος μιας συστάδας, εάν

η απόσταση μεταξύ δύο σημείων είναι μικρότερη ή ίση με την τιμή (eps), αυτά τα σημεία κρίνονται ως γειτονικά. Η παράμετρος minPoints είναι ο ελάχιστος αριθμός σημείων για να σχηματιστεί μια γειτονιά δηλαδή μια πυκνή περιοχή. Αν λόγου χάρη, θέσουμε την παράμετρο minPoints ίση με 5, τότε χρειαζόμαστε κατ' ελάχιστο 5 σημεία για να σχηματίσουμε μια πυκνή περιοχή.



Εικόνα 7: Παράδειγμα δημιουργίας συστάδων με τον DBSCAN, για MinPts=3 [83]

Ο αλγόριθμος ξεκινά με ένα τυχαίο στοιχείο p του συνόλου και ανακτά όλα τα στοιχεία τα οποία είναι πυκνά προσεγγίσιμα από το p. Αν το στοιχείο p είναι ένα αντικείμενο πυρήνα, ορίζεται μία συστάδα. Αν το στοιχείο p είναι ένα ακραίο στοιχείο, κανένα αντικείμενο δεν είναι πυκνά προσεγγίσιμο από το p και το p συμπεριλαμβάνεται στο θόρυβο και ο αλγόριθμος λαμβάνει το επόμενο στοιχείο της βάσης [2], [6], [80], [83].

Πλεονεκτήματα

1. Δεν απαιτεί να είναι εκ των προτέρων γνωστός ο αριθμός των συστάδων.
2. Έχει ανοχή στο θόρυβο.
3. Μπορεί να εντοπίσει ακόμη και συστάδες αυθαίρετου σχήματος.
4. Δεν επηρεάζεται από τη σειρά των δεδομένων στη βάση δεδομένων και απαιτεί μόνο δύο παραμέτρους [84].

Μειονεκτήματα

1. Όσο πιο αξιόπιστη είναι η μέτρηση της απόστασης τόσο καλύτερο είναι το αποτέλεσμα της συσταδοποίησης κάτι το οποίο είναι δύσκολο για δεδομένα υψηλών διαστάσεων.
2. Δεν μπορεί να ομαδοποιήσει αποτελεσματικά σύνολα δεδομένων με μεγάλες αποκλίσεις στις πυκνότητες, δεδομένου ότι σε αυτή την περίπτωση ο συνδυασμός MinPts δεν είναι δυνατόν να επιλεγεί κατάλληλα για όλες τις συστάδες [85].

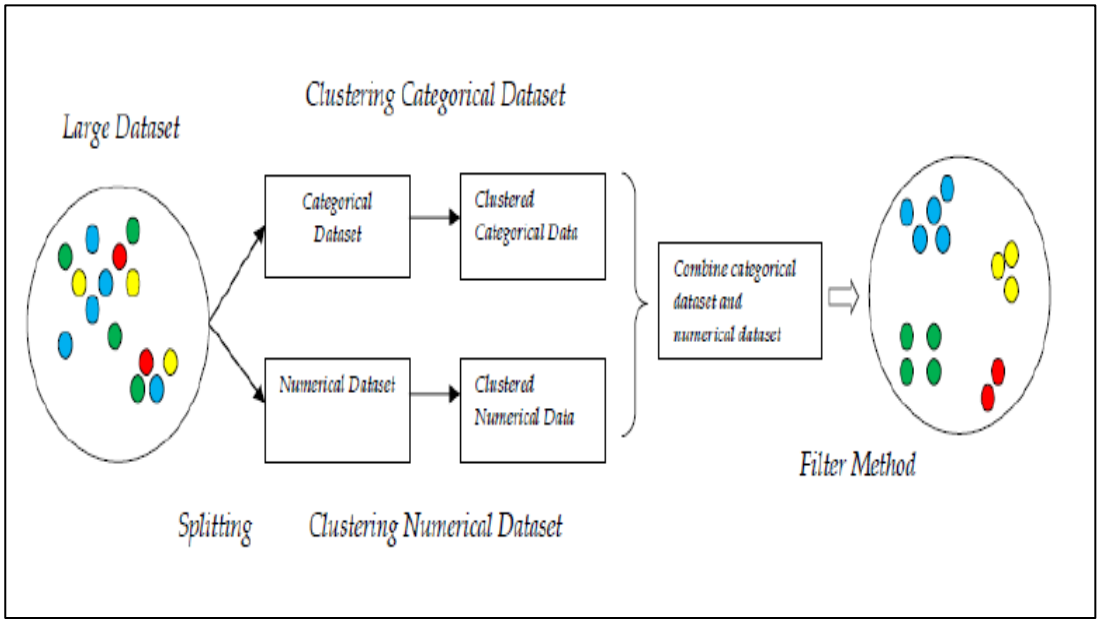
2.5 Αλγόριθμος Filtered Clusterer

Ο φιλτραρισμένος αλγόριθμος συστάδων βασίζεται κυρίως στην αποθήκευση πολυδιάστατων σημείων δεδομένων μέσα σε ένα δέντρο. Η διαδικασία σχετικά με το δέντρο είναι σαν μία μέθοδος δυαδικού δέντρου, καθώς αντιπροσωπεύει μια ιεραρχική υποδιαίρεση στο πλαίσιο οριοθέτησης του συνόλου σημείων δεδομένων, η χρήση του δικού τους άξονα, στη συνέχεια του οποίου η διαίρεση ευθυγραμμίζεται με τρόπο σε υπερ – επίπεδα. Κάθε κόμβος στο δέντρο σχετίζεται με ένα κλειστό πεδίο, που αναφέρεται ως κελί. Η ρίζα του κελιού είναι το πλαίσιο οριοθέτησης για το σύνολο δεδομένων. Εάν το κελί αποτελείται από το πολύ ένα σημείο, λαμβάνοντας υπόψη ότι είναι δηλωμένο σύμφωνα με το να παραμένει ένα φύλλο. Στη συνέχεια, τα σημεία εύρεσης που αφορούν το κελί διαιρούνται ανάλογα με τη μία πλευρά ή την ασήμαντη πλευρά του υπερεπιπέδου. Εντέλει, το αποτέλεσμα είναι τα παιδιά του αρχικού κελιού που οδηγεί σε μια δυαδική δομή δέντρου [86].

Βήματα αλγορίθμου:

1. Κάθε όριο μιας βάσης φίλτρου είναι επίσης ένα σημείο συστάδας της βάσης.
2. Μια βάση φίλτρου B που έχει ένα x ως σημείο συστάδας ενδέχεται να μην συγκλίνει με το x. Αλλά υπάρχει μια λεπτομερέστερη βάση φίλτρου που συγκλίνει.
3. Για μια βάση φίλτρου B, το σύνολο $\cap\{cl(B0) : B0 \in B\}$ είναι το σύνολο όλων των σημείων συστάδας του B.
4. Το χαμηλότερο όριο του συνόλου όλων των σημείων συστάδας του B είναι το κατώτερο όριο του B.

5. Το ανώτατο όριο του συνόλου όλων των σημείων συστάδας του B είναι το ανώτερο όριο του B.
6. Αν συμφωνούν το κατώτερο όριο και το ανώτερο όριο τότε μόνο το B είναι μια συγκλίνουσα βάση φίλτρου. Σε αυτή την περίπτωση, η τιμή στην οποία συμφωνούν είναι το όριο της βάσης του φίλτρου [87].



Εικόνα 8: Παράδειγμα εφαρμογής φίλτρου σε μικτά δεδομένα [88]**.**

Πλεονεκτήματα

1. Μπορεί να ταξινομήσει δεδομένα διαφορετικών τύπων σε μία συστάδα.
2. Είναι γρήγορος
3. Συνήθως δημιουργεί ομοιογενείς συστάδες.

Μειονεκτήματα

1. Απαιτεί εκ των προτέρων ορισμό αριθμού συστάδων.
2. Δυσκολία να προβλεφθεί ο βέλτιστος αριθμός συστάδων [89].

2.6 Αλγόριθμος Expectation – Maximation (EM)

Ο αλγόριθμος Expectation – Maximization EM (προσδοκίας - μεγιστοποίησης) αποτελεί μια επαναληπτική μέθοδος για την εύρεση εκτιμήσεων μέγιστης πιθανοφάνειας ή μέγιστης εκ των υστέρων (maximum a posteriori) πιθανοφάνειας των παραμέτρων σε στατιστικά μοντέλα, όπου το μοντέλο βρίσκεται σε εξάρτηση από μη παρατηρούμενες λανθάνουσες μεταβλητές. Η επανάληψη του ΕΜ εναλλάσσεται μεταξύ της εκτέλεσης ενός βήματος προσδοκίας (Ε), το οποίο υπολογίζει την προσδοκία της λογαριθμικής πιθανοφάνειας (log - likehood) που αποτιμάται με χρήση της τρέχουσας εκτίμησης για τις παραμέτρους και του βήματος μεγιστοποίησης (Μ), το οποίο υπολογίζει παραμέτρους που αυξάνουν την αναμενόμενη λογαριθμική πιθανοφάνεια που βρέθηκε στο βήμα (Ε). Αυτές οι εκτιμήσεις παραμέτρων χρησιμοποιούνται στη συνέχεια για τον προσδιορισμό της κατανομής των λανθάνοντων μεταβλητών στο επόμενο βήμα (Ε) [90].

Βήματα αλγορίθμου:

1. Αρχικοποίηση: Επιλογή των παραμέτρων των συστάδων (μΑ, σΑ και P(A)) ή υπόθεση των κλάσεων των στιγμιότυπων και μετά επανάληψη.
2. Expectation E-step: Για κάθε σημείο x και για κάθε συστάδα Α, υπολογισμός της πιθανότητας wi ότι το xi ανήκει στη συστάδα Α.

   Τύπος βήματος expectation:

$$\mu A = \frac{\sum_1^n W_i X_i}{\sum_1^n W_i}$$

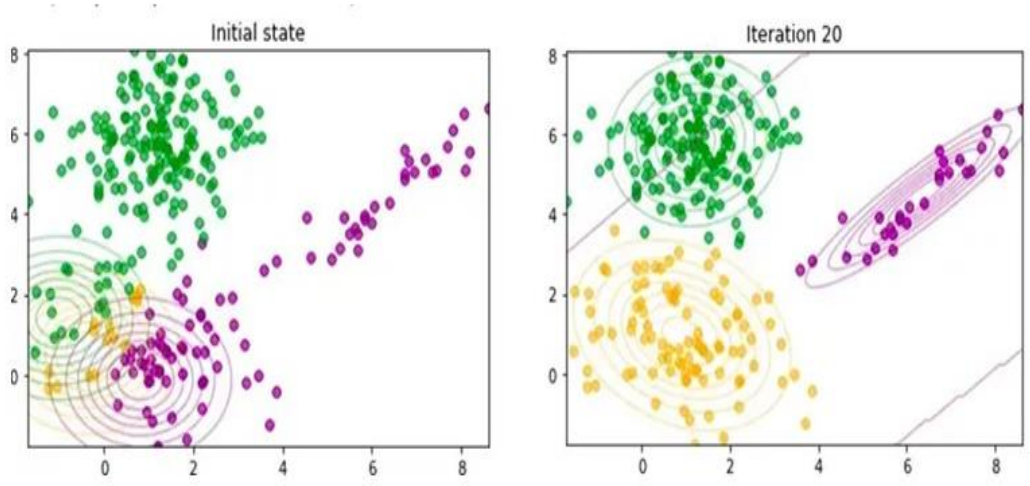
3. Τύπος βήματος Maximization M-step:

$$\sigma^2 A = \frac{\sum_1^n W_{i(X_{i-\mu})^2)}}{\sum_1^n W_i}$$

   Τύπος μεγιστοποίησης της πιθανότητας δεδομένων:

$$Log - likelihood = \Sigma_{i \log\left(\Sigma_A P(A) P\left(x|A\right)\right)}$$

4. Επανάληψη του βήματος 2 και του βήματος 3 μέχρι να συγκλίνουν οι παράμετροι (όταν η διαφορά μεταξύ δύο διαδοχικών επαναλήψεων γίνει αμελητέα) [91].



Εικόνα 9: Απεικόνιση συνόλου δεδομένων στο αρχικό στάδιο και μετά από 20 επαναλήψεις του αλγορίθμου ΕΜ [92]**.**

Πλεονεκτήματα

1. Δίνει εξαιρετικά χρήσιμα αποτέλεσμα για το σύνολο δεδομένων του πραγματικού κόσμου.
2. Μπορεί να είναι αποτελεσματικότερος σε σχέση με άλλους όταν πρόκειται για ανάλυση συστάδων μιας μικρής σκηνής ή μιας μικρής περιοχής ενδιαφέροντος.
3. Απλός και εύκολος στην υλοποίηση [93].

Μειονεκτήματα

1. Επιλογή των συστάδων εκ των προτέρων.
2. Εκ φύσεως πολύπλοκος αλγόριθμος
3. Η εφαρμογή του περιορίζεται σε μη φυσιολογικά ή υψηλών διαστάσεων δεδομένα.
4. Εμφανίζει αργή σύγκλιση σε ορισμένες περιπτώσεις [94].

2.7 Αλγόριθμος Cobweb

Ο αλγόριθμος Cobweb ομαδοποιεί αντικείμενα σε ένα σύνολο δεδομένων αντικειμένων - ιδιοτήτων. Αποδίδει ένα δενδρόγραμμα ομαδοποίησης που ονομάζεται δέντρο ταξινόμησης που χαρακτηρίζει κάθε συστάδα με μια πιθανολογική περιγραφή. Ο αλγόριθμος δημιουργεί μία ιεραρχική ομαδοποίηση, όπου οι συστάδες περιγράφονται πιθανολογικά. Κάθε κόμβος σε ένα δέντρο ταξινόμησης εκπροσωπεί μια κλάση (έννοια) και υποδεικνύεται με μια πιθανολογική έννοια που αθροίζει τις κατανομές χαρακτηριστικών - τιμών των αντικειμένων που ταξινομούνται κάτω από τον κόμβο. Για την πρόβλεψη χαρακτηριστικών που λείπουν ή της κλάσης ενός νέου αντικειμένου μπορεί να χρησιμοποιηθεί αυτό το δέντρο ταξινόμησης [95].

Βήματα αλγορίθμου:

COBWEB (Root,Instance)

Αρχή

1. Εάν ο ριζικός κόμβος (C0) δεν έχει δημιουργηθεί τότε
   Δημιουργία ριζικού κόμβου (C0) και
   Εισήγαγε τα στιγμιότυπα «a» και «b» ως ξεχωριστές συστάδες C1 και C2 στον ριζικό κόμβο (C0)
2. Διαφορετικά, αν ο ριζικός κόμβος (C0) έχει δημιουργηθεί τότε
   Κατά την εισαγωγή του επόμενου κόμβου
   Υπέθεσε την Εισαγωγή του επόμενου στιγμιότυπου σε όλες τις πιθανές θέσεις.
3. Υπολογίστε το CU για κάθε πιθανή θέση των συστάδων που χρησιμοποιώντας την παρακάτω συνάρτηση:

$$CU(C1, C2, \dots \dots Cn) = \sum_{k=1}^{m} P(C_k) \sum_{i} \sum_{j} P\left(A_i = V_{ij} \middle| C_k\right)^2 - P\left(A_i = V_{ij}\right)^2$$

Όπου P(A = V|C) είναι η πιθανότητα ότι ένα στιγμιότυπο έχει τιμή V για το χαρακτηριστικό του A, δεδομένου ότι ανήκει στην κατηγορία C και Vij είναι η πραγματική τιμή του στιγμιότυπου και m είναι ο αριθμός των συστάδων. CU η συνάρτηση Category Utility.

   a. Αν συγχώνευση CU > διαχωρισμός CU τότε

Συγχώνευση (root, instance)

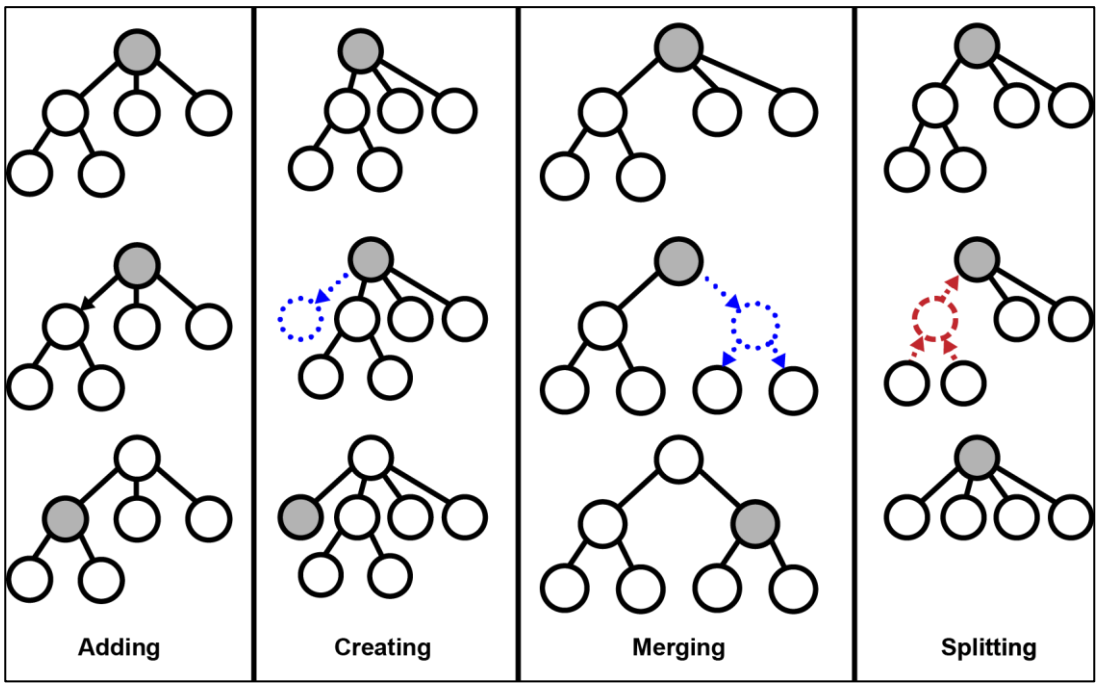
   b. Αλλιώς διαχωρισμός (root, instance)

Τέλος εάν

4. Μέχρι να εισαχθούν όλα τα στιγμιότυπα στο δέντρο ομαδοποίησης

   Τέλος εάν

   τέλος [96].



Εικόνα 10: Απεικόνιση βημάτων προσθήκης, δημιουργίας, συγχώνευσης και διαχωρισμού αλγορίθμου COBWEB [97]**.**

Πλεονεκτήματα

1. Χρησιμοποιεί ένα ευρετικό μέτρο αξιολόγησης που ονομάζεται Category Utility (CU) για να καθοδηγήσει την κατασκευή του δέντρου.
2. Πραγματοποιεί τόσο συγχώνευση όσο και διαχωρισμό κλάσεων έτσι μπορεί να κάνει αμφίδρομη αναζήτηση.
3. Εξαιρετικά ευαίσθητο σε ακραίες τιμές δεδομένων.

Μειονεκτήματα

1. Η υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο είναι συχνά ισχυρότερη από την ύπαρξη συσχέτισης μεταξύ τους.
2. Δεν είναι κατάλληλος για ομαδοποίηση δεδομένων μεγάλου όγκου.
3. Το δέντρο ταξινόμησης δεν είναι ισορροπημένο καθ' ύψος για αλλοιωμένα δεδομένα εισόδου, γεγονός που μπορεί να προκαλέσει δραματική υποβάθμιση της πολυπλοκότητας του χρόνου και του χώρου.

4. Η αναπαράσταση κατανομής πιθανότητας των συστάδων καθιστά αρκετά δαπανηρή την ενημέρωση και την αποθήκευση των συστάδων [98].

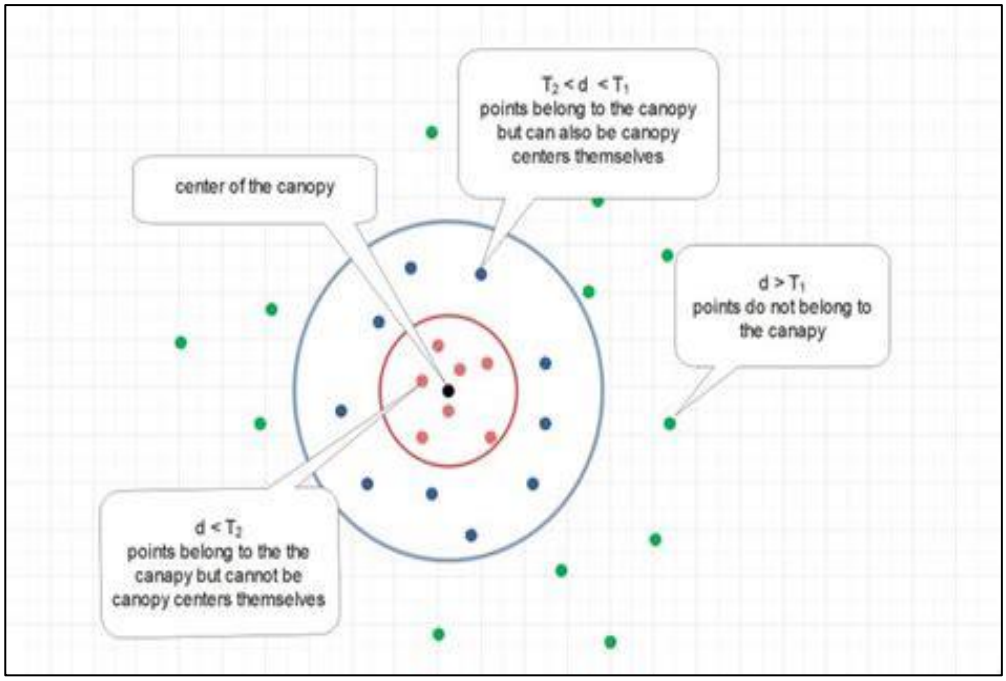
2.8 Αλγόριθμος Canopy

Είναι ένας αλγόριθμος που χρησιμοποιείται ως βήμα προ - ομαδοποίησης χωρίς επίβλεψη και εφαρμόζεται πριν από τον αλγόριθμο Ιεραρχικής ομαδοποίησης και τον αλγόριθμο K-means [99].

Βήματα αλγορίθμου:

Ο αλγόριθμος προχωρά ως εξής, χρησιμοποιώντας δύο κατώφλια απόστασης T1 και T2. Σε μια τέτοια κατάσταση, T1 > T2 (Αυστηρή απόσταση: T1, Χαλαρή απόσταση: T2)

1. Επιλογή ενός τυχαίου σημείου από το σύνολο των σημείων δεδομένων.
2. Αφαίρεση ενός σημείου από το σετ και μετά δημιουργία ενός νέου canopy («θόλου»).
3. Για κάθε σημείο, που έχει κρατηθεί στο σετ, του ανατίθεται νέος θόλος. Μετά υπολογίζεται η μικρότερη απόσταση d. Εάν d < T2, τοποθετείται το στο νέο θόλο.
4. Εάν d < T1, βγαίνει το από το αρχικό σετ.
5. Επανάληψη βημάτων 2 έως 4 μέχρι να αδειάσει το αρχικό σετ και να μην υπάρχουν άλλα σημεία για συσταδοποίηση [100].



Εικόνα 11: Διαδικασία ομαδοποίησης αλγορίθμου Canopy [101]**.**

Πλεονεκτήματα

1. Πολύ απλός και γρήγορος
2. Δεν απαιτείται ο καθορισμός του αριθμού (k) συστάδων.
3. Αποτελεσματική μέθοδος συσταδοποίησης για τη δημιουργία επικαλυπτόμενων συστάδων.

Μειονεκτήματα

1. Χαμηλή ακρίβεια [102].

3 Παρουσίαση λογισμικού Weka

Το λογισμικό Weka (Wekato Environment for knowledge Analysis) είναι ένα λογισμικό ανοικτού κώδικα που έχει αναπτυχθεί σε γλώσσα Java από το πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Αποτελεί ένα πολύ γνωστό εργαλείο μηχανικής μάθησης και χρησιμοποιείται ευρέως για εξόρυξη δεδομένων καθώς παρέχει στους χρήστες ελεύθερη πρόσβαση στον πηγαίο κώδικα, είναι συμβατό με διάφορες πλατφόρμες, μπορεί να χειριστεί πολλούς διαφορετικούς τύπους δεδομένων και περιλαμβάνει πλήθος αλγόριθμων μηχανικής μάθησης [103], [104].

Εικόνα 12: Λογότυπο λογισμικού Weka [103]

3.1 Περιβάλλον λογισμικού Weka

Το λογισμικό παρέχει ποικίλα εργαλεία για προεπεξεργασία των δεδομένων, ταξινόμηση, συσταδοποίηση και εύρεση κανόνων συσχέτισης, όπως είναι οι αλγόριθμοι διακριτοποίησης και δειγματοληψίας οι οποίοι παρέχουν τη δυνατότητα μετασχηματισμού συνόλων δεδομένων αλλά και δυνατότητα τροφοδότησης ενός συνόλου δεδομένων με ένα σχήμα εκμάθησης και ανάλυση του αποτελέσματος της ταξινόμησης και της δυνητικής απόδοσης του.

Εικόνα 13: Αρχική οθόνη λογισμικού Weka

Όπως φαίνεται στην παραπάνω εικόνα στην αρχική οθόνη της εφαρμογής Weka και στην περιοχή Applications εμφανίζονται οι επιλογές Explorer, Experimenter, KnowledgeFlow, Workbench και Simple CLI.

Μέσω της διεπαφής Explorer παρέχεται η δυνατότητα εισαγωγής των δεδομένων από εξωτερικό αρχείο μορφής .arff και η εφαρμογή των κυριών τεχνικών που παρέχονται από την εφαρμογή όπως προεπεξεργασία, κατηγοριοποίηση, ανάλυση συστάδων, επιλογή γνωρισμάτων και οπτικοποίηση.

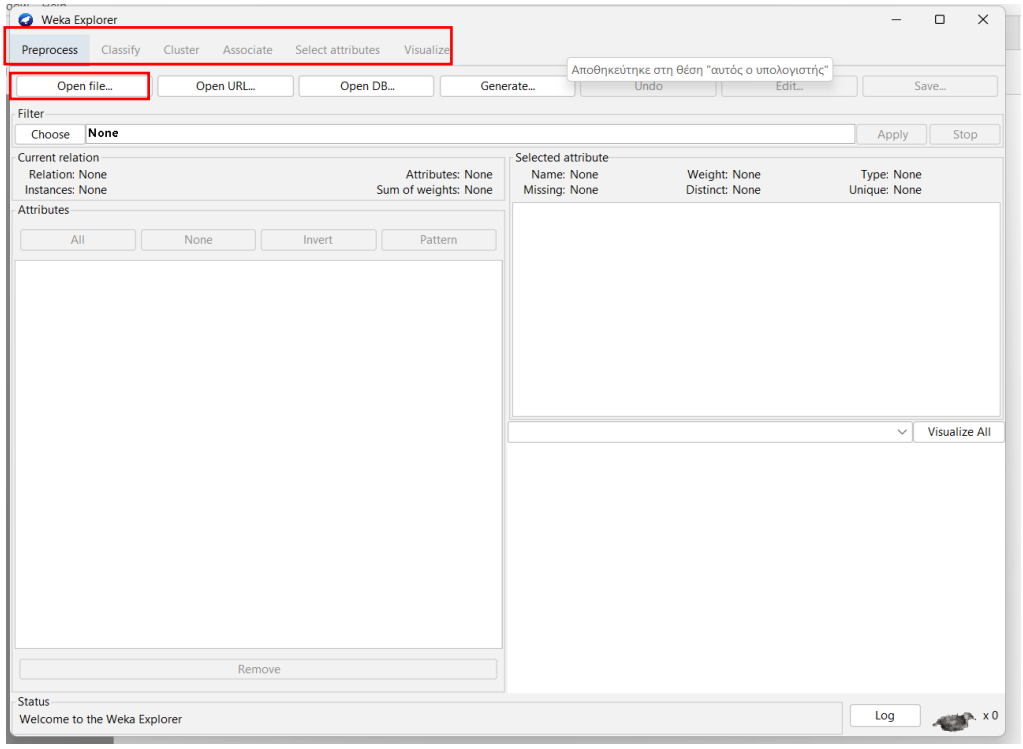
Μέσω της διεπαφής Experimenter παρέχεται η δυνατότητα διεξαγωγής πειραμάτων μεγάλης κλίμακας ώστε να γίνει αξιολόγηση μεθόδων κατηγοριοποίησης και παλινδρόμησης. Αποτελεί ένα ιδιαίτερα εύχρηστο τρόπο για σύγκριση και αποτίμηση της επίδοσης διαφορετικών μοντέλων με οπτικοποίηση των αποτελεσμάτων υπό μορφή πίνακα.

Μέσω της διεπαφής KnowledgeFlow παρέχεται η δυνατότητα εκτέλεσης εργασιών παρόμοιων με τις εργασίες οι οποίες υποστηρίζονται μέσω του Explorer με τη διαφορά ότι εμπεριέχονται κάποια γραφικά στοιχεία το οποία συνδέονται με τέτοιο τρόπο ώστε ορίζουν τη ροή εργασίας.

Μέσω της διεπαφής Workbench παρέχεται έναν περιβάλλον που περιλαμβάνει συνδυαστικά τις δυνατότητες των προηγούμενων διεπαφών Explorer, Experimenter και KnowledgeFlow. Αποτελεί ένα περιβάλλον μέσω του οποίου οι χρήστες έχουν τη δυνατότητα να συγκρίνουν μία ποικιλία τεχνικών εκμάθησης αλλά και να παραμετροποιήσουν τις ρυθμίσεις εφαρμογών και προσθέτων που τους παρέχονται.

Τέλος, η διεπαφή Simple CLI παρέχει ένα απλό περιβάλλον διεπαφής γραμμής εντολών που επιτρέπει την άμεση εκτέλεση εντολών WEKA για λειτουργικά συστήματα που δεν υποστηρίζουν ανάλογο περιβάλλον γραμμής εντολών [103], [104].

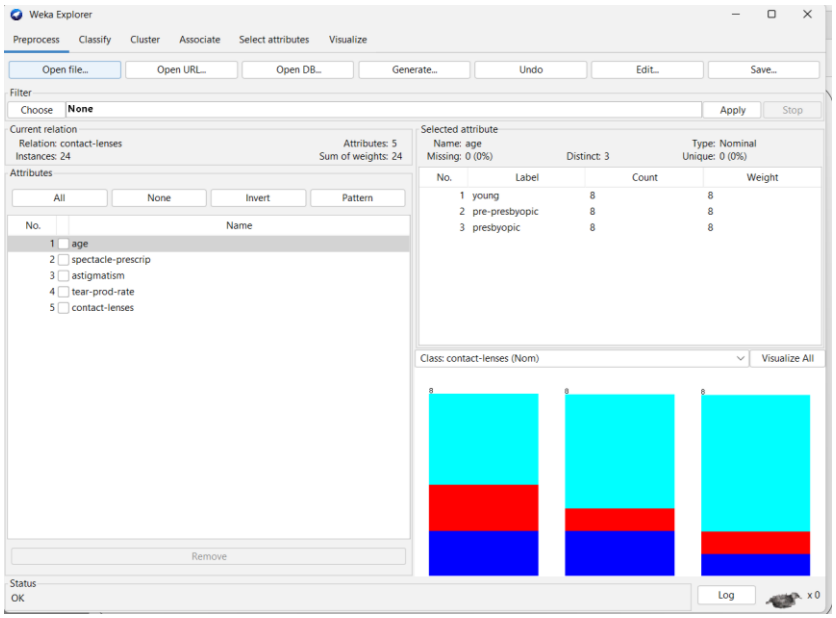
3.2 Διεπαφή Explorer



Εικόνα 14: Περιβάλλον Weka Explorer

3.3 Διεπαφή Προεπεξεργασίας

Όπως φαίνεται και στην παραπάνω εικόνα από το παράθυρο του Weka Explorer δίνεται η δυνατότητα μέσω της επιλογής “Open file” να ανεβάσουμε ένα εξωτερικό αρχείο δεδομένων τύπου .arff για προεπεξεργασία, η μόνη αρχικά διαθέσιμη καρτέλα είναι η Preprocess [105]. Στη συνέχεια όπως φαίνεται και στη παρακάτω εικόνα αφού επιλεγεί ένα σύνολο δεδομένων, εμφανίζονται γραφικά τα δεδομένα για καθένα από τα γνωρίσματα ξεχωριστά καθώς και στατιστικές πληροφορίες για αυτά. Εάν στο σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που ανήκουν στην ίδια κλάση εμφανίζονται με το ίδιο χρώμα [104].

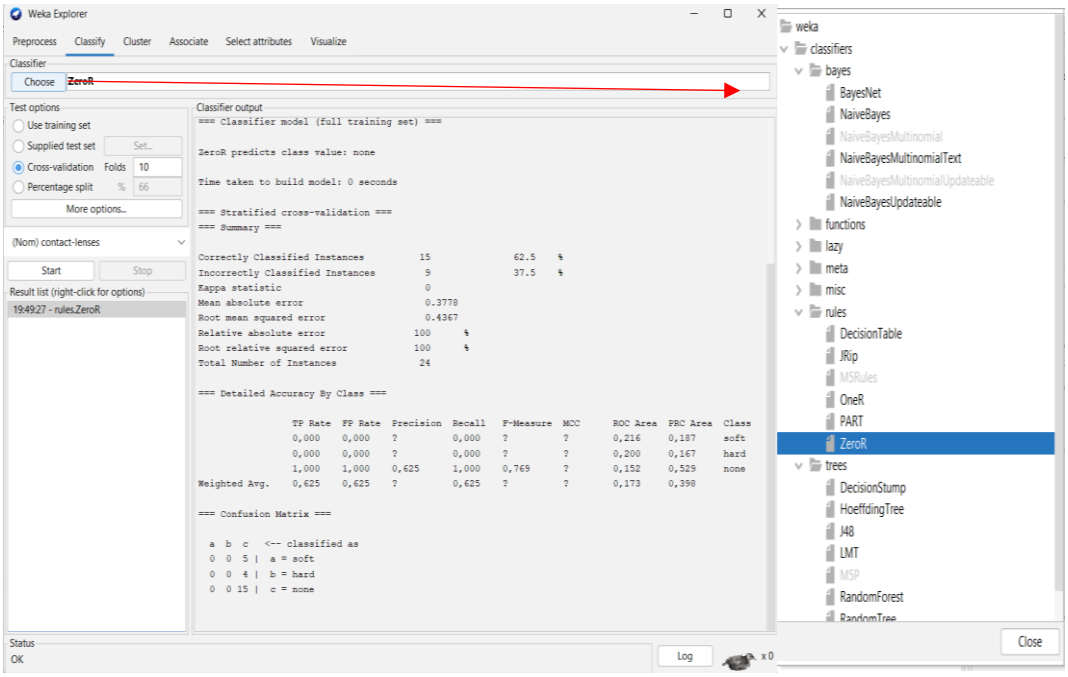


Εικόνα 15: Καρτέλα προεπεξεργασίας δεδομένων

3.4 Διεπαφή Classify

Επιπλέον της καρτέλας Preprocess όπως διακρίνεται στις εικόνες 16 και 17 υπάρχουν οι καρτέλες Classify, Cluster, Associate, Select Attributes, Visualize.

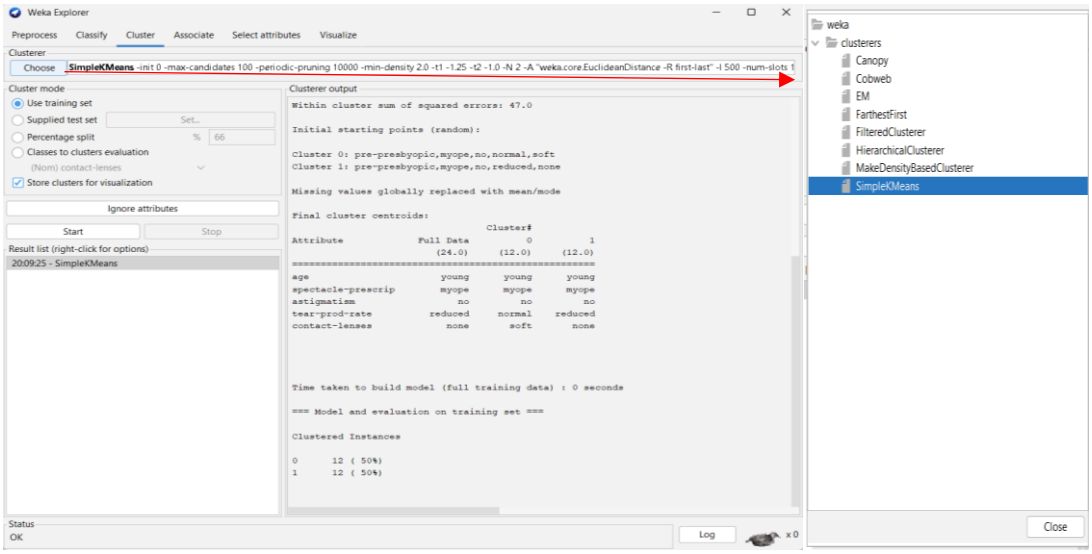
Η καρτέλα Classify παρέχει αλγόριθμους μηχανικής μάθησης υπο μορφή δέντρου για την κατηγοριοποίηση των δεδομένων κάποιοι από οποίους είναι η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα και πολλοί άλλοι ενώ προσφέρεται πληθώρα αλγόριθμων μηχανικής μάθησης τόσο με επίβλεψη όσο και χωρίς επίβλεψη. Μέσω της επιλογής “Choose” ορίζεται η μέθοδος κατηγοριοποίησης, στην περιοχή “Test options” ορίζεται η μέθοδος αξιολόγησης, στην περιοχή “Result list” περιλαμβάνεται η λίστα μοντέλων και στην περιοχή “Classifier output” εμφανίζονται τα αποτελέσματα βάση του μοντέλου κατηγοριοποίησης που έχει επιλεγεί όπως φαίνεται στην εικόνα 16.



Εικόνα 16: Καρτέλα Classify

3.5 Διεπαφή Cluster

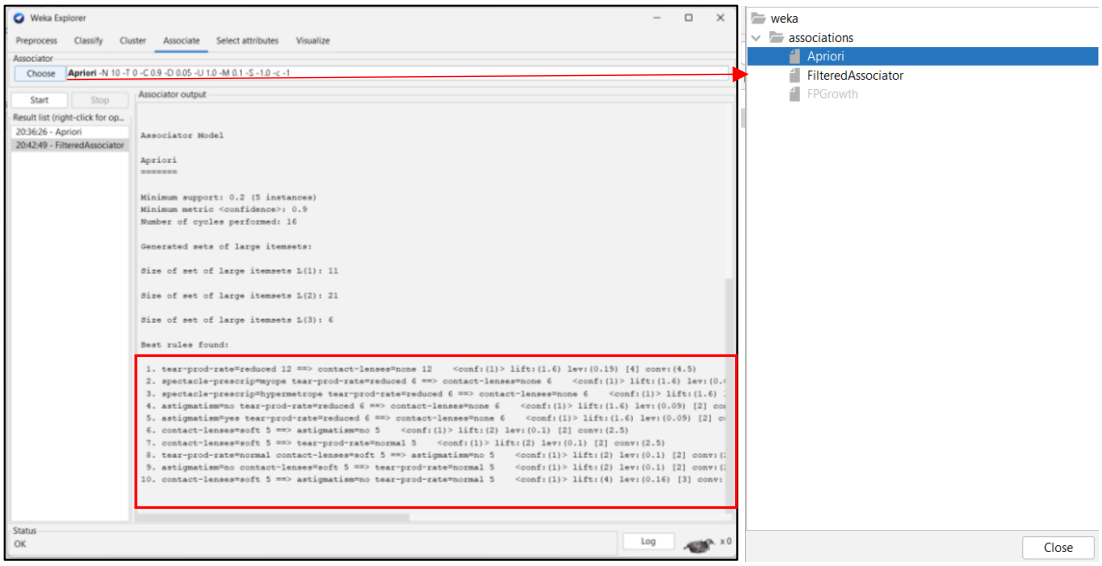
Η καρτέλα Cluster παρέχει αλγόριθμους μηχανικής μάθησης υπο μορφή δέντρου για την συσταδοποίηση των δεδομένων ενδεικτικά αναφέρονται οι K – means, DBscan, Ιεραρχική συσταδοποίηση και πολλοί άλλοι. Μέσω της επιλογής “Choose” ορίζεται ο επιθυμητός αλγόριθμος συσταδοποίησης, στην περιοχή “Cluster mode” ορίζονται κάποιες επιλογές αναφορικά με τα δεδομένα εκπαίδευσης, στην περιοχή “Result list” περιλαμβάνεται η λίστα μοντέλων και στην περιοχή “Clusterer output” εμφανίζονται τα αποτελέσματα βάση του μοντέλου συσταδοποίησης που έχει επιλεγεί όπως φαίνεται στην εικόνα 17.



Εικόνα 17: Καρτέλα Cluster

3.6 Διεπαφή Associate

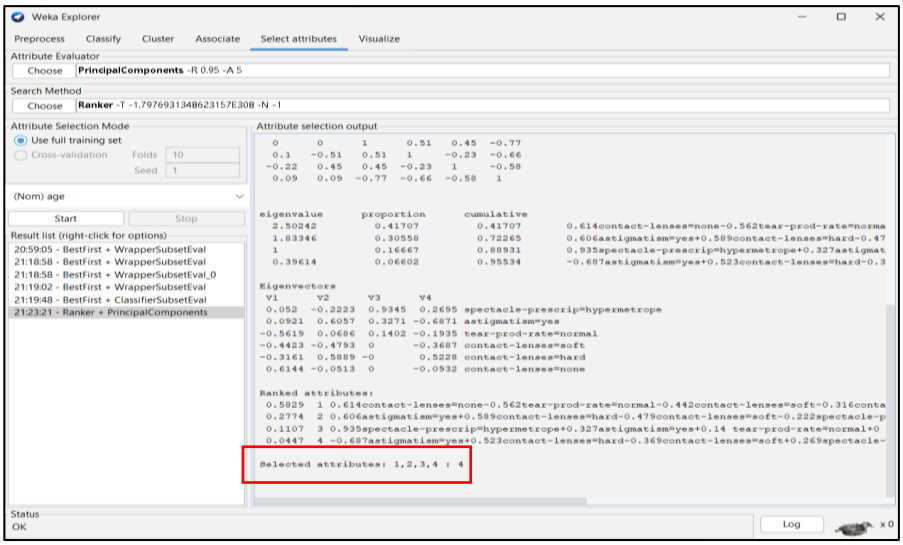
Η καρτέλα Accosiate παρέχει τη δυνατότητα στο χρήστη να εφαρμόσει αλγόριθμους και μεθόδους συσταδοποίησης ώστε να εντοπίσει κανόνες συσχέτισης και να τους αξιολογήσει. Μέσω της επιλογής “Choose” ορίζεται ο επιθυμητός accosiator από τους Apriori, FilteredAccosiator και FPGrowth ενώ μπορούν να οριστούν και διάφορες παράμετροι οι οποίες αφορούν διάφορες ιδιότητες. Μέσω της επιλογής “Start” εκκινεί ο επιλεγμένος αλγόριθμος ανακάλυψης κανόνων συσχέτισης και στην περιοχή “Accosiator Output” μια σειρά των καλύτερων κανόνων συσχέτισης που προέκυψαν.



Εικόνα 18: Καρτέλα Accosiate

3.7 Διεπαφή Select attributes

Η καρτέλα Select attributes παρέχει τη δυνατότητα πρόσβασης σε διάφορες μεθόδους επιλογής χαρακτηριστικών. Η περιοχή “Attribute Evaluator” αφορά στον αξιολογητή χαρακτηριστικών, ενδεικτικά αναφέρονται οι ClassifierSubsetEval, PrinicipalComponents κ.α. Η περιοχή “Search Method” αφορά τη μέθοδο αναζήτησης και περιλαμβάνει ενδεικτικά τις επιλογές BestFirst, GreedyStepwise και Ranker. Η περιοχή “Attribute Selection Mode” παρέχει επιλογές αναφορικά με τον τρόπο επιλογής χαρακτηριστικών από το σύνολο δεδομένων, ενώ στο κάτω πεδίο πρέπει να επιλεγεί το χαρακτηριστικό που θα χρησιμοποιηθεί ως κλάση. Τέλος, στην περιοχή “Attribute Selection Output” εμφανίζονται τα επιλεγμένα χαρακτηριστικά που προέκυψαν από την εφαρμογή μεθόδου και αξιολογητή.



Εικόνα 19: Καρτέλα Select attributes

3.8 Διεπαφή Visualize

Τέλος, στην καρτέλα Visualize παρέχεται η δυνατότητα οπτικοποιήσης ενός συνόλου δεδομένων. Πατώντας σε κάθε επιμέρους τετράγωνο της περιοχής “Plot Matrix” δίνεται η δυνατότητα γραφικής αναπαράστασης των δεδομένων για περαιτέρω ανάλυση κάθε ζεύγους ιδιοτήτων.

Εικόνα 20: Καρτέλα Visualize

4 Εξόρυξη εκπαιδευτικών δεδομένων στο περιβάλλον Weka

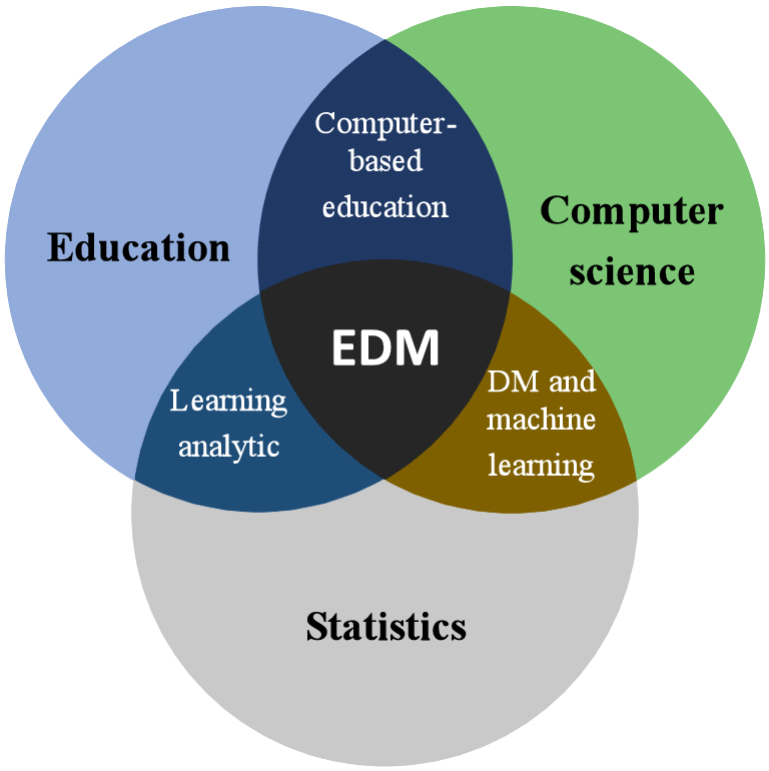
4.1 Εξόρυξη εκπαιδευτικών δεδομένων

Η διαδικασία εξόρυξης εκπαιδευτικών δεδομένων (Educational Data Mining - EDM) χρησιμοποιεί υπολογιστικές μεθόδους για να μετατρέψει ακατέργαστα δεδομένα από εκπαιδευτικά συστήματα σε χρήσιμες πληροφορίες ώστε να επιλύσει εκπαιδευτικά ζητήματα [106]. Οι κύριοι στόχοι της εξόρυξης εκπαιδευτικών δεδομένων είναι η πρόβλεψη της μελλοντικής μαθησιακής συμπεριφοράς των μαθητών, η προώθηση της επιστημονικής γνώσης, τα αποτελέσματα της εκπαιδευτικής υποστήριξης που παρέχεται στους εκπαιδευόμενους αλλά και ο αντίκτυπος της πανδημίας του COVID – 19 στην επίδοση και την ποιότητα μάθησης των μαθητών [107]. Ο απώτερος στόχος της εξόρυξης εκπαιδευτικών δεδομένων είναι να μετατρέψει τα εκπαιδευτικά δεδομένα σε γνώση που μπορεί να βελτιώσει τις εκπαιδευτικές διαδικασίες και αποφάσεις. Η εξόρυξη εκπαιδευτικών δεδομένων αποτελεί διεπιστημονικό τομέα που εμπλέκει διαφορετικούς τομείς όπως η Επιστήμη των Υπολογιστών, η Εκπαίδευση και η Στατιστική, η μεταξύ

τους αλληλοεπικάλυψη με τη σειρά της δημιουργεί νέες υποπεριοχές που σχετίζονται με τη διαδικασία εξόρυξης εκπαιδευτικών δεδομένων όπως οι εκπαιδευτικές αναλύσεις, η εξόρυξη δεδομένων και η μηχανική μάθηση και η ηλεκτρονική μάθηση [12].

Η εξόρυξη εκπαιδευτικών δεδομένων διαρθρώνεται σε τέσσερις φάσεις:

1. Η πρώτη φάση της διαδικασίας εξόρυξης εκπαιδευτικών δεδομένων (χωρίς την προεπεξεργασία) είναι η ανακάλυψη σχέσεων στα δεδομένα. Αυτό περιλαμβάνει την αναζήτηση μέσω μιας αποθήκης δεδομένων από ένα εκπαιδευτικό περιβάλλον με στόχο την εύρεση σταθερών σχέσεων μεταξύ των μεταβλητών.
2. Επιβεβαίωση σχέσεων που ανακαλύφθηκαν ώστε να αποφευχθεί η υπερπροσαρμογή.
3. Εφαρμογή επιβεβαιωμένων σχέσεων για να γίνουν προβλέψεις για μελλοντικά γεγονότα στο μαθησιακό περιβάλλον.
4. Χρήση των προβλέψεων για την υποστήριξη των διαδικασιών λήψης αποφάσεων και των αποφάσεων πολιτικής [108].



Εικόνα 21: Επιστημονικοί τομείς που εμπλέκονται στην εξόρυξη εκπαιδευτικών δεδομένων.

4.2 Προεπεξεργασία δεδομένων (data processing)

Η προεπεξεργασία δεδομένων είναι το πρώτο και πολύ σημαντικό βήμα στη διαδικασία εξόρυξης δεδομένων σε οποιοδήποτε επιστημονικό πεδίο. Στο επιστημονικό πεδίο της εκπαίδευσης είναι ιδιαίτερα κρίσιμο ώστε να εξασφαλίσουμε επαρκή σύνολα δεδομένα που δυνητικά περιλαμβάνουν όσο το δυνατόν πιο χρήσιμες πληροφορίες. Κατά τη διαδικασία της προεπεξεργασίας έχουμε δύο κύριες τεχνικές: τις τεχνικές ανίχνευσης για την εύρεση ατελειών σε σύνολα δεδομένων και τις τεχνικές μετατροπής που προσανατολίζονται στην απόκτηση πιο διαχειρίσιμων συνόλων δεδομένων [109]. Γενικά η διαδικασία της προεπεξεργασίας αποτελείται από τέσσερα βασικά στάδια: καθαρισμό, ενσωμάτωση, μείωση και μετασχηματισμό.

1. Καθαρισμός δεδομένων: Ο καθαρισμός δεδομένων είναι η διαδικασία καθαρισμού των συνόλων δεδομένων με την καταγραφή τιμών που λείπουν, την αφαίρεση των ακραίων τιμών, τη διόρθωση ασυνεπών σημείων δεδομένων και την εξομάλυνση των θορυβωδών δεδομένων.
2. Ενσωμάτωση δεδομένων: Δεδομένου ότι τα δεδομένα συλλέγονται από διάφορες πηγές, η συνένωση δεδομένων είναι κρίσιμη διαδικασία για την προετοιμασία των δεδομένων.
3. Μείωση δεδομένων: Η μείωση δεδομένων στοχεύει στη μείωση του όγκου των δεδομένων και επομένως στη μείωση του κόστους που αφορά την εξόρυξη δεδομένων ή την ανάλυση των δεδομένων αλλά και την εξάλειψη των περιττών χαρακτηριστικών.
4. Μετασχηματισμός δεδομένων: Ο μετασχηματισμός δεδομένων είναι η διαδικασία μετατροπής δεδομένων από μια μορφή σε άλλη. Στόχος είναι η βέλτιστη ακρίβεια και η αποδοτικότητα των αλγορίθμων εξόρυξης [110].

4.3 Μεθοδολογία

Η μεθοδολογία που χρησιμοποιήθηκε αφορά την εφαρμογή αλγορίθμων συσταδοποίησης που προσφέρονται από το περιβάλλον WEKA σε διαφορετικά σύνολα δεδομένων με απώτερο στόχο την προοδευτική σύγκριση και αποτίμηση της απόδοσης αυτών των αλγορίθμων. Στο πρώτο πείραμα έγινε αρχικά σύγκριση μεταξύ των αλγορίθμων Simple K –Means και Hierarchical για διαφορετικές λειτουργίες συσταδοποίησης και δευτερεύοντος σύγκριση με τον κατηγοριοποιητή J48. Στο δεύτερο πείραμα επίσης έγινε σύγκριση μεταξύ των αλγορίθμων Simple K –Means και Hierarchical για διαφορετικές λειτουργίες συσταδοποίησης και με διαφορετικές παραμέτρους και δευτερεύοντος σύγκριση με τον κατηγοριοποιητή J48. Στο τρίτο πείραμα εφαρμόζουμε τον αλγόριθμο Simple K – Means και κάνουμε σύγκριση μεταξύ διαφορετικών παραμέτρων του αλγορίθμου και διαφορετικών λειτουργιών συσταδοποίησης. Στο τέταρτο πείραμα πραγματοποιείται σύγκριση μεταξύ των αλγορίθμων Simple K – Means, Make Density Based Clusterer, Farthest First, Expectation Maximazation και Filtered σε δύο διαφορετικές λειτουργίες συσταδοποίησης και τέλος στο πέμπτο πείραμα γίνεται σύγκριση του συνόλου των αλγορίθμων συσταδοποίησης Canopy, Cobweb, Hierarchical, Simple K – Means, Make Density Based Clusterer, Farthest First, Expectation Maximazation και Filtered και εξαγωγή κανόνων συσχέτισης με εφαρμογή του αλγορίθμου Apriori.

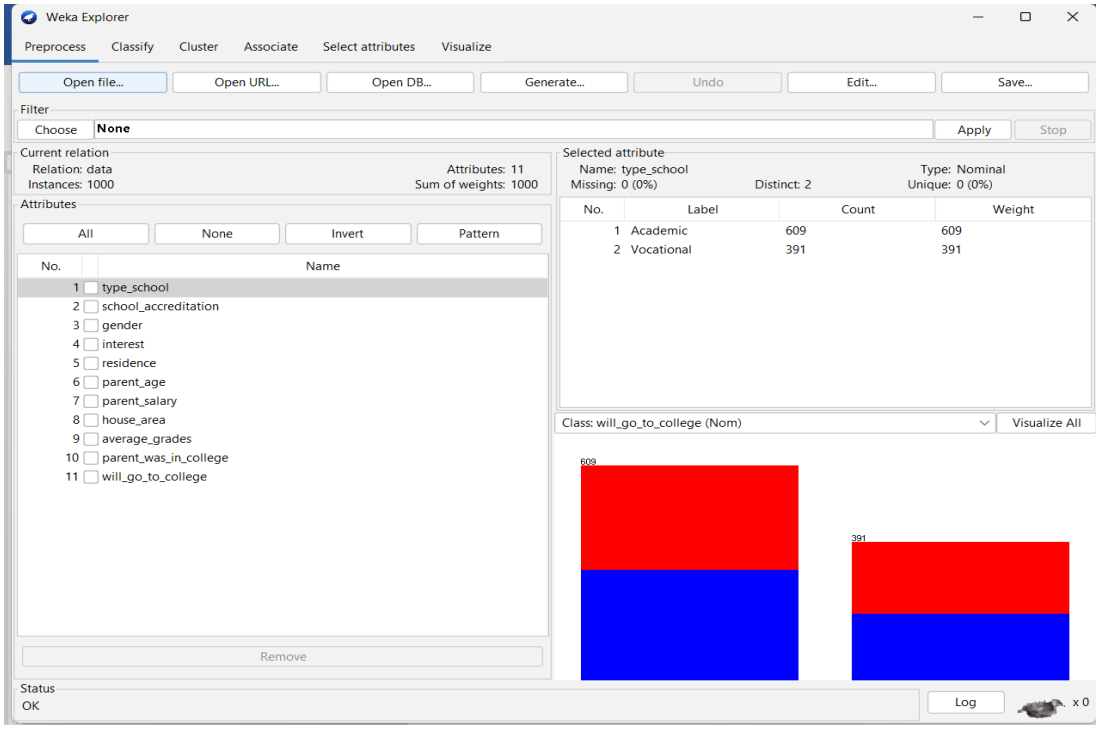
4.4 Περιγραφή δεδομένων

Τα δεδομένα που έχουμε χρησιμοποιήσει αφορούν εκπαιδευτικά δεδομένα τα οποία προέρχονται από το ψηφιακό αποθετήριο Kaggle και αποτελούνται από διάφορους τύπους δεδομένων κατηγορικά, αριθμητικά κτλ. ενώ ποικίλλουν και ως προς τον αριθμό των χαρακτηριστικών και των στιγμιότυπων. Αξιοποιήσαμε πέντε διαφορετικά σύνολα δεδομένων σχετικά με προβλέψεις των μαθητών που θα πάνε στο κολλέγιο, τη σχολική φοίτηση σε σχολεία στο Τέξας, τον αντίκτυπο της πανδημίας του Covid – 19 στην εκπαίδευση, τις βαθμολογίες του διαγωνισμού PISA μεταξύ των ετών 2006 – 2018 και της προσαρμοστικότητας μαθητών στη Βενεζουέλα στην διαδικτυακή εκπαίδευση.

4.5 Εκτέλεση πειραμάτων και σύγκριση

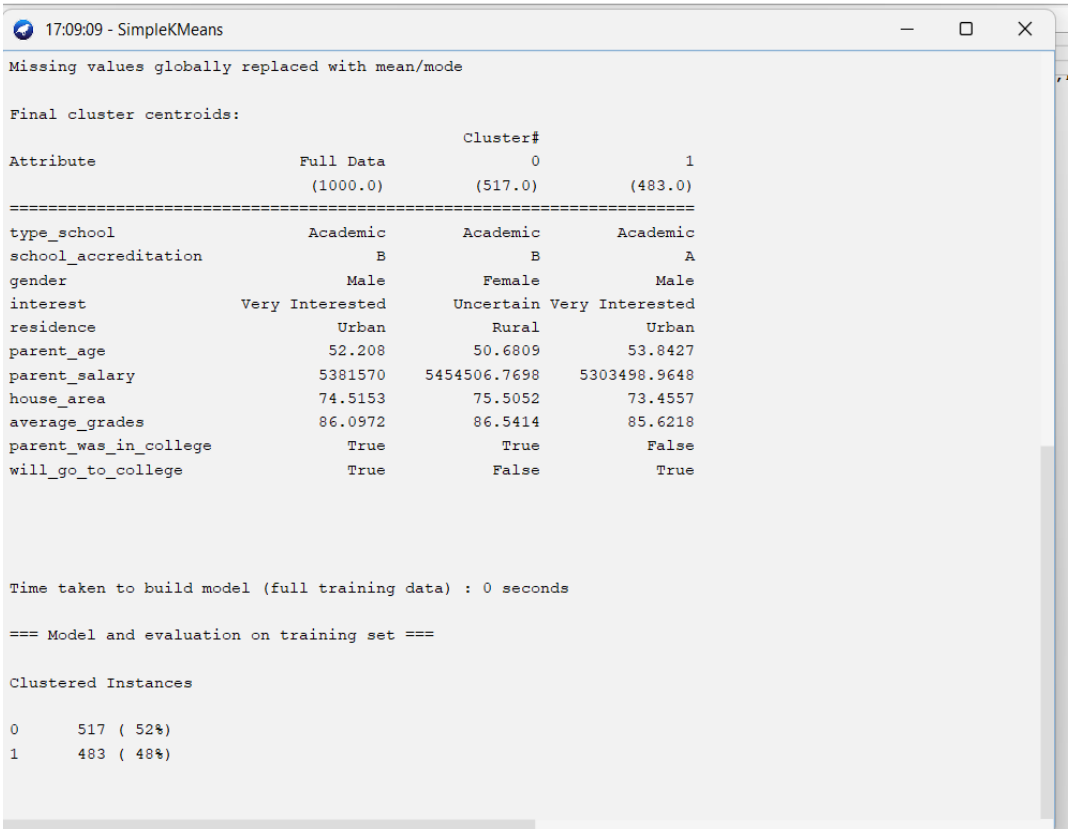
4.5.1 1ο πείραμα: Πρόβλεψη μαθητών που θα πάνε στο κολλέγιο

Στη συγκεκριμένη υλοποίηση χρησιμοποιήθηκε ένα σύνολο δεδομένων με όνομα “Go to College Dataset”. Το σύνολο δεδομένων αποτελείται από συνθετικά δεδομένα που δημιουργήθηκαν για ένα έργο κολεγίου όπου υποθετικά το υπουργείο Παιδείας της Ινδονησίας θέλει να αυξήσει το πλήθος των μαθητών που θα πάνε στο κολλέγιο και στοχεύουν στο να προβλέψουν εάν οι μαθητές θα συνεχίσουν την εκπαίδευση τους στο κολέγιο ή όχι. Παρουσιάζει το μέσο όρο βαθμολογίας, δημογραφικά, κοινωνικά και σχολικά χαρακτηριστικά των μαθητών. Συνολικά αποτελείται από 1000 στιγμιότυπα (instances) και 11 χαρακτηριστικά (attributes). Τα οποία είναι τα παρακάτω: type_school, school_accreditation, gender, interest, residence, parent_age, parent_salary, house_area, average_grades, parents_was_in_college, will_go_to_college. Το τελευταίο χαρακτηριστικό είναι προσημασμένο ως κλάση. Στην εικόνα 22 παρακάτω παρουσιάζεται το σύνολο δεδομένων του αρχείο data.arff που αποτελείται από αριθμητικά και κατηγορικά δεδομένα.



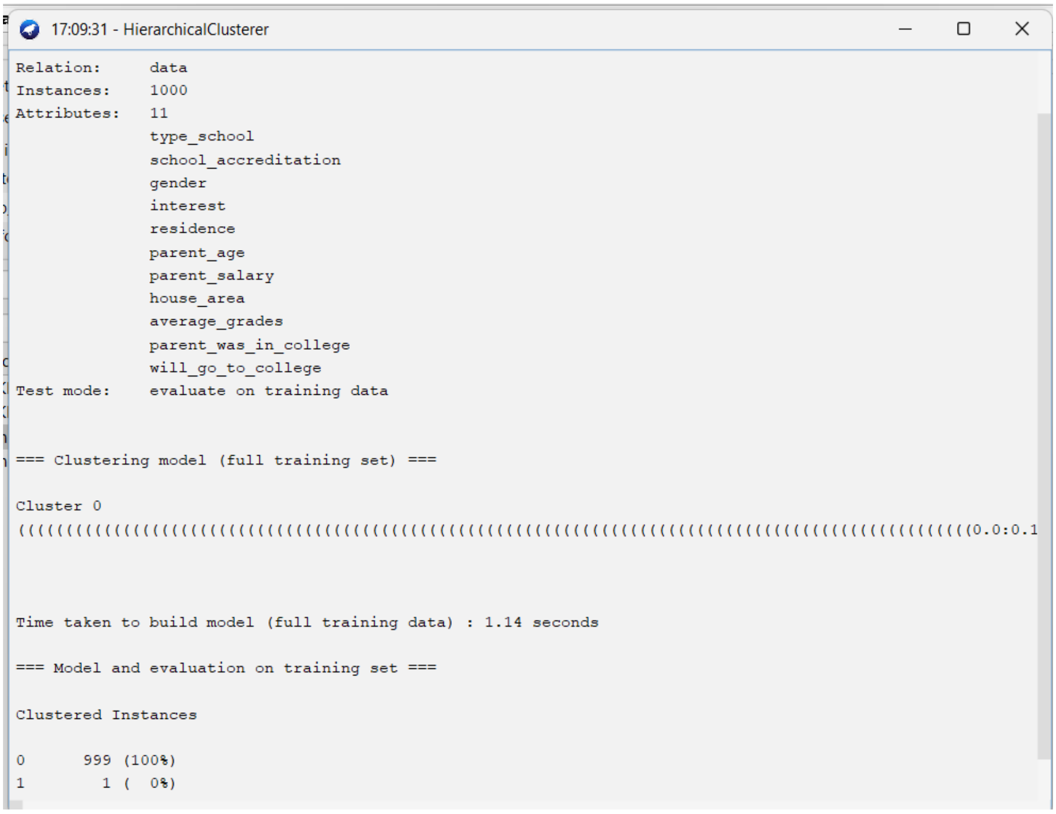
Εικόνα 22: Προεπεξεργασία του αρχείου εισόδου data.arff

Αρχικά, ξεκινάμε το πείραμα με εφαρμογή του αλγορίθμου Simple K-means με επιλεγμένο cluster mode “use training set” χωρίς να ορίσουμε μία συγκεκριμένη κλάση, σε αυτή την λειτουργία δεν χρησιμοποιούνται οι ετικέτες (labels) καθώς δεν υπάρχει δοκιμαστική φάση. Δημιουργείται μια λίστα με τον αριθμό των στιγμιότυπων που έχουν εκχωρηθεί σε κάθε συστάδα σε ποσοστό επι της εκατό, όπως φαίνεται στην εικόνα 23. Ως αριθμό k συστάδων αφήσαμε την προεπιλεγμένη τιμή 2 καθώς στόχος είναι να προβλέψουμε ποιοι θα πάνε (true) και ποιοι όχι (false) κολλέγιο. Επίσης, καταγράφεται ο αριθμός των επαναλήψεων όπου ανέρχεται σε 5, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) όπου ανέρχεται σε 2401.02 και ο χρόνος κατασκευής του μοντέλου όπου ανέρχεται σε 0 sec. Τέλος, παρατηρούμε ότι το 52% των στιγμιότυπων του συνόλου των δεδομένων ομαδοποιούνται στη συστάδα 0 (false).



Εικόνα 23: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means

Στη συνέχεια, με τις ίδιες παραμέτρους σταθερές δηλαδή πλήθος συστάδων ίσο με 2 και επιλεγμένο cluster mode “use training set” εφαρμόζουμε τον αλγόριθμο Hierarchical Clusterer. Όπως φαίνεται και στην εικόνα 24 ο χρόνος κατασκευής του μοντέλου αυξάνεται σημαντικά και ανέρχεται σε 1,14 sec. Τέλος, μια ακραία ομαδοποίηση στιγμιότυπων του συνόλου των δεδομένων σε ποσοστό 100% στη συστάδα 0.



Εικόνα 24: Αποτέλεσμα συσταδοποίησης Ιεραρχικού αλγορίθμου.

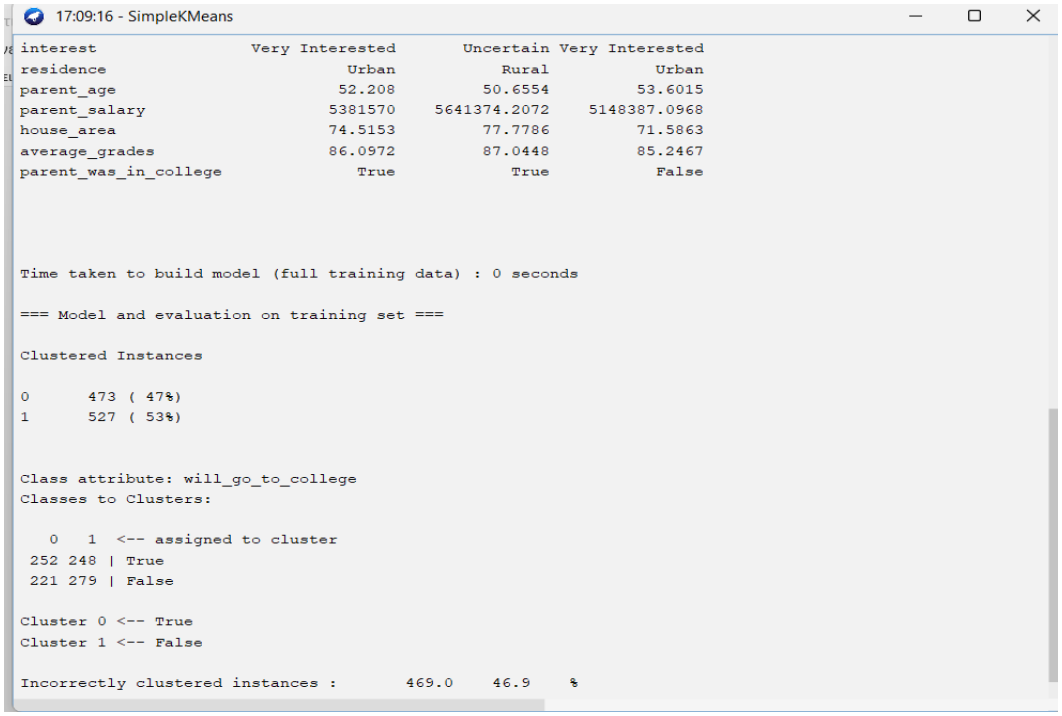
Όπως φαίνεται από τον παρακάτω πίνακα 1, ο αλγόριθμος Simple K - Means αποδίδει πιο ρεαλιστικά ποσοστά συσταδοποίησης μεταξύ των συστάδων του συνόλου των προσημασμένων στιγμιότυπων σε αντίθεση με τον Ιεραρχικό αλγόριθμο όπου παρατηρείται μια ακραία κατανομή. Συγκριτικά με τους χρόνους εκτέλεσης παρατηρείται ότι η ιεραρχική συσταδοποίηση δεν είναι τόσο αποτελεσματική στην υπολογιστική πολυπλοκότητα αφού έχει τον υψηλότερο χρόνο εκτέλεσης.

.

Αλγόριθμος	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων	
		0	1
Simple K - Means	0 sec.	52%	48%
Hierarchical	1,14 sec.	100%	0%

Πίνακας 1: Χρόνοι εκτέλεσης και ποσοστά ομαδοποιημένων στιγμιότυπων για κάθε υλοποίηση αλγορίθμων

Στο επόμενο στάδιο εφαρμόζεται ο αλγόριθμος Simple K – Means με επιλεγμένο cluster mode “classes to cluster evaluation” και ορίζεται ως κλάση το χαρακτηριστικό will_go_to_college, ενώ ο αριθμός k των συστάδων παραμένει 2. Σε αυτήν τη λειτουργία συσταδοποίησης, οι κλάσεις (ετικέτες) χρησιμοποιούνται στη δοκιμαστική φάση. Συγκρίνει πόσο καλά οι επιλεγμένες συστάδες ταιριάζουν με μια ήδη προσημασμένη κλάση (ετικέτα) στα δεδομένα. Όπως φαίνεται και παρακάτω στην εικόνα 25 το 47% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0 (true), Επίσης, καταγράφεται ο αριθμός των επαναλήψεων όπου ανέρχεται σε 5, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) όπου ανέρχεται σε 1936.90 και είναι εμφανώς μικρότερος σε σχέση με την προηγούμενη υλοποίηση και ο χρόνος κατασκευής του μοντέλου όπου ανέρχεται σε 0 sec και παραμένει ίδιος με την προηγούμενη υλοποίηση. Τέλος, το ποσοστό ακρίβειας που προκύπτει από τα στιγμιότυπα που δεν ομαδοποιήθηκαν σωστά ανέρχεται σε 53.1%.

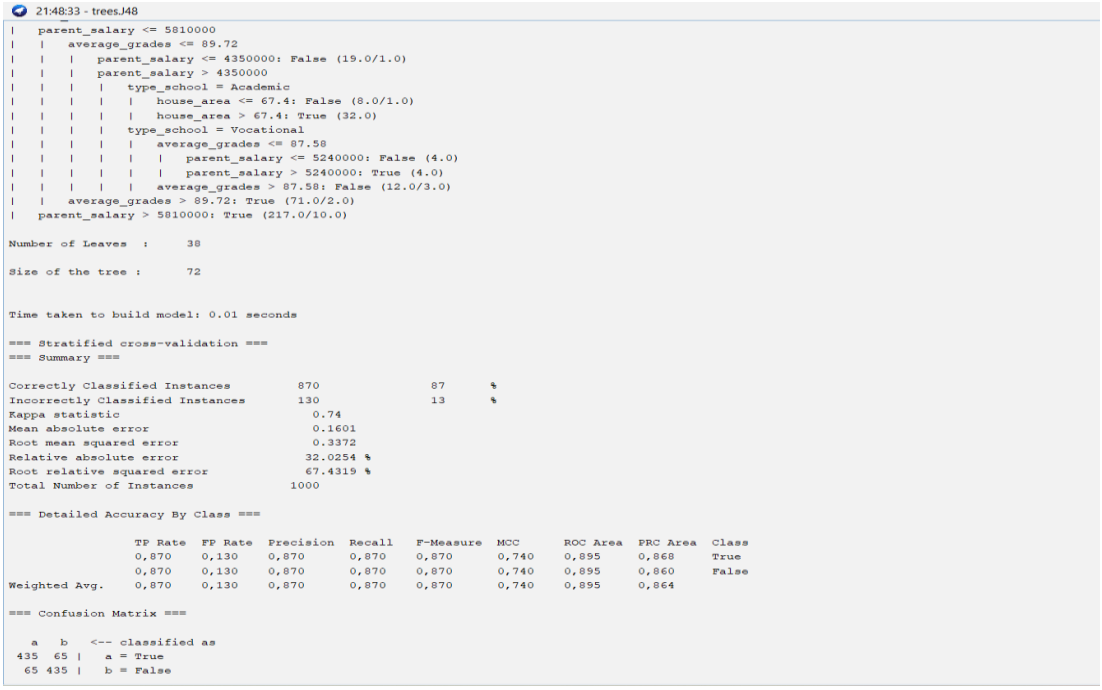


Εικόνα 25: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means (classes to cluster evaluation)

Στη συνέχεια υλοποιούμε με τις ίδιες παραμέτρους σταθερές δηλαδή πλήθος συστάδων ίσο με 2 και επιλεγμένο cluster mode “ classes to cluster evaluation ” και ορισμένη κλάση το χαρακτηριστικό will_go_to_college εφαρμόζουμε τον αλγόριθμο Hierarchical Clusterer. Παρατηρούμε τα εξής όπως φαίνεται και στην εικόνα 26, το 100% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0 (false) όπου υποδηλώνει μια ακραία κατανομή όπως και προηγουμένως, ο χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,99 sec. άρα είναι ελαφρά μειωμένος. Τέλος, το ποσοστό ακρίβειας που προκύπτει από τα στιγμιότυπα που δεν ομαδοποιήθηκαν σωστά ανέρχεται σε 50.1%.

Εικόνα 26: Αποτέλεσμα συσταδοποίησης αλγορίθμου Hierarchical Clusterer (classes to cluster evaluation)

Στο τελικό στάδιο δοκιμάζουμε να εκπαιδεύσουμε το ίδιο σύνολο δεδομένων με τον αλγόριθμο ομαδοποίησης J48 που ανήκει στους αλγορίθμους Δέντρων Απόφασης (Decision Tree) ώστε να εντοπίσουμε τους κανόνες ομαδοποίησης και να συγκρίνουμε την ακρίβεια των αποτελεσμάτων που προκύπτουν από τη σύγκριση μεταξύ των τριών αλγορίθμων καθώς θεωρείτε αποτελεσματικός στο να «κόβει» δεδομένα που μπορούν να χαρακτηριστούν ως outliers για την έρευνα μας. Στη συνέχεια στην επιλογή Test options επιλέγουμε “Cross – validation” και εφαρμόζουμε τον αλγόριθμο J48. Διαπιστώνουμε τα εξής όπως φαίνεται και στην εικόνα 27 παρακάτω, ο χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,01 sec. άρα προκύπτει ο επόμενος αποτελεσματικότερος από άποψη χρόνου μετά τον K – Means και το ποσοστό ακρίβειας ανέρχεται σε 87% εμφανώς μεγαλύτερο συγκριτικά με τους K – Means και Hierarchical Clusterer. Τέλος, και το ποσοστό σφαλμάτων (Root mean squared error) εμφανίζεται σημαντικά μικρότερο σε σχέση με τον K – Means και ανέρχεται σε 0.3372.



Εικόνα 27: Αποτέλεσμα ομαδοποίησης αλγορίθμου J48 (cross - validation)

Όπως φαίνεται από τον παρακάτω πίνακα 2, ο αλγόριθμος Simple K - Means αποδίδει καλύτερους χρόνους εκτέλεσης ενώ παρατηρείται ότι η ιεραρχική συσταδοποίηση δεν είναι τόσο αποτελεσματική στην υπολογιστική πολυπλοκότητα αφού έχει τον υψηλότερο χρόνο εκτέλεσης ενώ επίσης το ποσοστό ακρίβειας του J48 είναι σημαντικά καλύτερο σε σχέση με τον K – Means και τον Ιεραρχικό αλγόριθμο.

Αλγόριθμος	Υπολογιστικός χρόνος	Ακρίβεια
Simple K - Means	0 sec.	53,1%
Hierarchical	0,99 sec.	50,1%
J48	0,01 sec.	87%

Πίνακας 2: Ποσοστό ακρίβειας για κάθε αλγόριθμο.

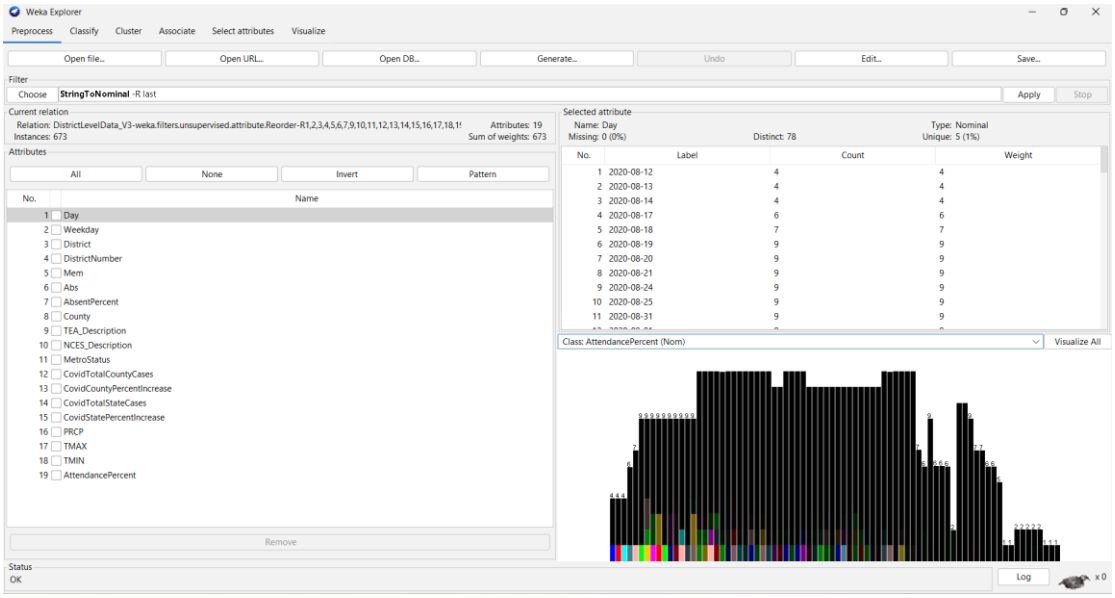
Συμπερασματικά, μπορούμε να πούμε ότι από την εφαρμογή δύο διαφορετικών αλγορίθμων σε δύο διαφορετικές λειτουργίες συσταδοποίησης (cluster mode) και όπως αποτυπώνεται από τους παραπάνω πίνακες ο αλγόριθμος Simple K – Means μοιάζει να

είναι καταλληλότερος και να εφαρμόζει πιο ρεαλιστική κατανομή στιγμιότυπων στις επιμέρους συστάδες. Ενώ ο J48 εμφανίζει πολύ μεγαλύτερη ακρίβεια και δημιουργεί ένα πιο εύρωστο μοντέλο σε σύγκριση με τους άλλους δύο αλγορίθμους.

4.5.2 2ο πείραμα: Σχολική φοίτηση σε σχολεία στο Τέξας

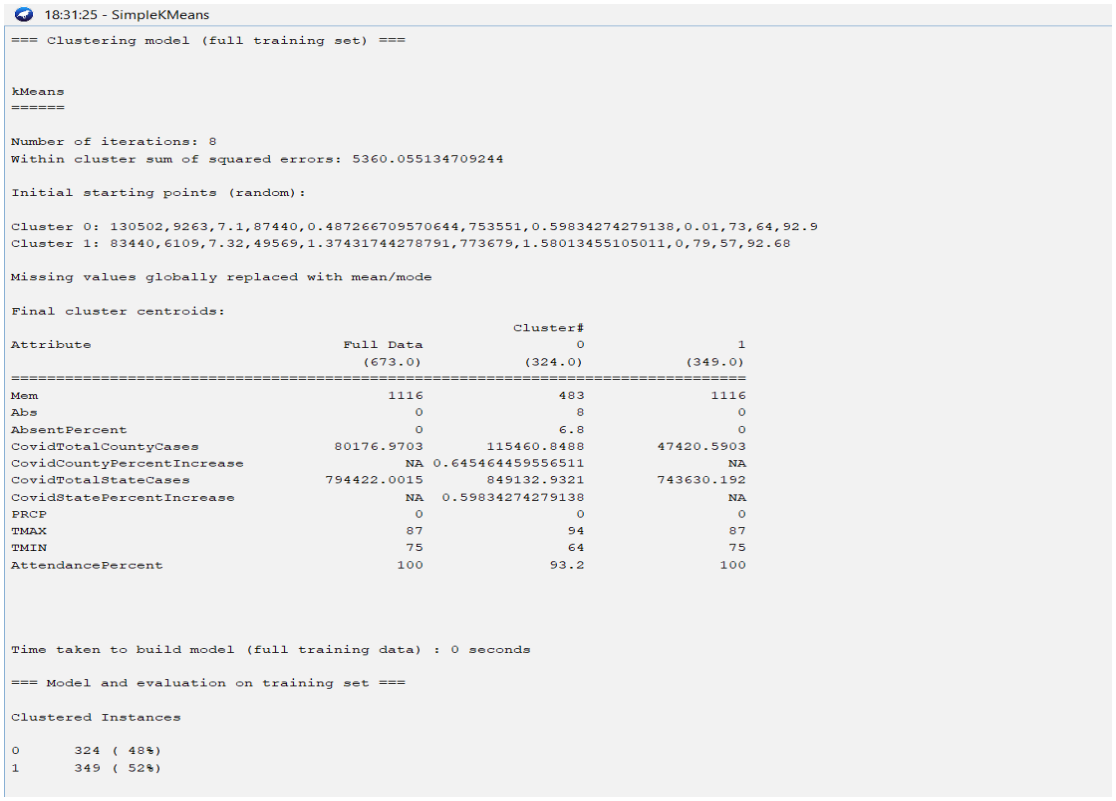
Στη συγκεκριμένη υλοποίηση επιλέχθηκε από μία ομάδα συνόλων δεδομένων με όνομα "Texas Attendance" ένα σύνολο δεδομένων από τρία συνολικά με όνομα DistrictLevelData_V3.csv καθώς τα δεδομένα που εμπεριέχει θεωρήθηκαν πιο πλήρη και σχετικά με το υπό έρευνα θέμα. Το σύνολο των δεδομένων αναφέρεται σε δεδομένα του πραγματικού κόσμου και συλλέχθηκε από σχολικές περιφέρειες στην πολιτεία του Τέξας, πληροφορίες δημόσιας απογραφής και δημόσια δεδομένα COVID 19. Σκοπός είναι να διερευνηθούν οι παράγοντες που σχετίζονται με τη συμμετοχή των μαθητών στην εκπαιδευτική διαδικασία στην πολιτεία του Τέξας για τους δύο πρώτους μήνες του ακαδημαϊκού σχολικού έτους 2020 – 2021 εξαιτίας και των ιδιαίτερων υγειονομικών συνθηκών. Παρουσιάζει δημογραφικά, υγειονομικά, σχολικά χαρακτηριστικά των μαθητών και χαρακτηριστικά των σχολείων.

Συνολικά αποτελείται από 673 στιγμιότυπα (instances) και 19 χαρακτηριστικά (attributes). Τα οποία είναι τα παρακάτω: No., Day, Weekday, District , DistrictNumber, Mem, Abs, AbsentPercent, County, TEA_Description, NCES_Description, MetroStatus, CovidTotalCountyCases, CovidCountyPercentIncrease, CovidTotalStateCases, CovidStatePercentIncrease, PRCP, TMAX, TMIN, AttendancePercent. Το τελευταίο χαρακτηριστικό είναι προσημασμένο ως κλάση. Στην φάση της προεπεξεργασίας κρίθηκε απαραίτητο να γίνει εφαρμογή του φίλτρου StringToNominal ώστε να μετατραπούν κάποια χαρακτηριστικά από string σε nominal ώστε να μπορεί να τα επεξεργαστεί ο K - Means. Στην εικόνα 28 παρακάτω παρουσιάζεται το σύνολο δεδομένων του αρχείου DistrictLevelData_V3.arff που αποτελείται από αριθμητικά και κατηγορικά δεδομένα.



Εικόνα 28: Προεπεξεργασία του αρχείου εισόδου DistrictLevelData_V3.arff

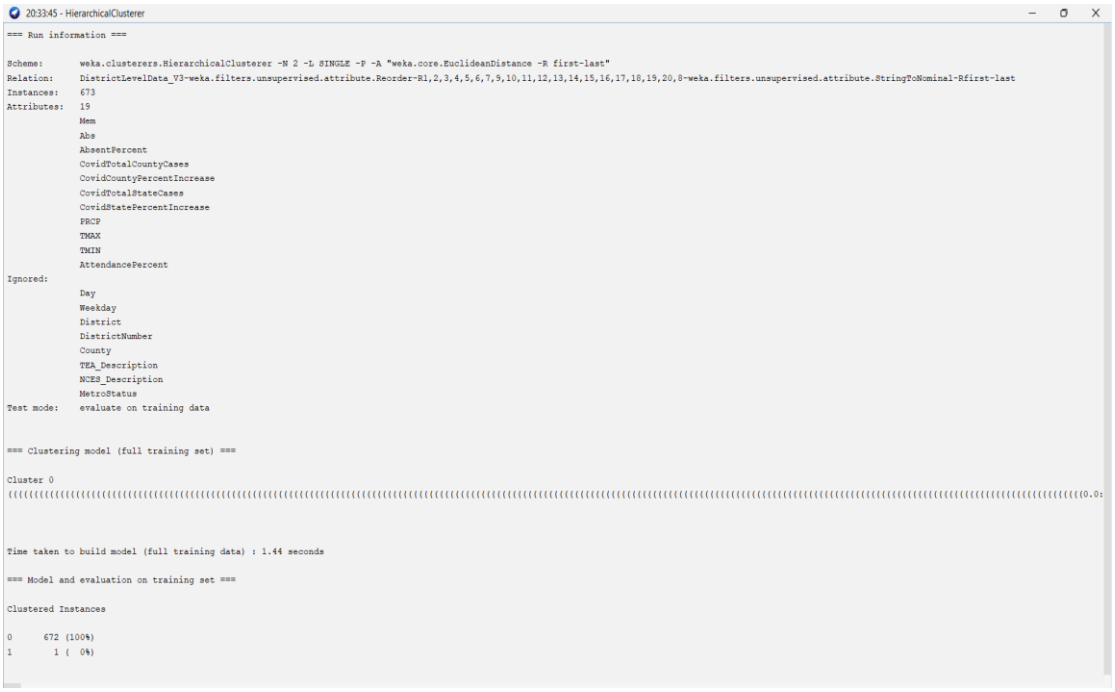
Αρχικά, ξεκινάμε το πείραμα με εφαρμογή του αλγορίθμου Simple K-means με επιλεγμένο cluster mode “use training set” χωρίς να ορίσουμε μία συγκεκριμένη κλάση, σε αυτή την λειτουργία δεν χρησιμοποιούνται οι ετικέτες (labels) καθώς δεν υπάρχει δοκιμαστική φάση. Δημιουργείται μια λίστα με τον αριθμό των στιγμιότυπων που έχουν εκχωρηθεί σε κάθε συστάδα σε ποσοστό επι της εκατό, όπως φαίνεται στην εικόνα 29. Ως αριθμό k συστάδων ορίσαμε διαδοχικά τις παρακάτω τιμές 2, 4, 6, 8, 10 και 12 ενώ από την επιλογή “Ignore attributes” επιλέξαμε να μην ληφθούν υπόψιν κάποια από τα χαρακτηριστικά του dataset που θεωρούμε ήσσονος σημασίας για την έρευνα μας που αφορά τους παράγοντες που επηρεάζουν την σχολική παρακολούθηση των μαθητών. Επίσης, καταγράφεται ο αριθμός των επαναλήψεων όπου ανέρχεται για k = 2, 8, 10 σε 8, για k = 4, 6 σε 6 και για k = 12 σε 12, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) παρατηρείται ότι μειώνεται σταδιακά ξεκινώντας από k = 2 σε 5360.05 και καταλήγοντας για k = 12 σε 4825.57 και ο χρόνος κατασκευής του μοντέλου όπου ανέρχεται σε 0 sec. για k = 2,4 και σε 0.01 sec. για k = 6, 8, 10, 12. Τέλος, παρατηρούμε την κατανομή στιγμιότυπων ανά πλήθος συστάδων στον πίνακα 3.



```
18:31:25 - SimpleKMeans
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 5360.055134709244

Initial starting points (random):

Cluster 0: 130502,9263,7.1,87440,0.487266709570644,753551,0.59834274279138,0.01,73,64,92.9
Cluster 1: 83440,6109,7.32,49569,1.37431744278791,773679,1.58013455105011,0,79,57,92.68

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                          Full Data           0                1
                                     (673.0)     (324.0)          (349.0)
=============================================================================
Mem                                  1116          483             1116
Abs                                     0            8                0
AbsentPercent                           0          6.8                0
CovidTotalCountyCases            80176.9703   115460.8488      47420.5903
CovidCountyPercentIncrease               NA  0.645464459556511           NA
CovidTotalStateCases            794422.0015    849132.9321      743630.192
CovidStatePercentIncrease                NA   0.59834274279138           NA
PRCP                                    0            0                0
TMAX                                   87           94               87
TMIN                                   75           64               75
AttendancePercent                     100         93.2              100



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       324 ( 48%)
1       349 ( 52%)
```

Εικόνα 29: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means για k = 2

Στη συνέχεια, με τις ίδιες παραμέτρους σταθερές δηλαδή πλήθος συστάδων ίσο με 2, 4, 6, 8, 10, 12 και επιλεγμένη την επιλογή “Ignore attributes” για παράληψη των χαρακτηριστικών που προαναφέρθηκαν ήδη παραπάνω ορίζουμε στην επιλογή cluster mode “use training set” και εφαρμόζουμε τον αλγόριθμο Hierarchical Clusterer. Όπως φαίνεται και στην εικόνα 30 ο χρόνος κατασκευής του μοντέλου αυξάνεται σημαντικά και ανέρχεται σε 1,44 sec. για k = 2 και κλιμακώνεται σε 1.46 sec. για k =12. Τέλος, παρατηρείται και σε αυτή την περίπτωση μια ακραία ομαδοποίηση στιγμιότυπων του συνόλου των δεδομένων σε μία συστάδα και συγκεκριμένα τη C0 σε ποσοστό 100% για k = 2, 4 σε ποσοστό 99% για k = 6, 8 και σε ποσοστό 98% για k = 10, και 12.



Εικόνα 30: Αποτέλεσμα συσταδοποίησης Ιεραρχικού αλγορίθμου για k = 2.

Όπως φαίνεται από τον παρακάτω πίνακα 3, ο αλγόριθμος Simple K - Means αποδίδει πιο ρεαλιστικά ποσοστά συσταδοποίησης μεταξύ των συστάδων του συνόλου των προσημασμένων στιγμιότυπων σε αντίθεση με τον Ιεραρχικό αλγόριθμο όπου παρατηρείται μια ακραία κατανομή. Συγκριτικά με τους χρόνους εκτέλεσης παρατηρείται ότι η ιεραρχική συσταδοποίηση δεν είναι τόσο αποτελεσματική στην υπολογιστική πολυπλοκότητα αφού έχει τον υψηλότερο χρόνο εκτέλεσης όπως παρατηρείται και στο 1^{ο} πείραμα.

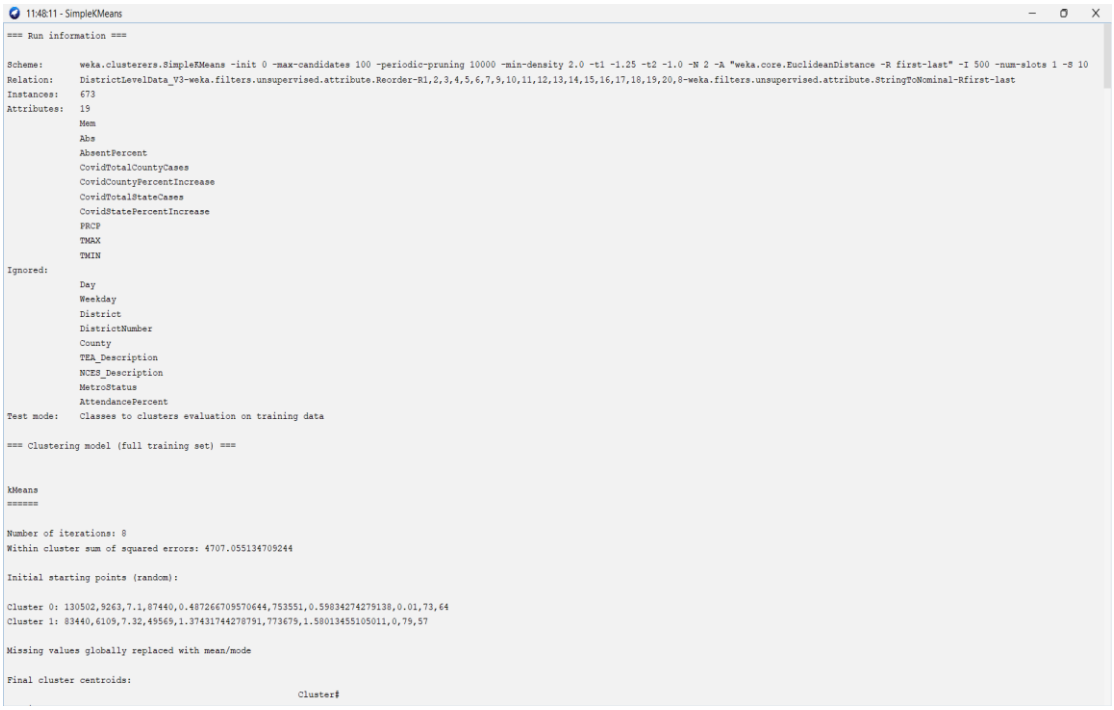
Αλγόριθμος	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων											
Simple K – Means k = 2	0 sec.	0	1										
		48%	52%										
Hierarchical k = 2	1,44 sec.	100%	0%										
Simple K – Means k = 4	0 sec.	0	1	2	3								
		18%	31%	28%	23%								
Hierarchical k = 4	1.53 sec.	100%	0%	0%	0%								
Simple K –Means k = 6	0,01 sec.	0	1	2	3	4	5						
		5%	19%	22%	14%	20%	19%						
Hierarchical k = 6	1,49 sec.	99%	0%	0%	0%	0%	0%						
Simple K –Means k = 8	0,01 sec.	0	1	2	3	4	5	6	7				
		5%	18%	12%	13%	19%	13%	15%	4%				
Hierarchical k = 8	2,16 sec.	99%	0%	0%	0%	0%	0%	0%	0%				
Simple K –Means k = 10	0,01 sec.	0	1	2	3	4	5	6	7	8	9		
		3%	16%	10%	12%	15%	12%	12%	9%	8%	2%		
Hierarchical k = 10	1,34 sec.	98%	0%	1%	0%	0%	0%	0%	0%	0%	0%		
Simple K – Means k = 12	0,01 sec.	0	1	2	3	4	5	6	7	8	9	10	11
		3%	15%	9%	8%	13%	10%	9%	3%	7%	2%	10%	11%
Hierarchical k = 12	1,46 sec.	98%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Πίνακας 3: Χρόνοι εκτέλεσης και ποσοστά ομαδοποιημένων στιγμιότυπων για κάθε υλοποίηση αλγορίθμων.

Στο επόμενο στάδιο εφαρμόζεται ο αλγόριθμος Simple K – Means με επιλεγμένο cluster mode “classes to cluster evaluation” και ορίζεται ως κλάση το χαρακτηριστικό AttendancePercent, ενώ από την επιλογή “Ignore attributes” επιλέξαμε να μην ληφθούν υπόψιν κάποια από τα χαρακτηριστικά του dataset που θεωρούμε ήσσονος σημασίας για την έρευνα μας που αφορά τους παράγοντες που επηρεάζουν την σχολική

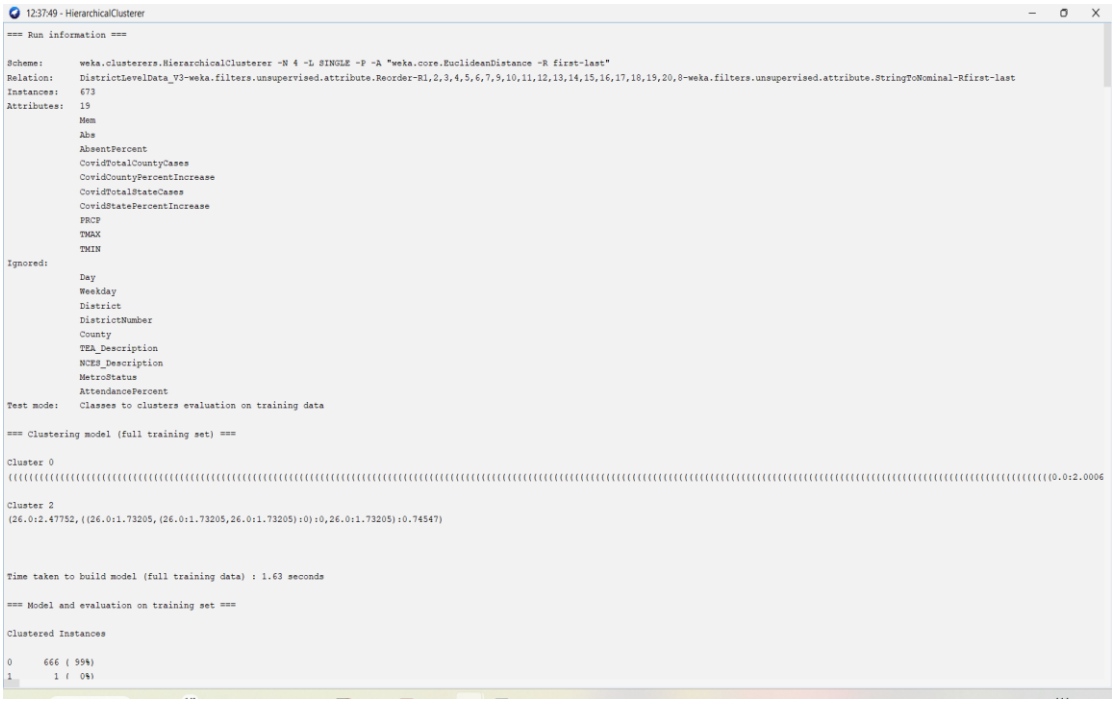
παρακολούθηση των μαθητών όπως και παραπάνω, τέλος, ο αριθμός k των συστάδων επιχειρήθηκε αρχικά να οριστεί σταδιακά σε 2, 4, 6, 8, 10, 12 όπως και προηγουμένως αλλά παρατηρήθηκε ότι όσο αυξανόταν ο αριθμός συστάδων αυξανόταν δυσανάλογα και η υπολογιστική πολυπλοκότητα καθώς μετά από ένα εύλογο χρονικό διάστημα της τάξης των 15 λεπτών ακόμα δεν είχε προκύψει κατανομή συστάδων επομένως και περιορίσαμε το πείραμα σε k = 2 και 4. Σε αυτήν τη λειτουργία συσταδοποίησης, οι κλάσεις (ετικέτες) χρησιμοποιούνται στη δοκιμαστική φάση. Συγκρίνει πόσο καλά οι επιλεγμένες συστάδες ταιριάζουν με μια ήδη προσημασμένη κλάση (ετικέτα) στα δεδομένα.

Όπως φαίνεται και παρακάτω στην εικόνα 31 για k = 2 το 48% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0 και το 52% στη συστάδα 1, επίσης ο αριθμός των επαναλήψεων ανέρχεται σε 8 και το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) ανέρχεται σε 4707.05 και ο χρόνος κατασκευής του μοντέλου όπου ανέρχεται σε 0,01 sec. Τέλος, το ποσοστό ακρίβειας που προκύπτει από τα στιγμιότυπα που δεν ομαδοποιήθηκαν σωστά ανέρχεται σε 2,9718% που είναι εξαιρετικά χαμηλό. Στο αντίποδα για k = 4 το 18% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0, το 31% στη συστάδα 1, το 28% στη συστάδα 2 και το 23% στη συστάδα 3, επίσης ο αριθμός των επαναλήψεων ανέρχεται σε 5 και το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) ανέρχεται σε 4549.74 και ο χρόνος κατασκευής του μοντέλου όπου ανέρχεται σε 0 sec. Τέλος, το ποσοστό ακρίβειας που προκύπτει από τα στιγμιότυπα που δεν ομαδοποιήθηκαν σωστά ανέρχεται σε 4,6062%. Όπως μπορούμε να διαπιστώσουμε τα αποτελέσματα που προκύπτουν είναι εμφανώς καλύτερα για k = 4 αλλά τα ποσοστά ακρίβειας συνεχίζουν να παραμένουν εξαιρετικά χαμηλά και στις δύο περιπτώσεις.



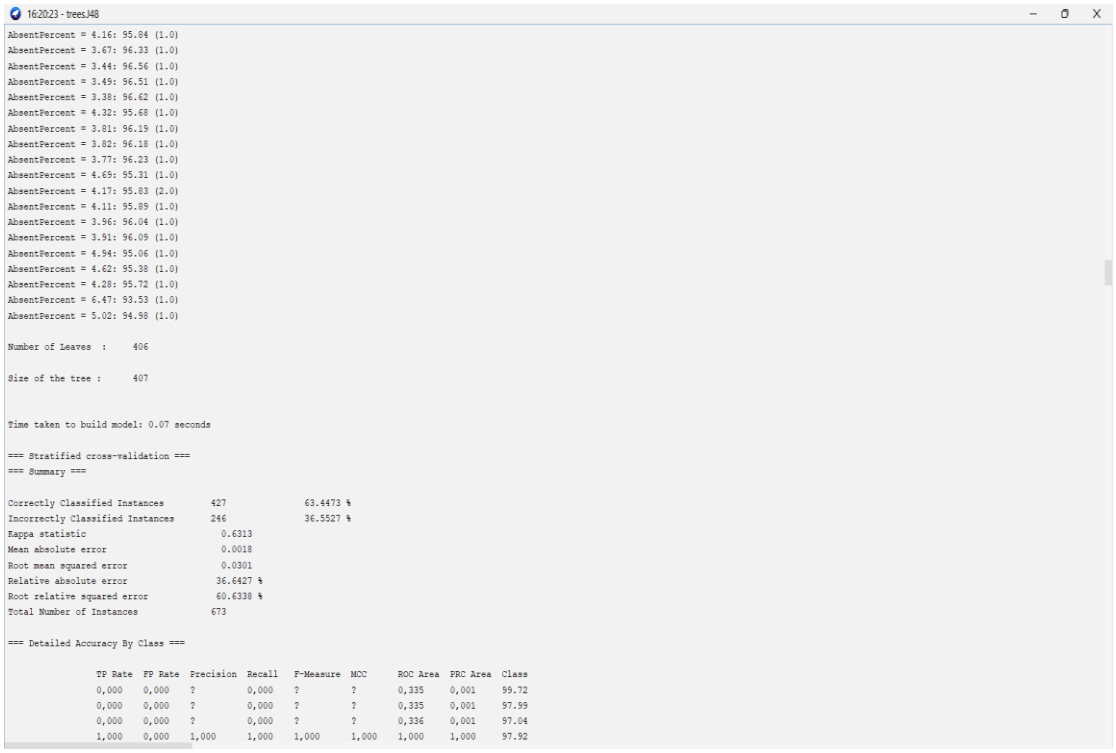
Εικόνα 31: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means (classes to cluster evaluation)

Στη συνέχεια υλοποιούμε με τις ίδιες παραμέτρους σταθερές, δηλαδή, πλήθος συστάδων ίσο με 2 και 4, επιλεγμένο cluster mode “classes to cluster evaluation”, ορισμένη κλάση το χαρακτηριστικό AttendancePercent και με τα χαρακτηριστικά Day, Weekday, District, DistrictNumber, County, TEA_Description, NCES_Description, MetroStatus να μην λαμβάνονται υπόψιν για την υλοποίηση μας εφαρμόζουμε τον αλγόριθμο Hierarchical Clusterer. Παρατηρούμε τα εξής όπως φαίνεται και στην εικόνα 32, για k = 2 το 100% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0 και 0% στη συστάδα 1, ενώ για k = 4 το 99% των στιγμιότυπων ομαδοποιούνται στη συστάδα 0, 0% στη συστάδα 1, 1% στη συστάδα 2 και 0% στη συστάδα 3 όπου όπως φαίνεται πάλι έχουμε μια ακραία κατανομή στη συστάδα 0, ο χρόνος κατασκευής του μοντέλου για k = 2 ανέρχεται σε 2,14 sec. και για k = 4 ανέρχεται σε 1,63 sec. άρα είναι ελαφρά μειωμένος. Τέλος, το ποσοστό ακρίβειας που προκύπτει από τα στιγμιότυπα που δεν ομαδοποιήθηκαν σωστά για k = 2 ανέρχεται σε 1,6345% και για k = 4 ανέρχεται σε 1,9316.



Εικόνα 32: Αποτέλεσμα συσταδοποίησης αλγορίθμου Hierarchical Clusterer (classes to cluster evaluation) για k = 2

Στο τελικό στάδιο δοκιμάζουμε να εκπαιδεύσουμε το ίδιο σύνολο δεδομένων με τον αλγόριθμο ομαδοποίησης J48 που ανήκει στους αλγορίθμους Δέντρων Απόφασης (Decision Tree) ώστε να εντοπίσουμε τους κανόνες ομαδοποίησης και να συγκρίνουμε την ακρίβεια των αποτελεσμάτων που προκύπτουν από τη σύγκριση μεταξύ των τριών αλγορίθμων καθώς θεωρείτε αποτελεσματικός στο να «κόβει» δεδομένα που μπορούν να χαρακτηριστούν ως outliers για την έρευνα μας. Στη συνέχεια στην επιλογή Test options επιλέγουμε “Cross – validation” και εφαρμόζουμε τον αλγόριθμο J48. Διαπιστώνουμε τα εξής όπως φαίνεται και στην εικόνα 33 παρακάτω, ο χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,07 sec. άρα προκύπτει ο επόμενος αποτελεσματικότερος από άποψη χρόνου μετά τον K – Means και το ποσοστό ακρίβειας ανέρχεται σε 63.4473% εμφανώς μεγαλύτερο συγκριτικά με τους K – Means και Hierarchical Clusterer. Τέλος, και το ποσοστό σφαλμάτων (Root mean squared error) εμφανίζεται σημαντικά μικρότερο σε σχέση με τον K – Means και ανέρχεται σε 0.0301.



Εικόνα 33: Αποτέλεσμα ομαδοποίησης αλγορίθμου J48 (cross - validation)

Όπως φαίνεται από τον παρακάτω πίνακα 4, ο αλγόριθμος Simple K - Means αποδίδει καλύτερους χρόνους εκτέλεσης ενώ παρατηρείται ότι η ιεραρχική συσταδοποίηση δεν είναι τόσο αποτελεσματική στην υπολογιστική πολυπλοκότητα αφού έχει τον υψηλότερο χρόνο εκτέλεσης ενώ επίσης το ποσοστό ακρίβειας του J48 είναι σημαντικά καλύτερο σε σχέση με τον K – Means και τον Ιεραρχικό αλγόριθμο.

Αλγόριθμος	Υπολογιστικός χρόνος	Ακρίβεια
Simple K – Means k = 2	0,01 sec.	2,9718%
Simple K – Means k = 4	0 sec.	4,6062%
Hierarchical k = 2	2,14 sec.	1,6345%
Hierarchical k = 4	1,63 sec.	1,9316%
J48	0,07 sec.	63,4473%

Πίνακας 4: Ποσοστό ακρίβειας για κάθε αλγόριθμο.

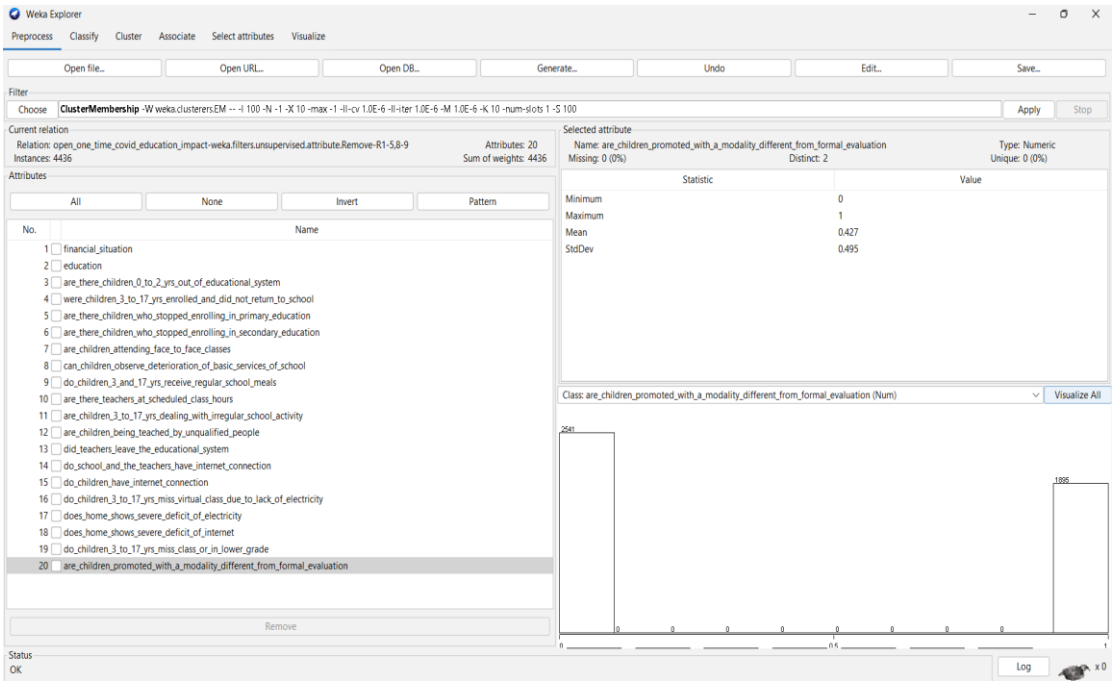
Συμπερασματικά, μπορούμε να πούμε ότι από την εφαρμογή δύο διαφορετικών αλγορίθμων σε δύο διαφορετικές λειτουργίες συσταδοποίησης (cluster mode) και όπως αποτυπώνεται από τους παραπάνω πίνακες ο αλγόριθμος Simple K – Means μοιάζει να είναι καταλληλότερος και να εφαρμόζει πιο ρεαλιστική κατανομή στιγμιότυπων στις επιμέρους συστάδες, παρόλα αυτά η ακρίβεια παραμένει εξαιρετικά χαμηλή γεγονός που υποδηλώνει πολυπλοκότητα των δεδομένων ενώ παρατηρείται ότι για k = 4 στην περίπτωση εκπαίδευσης του μοντέλου στη λειτουργία “Classes to cluster” και για τους δύο αλγορίθμους εμφανίζεται μεγαλύτερη ακρίβεια σε συνάρτηση με μικρότερη υπολογιστική πολυπλοκότητα. Εν τέλει και σε αυτό το πείραμα ο J48 εμφανίζει πολύ μεγαλύτερη ακρίβεια και δημιουργεί ένα πιο εύρωστο μοντέλο σε σύγκριση με τους άλλους δύο αλγορίθμους και όλες τις υλοποιήσεις που εφαρμόστηκαν.

4.5.3 3ο πείραμα: Έρευνα για τον αντίκτυπο της Covid – 19 στην εκπαίδευση

Στη συγκεκριμένη υλοποίηση επιλέχθηκε από μία ομάδα συνόλων δεδομένων με όνομα “COVID-19 Pandemic” το σύνολο δεδομένων με όνομα open_one_time_covid_education_impact.csv. Το σύνολο των δεδομένων αναφέρεται σε δεδομένα του πραγματικού κόσμου και συλλέχθηκε μέσω ερωτηματολογίου από μία εφαρμογή για κινητές συσκευές σε πολίτες της Βενεζουέλας και στοχεύει να αποτυπώσει τον αντίκτυπο της COVID - 19 στην εκπαίδευση των παιδιών. Παρουσιάζει δημογραφικά, κοινωνικά, γεωγραφικά και εκπαιδευτικά χαρακτηριστικά μαθητών.

Συνολικά αποτελείται από 4435 στιγμιότυπα (instances) και 27 χαρακτηριστικά (attributes). Τα οποία είναι τα παρακάτω: submission_id, submission_date, gender, age, geography, financial_situation, education, employment_status, submission_state, are_there_children_0_to_2_yrs_out_of_educational_system,
were_children_3_to_17_yrs_enrolled_and_did_not_return_to_school,
are_there_children_who_stopped_enrolling_in_primary_education,
are_there_children_who_stopped_enrolling_in_secondary_education,
are_children_attending_face_to_face_classes,
can_children_observe_deterioration_of_basic_services_of_school,
do_children_3_and_17_yrs_receive_regular_school_meals,
are_there_teachers_at_scheduled_class_hours,

are_children_3_to_17_yrs_dealing_with_irregular_school_activity,
are_children_being_teached_by_unqualified_people,
did_teachers_leave_the_educational_system,
do_school_and_the_teachers_have_internet_connection,
do_children_have_internet_connection,
do_children_3_to_17_yrs_miss_virtual_class_due_to_lack_of_electricity,
does_home_shows_severe_deficit_of_electricity,
does_home_shows_severe_deficit_of_internet,
do_children_3_to_17_yrs_miss_class_or_in_lower_grade,
are_children_promoted_with_a_modality_different_from_formal_evaluation. Το τελευταίο χαρακτηριστικό είναι προσημασμένο ως κλάση. Στην φάση της προεπεξεργασίας κρίθηκε θεμιτό να απορριφθούν κάποια χαρακτηριστικά που θεωρήθηκε ότι δεν εισφέρουν προστιθέμενη αξία στο ζητούμενο της έρευνας που αφορά τον αντίκτυπο της πανδημίας στην εκπαίδευση των παιδιών και αυτά είναι τα: submission_id, submission_date, gender, age, geography, employment_status, submission_state, επομένως συνολικά έχουμε 20 χαρακτηριστικα (attributes). Στην εικόνα 34 παρακάτω παρουσιάζεται το σύνολο δεδομένων του αρχείου open_one_time_covid_education_impact.arff που αποτελείται από αριθμητικά και κατηγορικά δεδομένα.

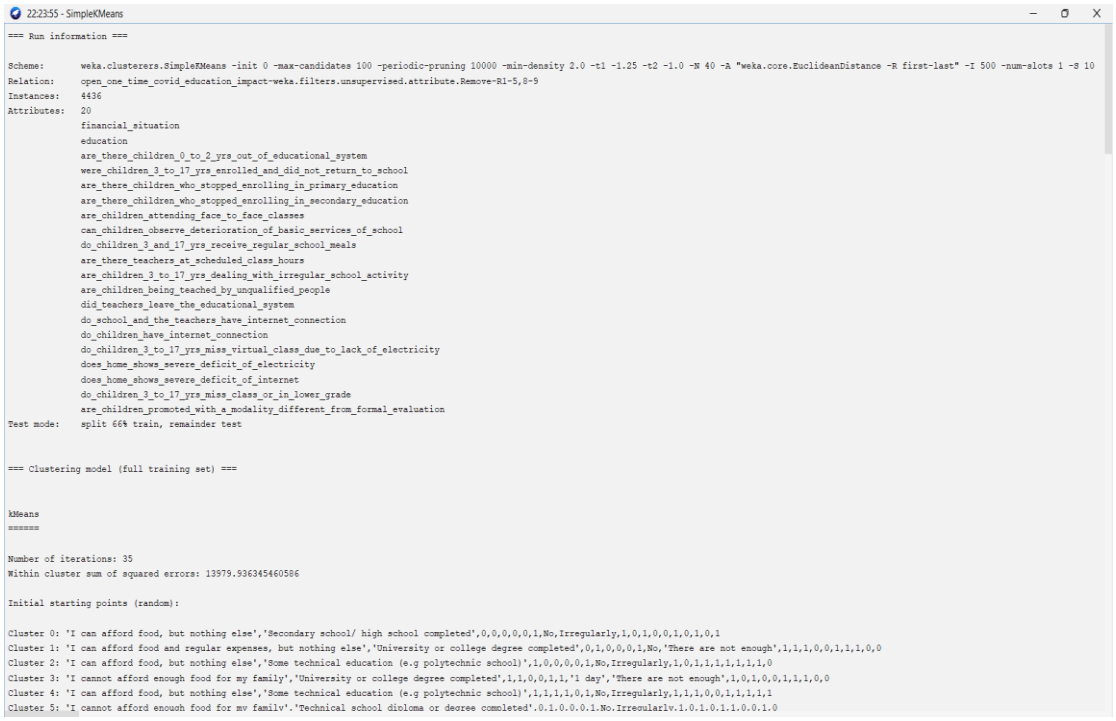


Εικόνα 34: Προεπεξεργασία του αρχείου εισόδου open_one_time_covid_education_impact.arff

Αρχικά, ξεκινάμε το πείραμα με εφαρμογή του αλγορίθμου Simple K-means με επιλεγμένο cluster mode “use training set” χωρίς να ορίσουμε μία συγκεκριμένη κλάση, καθώς σκοπός μας είναι να εξετάσουμε ένα περιγραφικό μοντέλο ώστε να γίνει καλύτερα κατανοητό το ζητούμενο της έρευνας. Δημιουργείται μια λίστα με τον αριθμό των στιγμιότυπων που έχουν εκχωρηθεί σε κάθε συστάδα σε ποσοστό επι της εκατό, όπως φαίνεται στην εικόνα 35. Ως αριθμό k συστάδων ορίσαμε διαδοχικά τις παρακάτω τιμές 2, 5, 10, 20, 30 και 40 ενώ από την επιλογή distanceFunction ορίσαμε ως συνάρτηση απόστασης την Ευκλείδεια απόσταση (EuclideanDistance). Επίσης, καταγράφεται ο αριθμός των επαναλήψεων όπου ανέρχεται για k = 2 σε 11, για k = 5 σε 35, για k = 10 σε 61, για k = 20 σε 52, για k = 30 σε 46 και για k = 40 σε 58, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) παρατηρείται ότι μειώνεται σταδιακά ξεκινώντας για k = 2 σε 21358.81, για k = 5 σε 18657.50, για k = 10 σε 17208.94, για k = 20 σε 15452.28, για k = 30 σε 14592.67 και για k = 40 σε 13995.46 ανάλογα, ο χρόνος κατασκευής του μοντέλου αυξάνεται και ανέρχεται για k = 2 σε 0.05 sec., για k = 5 σε 0.17 sec., για k = 10 σε 0.51 sec., για k = 20 σε 0.73 sec., για k = 30 σε 1.04 sec. και για k = 40 σε 1.29 sec. Τέλος, η κατανομή στιγμιότυπων ανά πλήθος συστάδων αποτυπώνεται στον πίνακα 5.

Εικόνα 35: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means για k = 40 και Euclidean Distance.

Στη συνέχεια, με τις ίδιες παραμέτρους δηλαδή ίδιο πλήθος συστάδων και Ευκλείδεια απόσταση δοκιμάζουμε να εκπαιδεύσουμε το μοντέλο μας επιλέγοντας ως cluster mode την επιλογή “Percentage split” ώστε να συγκρίνουμε την απόδοση του μοντέλου τόσο σε επίπεδο εκπαίδευσης όσο και σε επίπεδο δοκιμής του. Υλοποιούμε το πείραμα με το προεπιλεγμένο ποσοστό διαχωρισμού 66% και αφήνουμε το υπόλοιπο 34% ως test set. Από το παράθυρο του αποτελέσματος συσταδοποίησης (clusterer output) παίρνουμε τις εξής πληροφορίες ο αριθμός των επαναλήψεων ανέρχεται για k = 2 σε 8, για k = 5 σε 32, για k = 10 σε 35, για k = 20 σε 31, για k = 30 σε 34 και για k = 40 σε 30, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) παρατηρείται ότι μειώνεται σταδιακά ξεκινώντας για k = 2 σε 14110.69, για k = 5 σε 18450.25, για k = 10 σε 11180.24, για k = 20 σε 10285.59, για k = 30 σε 9700.04 και για k = 40 σε 9317.94 ανάλογα, ο χρόνος κατασκευής του μοντέλου ανέρχεται για k = 2 σε 0.04 sec., για k = 5 σε 0.1 sec., για k = 10 σε 0.16 sec., για k = 20 σε 0.36 sec., για k = 30 σε 0.4 sec. και για k = 40 σε 0.53 sec. Τέλος, σε ότι αφορά την κατανομή των στιγμιότυπων ανά πλήθος συστάδων αποτυπώνεται στον πίνακα 5.



Εικόνα 36: : Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means σε cluster mode "Percentage split"

Όπως φαίνεται από τον παρακάτω πίνακα 5, και στις δύο λειτουργίες συσταδοποίησης (cluster mode) έχουμε ομαλή κατανομή ποσοστών συσταδοποίησης μεταξύ των συστάδων του συνόλου γεγονός που υποδηλώνει ότι τα κέντρα των συστάδων δεν μεταβάλλονται σημαντικά, επιπλέον παρατηρείται ότι στη συγκεκριμένη λειτουργία το μεγαλύτερο ποσοστό συγκέντρωσης στοιχείων απαντάτε μεταξύ των συστάδων C0 ή C1 αναλογικά και του πλήθους συστάδων, ενώ παρατηρούμε ότι όσο αυξάνουμε των αριθμό συστάδων τόσο μειώνεται η πυκνότητα. Τέλος, στη λειτουργία "Percentage Split" έχουμε εμφανώς καλύτερους χρόνους εκτέλεσης αλλά και λιγότερες επαναλήψεις όπως έχει παρουσιαστεί και παραπάνω.

Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων
Use training set	0,05 sec.	0 1 48% 52%
Percentage Split	0,03 sec.	65% 35%
Use training set	0,17 sec.	0 1 2 3 4 20% 20% 23% 17% 20%
Percentage Split	0,1 sec.	25% 22% 19% 20% 15%
Use training set	0,51 sec.	0 1 2 3 4 5 6 7 8 9 8% 11% 10% 11% 13% 6% 11% 10% 15% 6%
Percentage Split	0,16 sec.	16% 16% 9% 9% 6% 10% 9% 9% 9% 7%
Use training set	0,73 sec.	0 1 2 3 4 5 6 7 8 9 4% 5% 5% 4% 7% 4% 3% 7% 8% 4% 10 11 12 13 14 15 16 17 18 19 6% 3% 5% 6% 4% 5% 6% 5% 4% 4%
Percentage Split	0,36 sec.	0 1 2 3 4 5 6 7 8 9 5% 13% 6% 6% 3% 5% 4% 6% 5% 3% 10 11 12 13 14 15 16 17 18 19 5% 5% 3% 6% 4% 5% 6% 4% 4% 5%
Use training set	1,04 sec.	0 1 2 3 4 5 6 7 8 9 3% 2% 5% 3% 3% 3% 3% 5% 6% 3% 10 11 12 13 14 15 16 17 18 19 4% 2% 2% 4% 3% 2% 4% 4% 3% 4%

Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων									
		20	21	22	23	24	25	26	27	28	29
		3%	3%	2%	3%	3%	3%	3%	3%	4%	2%
Percentage Split	0,36 sec.	0	1	2	3	4	5	6	7	8	9
		4%	9%	3%	4%	2%	3%	2%	6%	3%	3%
		10	11	12	13	14	15	16	17	18	19
		2%	4%	3%	4%	4%	5%	3%	2%	3%	3%
		20	21	22	23	24	25	26	27	28	29
		3%	2%	2%	4%	2%	3%	2%	4%	5%	3%
Use training set	1,29 sec.	0	1	2	3	4	5	6	7	8	9
		2%	2%	3%	2%	3%	1%	2%	4%	5%	2%
		10	11	12	13	14	15	16	17	18	19
		4%	1%	2%	4%	2%	3%	3%	3%	2%	3%
		20	21	22	23	24	25	26	27	28	29
		3%	3%	2%	2%	2%	2%	3%	2%	3%	2%
		30	31	32	33	34	35	36	37	38	39
		3%	3%	1%	4%	1%	3%	3%	2%	2%	2%
Percentage Split	0,53 sec.	0	1	2	3	4	5	6	7	8	9
		3%	8%	3%	2%	3%	3%	2%	2%	3%	2%
		10	11	12	13	14	15	16	17	18	19
		2%	2%	2%	4%	3%	3%	3%	2%	2%	3%

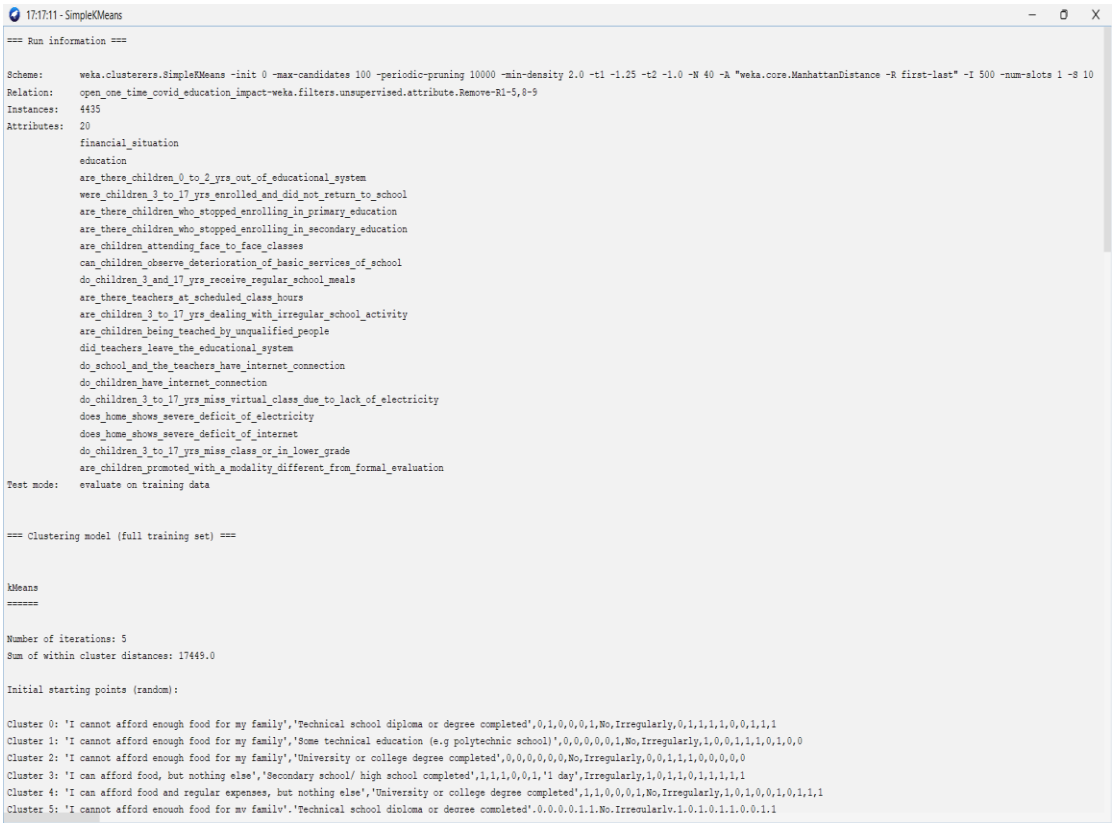
Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων									
		20	21	22	23	24	25	26	27	28	29
		3%	2%	2%	3%	1%	2%	2%	3%	2%	3%
		30	31	32	33	34	35	36	37	38	39
		2%	1%	2%	4%	3%	2%	2%	3%	0%	3%

Πίνακας 5: Χρόνοι εκτέλεσης και ποσοστά ομαδοποιημένων στιγμιότυπων για κάθε cluster mode.

Στο επόμενο στάδιο υλοποιούμε και πάλι το πείραμα μας χρησιμοποιώντας τον ίδιο αλγόριθμο Simple K – Means, στις δύο διαφορετικές λειτουργίες συσταδοποίησης “Use training set” και “Percentage split” και ορίζοντας των αριθμό k σταδιακά σε 2, 5 10, 20, 30 και 40 συστάδες. Στη συγκεκριμένη υλοποίηση θα τροποποιήσουμε την παράμετρο της συνάρτησης απόστασης από EuclideanDistance σε ManhattanDistance με σκοπό να διερευνήσουμε την ενδεχόμενη διαφοροποίηση που προκύπτει εξαιτίας του διαφορετικού τρόπου υπολογισμού της απόστασης.

Στην πρώτη εφαρμογή έχουμε ορίσει τις εξής παραμέτρους, cluster mode: “Use training set”, distanceFunction: ManhattanDistance και ξεκινάμε με k = 2 και έπειτα συνεχίζουμε με k = 5, k = 10, k = 20, k = 30 και k = 40. Τα αποτελέσματα που προκύπτουν έχουν ως εξής, ο αριθμός των επαναλήψεων ανέρχεται για k = 2 σε 6, για k = 5 σε 4, για k = 10 σε 5, για k = 20 σε 6, για k = 30 σε 5 και για k = 40 σε 5, οι οποίες είναι αρκετά λιγότερες συγκριτικά με την προηγούμενη υλοποίηση. Το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) παρατηρείται αυξημένο ξεκινώντας για k = 2 σε 27266.0, για k = 5 σε 23090.0, για k = 10 σε 20999.0, για k = 20 σε 19284.0, για k = 30 σε 18235.0 και για k = 40 σε 17449.0 ανάλογα, ο χρόνος

κατασκευής του μοντέλου ανέρχεται για k = 2 σε 0.13 sec., για k = 5 σε 0.07 sec., για k = 10 σε 0.09 sec., για k = 20 σε 0.14 sec., για k = 30 σε 0.22 sec. και για k = 40 σε 0.2 sec. Τέλος, παρατηρούμε την κατανομή στιγμιότυπων ανά πλήθος συστάδων στον πίνακα 6. Παρακάτω στην εικόνα 37 μπορούμε να παρατηρήσουμε κάποια από τα αποτελέσματα.



Εικόνα 37: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means για k = 40 και Manhattan Distance.

Στη συνέχεια, με τις ίδιες παραμέτρους δηλαδή ίδιο πλήθος συστάδων και Manhattan Distance δοκιμάζουμε να εκπαιδεύσουμε το μοντέλο μας επιλέγοντας ως cluster mode την επιλογή “Percentage split” ώστε να συγκρίνουμε την απόδοση του μοντέλου τόσο σε επίπεδο εκπαίδευσης όσο και σε επίπεδο δοκιμής του. Υλοποιούμε το πείραμα με το προεπιλεγμένο ποσοστό διαχωρισμού 66% και αφήνουμε το υπόλοιπο 34% ως test set. Από το παράθυρο του αποτελέσματος συσταδοποίησης (clusterer output) παίρνουμε τις εξής πληροφορίες ο αριθμός των επαναλήψεων ανέρχεται για k = 2 σε 3, για k = 5 σε 5, για k = 10 σε 3, για k = 20 σε 3, για k = 30 σε 5 και για k = 40 σε 7, το ποσοστό σφαλμάτων (μέσος όρος σφαλμάτων στο τετράγωνο) παρατηρείται ότι μειώνεται σταδιακά ξεκινώντας για k = 2 σε 18024.0, για k = 5 σε 15272.0, για k = 10

σε 13939.0, για k = 20 σε 12780.0, για k = 30 σε 11956.0 και για k = 40 σε 11501.0 ανάλογα, ο χρόνος κατασκευής του μοντέλου ανέρχεται για k = 2 σε 0.02 sec., για k = 5 σε 0.03 sec., για k = 10 σε 0.03 sec., για k = 20 σε 0.04 sec., για k = 30 σε 0.11 sec. και για k = 40 σε 0.33 sec. Τέλος, παρατηρούμε την κατανομή στιγμιότυπων ανά πλήθος συστάδων στον πίνακα 6.

Εικόνα 38: Αποτέλεσμα συσταδοποίησης αλγορίθμου Simple K - Means σε cluster mode "Percentage split" και Manhattan Distance

Όπως φαίνεται από τον παρακάτω πίνακα 6, και στις δύο λειτουργίες συσταδοποίησης (cluster mode) έχουμε λιγότερο ομαλή κατανομή ποσοστών συσταδοποίησης μεταξύ των συστάδων του συνόλου σε σχέση με την Ευκλείδεια απόσταση, επιπλέον παρατηρείται ότι στη λειτουργία "Percentage split" το μεγαλύτερο ποσοστό συγκέντρωσης στοιχείων απαντάτε μεταξύ των συστάδων C0 ή C1 αναλογικά και του πλήθους συστάδων κάτι που στη συγκεκριμένη περίπτωση δεν ισχύει και για τη λειτουργία «Use training set", ενώ παρατηρούμε ότι όταν ορίζουμε μεγάλο αριθμό συστάδων k = 30 και k = 40 αρχίζουν να εμφανίζονται μηδενικές κατανομές σε συστάδες. Τέλος, στη λειτουργία "Percentage Split" έχουμε εμφανώς καλύτερους χρόνους εκτέλεσης αλλά και λιγότερες επαναλήψεις που συνεπάγονται όμως αύξηση του ποσοστού σφαλμάτων όπως έχει παρουσιαστεί και παραπάνω.

Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων
Use training set	0,05 sec.	0 1 70% 30%
Percentage Split	0,03 sec.	66% 34%
Use training set	0,17 sec.	0 1 2 3 4 24% 21% 21% 11% 22%
Percentage Split	0,1 sec.	37% 18% 28% 5% 12%
Use training set	0,51 sec.	0 1 2 3 4 5 6 7 8 9 12% 13% 13% 9% 19% 3% 3% 10% 13% 5%
Percentage Split	0,16 sec.	24% 18% 15% 3% 2% 12% 5% 6% 4% 10%
Use training set	0,73 sec.	0 1 2 3 4 5 6 7 8 9 9% 10% 10% 7% 9% 2% 3% 6% 11% 5% 10 11 12 13 14 15 16 17 18 19 4% 1% 2% 7% 2% 1% 4% 4% 1% 2%
Percentage Split	0,36 sec.	0 1 2 3 4 5 6 7 8 9 13% 16% 8% 3% 2% 6% 3% 5% 3% 7% 10 11 12 13 14 15 16 17 18 19 2% 4% 2% 4% 3% 6% 5% 2% 3% 3%
Use training set	1,04 sec.	0 1 2 3 4 5 6 7 8 9 6% 7% 11% 5% 4% 2% 2% 7% 11% 4% 10 11 12 13 14 15 16 17 18 19 2% 2% 2% 6% 2% 1% 3% 4% 1% 2%

Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων
		20 21 22 23 24 25 26 27 28 29
		3% 1% 1% 1% 3% 2% 2% 2% 1% 1%
Percentage Split	0,36 sec.	0 1 2 3 4 5 6 7 8 9
		9% 13% 8% 2% 2% 6% 3% 4% 2% 6%
		10 11 12 13 14 15 16 17 18 19
		1% 3% 1% 3% 3% 5% 5% 2% 3% 3%
		20 21 22 23 24 25 26 27 28 29
		3% 0% 1% 3% 2% 1% 0% 1% 3% 2%
Use training set	1,29 sec.	0 1 2 3 4 5 6 7 8 9
		5% 6% 5% 2% 1% 2% 1% 5% 10% 4%
		10 11 12 13 14 15 16 17 18 19
		3% 1% 2% 6% 2% 1% 3% 3% 1% 2%
		20 21 22 23 24 25 26 27 28 29
		3% 1% 1% 1% 2% 1% 2% 2% 1% 1%
		30 31 32 33 34 35 36 37 38 39
		2% 1% 1% 6% 0% 1% 3% 1% 2% 1%
Percentage Split	0,53 sec.	0 1 2 3 4 5 6 7 8 9
		8% 11% 6% 1% 1% 4% 2% 3% 2% 5%
		10 11 12 13 14 15 16 17 18 19

Cluster mode	Υπολογιστικός χρόνος	Ποσοστό ομαδοποιημένων στιγμιότυπων									
		1%	4%	1%	3%	4%	4%	5%	2%	3%	3%
		20	21	22	23	24	25	26	27	28	29
		3%	0%	1%	3%	0%	1%	1%	1%	3%	2%
		30	31	32	33	34	35	36	37	38	39
		1%	1%	2%	2%	3%	2%	1%	0%	0%	1%

Πίνακας 6: Χρόνοι εκτέλεσης και ποσοστά ομαδοποιημένων στιγμιότυπων για κάθε cluster mode.

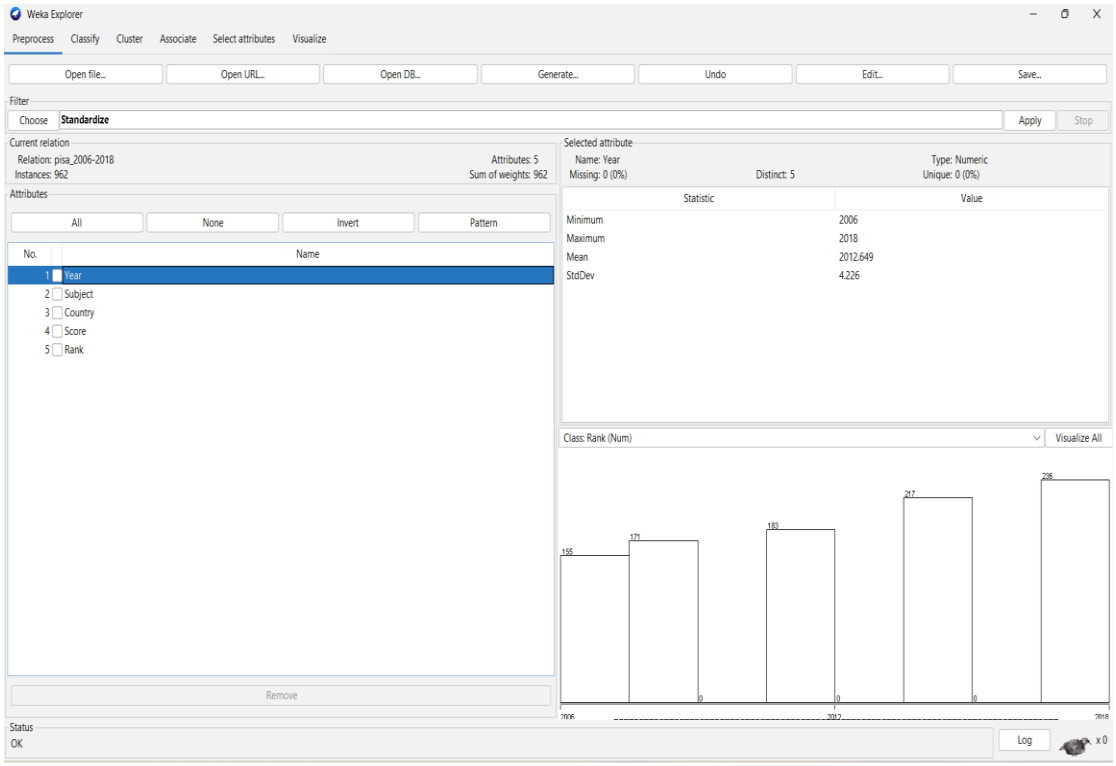
Συμπερασματικά, μπορούμε να πούμε ότι από την εφαρμογή του αλγορίθμου Simple K – Means σε δύο διαφορετικές λειτουργίες συσταδοποίησης (cluster mode) σε επίπεδο εκπαίδευσης (Use training test) και σε επίπεδο δοκιμής (Percentage split) με διαφορετικές μετρικές απόστασης τόσο σε Euclidean Distance όσο και σε Manhattan Distance αλλά και για διαφορετικό αριθμό συστάδων, παρότι η Ευκλείδεια απόσταση παρουσιάζει μεγαλύτερη υπολογιστική πολυπλοκότητα από άποψη χρόνου και επαναλήψεων μοιάζει να δημιουργεί ένα πιο εύρωστο μοντέλο με λιγότερα outliers όπως φαίνεται να συμβαίνει στην περίπτωση της απόστασης Μανχάταν σε λειτουργία "Percentage split" όπου βλέπουμε ότι εμφανίζονται συστάδες με ποσοστά ομαδοποίησης 0%.

Τέλος, διαπιστώνεται από τις συστάδες που συγκεντρώνουν τα υψηλότερα ποσοστά ομαδοποιήσεων ότι οι σημαντικότεροι παράγοντες επίδρασης εξαιτίας της πανδημίας αφορούν παιδιά ηλικίας από 3 έως 17 χρονών που έχουν γραφτεί στο σχολείο αλλά δε επέστρεψαν για παρακολούθηση, ότι τα παιδιά έχουν παρατηρήσει υποβάθμιση των βασικών υπηρεσιών του σχολείου, υπάρχει σχετική έλλειψη δασκάλων τις

προγραμματισμένες ώρες μαθημάτων, παιδιά με παραβατική συμπεριφορά μεταξύ 3 έως 17 ετών τείνουν να εγκαταλείπουν το σχολείο, υπάρχουν διαθέσιμοι εκπαιδευτικοί στο εκπαιδευτικό σύστημα και ότι δεν χάνονται διαδικτυακά μαθήματα λόγω διακοπής ρεύματος. Οι κοινωνικοί παράγοντες μοιάζουν να επηρεάζουν περισσότερο την εκπαίδευση των παιδιών παρά οι υγειονομικοί.

4.5.4 4ο πείραμα: Βαθμολογίες PISA 2006 – 2018

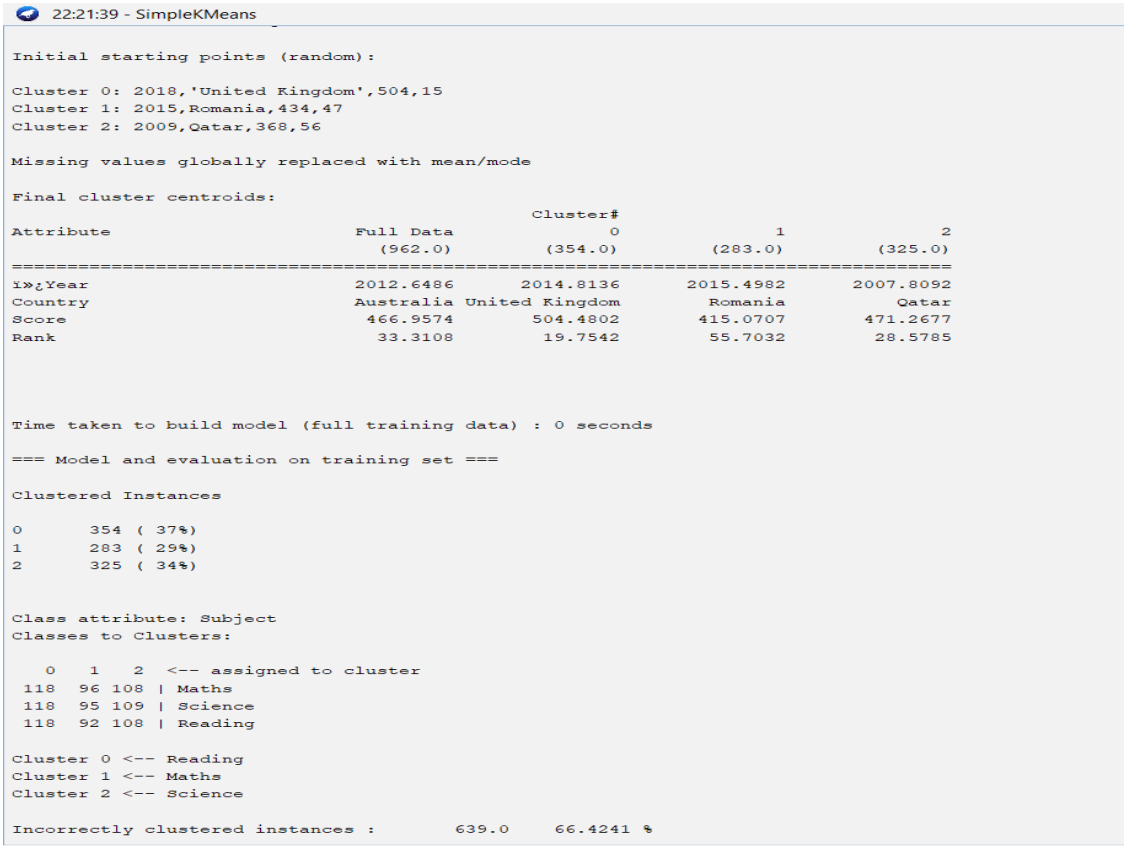
Στη συγκεκριμένη υλοποίηση χρησιμοποιήθηκε ένα σύνολο δεδομένων με όνομα pisa_2006-2018.csv. Το σύνολο δεδομένων αποτελείται από πραγματικά δεδομένα που έχουν συλλεχθεί μέσω του Προγράμματος Διεθνούς Αξιολόγησης Μαθητών (**P**rogramme for **I**nternational **S**tudent **A**ssessmen) και πραγματοποιείται ανά 3 χρόνια ενώ αποτελεί παγκόσμια μελέτη από τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ) σε κράτη μέλη και τρίτες χώρες με σκοπό την αξιολόγηση των εκπαιδευτικών συστημάτων με τη μέτρηση της σχολικής επίδοσης μαθητών 15 ετών. για τα μαθήματα των μαθηματικών, της επιστήμης και της ανάγνωσης. Συνολικά αποτελείται από 962 στιγμιότυπα (instances) και 5 χαρακτηριστικά (attributes) καθώς κατά την φάση της προεπεξεργασίας εξαλείφθηκαν κάποια κενά πεδία τιμών. Τα χαρακτηριστικά είναι τα παρακάτω: Year, Subject, Country, Score, Rank. Το τελευταίο χαρακτηριστικό είναι προσημασμένο ως κλάση. Στην εικόνα 39 παρακάτω παρουσιάζεται το σύνολο δεδομένων του αρχείο pisa_2006-2018.arff που αποτελείται από αριθμητικά και κατηγορικά δεδομένα.



Εικόνα 39: Προεπεξεργασία του αρχείου εισόδου pisa_2006-2018.arff

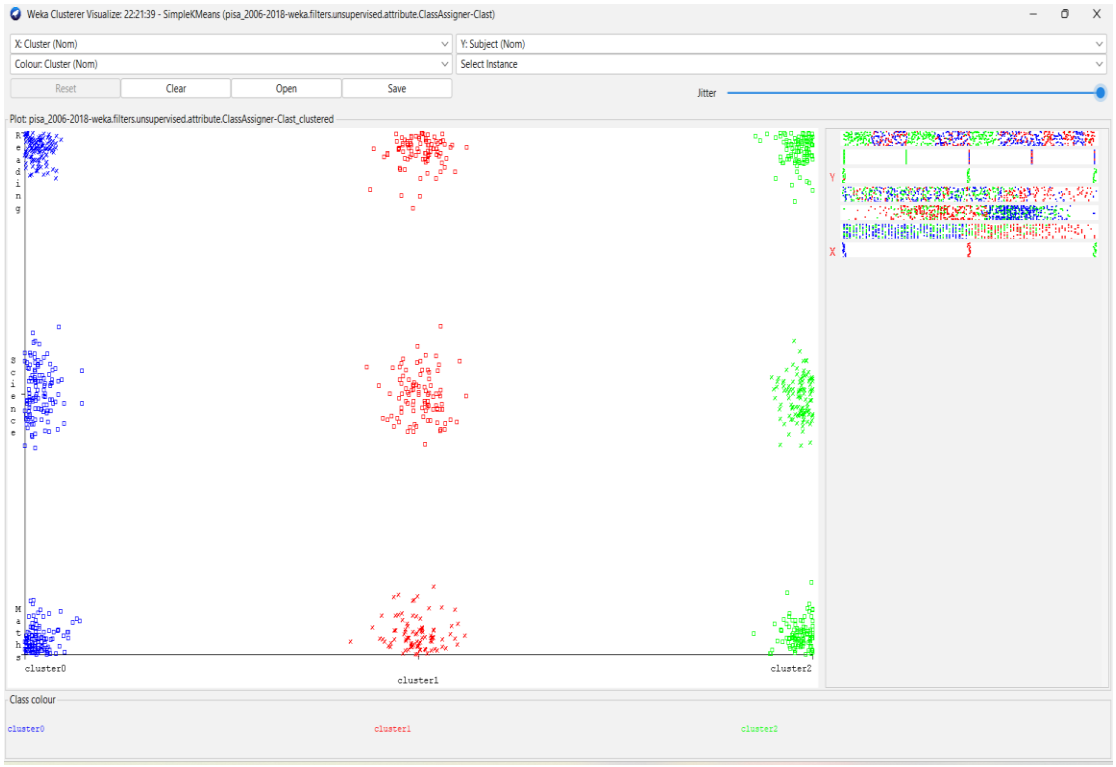
Στη συγκεκριμένη υλοποίηση θα επιχειρήσουμε να αξιολογήσουμε τα αποτελέσματα της εκπαίδευσης των αλγορίθμων συσταδοποίησης Simple K – Means, Make Density Based Clusterer, Farthest First, Expectation Maximazation και Filtered Clusterer. Θα εκτελέσουμε τα πειράματα μας σε δύο διαφορετικές λειτουργίες συσταδοποίησης “Use training set” και “Classes to clusters evaluation”. Τα αποτελέσματα αποτιμώνται ως προς των αριθμό των συστάδων, το ποσοστό συσταδοποίησης στιγμιότυπων για κάθε συστάδα, το άθροισμα τετραγωνισμένου σφάλματος των αποστάσεων των στοιχείων από το κέντρο της συστάδας (sum of squared errors), τον υπολογιστικό χρόνο κατασκευής του μοντέλου, τον αριθμό επαναλήψεων και για ορισμένους από τους αλγορίθμους τη λογαριθμική πιθανότητα (log likehood). Στην εικόνα 40 βλέπουμε το αποτέλεσμα συσταδοποίησης για τον αλγόριθμο Simple K - Means για k = 3 και λειτουργία συσταδοποίησης “Classes to clusters evaluation” με προσημασμένη κλάση το χαρακτηριστικό Subject ώστε να διαπιστώσουμε πόσο καλά ταιριάζουν οι συστάδες που σχηματίζονται στην επιλεγμένη κλάση. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 9, το άθροισμα τετραγωνισμένου σφάλματος σε 999.57, ο υπολογιστικός χρόνος κατασκευής του μοντέλου σε 0 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 37%, C1 = 29% και C3 = 34% και τέλος το ποσοστό ακρίβειας

ανέρχεται σε 33,5759% αρκετά χαμηλό για να θεωρηθεί εύρωστο και ακριβές το μοντέλο.



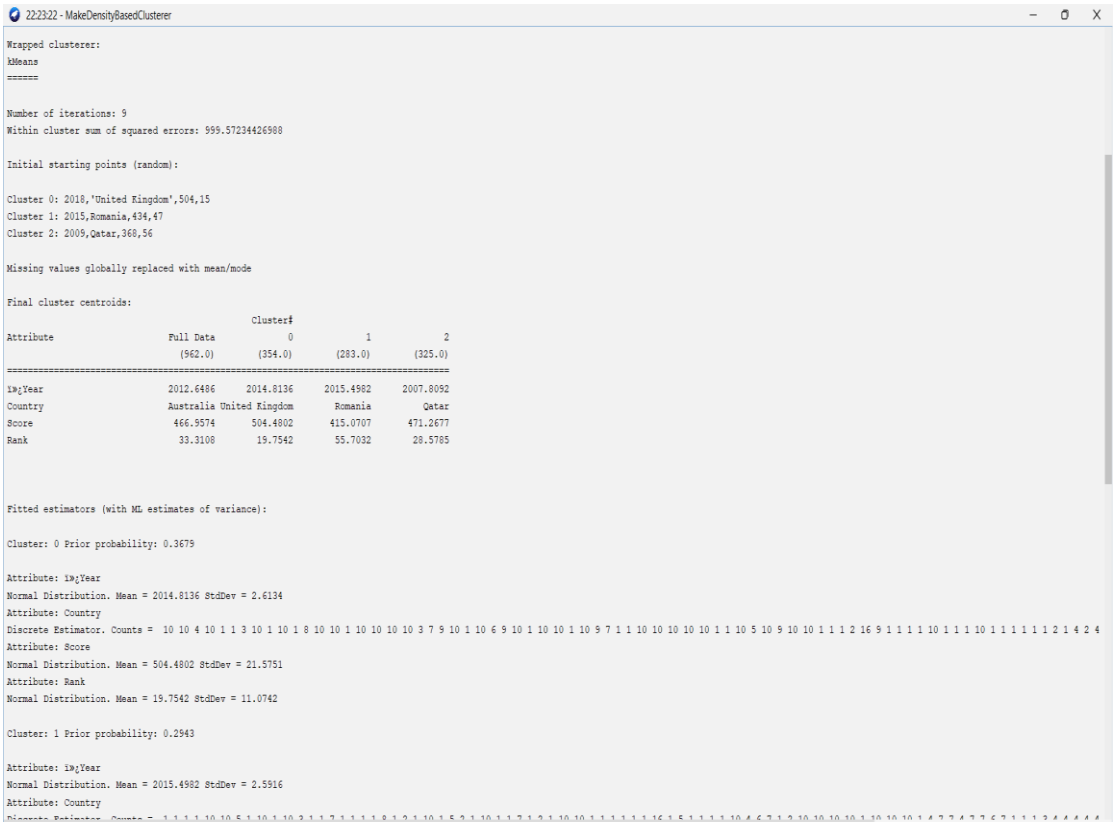
```
22:21:39 - SimpleKMeans

Initial starting points (random):

Cluster 0: 2018,'United Kingdom',504,15
Cluster 1: 2015,Romania,434,47
Cluster 2: 2009,Qatar,368,56

Missing values globally replaced with mean/mode

Final cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                            (962.0)       (354.0)       (283.0)       (325.0)
=============================================================================
ï»¿Year                  2012.6486     2014.8136     2015.4982     2007.8092
Country                 Australia United Kingdom       Romania         Qatar
Score                    466.9574      504.4802      415.0707      471.2677
Rank                      33.3108       19.7542       55.7032       28.5785



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       354 ( 37%)
1       283 ( 29%)
2       325 ( 34%)


Class attribute: Subject
Classes to Clusters:

   0   1   2  <-- assigned to cluster
 118  96 108 | Maths
 118  95 109 | Science
 118  92 108 | Reading

Cluster 0 <-- Reading
Cluster 1 <-- Maths
Cluster 2 <-- Science

Incorrectly clustered instances :       639.0     66.4241 %
```

Εικόνα 40: Αποτελέσματα συσταδοποίησης αλγορίθμου Simple K - Means για k = 3 σε λειτουργία "Classes to cluster evaluation”

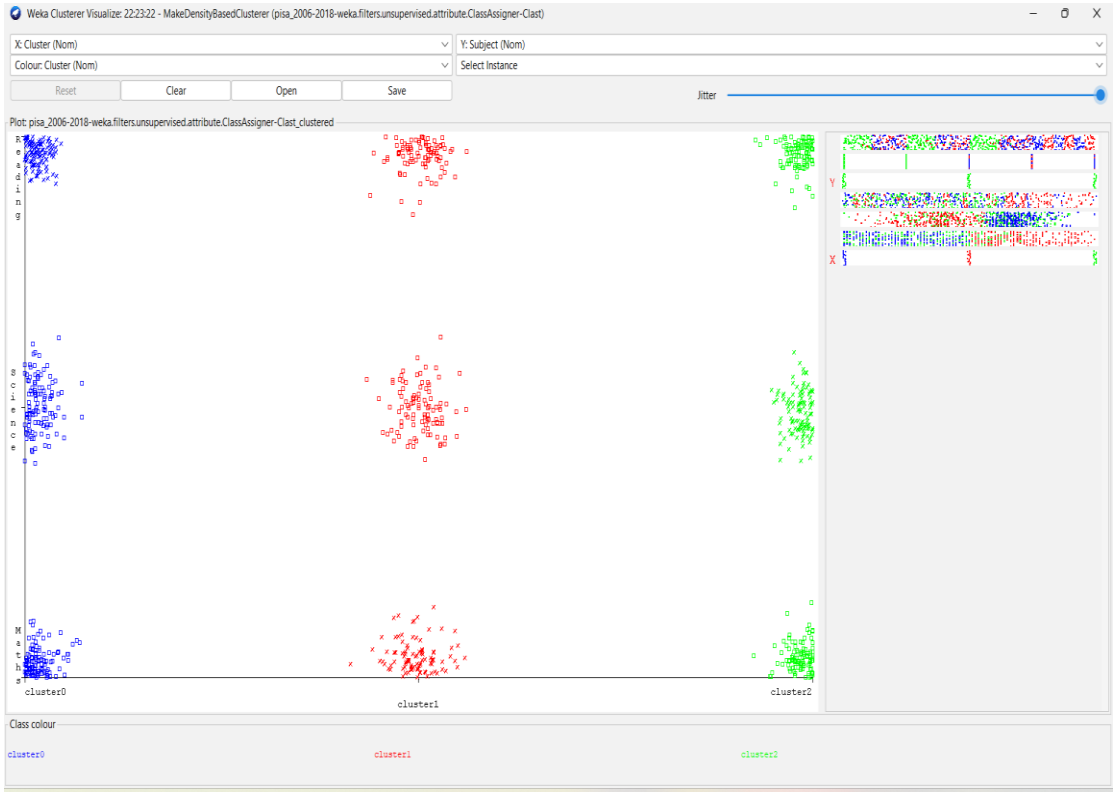


Εικόνα 41: Γραφική αναπαράσταση συστάδων αλγορίθμου Simple K - Means ανά θεματική ενότητα.

Στη συνέχεια, στην εικόνα 42 βλέπουμε το αποτέλεσμα συσταδοποίησης για τον αλγόριθμο Make Density Based Clusterer για k = 3 και λειτουργία συσταδοποίησης “Classes to clusters evaluation” με προσημασμένη κλάση το χαρακτηριστικό Subject. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 9, το άθροισμα τετραγωνισμένου σφάλματος σε 999.57, ο υπολογιστικός χρόνος κατασκευής του μοντέλου σε 0,01 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 36%, C1 = 30% και C3 = 33%, το ποσοστό ακρίβειας ανέρχεται σε 33,7838% και τέλος η λογαριθμική πιθανότητα κατανομής των δεδομένων στη σωστή ομάδα ανέρχεται σε -15.97207, αν αναλογιστούμε το χαμηλό ποσοστό ακρίβειας συνάρτηση της υψηλής λογαριθμικής πιθανότητας το μοντέλο μας δεν θεωρείτε ικανοποιητικά ακριβές και εύρωστο.

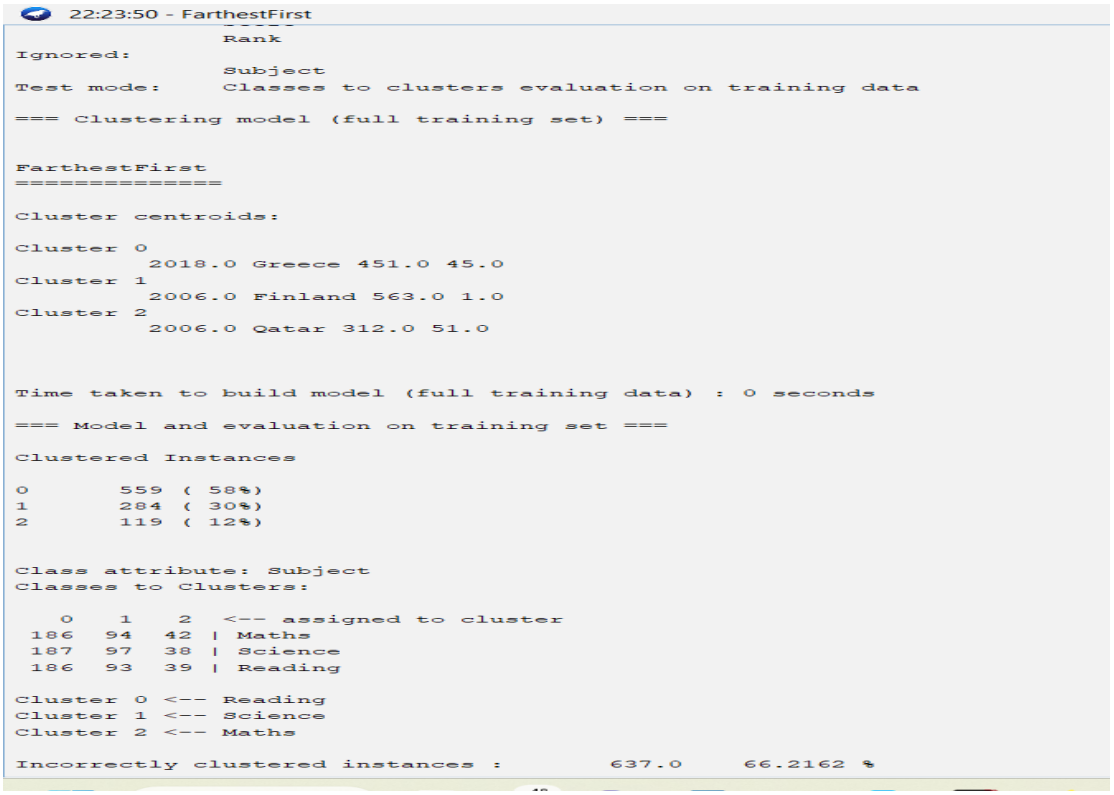


Εικόνα 42: Αποτελέσματα συσταδοποίησης αλγορίθμου Make Density Based Clusterer για k = 3 σε λειτουργία "Classes to cluster evaluation".

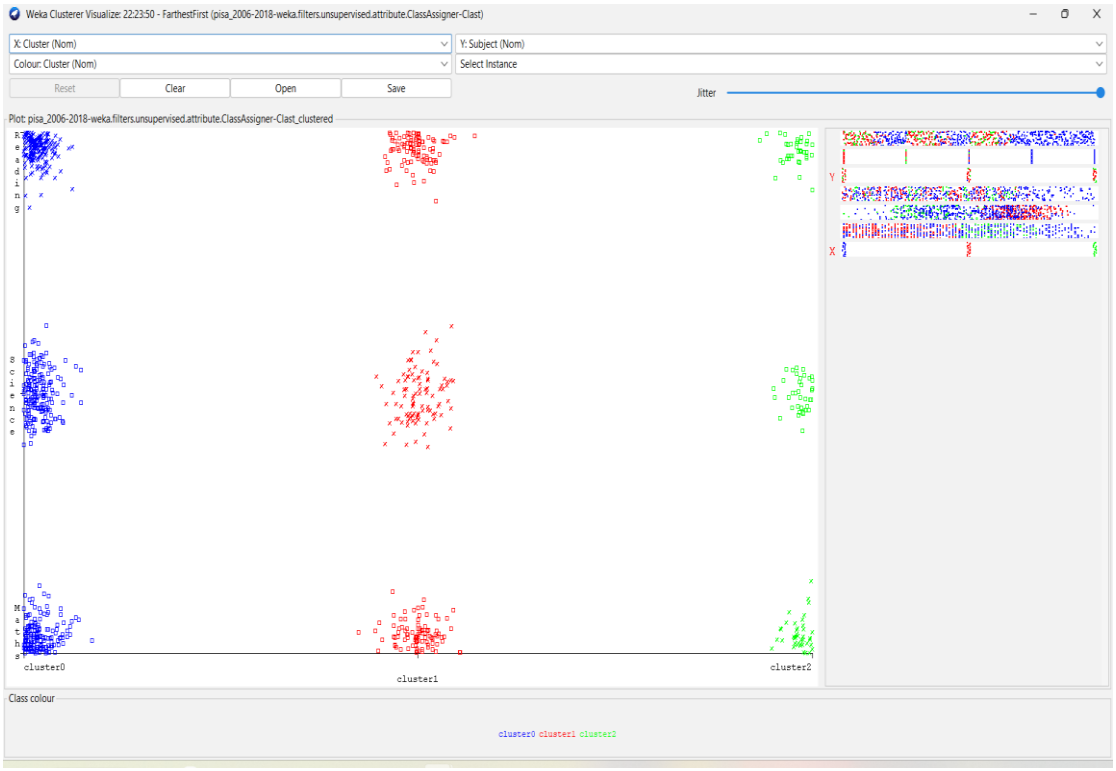


Εικόνα 43: Γραφική αναπαράσταση συστάδων αλγορίθμου Make Density Based ανά θεματική ενότητα.

Έπειτα, συνεχίζουμε με την εφαρμογή του αλγορίθμου Farthest First, με τις ίδιες παραμέτρους σταθερές, στην εικόνα 44 μπορούμε να δούμε τα αποτελέσματα συσταδοποίησης. Όπως φαίνεται ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 58%, C1 = 30% και C3 = 12% και το ποσοστό ακρίβειας ανέρχεται σε 33,7838%.



Εικόνα 44: Αποτελέσματα συσταδοποίησης αλγορίθμου Farthest First για k = 3 σε λειτουργία "Classes to cluster evaluation".



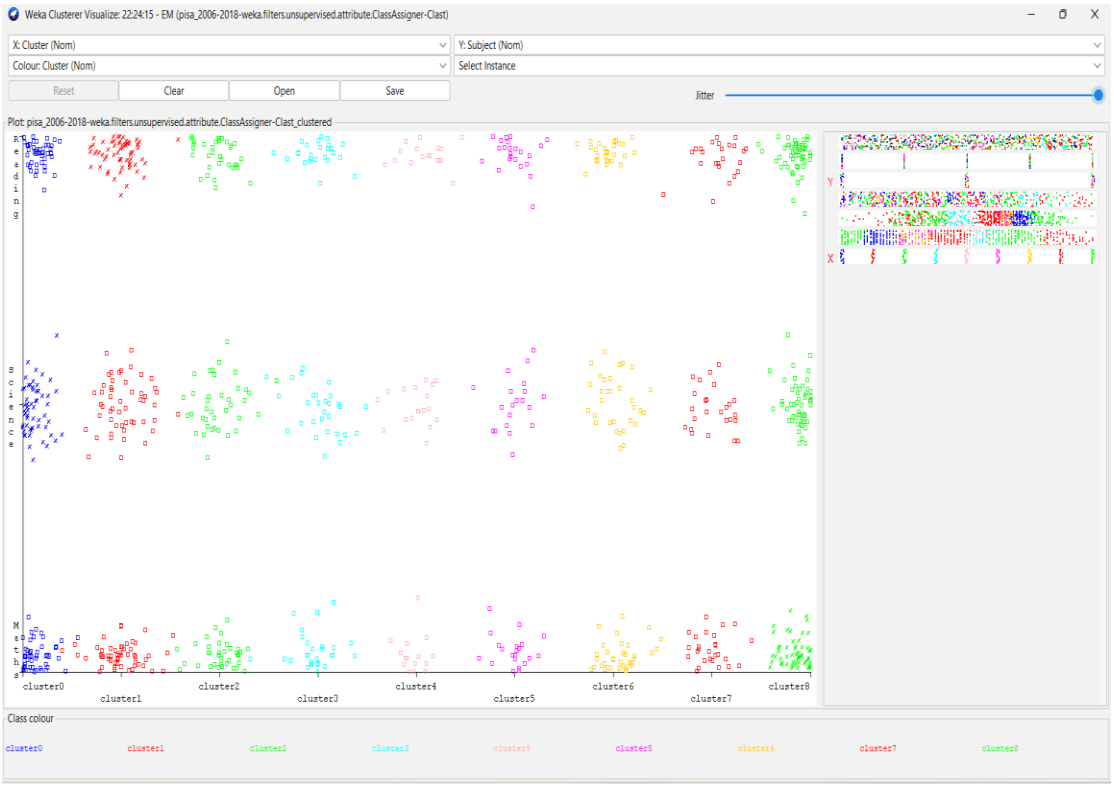
Εικόνα 45: Γραφική αναπαράσταση συστάδων αλγορίθμου Farthest First ανά θεματική ενότητα.

Ακολούθως, έχουμε την εφαρμογή του αλγορίθμου Expectation Maximization σε λειτουργία συσταδοποίησης “Classes to clusters evaluation” με τη διαφορά ότι στην συγκεκριμένη υλοποίηση ο τελικός αριθμός συστάδων προκύπτει αυτόματα μέσω της μεθόδου διασταυρούμενης επικύρωσης (cross validation) όπου έχουμε εκπαίδευση του μοντέλου διαδοχικά στις επιμέρους διαμερισματοποίησης εκτός μίας η οποία χρησιμοποιείται ως δοκιμαστικό σετ, τα αποτελέσματα του οποίου διακρίνουμε στην εικόνα 46. Όπως φαίνεται ο αριθμός των συστάδων που προκύπτουν ανέρχεται σε 9, ο αριθμός των επαναλήψεων ανέρχεται σε 76, ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 10.78 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 16%, C1 = 18% και C2 = 11%, C3 = 9%, C4 = 5%, C5 = 7%, C6 = 10%, C7 = 8% και C8 = 17%, το ποσοστό ακρίβειας ανέρχεται σε 18,1913% και τέλος η λογαριθμική πιθανότητα κατανομής των δεδομένων στη σωστή ομάδα ανέρχεται σε -14.28902, αν αναλογιστούμε το χαμηλό ποσοστό ακρίβειας συνάρτηση της υψηλής λογαριθμικής πιθανότητας το μοντέλο μας δεν αποτιμάται ικανοποιητικά ακριβές και εύρωστο.

22:24:15 - EM

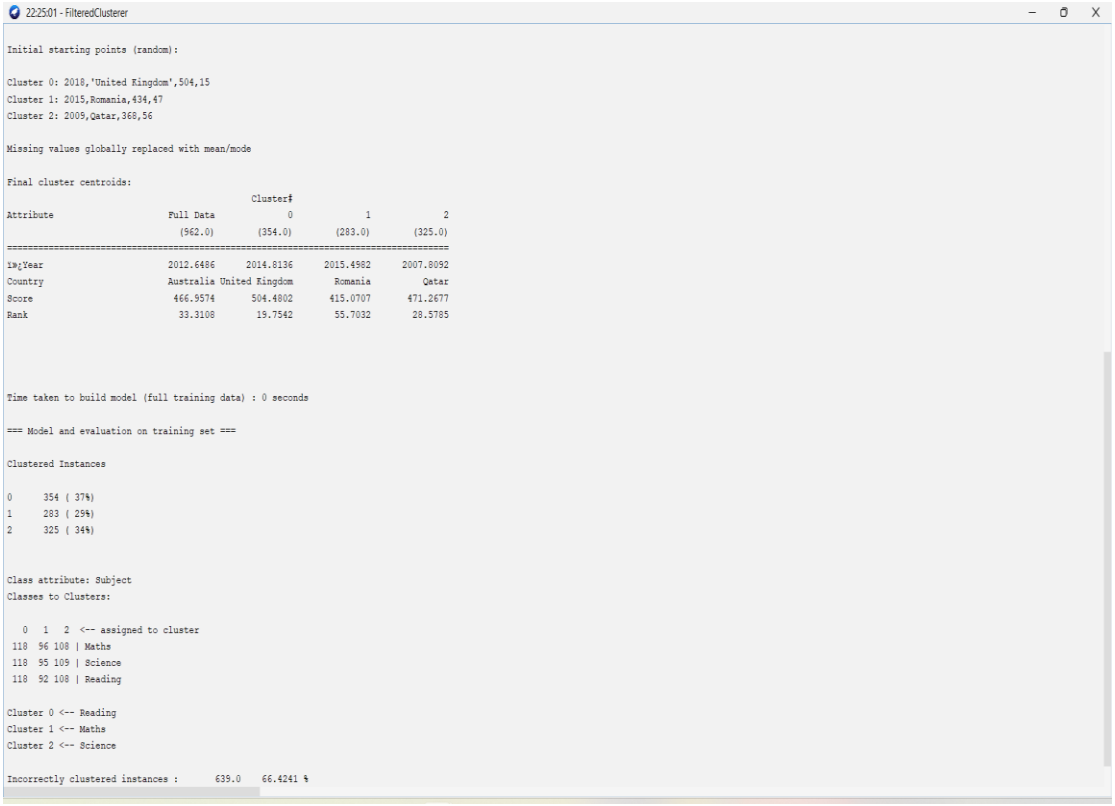
```
Rank
  mean                      12.9327   34.1821    4.2439   45.874   42.2444   22.8393   23.4803   70.1716   54.9554
  std. dev.                  3.2677    3.7136     2.245    3.0345   2.8161    3.8639    3.3065    4.3674     4.68


Time taken to build model (full training data) : 10.78 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       153 ( 16%)
1       171 ( 18%)
2       109 ( 11%)
3        83 (  9%)
4        51 (  5%)
5        64 (  7%)
6        94 ( 10%)
7        76 (  8%)
8       161 ( 17%)


Log likelihood: -14.28902


Class attribute: Subject
Classes to Clusters:

  0  1  2  3  4  5  6  7  8  <-- assigned to cluster
 50 54 37 29 16 21 34 27 54 | Maths
 56 52 38 29 18 20 31 25 53 | Science
 47 65 34 25 17 23 29 24 54 | Reading

Cluster 0 <-- Science
Cluster 1 <-- Reading
Cluster 2 <-- No class
Cluster 3 <-- No class
Cluster 4 <-- No class
Cluster 5 <-- No class
Cluster 6 <-- No class
Cluster 7 <-- No class
Cluster 8 <-- Maths

Incorrectly clustered instances :       787.0    81.8087 %
```

Εικόνα 46: Αποτελέσματα συσταδοποίησης αλγορίθμου Expectation Maximization σε λειτουργία "Classes to cluster evaluation".

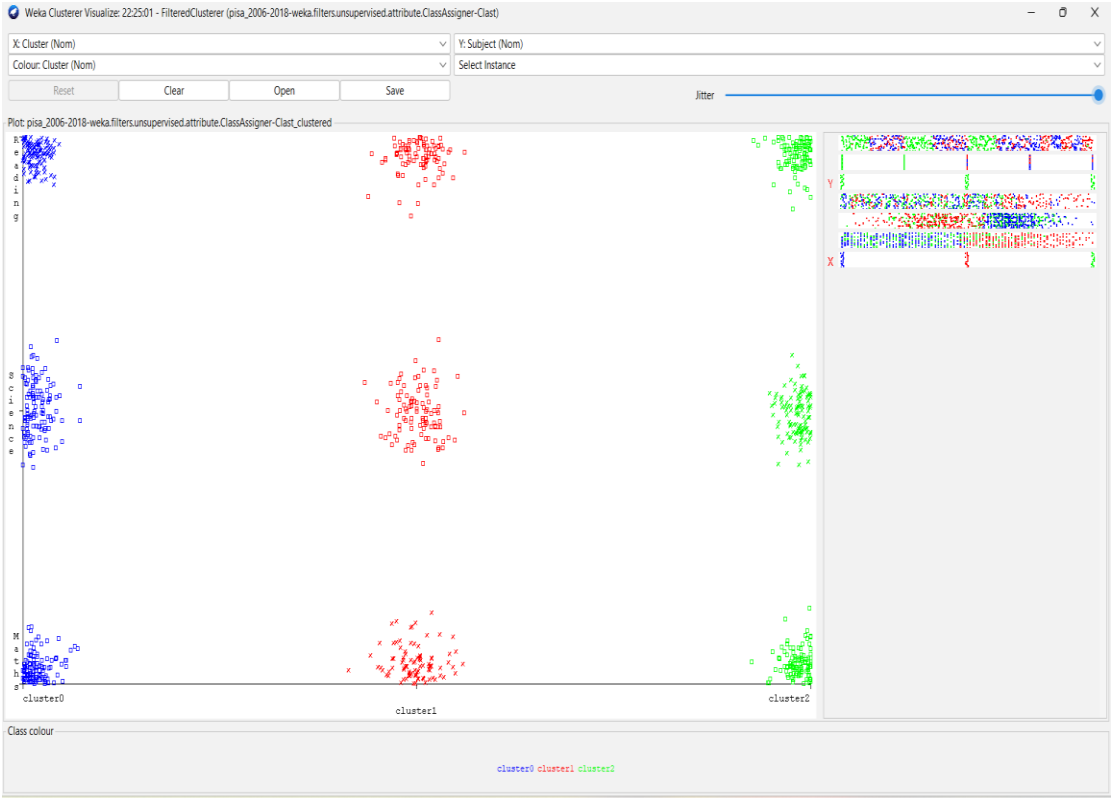


Εικόνα 47: Γραφική αναπαράσταση συστάδων αλγορίθμου Expectation Maximization ανά θεματική ενότητα.

Στην τελευταία υλοποίηση σε λειτουργία “Classes to clusters evaluation” εφαρμόζουμε τον αλγόριθμο Filtered Clusterer με αριθμό k = 3, τα αποτελέσματα παρουσιάζονται στην εικόνα 48. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 9, ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0 sec., το άθροισμα τετραγωνισμένου σφάλματος ανέρχεται σε 999.57, το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 37%, C1 = 29% και C2 = 34% και τέλος το ποσοστό ακρίβειας ανέρχεται σε 33,5759%.



Εικόνα 48: Αποτελέσματα συσταδοποίησης αλγορίθμου Filtered Clusterer για k = 3 σε λειτουργία "Classes to cluster evaluation”.



Εικόνα 49: Γραφική αναπαράσταση συστάδων αλγορίθμου Filtered clusterer ανά θεματική ενότητα.

Στον παρακάτω πίνακα 7 απεικονίζονται συγκεντρωτικά τα αποτελέσματα όλων των προηγούμενων αλγορίθμων. Όπως μπορούμε να διαπιστώσουμε οι αλγόριθμοι Simple K – Means, Make Density Based και Filtered εμφανίζουν ακριβώς την ίδια απόδοση αναφορικά με των αριθμό επαναλήψεων, το άθροισμα των τετραγωνικών σφαλμάτων και την ακρίβεια τους. Σε ότι αφορά τον υπολογιστικό χρόνο ο αλγόριθμος Expectation Maximization εμφανίζει το μεγαλύτερο χρόνο γεγονός που σχετίζεται με την υψηλή πολυπλοκότητα του. Σε ότι αφορά την ακρίβεια που χαρακτηρίζει τη διαδικασία συσταδοποίησης ο αλγόριθμος Farthest First εμφανίζει οριακά καλύτερη απόδοση αν και γενικά τα ποσοστά ακρίβειας χαρακτηρίζονται ως σχετικά χαμηλά. Συγκριτικά με τη λογαριθμική πιθανότητα των αλγορίθμων Make Density Based και EM διαπιστώνουμε ότι έχουμε αρνητικές τιμές γεγονός που υποδηλώνει ατέλειες στη συσταδοποίηση. Τέλος, σε ότι αφορά τα ποσοστά των ομαδοποιημένων στιγμιότυπων ανά συστάδα οι αλγόριθμοι K – Means, Make Density Based και Filtered εμφανίζουν παρόμοια κατανομή συστάδων. Συμπερασματικά μπορούμε να πούμε ότι ο αλγόριθμος Farthest First εμφανίζει τη μεγαλύτερη ακρίβεια στον βέλτιστο υπολογιστικό χρόνο.

	K - Means	Make Density Based	Farthest First	EM	Filtered
Συστάδες	3	3	3	9	3
Επαναλήψεις	9	9	-	76	9
SSE	999.57	999.57	-	-	999.57
Υπολογιστικός χρόνος	0 sec.	0,01 sec.	0 sec.	10,78 sec	0 sec.
Ακρίβεια	33,5759	33,5759	33,7838	18,1913	33,5759
logLikehood	-	-15.97207	-	-14.28902	-

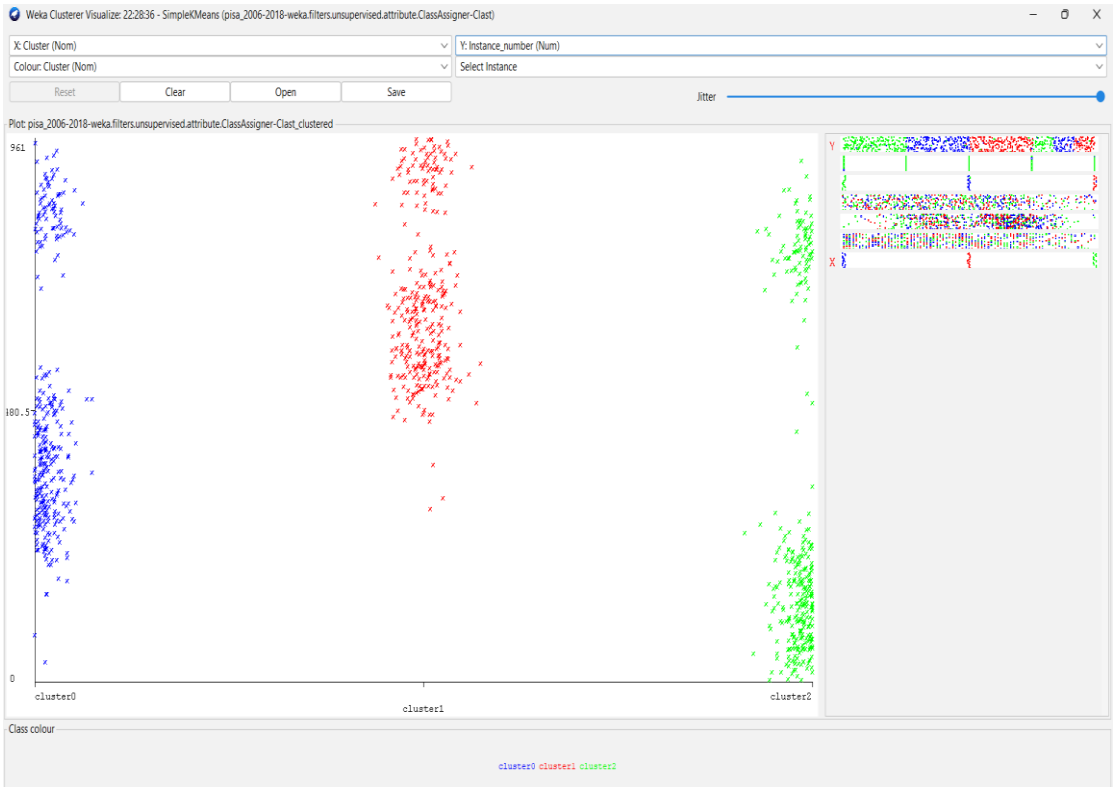
	0	1	2	3	4	5	6	7	8
K - Means	37%	29%	34%	-	-	-	-	-	-
Make Density Based	36%	30%	33%	-	-	-	-	-	-
Farthest First	58%	30%	12%	-	-	-	-	-	-
EM	16%	18%	11%	9%	5%	7%	10%	8%	17%
Filtered	37%	29%	34%						

Πίνακας 7: Συγκεντρωτικά αποτελέσματα για όλους τους αλγορίθμους.

Στο επόμενο στάδιο προχωράμε με την εφαρμογή των ίδιων αλγορίθμων σε λειτουργία συσταδοποίησης “Use training set” και αριθμό k = 3. Στην εικόνα 50 βλέπουμε το αποτέλεσμα συσταδοποίησης για τον αλγόριθμο Simple K – Means. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 7, το άθροισμα τετραγωνισμένου σφάλματος ανέρχεται σε 1161.06, ο υπολογιστικός χρόνος κατασκευής του μοντέλου σε 0,01 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 33%, C1 = 33% και C3 = 34%. Τέλος, όπως διαπιστώνεται από τη γραφική αναπαράσταση (εικόνα 51) δεν σχηματίζονται ιδιαίτερα συμπαγείς συστάδες.

```
22:28:36 - SimpleKMeans
                Country
                Score
                Rank
Test mode:      evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 1161.0617657060186

Initial starting points (random):

Cluster 0: 2018,Reading,'United Kingdom',504,15
Cluster 1: 2015,Reading,Romania,434,47
Cluster 2: 2009,Maths,Qatar,368,56

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                          Full Data           0           1           2
                                     (962.0)     (322.0)     (313.0)     (327.0)
=============================================================================
ï»¿Year                          2012.6486   2012.6522   2012.6518   2012.6422
Subject                              Maths     Science     Reading       Maths
Country                          Australia United Kingdom     Romania       Qatar
Score                            466.9574    473.0621    464.5911    463.211
Rank                              33.3108     32.4783     33.1054     34.3272



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       322  ( 33%)
1       313  ( 33%)
2       327  ( 34%)
```

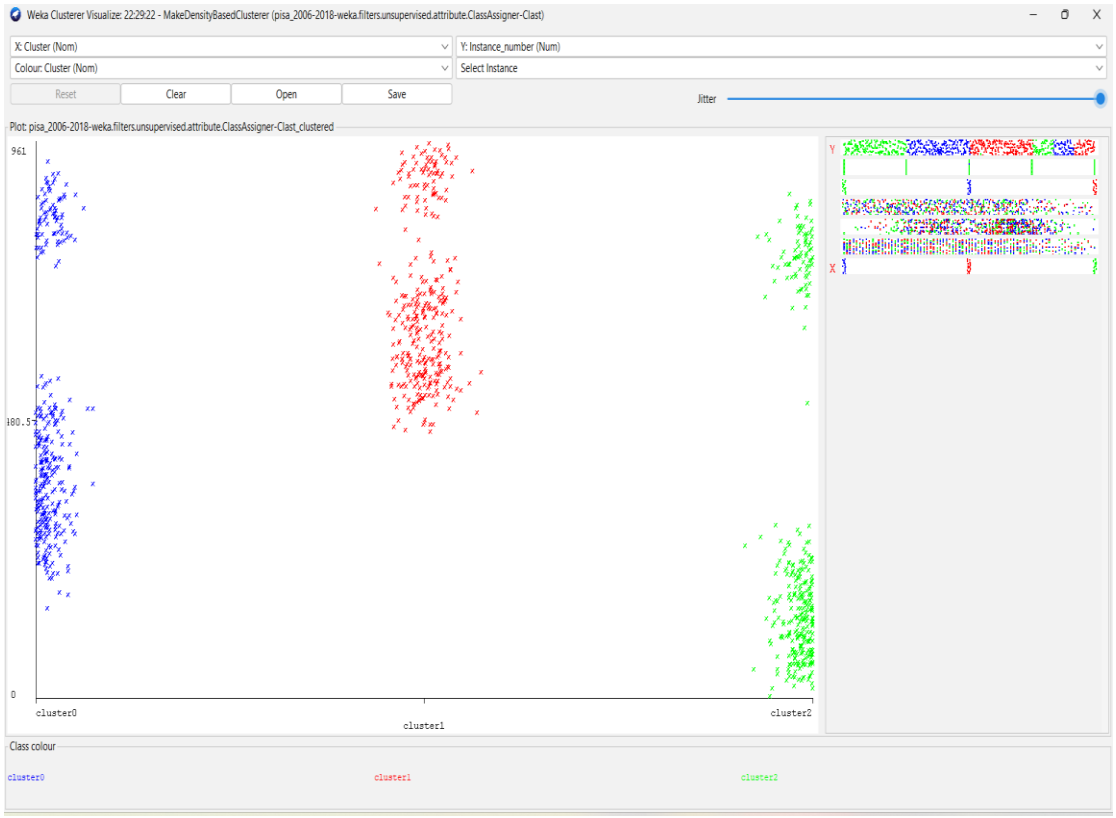
Εικόνα 50: Αποτελέσματα συσταδοποίησης αλγορίθμου Simple K - Means για k = 3 σε λειτουργία "Use training set".



Εικόνα 51: Γραφική αναπαράσταση συστάδων αλγορίθμου Simple K – Means.

Στη συνέχεια, στην εικόνα 52 βλέπουμε το αποτέλεσμα συσταδοποίησης για τον αλγόριθμο Make Density Based Clusterer για k = 3 και λειτουργία συσταδοποίησης “Use training set”. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 7, το άθροισμα τετραγωνισμένου σφάλματος σε 1161.06, ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,01 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 33%, C1 = 33% και C3 = 34%, και τέλος η λογαριθμική πιθανότητα κατανομής των δεδομένων στη σωστή ομάδα ανέρχεται σε -18.10791. Όπως φαίνεται και από την γραφική αναπαράσταση (εικόνα 53) και σε αυτή την περίπτωση οι σχηματιζόμενες συστάδες δεν είναι ιδιαίτερα συμπαγής και εμφανίζονται και ορισμένα ακραία στοιχεία.

Εικόνα 52: Αποτελέσματα συσταδοποίησης αλγορίθμου Make Density Based για k = 3 σε λειτουργία "Use training set”.

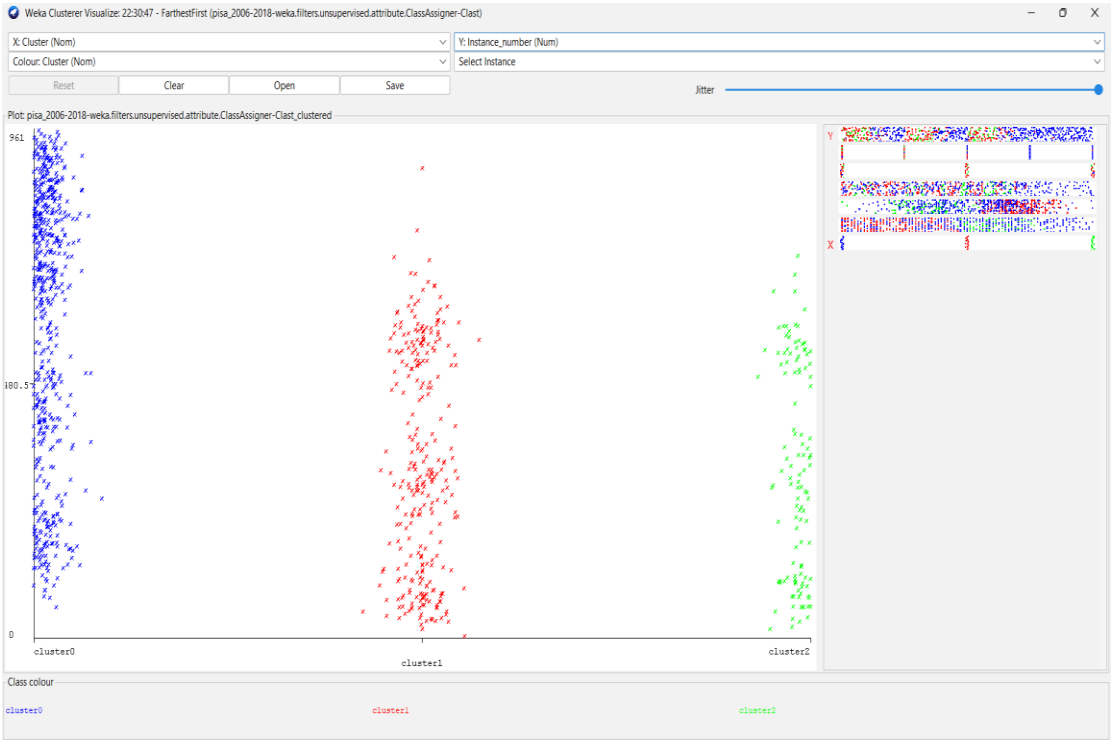


Εικόνα 53: Γραφική αναπαράσταση συστάδων αλγορίθμου Make Density Based.

Έπειτα, συνεχίζουμε με την εφαρμογή του αλγορίθμου Farthest First, με τις ίδιες παραμέτρους σταθερές, στην εικόνα 54 μπορούμε να δούμε τα αποτελέσματα συσταδοποίησης. Όπως φαίνεται ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 58%, C1 = 30% και C3 = 12%. Ενώ σε αυτή την υλοποίηση όπως διαπιστώνουμε από τη γραφική αναπαράσταση (εικόνα 55) έχουμε καλύτερη ποιότητα συστάδων.

```
22:30:47 - FarthestFirst
=== Run information ===

Scheme:       weka.clusterers.FarthestFirst -N 3 -S 1
Relation:     pisa_2006-2018-weka.filters.unsupervised.attribute.ClassAssigner-Clast
Instances:    962
Attributes:   5
              ï»¿Year
              Country
              Score
              Rank
Ignored:
              Subject
Test mode:    evaluate on training data

=== Clustering model (full training set) ===


FarthestFirst
==============

Cluster centroids:

Cluster 0
          2018.0 Greece 451.0 45.0
Cluster 1
          2006.0 Finland 563.0 1.0
Cluster 2
          2006.0 Qatar 312.0 51.0


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       559  ( 58%)
1       284  ( 30%)
2       119  ( 12%)
```

Εικόνα 54: Αποτελέσματα συσταδοποίησης αλγορίθμου Farthest First για k = 3 σε λειτουργία "Use training set".

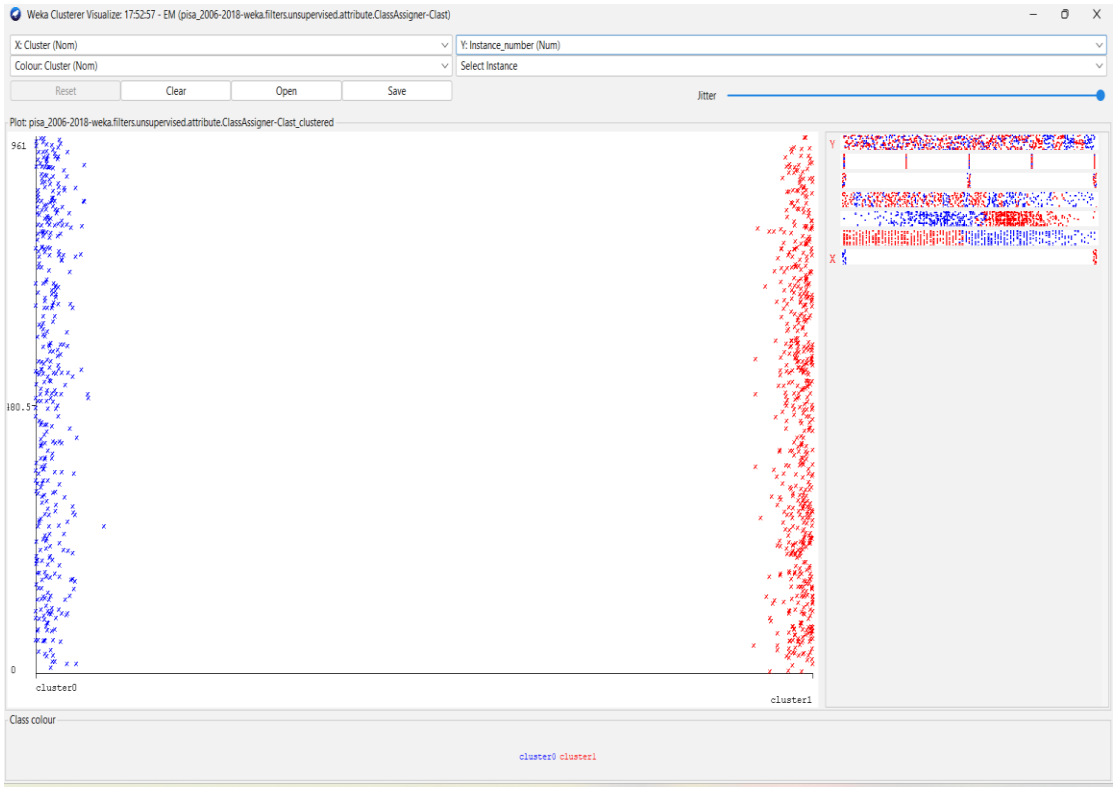


Εικόνα 55: Γραφική αναπαράσταση συστάδων αλγορίθμου Farthest First.

Ακολούθως, έχουμε την εφαρμογή του αλγορίθμου Expectation Maximization σε λειτουργία συσταδοποίησης “Use training set” με τη διαφορά ότι στην συγκεκριμένη υλοποίηση ο τελικός αριθμός συστάδων προκύπτει αυτόματα μέσω της μεθόδου διασταυρούμενης επικύρωσης (cross validation) όπου έχουμε εκπαίδευση του μοντέλου διαδοχικά στις επιμέρους διαμερισματοποιήσεις εκτός μίας η οποία χρησιμοποιείται ως δοκιμαστικό σετ, τα αποτελέσματα του οποίου διακρίνουμε στην εικόνα 56. Όπως φαίνεται ο αριθμός των συστάδων που προκύπτουν ανέρχεται σε 2, ο αριθμός των επαναλήψεων ανέρχεται σε 16, ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,91 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 40% και C1 = 60% και τέλος η λογαριθμική πιθανότητα κατανομής των δεδομένων στη σωστή ομάδα ανέρχεται σε -16.83595. Όπως είναι προφανές από την γραφική απεικόνιση (εικόνα 57) οι συστάδες που σχηματίζονται φαίνεται να έχουν πιο συμπαγή κατανομή στιγμιοτύπων σε σχέση με την προηγούμενες υλοποιήσεις.

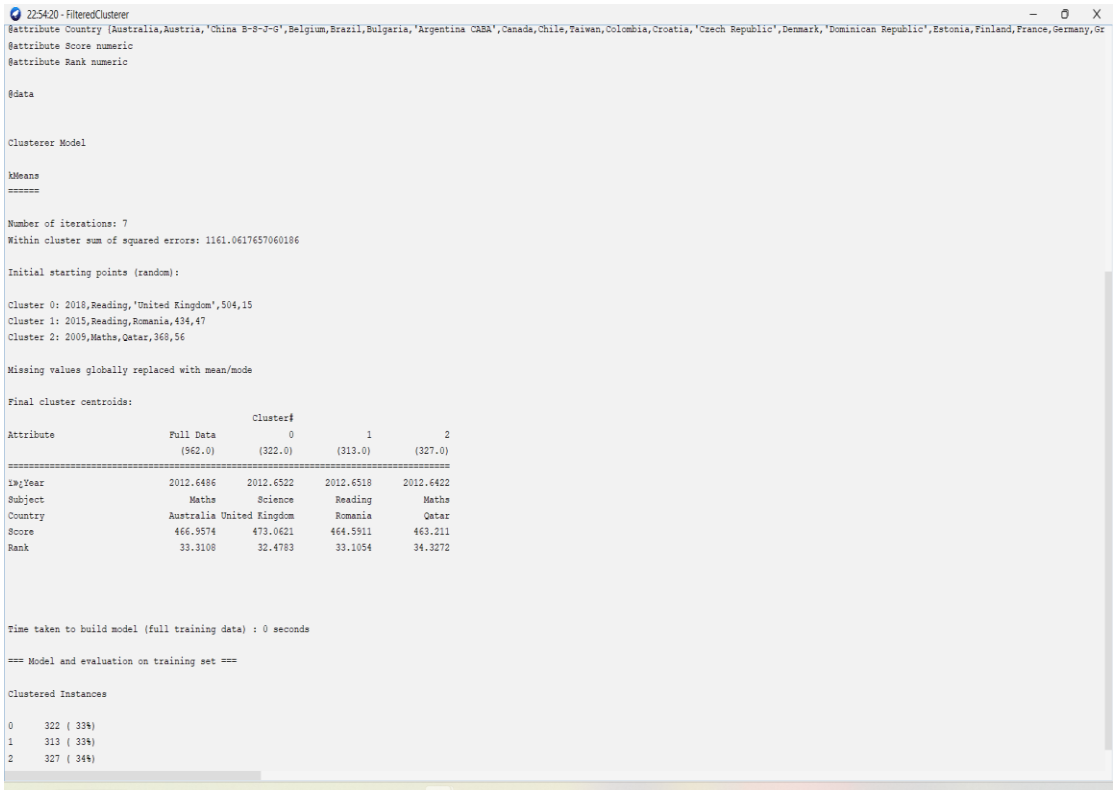
```
17:52:57 - EM
   Vietnam                          1.0014     9.9986
   Algeria                               4          1
   Argentina                             7          1
   Cyprus                           6.9973     1.0027
   Macedonia                             4          1
   Georgia                               7          1
   Kosovo                                7          1
   Malta                            5.6936     2.3064
   Moldova                               7          1
   China (B-S-J-Z)                       1          4
   Macau(China)                          1          2
   Hong Kong (China)                     1          4
   Belarus                          1.6493     3.3507
   Ukraine                          3.5982     1.4018
   Serbia                           3.9982     1.0018
   Brunei                                4          1
   Azerbaijan                            4          1
   Bosnia and Herzegovina                4          1
   North Macedonia                       4          1
   Saudi Arabia                          4          1
   Morocco                               4          1
   Panama                                4          1
   Philippines                           4          1
   Macau (China)                         1          3
   [total]                        476.7864   661.2136
Score
   mean                           412.7291   503.7381
   std. dev.                       30.4893    22.1941

Rank
   mean                            53.5027    19.6155
   std. dev.                       10.4162    11.1074


Time taken to build model (full training data) : 0.91 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       388 ( 40%)
1       574 ( 60%)


Log likelihood: -16.83595
```

Εικόνα 56: Αποτελέσματα συσταδοποίησης αλγορίθμου Expectation Maximization σε λειτουργία "Use training set”.

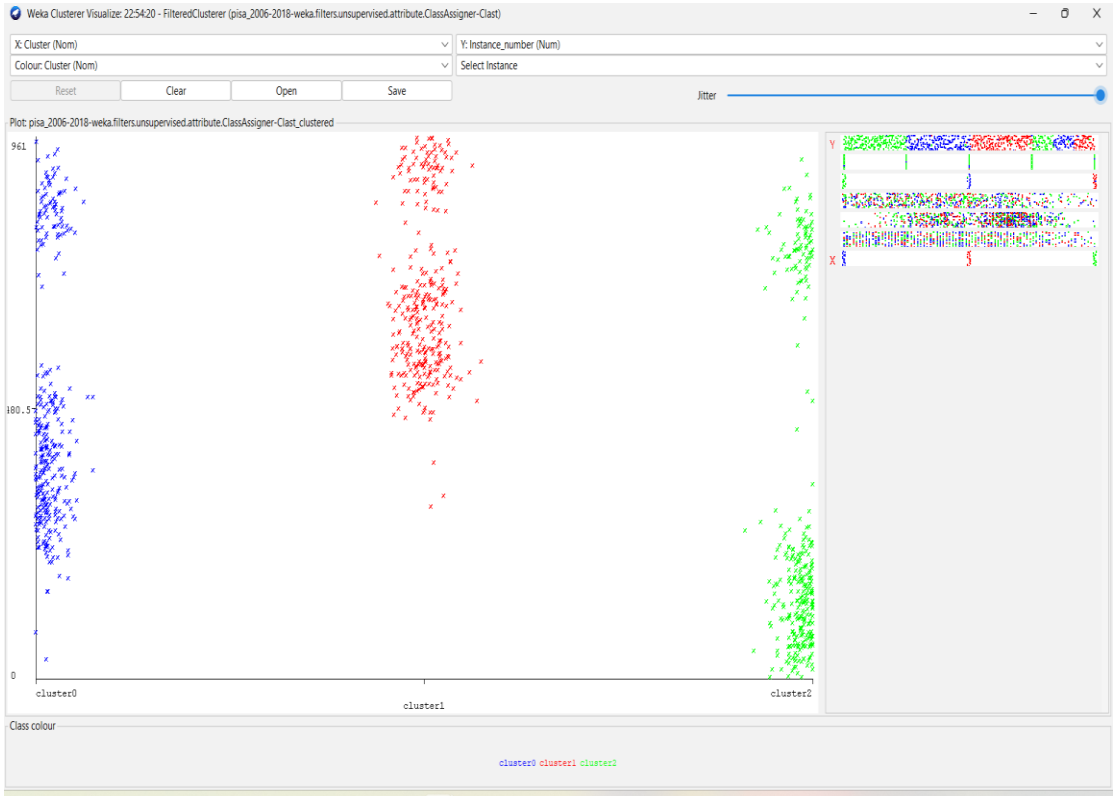


Εικόνα 57: Γραφική αναπαράσταση συστάδων αλγορίθμου Expectation Maximization.

Τέλος, σε λειτουργία “Use training set” εφαρμόζουμε τον αλγόριθμο Filtered Clusterer με αριθμό k = 3, τα αποτελέσματα παρουσιάζονται στην εικόνα 58. Όπως φαίνεται ο αριθμός των επαναλήψεων ανέρχεται σε 7, ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0 sec., το άθροισμα τετραγωνισμένου σφάλματος ανέρχεται σε 1161.06, το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 33%, C1 = 33% και C2 = 34% τέλος και σε αυτή την περίπτωση δεν παρατηρείται ιδιαίτερα συμπαγής κατανομή των συστάδων (εικόνα 59).



Εικόνα 58:Αποτελέσματα συσταδοποίησης αλγορίθμου Filtered Clusterer για k = 3 σε λειτουργία "Use training set".



Εικόνα 59: Γραφική αναπαράσταση συστάδων αλγορίθμου Filtered Clusterer.

Στον παρακάτω πίνακα 8 απεικονίζονται συγκεντρωτικά τα αποτελέσματα όλων των προηγούμενων αλγορίθμων. Όπως μπορούμε να διαπιστώσουμε οι αλγόριθμοι Simple K – Means, Make Density Based και Filtered εμφανίζουν ακριβώς την ίδια απόδοση αναφορικά με των αριθμό επαναλήψεων και το άθροισμα των τετραγωνικών σφαλμάτων ενώ από άποψη υπολογιστικού χρόνου ο Filtered εμφανίζεται ελάχιστα πιο γρήγορος. Σε ότι αφορά τον υπολογιστικό χρόνο ο αλγόριθμος Expectation Maximization εμφανίζει το μεγαλύτερο χρόνο όπως και στη λειτουργία συσταδοποίησης “Classes to clusters evaluation”. Συγκριτικά με τη λογαριθμική πιθανότητα των αλγορίθμων Make Density Based και EM διαπιστώνουμε ότι έχουμε αρνητικές τιμές γεγονός που υποδηλώνει ατέλειες στη συσταδοποίηση με τον Expectation Maximization να εμφανίζει σχετικά καλύτερη απόδοση. Τέλος, σε ότι αφορά τα ποσοστά των ομαδοποιημένων στιγμιότυπων ανά συστάδα οι αλγόριθμοι K – Means, Make Density Based και Filtered εμφανίζουν πανομοιότυπη κατανομή συστάδων, ενώ ο Farthest First εμφανίζει ακριβώς τα ίδια αποτελέσματα και για τις δύο λειτουργίες συσταδοποίησης. Τέλος, μπορούμε να πούμε ότι ο αλγόριθμος Expectation Maximization σχηματίζει καλύτερα ομαδοποιημένες συστάδες αν και από υπολογιστική άποψη βέλτιστος είναι ο Farthest First.

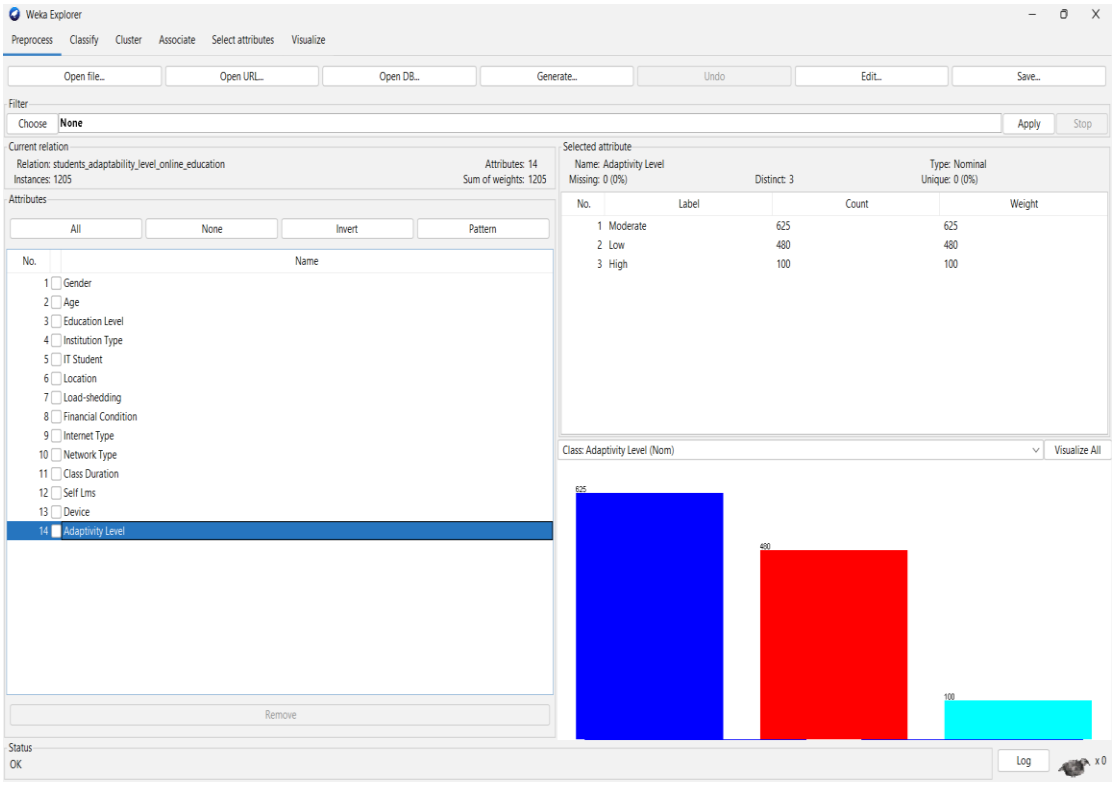
	K - Means	Make Density Based	Farthest First	EM	Filtered
Συστάδες	3	3	3	2	3
Επαναλήψεις	7	7	-	16	7
SSE	1161.06	1161.06	-	-	1161.06
Υπολογιστικός χρόνος	0,01 sec.	0,01 sec.	0 sec.	0,91 sec	0 sec.
logLikehood	-	-18.10791	-	-16.83595	-
		0 1 2			
K - Means		33% 33% 34%			
Make Density Based		33% 33% 34%			
Farthest First		58% 30% 12%			
EM		40% 60% -			
Filtered		33% 33% 34%			

Πίνακας 8: Συγκεντρωτικά αποτελέσματα για όλους τους αλγορίθμους.

Συμπερασματικά μπορούμε να πούμε ότι οι αλγόριθμοι K – Means, Make Density Based και Filtered εμφανίζουν όμοια αποτελέσματα και στις δύο λειτουργίες συσταδοποίησης "Use training set" και “Classes to clusters evaluation”. Στην πρώτη περίπτωση όπου χρησιμοποιούμε τις ετικέτες στη φάση της δοκιμής και αξιολογούμε την ακρίβεια συσταδοποίησης βάση της ορισμένης κλάσης ο Farthest First εμφανίζει την μεγαλύτερη ακρίβεια στον καλύτερο χρόνο. Στη δεύτερη φάση όπου δεν χρησιμοποιούνται ετικέτες για την εκπαίδευση του μοντέλου διαπιστώνουμε από τη ομαδοποίηση των στιγμιότυπων στις συστάδες ότι και πάλι ο Farthest First είναι πιο ακριβής καθώς μας εμφανίζει ακριβώς όπως και στην πρώτη περίπτωση τα ίδια αποτελέσματα.

4.5.5 5ο πείραμα: Προσαρμοστικότητα μαθητών στην διαδικτυακή εκπαίδευση

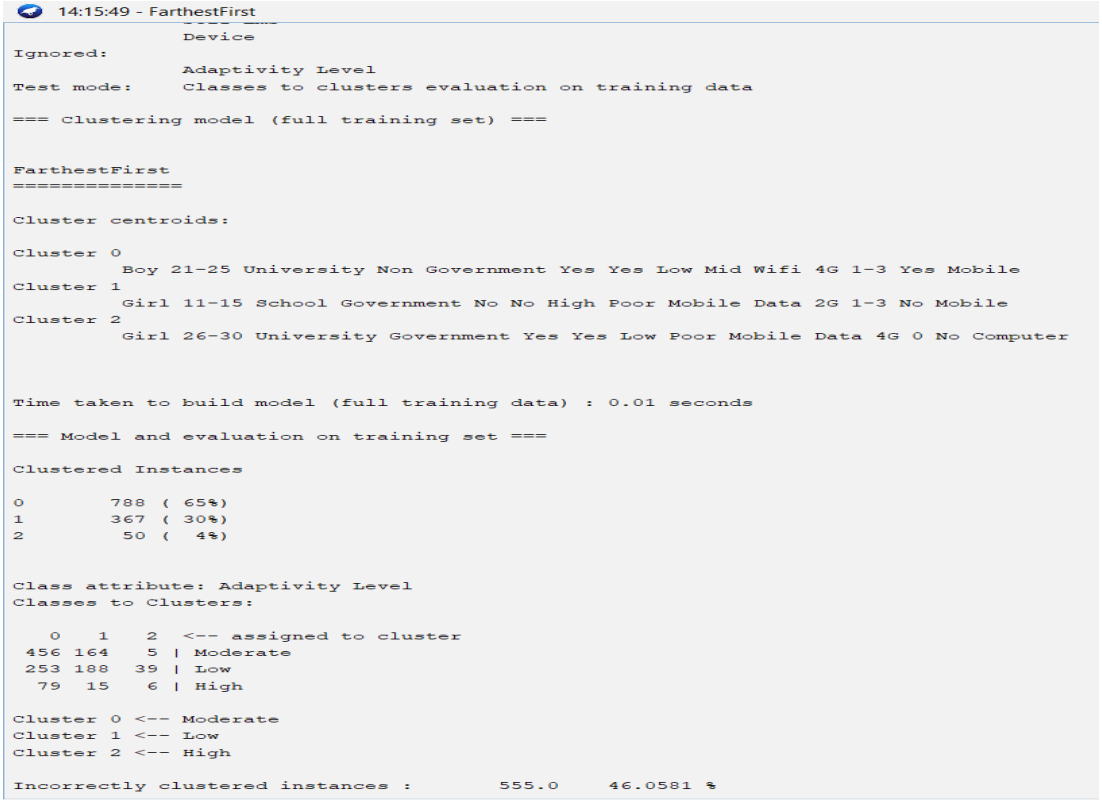
Στη συγκεκριμένη υλοποίηση χρησιμοποιήθηκε ένα σύνολο δεδομένων με όνομα students_adaptability_level_online_education.csv και αφορά την πρόβλεψη του επιπέδου προσαρμοστικότητας των μαθητών στη διαδικτυακή εκπαίδευση. Το σύνολο δεδομένων αποτελείται από πραγματικά δεδομένα που έχουν συλλεχθεί μέσω ηλεκτρονικών και έντυπων ερωτηματολογίων σε μαθητές σχολείων, κολλεγίων και πανεπιστημίων στο Μπαγκλαντές μεταξύ της περιόδου 10 Δεκεμβρίου 2020 έως 5 Φεβρουαρίου 2021 και αποτελούνται από κοινωνικά και δημογραφικά χαρακτηριστικά. Συνολικά περιλαμβάνονται 1205 στιγμιότυπα (instances) και 14 χαρακτηριστικά (attributes). Τα χαρακτηριστικά είναι τα παρακάτω: Gender, Age, Education Level, Institution Type, IT Student, Location, Load-shedding, Financial Condition, Internet Type, Network Type, Class Duration, Self Lms, Device, Adaptivity Level. Το τελευταίο χαρακτηριστικό είναι προσημασμένο ως κλάση. Στην εικόνα 60 παρακάτω παρουσιάζεται το σύνολο δεδομένων του αρχείο students_adaptability_level_online_education.arff που αποτελείται από κατηγορικά δεδομένα.



Εικόνα 60: Προεπεξεργασία του αρχείου εισόδου students_adaptability_level_online_education.arff

Στη συγκεκριμένη υλοποίηση θα επιχειρήσουμε να αξιολογήσουμε τα αποτελέσματα της εκπαίδευσης των αλγορίθμων συσταδοποίησης που προσφέρονται από τη διεπαφή Cluster του Weka και είναι οι Canopy, Cobweb, Hierarchical, Simple K – Means, Make Density Based Clusterer, Farthest First, Expectation Maximazation και Filtered Clusterer. Σκοπός είναι να αξιολογήσουμε για κάθε ένα αλγόριθμο σε λειτουργία συσταδοποίησης “Classes to clusters evaluation” τα ποσοστά ακρίβειας που επιτυγχάνονται ως προς το κατά πόσο οι επιλεγμένες συστάδες ταιριάζουν με την προσημασμένη κλάση Adaptivity Level ενώ ορίζουμε τον αριθμό συστάδων k = 3 όσο και το πλήθος των τιμών που δέχεται το χαρακτηριστικό Adaptivity Level. Τέλος, εφαρμόσουμε τον αλγόριθμο A priori ώστε να εντοπίσουμε τους κανόνες συσχέτισης και τα πιθανά πρότυπα στα δεδομένα σε ότι αφορά την προσαρμοστικότητα των μαθητών στη διαδικτυακή εκπαίδευση. Στην εικόνα 61 παρουσιάζονται τα αποτελέσματα του αλγορίθμου Farthest First ο οποίος εμφανίζει και το μεγαλύτερο ποσοστό ακρίβειας ενώ τα συνολικά αποτελέσματα παρουσιάζονται στον πίνακα 9. Όπως φαίνεται ο υπολογιστικός χρόνος κατασκευής του μοντέλου ανέρχεται σε 0,01 sec., το ποσοστό ομαδοποίησης των στιγμιότυπων ανά συστάδα ανέρχεται σε C0 = 65% που αντιστοιχεί στη μέτρια προσαρμοστικότητα, C1 = 30% που αντιστοιχεί στη χαμηλή

προσαρμοστικότητα και C2 = 4% που αντιστοιχεί στην υψηλή προσαρμοστικότητα και τέλος το ποσοστό ακρίβειας ανέρχεται σε 53,9419% πού είναι και το υψηλότερο. Όπως μπορούμε να αντιληφθούμε το μεγαλύτερο ποσοστό στιγμιότυπων ομαδοποιείται στη συστάδα που αφορά το μέτριο επίπεδο προσαρμοστικότητας.



Εικόνα 61: Αποτελέσματα συσταδοποίησης αλγορίθμου Farthest First για k = 3 και επιλεγμένη κλάση "Adaptivity Level".

Όπως μπορούμε να διακρίνουμε από τον πίνακα 9 ο αλγόριθμος Cobweb εμφανίζει τη μικρότερη ακρίβεια αποτελεσμάτων ενώ σχηματίζει δυσανάλογα μεγάλο πλήθος συστάδων γεγονός που οφείλεται στην υψηλή του πολυπλοκότητα. Οι αλγόριθμοι Simple K – Means, Make Density Based και Filtered εμφανίζουν όμοια αποτελέσματα ως προς το πλήθος συστάδων, τις επαναλήψεις, το άθροισμα τετραγωνικών σφαλμάτων, τον υπολογιστικό χρόνο και την ακρίβεια με τον Make Density Based να εμφανίζει ελάχιστα καλύτερη ακρίβεια. Ο αλγόριθμος Expectation Maximization είναι λιγότερο αποδοτικός από υπολογιστική άποψη και αριθμό επαναλήψεων καθώς επίσης και ο Hierarchical εμφανίζει μετά τον EM χαμηλή απόδοση από άποψη υπολογιστικού χρόνου. Τέλος, ο αλγόριθμος Farthest First εμφανίζει την

μεγαλύτερη ακρίβεια ως προς το πόσα στιγμιότυπα της κλάσης ομαδοποιούνται στις αντίστοιχες συστάδες.

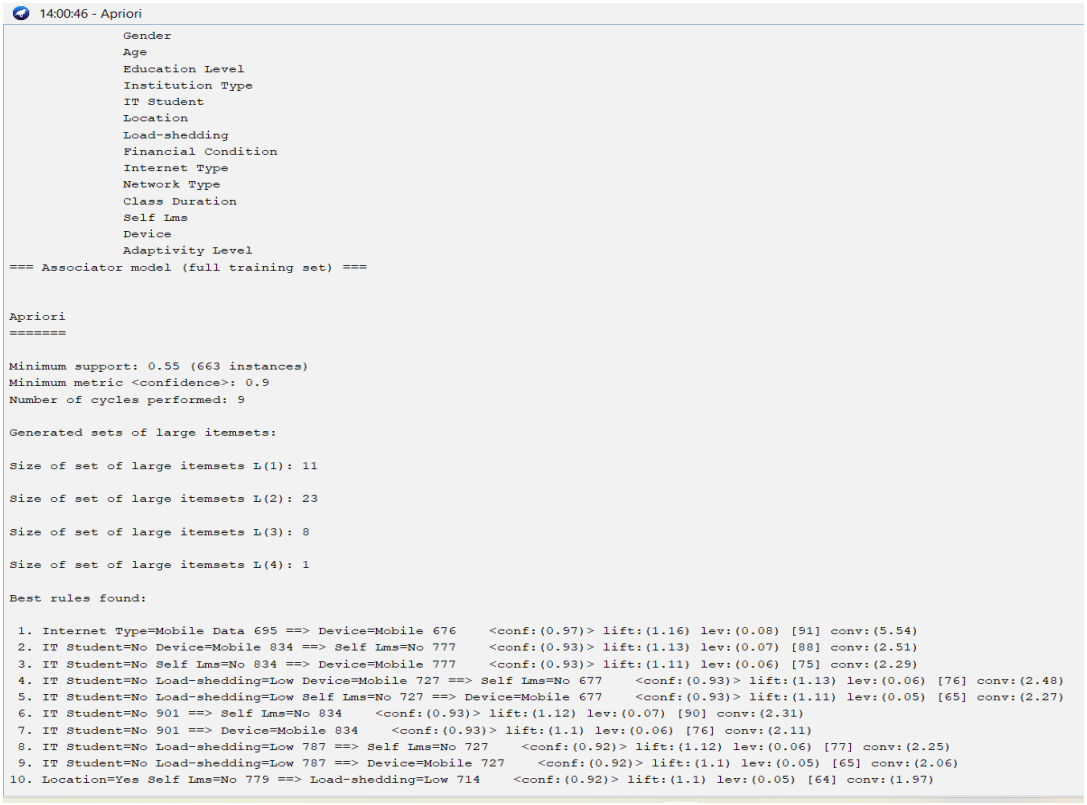
	Συστάδες	Επαναλήψεις	SSE	Υπολογιστικός χρόνος	logLikehood	Ακρίβεια
Canopy	32	-	-	0,01 sec.	-	19,0871%
Cobweb	478	-	-	0,3 sec.	-	5,3112%
Hierarchical	3	-	-	2,22 sec.	-	52,1162%
K - Means	3	5	3846	0 sec.	-	35,8506%
Make Density Based	3	5	3846	0,01 sec.	-8.31853	36,6805%
Farthest First	3	-	-	0,01 sec.	-	53,9419%
EM	10	42	-	31,97 sec	-6.99041	23,0705%
Filtered	3	5	3846	0 sec.	-	35,8506%

Πίνακας 9: Συγκεντρωτικός πίνακας αποτελεσμάτων αλγορίθμων clustering του Weka

Στη συνέχεια, εφαρμόζουμε στα δεδομένα μας τον αλγόριθμο A priori για τη δημιουργία κανόνων, εξάγονται οι δέκα καλύτεροι κανόνες βάση της μέσης εμπιστοσύνης που υπολογίζεται σε 0,9 όπως φαίνεται και στην εικόνα 62. Οι κανόνες που προκύπτουν είναι οι εξής:

1. Αν τύπος ίντερνετ = Δεδομένα κινητής (υποστήριξη = 695 στιγμιότυπα) τότε Συσκευή = Κινητό (υποστήριξη = 676 στιγμιότυπα). Εμπιστοσύνη 97%.
2. Αν Σπουδαστής Πληροφορικής = Όχι και Συσκευή = Κινητό (υποστήριξη = 834) τότε Ιδιόκτητο LMS = Όχι (υποστήριξη = 777 στιγμιότυπα). Εμπιστοσύνη = 93%.
3. Αν Σπουδαστής Πληροφορικής = Όχι και Ιδιόκτητο LMS = Όχι (υποστήριξη = 834 στιγμιότυπα) τότε Συσκευή = Κινητό (υποστήριξη = 777 στιγμιότυπα). Εμπιστοσύνη = 93%.
4. Αν Σπουδαστής Πληροφορικής = Όχι και Πτώση τάσης = Χαμηλή και Συσκευή = Κινητό (υποστήριξη = 727 στιγμιότυπα) τότε Ιδιόκτητο LMS = Όχι (υποστήριξη = 677 στιγμιότυπα). Εμπιστοσύνη = 93%.

5. Αν Σπουδαστής Πληροφορικής = Όχι και Πτώση τάσης = Χαμηλή και Ιδιόκτητο LMS = Όχι (υποστήριξη = 727 στιγμιότυπα) τότε Συσκευή = Κινητό (υποστήριξη = 677 στιγμιότυπα). Εμπιστοσύνη = 93%.
6. Αν Σπουδαστής Πληροφορικής = Όχι (υποστήριξη = 901 στιγμιότυπα) τότε Ιδιόκτητο LMS = Όχι (υποστήριξη = 834 στιγμιότυπα). Εμπιστοσύνη = 93%.
7. Αν Σπουδαστής Πληροφορικής = Όχι (υποστήριξη = 901 στιγμιότυπα) τότε Συσκευή = Κινητό (υποστήριξη = 834 στιγμιότυπα). Εμπιστοσύνη = 93%.
8. Αν Σπουδαστής Πληροφορικής = Όχι και Πτώση τάσης = Χαμηλή (υποστήριξη = 787 στιγμιότυπα) τότε Ιδιόκτητο LMS = Όχι (υποστήριξη = 727 στιγμιότυπα). Εμπιστοσύνη = 92%.
9. Αν Σπουδαστής Πληροφορικής = Όχι και Πτώση τάσης = Χαμηλή (υποστήριξη = 787 στιγμιότυπα) τότε Συσκευή = Κινητό (υποστήριξη = 727 στιγμιότυπα). Εμπιστοσύνη = 92%.
10. Αν Τοποθεσία = Ναι και Ιδιόκτητο LMS = Όχι (υποστήριξη = 779 στιγμιότυπα) τότε Πτώση Τάσης = Χαμηλή (υποστήριξη = 714 στιγμιότυπα). Εμπιστοσύνη = 92%.



Εικόνα 62: Αποτελέσματα κανόνων συσχέτισης αλγορίθμου Apriori.

Συμπερασματικά, μπορούμε να διαπιστώσουμε τα εξής, από τη διαδικασία συσταδοποίησης και την εφαρμογή του αλγορίθμου Farthest First ο οποίος εμφανίζει το υψηλότερο ποσοστό ακρίβειας αποτυπώνεται μέτρια προσαρμοστικότητα των μαθητών στην διαδικτυακή εκπαίδευση γεγονός το οποίο επιβεβαιώνεται και στην βασική δομή του συνόλου δεδομένων. Παράλληλα, σε συνδυασμό με τους κανόνες συσχέτισης που πρόκυψαν από την εφαρμογή του Αλγορίθμου Apriori μπορούμε να προχωρήσουμε στη συσχέτιση ότι η μέτρια προσαρμοστικότητα στην διαδικτυακή εκπαίδευση σχετίζεται με το γεγονός ότι μερίδα μαθητών με κινητό και χρήση δεδομένων που δεν σπουδάζει στον τομέα της πληροφορικής και το εκπαιδευτικό ίδρυμα που φοιτούν δεν διαθέτει ιδιόκτητο Σύστημα διαχείρισης μάθησης ενώ ζουν σε πόλη με προβλήματα χαμηλής πτώσης τάσης ηλεκτρισμού αποτελούν χαρακτηριστικά που συγκέντρωσαν τα υψηλότερα ποσοστά εμπιστοσύνης στα αποτελέσματα των κανόνων συσχέτισης.

4.6 Συμπεράσματα

Στην παρούσα διπλωματική εργασία ασχοληθήκαμε με την παρουσίαση των αλγορίθμων συσταδοποίησης και την υλοποίηση τους μέσω του ανοιχτού λογισμικού εξόρυξης δεδομένων WEKA. Επίσης, παρουσιάστηκαν κάποια ενδεικτικά παραδείγματα εφαρμογής των αλγορίθμων αυτών σε διάφορους τομείς μεταξύ των οποίων και του τομέα της εξόρυξης εκπαιδευτικών δεδομένων. Παρουσιάστηκε το περιβάλλον του λογισμικού WEKA και έγινε σύγκριση των διαφορετικών αλγορίθμων συσταδοποίησης με σκοπό να προσδιοριστούν συστάδες στα διαφορετικά αρχεία δεδομένων με τα οποία πραγματοποιήσαμε τα πειράματα μας.

Από τα αποτελέσματα των πειραμάτων μπορούμε να συμπεράνουμε ότι δεν υπάρχει κάποιος συγκεκριμένος κανόνας που να προσδιορίζει ποιος είναι ο καταλληλότερος αλγόριθμος συσταδοποίησης αλλά αυτό ανακαλύπτεται εμπειρικά και είχε σχέση με το είδος και την ποιότητα των δεδομένων μας αλλά και με το ζητούμενο της μελέτης μας που μπορεί να αφορά είτε την πρόβλεψη, είτε την ανακάλυψη προτύπων μέσα στα δεδομένα μας, είτε την σύγκριση της απόδοσης συγκεκριμένων αλγορίθμων για ένα συγκεκριμένο σύνολο δεδομένων.

Τέλος, από τα πειράματα μας μπορούμε να συμπεράνουμε ότι ο αλγόριθμος Simple K – Means είναι αποτελεσματικός και μπορεί να εφαρμοστεί στις περισσότερες περιπτώσεις. Όλοι οι αλγόριθμοι διακρίνονται από σχετική ασάφεια γεγονός που συνδέεται με την ποιότητα των δεδομένων. Οι αλγόριθμοι EM και Canopy διακρίνονται από υψηλή πολυπλοκότητα ενώ τα αποτελέσματα του Hierarchical δείχνουν να επηρεάζονται από θορυβώδη δεδομένα. Επίσης, καλά αποτελέσματα προκύπτουν και από τον αλγόριθμο Farthest First ως προς την ποιότητα συσταδοποίησης και τον χρόνο εκτέλεσης.

Βιβλιογραφία

[1] Ι. Βλαχάβας, Π. Κεφαλάς, Βασιλειάδης Νικόλαος, Φ. Κόκκορας, και Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*, Τρίτη Έκδοση. Αθήνα: Γκιούρδας Β., 2006.

[2] Μ. Χαλκίδη και Μ. Βαζιργιάννης, *Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό*, Δεύτερη Έκδοση. Αθήνα: Τυπωθήτω, 2005.

[3] Β. Βερύκιος, Β. Καγκλής, και Η. Σταυρόπουλος, 'Κεφάλαιο 6: Συσταδοποίηση Σύνοψη', στο *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R* , Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις, 2015, σσ. 135–152.

[4] Σ. Ζήμερας, 'Συσταδοποίηση', Σάμος, 2021.

[5] Ε. Πιτουρά, 'Συσταδοποίηση Ι', Ιωάννινα, 2010.

[6] P.-N. Tan, M. Steinbach, και V. Kumar, *Introduction to Data Mining*, First Edision. Pearson , 2005.

[7] IBM Cloud Education, 'Supervised Learning', Αυγούστου 19, 2020.

[8] T. Wood, 'Unsupervised Learning', *DeepAI*.

[9] P. Schneider και F. Xhafa, 'Machine learning: ML for eHealth systems', *Anomaly Detection and Complex Event Processing over IoT Data Streams*, σσ. 149–191, Ιανουαρίου 2022, doi: 10.1016/B978-0-12-823818-9.00019-5.

[10] K. Cawley, 'When To Use Supervised And Unsupervised Data Mining'. https://cloudtweaks.com/2014/09/supervised-unsupervised-data-mining/ (ημερομηνία πρόσβασης Δεκεμβρίου 07, 2022).

[11] N. U. Sati, 'A Comparative Study of WEKA Clustering Algorithms on Evaluation of Students' Performance', Ιουνίου 2021, σσ. 289–296. [Έκδοση σε ψηφιακή μορφή]. Available: https://www.researchgate.net/publication/354068076

[12] A. Peña - Ayala, *Educational Data Mining: Applications and Trends*, τ. 524. Springer, 2014.

[13] Y. Peng, G. Kou, Y. Shi, και Z. Chen, 'A descriptive framework for the field of data mining and knowledge discovery', στο *International Journal of Information Technology and Decision Making*, Δεκεμβρίου 2008, τ. 7, τχ. 4, σσ. 639–682. doi: 10.1142/S0219622008003204.

[14] Κ.-Μ. Κοντούλη, 'Σύγκριση των μεθόδων συσταδοποίησης - Compare the types of clustering', Διπλωματική εργασία, Πανεπιστήμιο Θεσσαλίας, 2017.

[15] G. Siemens και R. S. J. D. Baker, ‘Learning analytics and educational data mining: towards communication and collaboration’, *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, σσ. 252–254, 2012, doi: 10.1145/2330601.2330661.

[16] A. Villanueva και L. G. Moreno, ‘Data mining techniques applied in educational environments: Literature review Andrés Villanueva Manjarres Data mining techniques applied in educational environments: Literature review’, 2018. [Έκδοση σε ψηφιακή μορφή]. Available: http://greav.ub.edu/der/

[17] S. Hussain, R. Atallah, A. Kamsin, και J. Hazarika, ‘Classification, clustering and association rule mining in educational datasets using data mining tools: A case study’, *Advances in Intelligent Systems and Computing*, τ. 765, σσ. 196–211, 2019, doi: 10.1007/978-3-319-91192-2_21/COVER.

[18] U. Kutbay, ‘Partitional Clustering. Recent Applications in Data Clustering’, Ankara, 2018. Ημερομηνία πρόσβασης: Δεκεμβρίου 13, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: https://sci-hub.ru/10.5772/intechopen.75836

[19] D. S. B. Everitt, S. Landau, M. Leese, *Cluster analysis*, τ. 5, τχ. 1. Chichester, West Sussex, U.K: Wiley, 2011.

[20] M. E. Celebi, H. A. Kingravi, και P. A. Vela, ‘A comparative study of efficient initialization methods for the k-means clustering algorithm’, *Expert Syst Appl*, τ. 40, τχ. 1, σσ. 200–210, Ιανουαρίου 2013, doi: 10.1016/j.eswa.2012.07.021.

[21] S. Maraggi, ‘Where would you open a new Pizza Restaurant in Buenos Aires?’, 2020, [Έκδοση σε ψηφιακή μορφή]. Available: https://www.linkedin.com/in/santiagomaraggi/

[22] A. Dutt, S. Aghabozrgi, M. Akmal, B. Ismail, και H. M., ‘Clustering Algorithms Applied in Educational Data Mining’, *International Journal of Information and Electronics Engineering*, τ. 5, τχ. 2, 2015, doi: 10.7763/IJIEE.2015.V5.513.

[23] ‘Clustering - Fuzzy C-means’, *home.deib.polimi.it*, Ημερομηνία πρόσβασης: Δεκεμβρίου 11, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html

[24] ‘Fuzzy Clustering’, *reference.wolfram.com*, Ημερομηνία πρόσβασης: Δεκεμβρίου 11, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: http://reference.wolfram.com/applications/fuzzylogic/Manual/12.html

[25] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, και T. Moriarty, ‘A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation

of MRI Data', *IEEE Trans Med Imaging*, τ. 21, τχ. 3, σσ. 193–199, Μαρτίου 2002, doi: 10.1109/42.996338.

[26] F. Valafar, 'Pattern Recognition Techniques in Microarray Data Analysis', *Ann N Y Acad Sci*, τ. 980, τχ. 1, σσ. 41–64, Δεκεμβρίου 2002, doi: 10.1111/j.1749-6632.2002.tb04888.x.

[27] J. C. Dunn, 'A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters', *Journal of Cybernetics*, τ. 3, τχ. 3, σσ. 32–57, Ιανουαρίου 1973, doi: 10.1080/01969727308546046.

[28] M. Popescu, J. Keller, J. Bezdek, και A. Zare, 'Random projections fuzzy c-means (RPFCM) for big data clustering', στο *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Αυγούστου 2015, τ. 2015-November, σσ. 1–6. doi: 10.1109/FUZZ-IEEE.2015.7337933.

[29] J. Li και H. W. Lewis, 'Fuzzy Clustering Algorithms - Review of the Applications', *Proceedings - 2016 IEEE International Conference on Smart Cloud, SmartCloud 2016*, σσ. 282–288, Δεκεμβρίου 2016, doi: 10.1109/SMARTCLOUD.2016.14.

[30] V. Moertini, 'Introduction to Five DataClustering Algorithms Clustering Algorithm', *INTEGRAL*, τ. 7, τχ. 2, 2002, Ημερομηνία πρόσβασης: Δεκεμβρίου 17, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: https://www.researchgate.net/profile/Olga-Quintero/post/What_software_can_I_use_for_ANFIS_calibration/attachment/59d62d80c49f478072e9e8d2/AS%3A273562273812480%401442233737943/download/Introduction+to+Five+Data+Clustering.PDF

[31] L. F. Lago-Fernández και F. Corbacho, 'Normality-based validation for crisp clustering', *Pattern Recognit*, τ. 43, τχ. 3, σσ. 782–795, Μαρτίου 2010, doi: 10.1016/J.PATCOG.2009.09.018.

[32] J. Carr, 'An Introduction to Genetic Algorithms', 2014.

[33] P. Grznár *κ.ά.*, 'The Use of a Genetic Algorithm for Sorting Warehouse Optimisation', 2021, doi: 10.3390/pr9071197.

[34] A. H. Blasi και M. A. Alsuwaiket, 'Analysis of Students' Misconducts in Higher Education Institutions using Decision Tree and ANNs', *Technology & Applied Science Research*, τ. 10, τχ. 6, σσ. 6510–6514, 2020, [Έκδοση σε ψηφιακή μορφή]. Available: www.etasr.com

[35] W. Snyder, D. Nissman, D. van den Bout, και G. Bilbro, 'Kohonen Networks and Clustering | Enhanced Reader', στο *NIPS*, 1990.

[36] E. C. K. Tsao, J. C. Bezdek, και N. R. Pal, ‘Fuzzy Kohonen clustering networks’, *Pattern Recognit*, τ. 27, τχ. 5, σσ. 757–764, Μαΐου 1994, doi: 10.1016/0031-3203(94)90052-3.

[37] T. Kohonen και T. Honkela, ‘Kohonen network’, *Scholarpedia*, τ. 2, τχ. 1, σ. 1568, 2007, doi: 10.4249/SCHOLARPEDIA.1568.

[38] P. J. (Petrus J.) Braspenning, Thuijsman F., και A. J. M. M. Weijters, ‘Artificial neural networks : an introduction to ANN theory and practice’, σ. 293, 1995.

[39] E. Budi Perkasa, H. Santoso, Supardi, και A. A. Alkodri, ‘Kohonen Network Modeling for Lemma Recognition in Bangka Dialect of Malay’, στο *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 2020, σσ. 1–4. doi: 10.1109/CITSM50537.2020.9268929.

[40] S. D. de Carvalho, F. R. de Melo, E. L. Flôres, S. R. Pires, και L. F. B. Loja, ‘Intelligent tutoring system using expert knowledge and Kohonen maps with automated training’, *Neural Comput Appl*, τ. 32, τχ. 17, σσ. 13577–13589, 2020, doi: 10.1007/s00521-020-04767-0.

[41] T. Kohonen και T. Honkela, ‘Kohonen network’, *Scholarpedia*, τ. 2, τχ. 1, σ. 1568, 2007, doi: 10.4249/SCHOLARPEDIA.1568.

[42] S. C. Johnson, ‘Hierarchical clustering schemes’, *Psychometrika*, τ. 32, τχ. 3, σσ. 241–254, Σεπτεμβρίου 1967, doi: 10.1007/BF02289588.

[43] E. B. Fowlkes και C. L. Mallows, ‘A Method for Comparing Two Hierarchical Clusterings’, *J Am Stat Assoc*, τ. 78, τχ. 383, σσ. 553–569, 1983, doi: 10.1080/01621459.1983.10478008.

[44] A. Kraskov, H. St�gbauer, R. G. Andrzejack, και P. Grassberger, ‘Hierarchical clustering based on mutual information’, *Europhys Lett*, τ. 70, τχ. 2, σ. 278, 2008.

[45] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, και A. Song, ‘Efficient agglomerative hierarchical clustering’, *Expert Syst Appl*, τ. 42, τχ. 5, σσ. 2785–2797, Απριλίου 2015, doi: 10.1016/J.ESWA.2014.09.054.

[46] H. P. Kriegel, P. Kröger, J. Sander, και A. Zimek, ‘Density-based Clustering’, *WIREs Data Mining and Knowledge Discovery*, τ. 1, τχ. 3, σσ. 231–240, Μαΐου 2011, doi: 10.1002/widm.30.

[47] D. Birant και A. Kut, ‘ST-DBSCAN: An algorithm for clustering spatial-temporal data’, *Data Knowl Eng*, τ. 60, τχ. 1, σσ. 208–221, Ιανουαρίου 2007, doi: 10.1016/J.DATAK.2006.01.013.

[48] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, και X. Xu, 'DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN', *ACM Trans. Database Syst.*, τ. 42, τχ. 3, σσ. 19:1-19:21, Ιουλίου 2017, doi: 10.1145/3068335.

[49] J. Sander, M. Ester, H.-P. Kriegel, και X. Xu, 'Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications', *Data Min Knowl Discov*, τ. 2, τχ. 2, σσ. 169–194, 1998, doi: 10.1023/A:1009745219419.

[50] P. and W. Z. and R. Y. and W. F. L. Wang Zitong and Kang, 'A Density-Based Clustering Algorithm with Educational Applications', στο *Current Developments in Web Based Learning*, 2016, σσ. 118–127.

[51] D. S. B. Everitt, S. Landau, M. Leese, 'Cluster Analysis, 5th Edition', *John Wiley & Sons Ltd*, τ. 5, τχ. 1, σσ. 75–100, 2011.

[52] C. C. Aggarwal και C. K. Reddy, 'Data Clustering : Algorithms and Applications'.

[53] R. Sibson, 'SLINK: an optimally efficient algorithm for the single-link cluster method', *Comput J*, τ. 16, τχ. 1, σσ. 30–34, Ιανουαρίου 1973, doi: 10.1093/comjnl/16.1.30.

[54] P. Grabusts και A. Borisov, 'Using grid-clustering methods in data classification', στο *Proceedings. International Conference on Parallel Computing in Electrical Engineering*, 2002, σσ. 425–426. doi: 10.1109/PCEE.2002.1115319.

[55] X. Chi, 'A multi-protocol network log clustering method based on grid', στο *2011 IEEE International Symposium on IT in Medicine and Education*, 2011, τ. 2, σσ. 524–527. doi: 10.1109/ITiME.2011.6132164.

[56] Y. Ma, Q. Yang, J. Zhang, και Y. Qiu, 'A research on a grid community group model based education resource', στο *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, 2009, τ. 1, σσ. 161–164. doi: 10.1109/CCCM.2009.5268126.

[57] R. Vidal, 'Subspace Clustering', *IEEE Signal Processing Magazine*, σσ. 52–68, Μαρτίου 2011.

[58] R. Vidal, 'A TUTORIAL ON SUBSPACE CLUSTERING'.

[59] L. Lu και R. Vidal, 'Combined Central and Subspace Clustering for Computer Vision Applications', στο *Proceedings of the 23rd International Conference on Machine Learning*, 2006, σσ. 593–600. doi: 10.1145/1143844.1143919.

[60] D. v Paul, C. Nayagam, και J. D. Pawar, 'Modeling Academic Performance using Subspace Clustering Algorithm', στο *2016 IEEE Eighth International Conference*

on Technology for Education (T4E), 2016, σσ. 254–255. doi: 10.1109/T4E.2016.066.

[61] E. Elhamifar και R. Vidal, 'Sparse Subspace Clustering: Algorithm, Theory, and Applications', *IEEE Trans Pattern Anal Mach Intell*, τ. 35, τχ. 11, σσ. 2765–2781, 2013, doi: 10.1109/TPAMI.2013.57.

[62] W. M. Rand, 'Objective criteria for the evaluation of clustering methods', *J Am Stat Assoc*, τ. 66, τχ. 336, σσ. 846–850, Δεκεμβρίου 1971, doi: 10.2307/2284239.

[63] D. Akume, G. W.-B. технологии, και undefined 2002, 'Cluster algorithms: theory and methods', *cyberleninka.ru*, Ημερομηνία πρόσβασης: Δεκεμβρίου 30, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: https://cyberleninka.ru/article/n/cluster-algorithms-theory-and-methods

[64] C. Perrotta και B. Williamson, 'The social life of Learning Analytics: cluster analysis and the 'performance' of algorithmic education', *Learn Media Technol*, τ. 43, τχ. 1, σσ. 3–16, 2018, doi: 10.1080/17439884.2016.1182927.

[65] J. Grim, 'EM cluster analysis for categorical data', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, τ. 4109 LNCS, σσ. 640–648, 2006, doi: 10.1007/11815921_70/COVER.

[66] H. Řezanková, 'Cluster analysis and categorical data', Praha, 2009. [Έκδοση σε ψηφιακή μορφή]. Available: https://www.researchgate.net/publication/228758935

[67] S. Naouali, S. ben Salem, και Z. Chtourou, 'Clustering Categorical Data: A Survey. ', *Int J Inf Technol Decis Mak*, Δεκεμβρίου 2019.

[68] A. Nagpal, A. Jatain, και D. Gaur, 'Review based on data clustering algorithms. ', στο *2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES,* 2013. Ημερομηνία πρόσβασης: Δεκεμβρίου 13, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: https://sci-hub.ru/10.1109/CICT.2013.6558109

[69] L. A. Rasyid και S. Andayani, 'Review on Clustering Algorithms Based on Data Type: Towards the Method for Data Combined of Numeric-Fuzzy Linguistics', *J Phys Conf Ser*, τ. 1097, τχ. 1, σ. 012082, Σεπτεμβρίου 2018, doi: 10.1088/1742-6596/1097/1/012082.

[70] S. Guha, R. Rastogi, και K. Shim, 'Rock: A robust clustering algorithm for categorical attributes', *Inf Syst*, τ. 25, τχ. 5, σσ. 345–366, Ιουλίου 2000, doi: 10.1016/S0306-4379(00)00022-3.

[71] V. Kumar, ‘What is K-Means algorithm and how it works’.

[72] M. San, ‘Changelog: 12 Dec 2016 Advantages & Disadvantages of k--Means and Hierarchical clustering (Unsupervised Learning) Machine Learning for Language Technology ML4LT (2016) 2016 Advantages & Disadvantages of k--Means and Hierarchical Clustering’.

[73] N. Sharma, S. S. Sai, R. Litoriya, W. : Www, A. Bajpai, και M. R. Litoriya, ‘Comparison the various clustering algorithms of weka tools International Journal of Emerging Technology and Advanced Engineering Comparison the various clustering algorithms of weka tools’, 2012. [Έκδοση σε ψηφιακή μορφή]. Available: www.ijetae.com

[74] S. Sharmila και M. Kumar, ‘An optimized farthest first clustering algorithm’, *2013 Nirma University International Conference on Engineering, NUiCONE 2013*, 2013, doi: 10.1109/NUICONE.2013.6780070.

[75] D. J. Rosenkrantz, R. E. Stearns, και P. M. Lewis, II, ‘An analysis of several heuristics for the traveling salesman problem’, *SIAM Journal on Computing*, τ. 6, τχ. 3, σσ. 563–581, Σεπτεμβρίου 1977, doi: 10.1137/0206041.

[76] I. J. of W. & S. T. (IJWesT), ‘Farthest First Clustering in Links Reorganization’, *International journal of Web & Semantic Technology*, τ. 5, τχ. 3, σσ. 17–24, Ιουλίου 2014, doi: 10.5121/IJWEST.2014.5302.

[77] Y. Eldar, M. Lindenbaum, M. Porat, και Y. Y. Zeevi, ‘The farthest point strategy for progressive image sampling’, *IEEE Transactions on Image Processing*, τ. 6, τχ. 9, σσ. 1305–1315, 1997, doi: 10.1109/83.623193.

[78] S. S. Ravi, D. J. Rosenkrantz, και G. K. Tayi, ‘Heuristic and special case algorithms for dispersion problems’, *Oper Res*, τ. 42, τχ. 2, σσ. 299–310, 1994, doi: 10.1287/opre.42.2.299.

[79] Y. Rani και H. Rohil, ‘A Study of Hierarchical Clustering Algorithm’, 2013. [Έκδοση σε ψηφιακή μορφή]. Available: http://www.irphouse.com/ijict.htm

[80] J. Han, J. Pei, και M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition. Morgan Kaufmann, Elsevier., 2012.

[81] F. Murtagh και P. Contreras, ‘Algorithms for hierarchical clustering: An overview’, *Wiley Interdiscip Rev Data Min Knowl Discov*, τ. 2, τχ. 1, σσ. 86–97, Ιανουαρίου 2012, doi: 10.1002/WIDM.53.

[82] F. Murtagh και P. Contreras, ‘Algorithms for hierarchical clustering: an overview, II’, *Wiley Interdiscip Rev Data Min Knowl Discov*, τ. 7, τχ. 6, Νοεμβρίου 2017, doi: 10.1002/WIDM.1219.

[83] N.-S. Chauhan, ‘DBSCAN Clustering Algorithm in Machine Learning’, Μαρτίου 04, 2022.

[84] H. P. Kriegel, P. Kröger, J. Sander, και A. Zimek, ‘Density-based clustering’, *Wiley Interdiscip Rev Data Min Knowl Discov*, τ. 1, τχ. 3, σσ. 231–240, Μαΐου 2011, doi: 10.1002/WIDM.30.

[85] P. Bhattacharjee και P. Mitra, ‘A survey of density based clustering algorithms’, *Front Comput Sci*, τ. 15, τχ. 1, Φεβρουαρίου 2021, doi: 10.1007/S11704-019-9059-3.

[86] S. Kodati, R. Vivekanandam, και G. Ravi, ‘Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool’, *Advances in Intelligent Systems and Computing*, τ. 900, σσ. 111–117, 2019, doi: 10.1007/978-981-13-3600-3_11.

[87] S. Gnanapriya, R. A. Freeda, και M. Sowmiya, ‘Evaluation of Clustering Capability Using Weka Tool’, doi: 10.21172/ijiet.81.025.

[88] M. jagannatha reddy και B. Kavitha, ‘Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method’, τ. 5, Ιανουαρίου 2012.

[89] J. Khalfallah και J. B. H. Slama, ‘A comparative study of the various clustering algorithms in e-learning systems using weka tools’, *Proceedings of 2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing, JCCO: TICET-ICCA-GECO 2018*, σσ. 76–82, Νοεμβρίου 2018, doi: 10.1109/ICCA-TICET.2018.8726188.

[90] E. Frank, M. Hall, G. Holmes, … R. K.-D. mining and, και undefined 2009, ‘Weka-a machine learning workbench for data mining’, *Springer*, Ημερομηνία πρόσβασης: Ιανουαρίου 19, 2023. [Έκδοση σε ψηφιακή μορφή]. Available: https://link.springer.com/chapter/10.1007/978-0-387-09823-4_66

[91] F. N. Ali, A. M. H.-J. of I. and, και undefined 2018, ‘Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents’, *Taylor & Francis*, Ημερομηνία πρόσβασης: Ιανουαρίου 19, 2023. [Έκδοση σε ψηφιακή μορφή]. Available: https://www.tandfonline.com/doi/abs/10.1080/24751839.2018.1448205

[92] S. K. Ng, T. Krishnan, και G. J. McLachlan, ‘The EM Algorithm’, *Handbook of Computational Statistics*, σσ. 139–172, Δεκεμβρίου 2011, doi: 10.1007/978-3-642-21551-3_6.

[93] E. Frank *κ.ά.*, 'Weka-A Machine Learning Workbench for Data Mining', *Data Mining and Knowledge Discovery Handbook*, σσ. 1269–1277, 2009, doi: 10.1007/978-0-387-09823-4_66.

[94] F. M. Nafie Ali και A. A. Mohamed Hamed, 'Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents', *Journal of Information and Telecommunication*, τ. 2, τχ. 3, σσ. 231–245, Ιουλίου 2018, doi: 10.1080/24751839.2018.1448205.

[95] J. Agrawal *κ.ά.*, 'Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset Use of Machine Learning with Data Mining View project Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset', *Article in International Journal of Computer Applications*, τ. 168, τχ. 13, σσ. 975–8887, 2017, doi: 10.5120/ijca2017914522.

[96] N. Mulani, A. Pawar, P. Mulay, και A. Dani, 'Variant of COBWEB Clustering for Privacy Preservation in Cloud DB Querying', *Procedia Comput Sci*, τ. 50, σσ. 363–368, Ιανουαρίου 2015, doi: 10.1016/J.PROCS.2015.04.034.

[97] 'To Construct This Tree Hierarchy, Cobweb Sorts Each - Cobweb Algorithm - 1698x1050 PNG Download - PNGkit'. https://www.pngkit.com/view/u2q8y3a9q8e6r5u2_to-construct-this-tree-hierarchy-cobweb-sorts-each/ (ημερομηνία πρόσβασης Ιανουαρίου 20, 2023).

[98] M. Theodorakis, A. Vlachos, και T. Z. Kalamboukis, 'Using Hierarchical Clustering to Enhance Classification Accuracy'.

[99] A. McCallum, K. Nigam, και L. H. Ungar, 'Efficient clustering of high-dimensional data sets with application to reference matching', *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, σσ. 169–178, 2000, doi: 10.1145/347090.347123.

[100] 'The Canopies Algorithm', *courses.cs.washington.edu*, Ημερομηνία πρόσβασης: Ιανουαρίου 20, 2023. [Έκδοση σε ψηφιακή μορφή]. Available: http://courses.cs.washington.edu/courses/cse590q/04au/slides/DannyMcCallumKDD00.ppt

[101] Y. Zhang, P. Ruan, και J. Zhao, 'Design of digital economy consumer psychology prediction model based on canopy clustering algorithm', *Front Psychol*, τ. 13, Αυγούστου 2022, doi: 10.3389/FPSYG.2022.939283/FULL.

[102] N. S. Sagheer και S. A. Yousif, 'Canopy with k-means clustering algorithm for big data analytics', τ. 2334, σ. 70006, 2021, doi: 10.1063/5.0042398.

[103] E. Frank, M.-A. Hall, I.-H. Witten, και C.-J. Pal, *Online Appendix for Data Mining: 'Practical Machine Learning Tools and Techniques'*, Fourth Edition. Morgan Kaufmann, 2016.

[104] Σ. Ζήμερας, 'WEKA ΠΑΡΑΔΕΙΓΜΑΤΑ', Σάμος, 2021.

[105] 'WEKA'. Tutorials Point Pvt. Ltd., 2019. Ημερομηνία πρόσβασης: Οκτωβρίου 25, 2022. [Έκδοση σε ψηφιακή μορφή]. Available: http://www.tutorialspoint.com

[106] E. Alshehri, H. Alhakami, A. Baz, και T. Alsubait, 'A Comparison of EDM Tools and Techniques', 2020. [Έκδοση σε ψηφιακή μορφή]. Available: www.ijacsa.thesai.org

[107] R. Saxena, 'Educational Data Mining: Performance Evaluation of Decision Tree and Clustering Techniques Using WEKA Platform'.

[108] R. Baker, *Data Mining for Education*. oxford, UK: Elsevier. Ημερομηνία πρόσβασης: Ιανουαρίου 28, 2023. [Έκδοση σε ψηφιακή μορφή]. Available: http://www.columbia.edu/~rsb2162/Encyclopedia%20Chapter%20Draft%20v10%20-fw.pdf

[109] E. A. Amrieh, T. Hamtini, και I. Aljarah, 'Preprocessing and analyzing educational data set using X-API for improving student's performance', *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015*, Δεκεμβρίου 2015, doi: 10.1109/AEECT.2015.7360581.

[110] A. Dutt, M. Akmar Ismail, και T. Herawan, 'A Systematic Review on Educational Data Mining', doi: 10.1109/ACCESS.2017.2654247.