ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΊΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)

ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΤΟΥΣ ΣΤΗΝ ΠΡΟΩΘΗΣΗ ΑΓΑΘΩΝ

DATA MINING TECHNIQUES AND APPLICATIONS IN MARKETING

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ των
ΑΡΚΑΔΙΝΟΥ ΚΩΝΣΤΑΝΤΙΝΑ (ΑΜ:12625)
ΚΩΝΣΤΑ ΕΥΘΥΜΙΑ (ΑΜ:12535)

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ Δρ. Ηλίας Κ. Σταυρόπουλος

ΠΑΤΡΑ 2015

ΠΕΡΙΛΗΨΗ

Στην παρούσα πτυχιακή, περιγράφουμε τεχνικές εξόρυξης δεδομένων και πως εφαρμόζονται στον κλάδο του μάρκετινγκ. Το μάρκετινγκ ή αλλιώς προώθηση αγαθών αποτελεί έναν σημαντικό τομέα για την ανάπτυξη των επιχειρήσεων και των προϊόντων-υπηρεσιών τους. Χωρίς αυτό, δεν μπορεί να γίνει μια σωστή διαφήμιση των προϊόντων, αλλά και να υπάρξει ένας καλός χειρισμός από την παραγωγή τους μέχρι να φτάσουν στην κατανάλωσή τους από τον πελάτη. Σαν μάρκετινγκ ορίζεται η ικανοποίηση των αναγκών των καταναλωτών. Πως όμως θα αντιληφθεί η επιχείρηση τι ακριβώς αναζητά ο πελάτης, ώστε να τον ικανοποιήσει;

Εδώ καταλαβαίνουμε την σημασία της εξόρυξης δεδομένων, γιατί από τις συναλλαγές που πραγματοποιούνται καθημερινά, οι επιχειρήσεις αντλούν στοιχεία, για το ποιες είναι οι προτιμήσεις των καταναλωτών ή ποιες ανάγκες τους πρέπει να ικανοποιηθούν. Τα στοιχεία αυτά πρέπει με κάποιο τρόπο να επεξεργασθούν και να βγουν χρήσιμα συμπεράσματα τα οποία είναι σημαντικά για την στρατηγική που θα ακολουθήσει μια επιχείρηση. Έτσι με την εξόρυξη δεδομένων οι επιχειρήσεις έχουν μια ολοκληρωμένη εικόνα για τις πωλήσεις τους. Για παράδειγμα, μια επιχείρηση η οποία παράγει και πουλάει στην αγορά σοκολάτες, θέλει να μάθει ανάλογα με την ηλικία ή το φύλο των πελατών της, ποια κατηγορία πελατών αγοράζει περισσότερο, ώστε να καταλάβει σε ποια κατηγορία υπερτερεί και που χρειάζεται να αυξήσει τις πωλήσεις της.

Για την εξόρυξη δεδομένων χρησιμοποιούνται τεχνικές οι οποίες επεξεργάζονται τα δεδομένα. Οι πιο συνηθισμένες είναι:

- Ταξινόμηση
- Κανόνες Συσχέτισης
- Συσταδοποίηση
- Παλινδρόμηση

Για την εκτέλεση των τεχνικών αυτών χρησιμοποιούνται αρκετά λογισμικά, όπως:

- Sharky Neural Network
- Matlab
- Weka.

Στα κεφάλαια που θα ακολουθήσουν, αρχικά γίνεται μια περιγραφή για το τι είναι το μάρκετινγκ και η εξόρυξη γνώσης. Στην συνέχεια, θα αναφέρουμε τρεις τεχνικές εξόρυξης γνώσης οι οποίες εφαρμόζονται στο μάρκετινγκ, την ταξινόμηση, τους κανόνες συσχέτισης και την συσταδοποίηση. Και τέλος, στο τελευταίο κεφάλαιο θα παρουσιάσουμε λογισμικά που χρησιμοποιούνται στις τρείς παραπάνω τεχνικές.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΑΛΓΟΡΙΘΜΩΝ

ΕΙΣΑΓΩΓΗ

Ο ανταγωνισμός που υπάρχει ανάμεσα στις επιχειρήσεις, ώστε να πουλήσουν περισσότερα προϊόντα και η ανάπτυξη της τεχνολογίας, οδηγούν στην εφαρμογή νέων τεχνικών, όπως είναι οι τεχνικές εξόρυξη γνώσης. Οι τεχνικές εξόρυξης γνώσης εφαρμόζονται σε θέματα μάρκετινγκ, όπως στην τμηματοποίηση της αγοράς, στο καλάθι της νοικοκυράς, στο τι αγοράζουν περισσότερο οι διάφορες ηλικίες πελατών και άλλα παρόμοια θέματα. Οι επιχειρήσεις σήμερα χρησιμοποιούν τις τεχνικές αυτές για να βγάλουν συμπεράσματα τα οποία τους είναι χρήσιμα. Έτσι οργανώνουν καλύτερα τους στόχους τους και εκμεταλλεύονται κάποια στοιχεία στα οποία υπερτερούν Επομένως οι τεχνικές είναι σημαντικές για τις διάφορες επιχειρήσεις, ώστε να προωθήσουν τα προϊόντα τους, να κατανοήσουν τι θέλουν οι αγοραστές και πως το θέλουν.

Στα κεφάλαια που θα ακολουθήσουν, θα αναπτύξουμε την εφαρμογή της εξόρυξης γνώσης στο μάρκετινγκ. Στο μάρκετινγκ, η εξόρυξη γνώσης εφαρμόζεται, ώστε να αντιληφθεί η κάθε επιχείρηση την πορεία των πωλήσεών της, το αν έχει καλή φήμη στην αγορά και γενικά τι γνώμη έχουν οι πελάτες της για αυτήν. Αυτά τα χαρακτηριστικά επιδιώκει να δει η κάθε επιχείρηση, ώστε να βελτιωθεί, να αυξήσει την φήμη της και άρα τις πωλήσεις της. Τρείς από τις τεχνικές εξόρυξης γνώσης που εφαρμόζονται στο μάρκετινγκ είναι: η ταξινόμηση, οι κανόνες συσχέτισης και η συσταδοποίηση.

Η εξόρυξη δεδομένων δεν εφαρμόζεται μόνο στο μάρκετινγκ, αλλά τα τελευταία χρόνια συναντάμε τις τεχνικές της σχεδόν σε όλους τους τομείς. Ένα χαρακτηριστικό παράδειγμα εφαρμογής της εξόρυξης δεδομένων είναι στην ιατρική. Σε αυτόν τον τομέα η εξόρυξη δεδομένων εφαρμόζεται ώστε οι ιατροί να παρακολουθήσουν την πορεία μιας θεραπείας, να διαγνώσουν μια ασθένεια ή ακόμη να δουν ποια θεραπεία θα εφαρμόσουν στον ασθενή ανάλογα με την ηλικία, το φύλο και την ομάδα αίματος. Επίσης, εφαρμογή εξόρυξης δεδομένων υπάρχει και στον αθλητισμό, για την επίδοση των παικτών ανάλογα με την φυσική κατάσταση τους ή την ηλικία τους. Όπως είδαμε, οι εφαρμογές εξόρυξης δεδομένων είναι αρκετές και σίγουρα υπάρχουν και άλλες, για αυτό οι πηγές που μπορεί να ψάξει και να ανατρέξει κάποιος είναι πολλές και πρόσφατες, αφού η εξόρυξη δεδομένων πρωτοεμφανίζεται στο άρθρο [1] την δεκαετία του 80.

Σκοπός της εργασίας, είναι να αναπτυχθούν οι σημαντικότερες τεχνικές εξόρυξης δεδομένων που βρίσκουν εφαρμογή στο χώρο του μάρκετινγκ σύμφωνα με βιβλιογραφικές αναφορές. Στο τέλος, ο αναγνώστης θα πρέπει να έχει κατανοήσει τι είναι η εξόρυξη δεδομένων, πως εφαρμόζεται στο μάρκετινγκ και κυρίως ποιες είναι οι κύριες τεχνικές εξόρυξης δεδομένων.

1 Ο ΚΛΑΔΟΣ ΤΟΥ ΜΑΡΚΕΤΙΝΓΚ

1.1 ΟΡΙΣΜΟΣ ΜΑΡΚΕΤΙΝΓΚ

Το μάρκετινγκ ή αλλιώς προώθηση αγαθών (marketing) είναι μια θεωρία, η οποία έχει γίνει μια από τις σημαντικότερες πρακτικές σε μια επιχείρηση. Η διαδικασία του μάρκετινγκ αρχίζει από την παραγωγή του προϊόντος μέχρι την κατανάλωσή του από τον τελικό καταναλωτή.

Υπάρχουν αρκετοί ορισμοί που προσδιορίζουν τι είναι το μάρκετινγκ. Όμως κανένας δεν μπορεί να προσεγγίσει ακριβώς αυτήν την θεωρία. Έτσι ο όρος μάρκετινγκ δεν μπορεί να καθοριστεί και να μεταφραστεί επ' ακριβώς. Όμως όλοι οι ορισμοί αναφέρονται σε κάποια χαρακτηριστικά. Όπως, ότι στο μάρκετινγκ υπάρχουν συναλλαγές σε όλη την εφοδιαστική αλυσίδα, μεταξύ της επιχείρησης και του προμηθευτή ή της επιχείρησης με τον καταναλωτή. Κατά τον Philip Kotler: «μάρκετινγκ είναι οι διάφορες ανθρώπινες δραστηριότητες που έχουν σκοπό τη διευκόλυνση και ολοκλήρωση των συναλλαγών» [2] ή εναλλακτικά, «η επιχειρηματική δραστηριότητα που κατευθύνει τη ροή των αγαθών και των υπηρεσιών από την προσφορά στην ζήτηση, δηλαδή από τον παραγωγό, απ' ευθείας ή διαμέσου του μεταπωλητή, στον καταναλωτή ή χρήστη» [3] .

Πρωταρχικός στόχος είναι η αναγνώριση των αναγκών του καταναλωτή και έπειτα η ικανοποίησή του, η ανάπτυξη των προϊόντων και υπηρεσιών που τις ικανοποιούν, η δημιουργία των προϋποθέσεων ζήτησης, οι οποίες θα οδηγήσουν σε επιτυχείς πωλήσεις και η πραγματοποίηση των στόχων που έχει θέσει η επιχείρηση. Δηλαδή με την ικανοποίηση και της επιχείρησης όσον αφορά τους στόχους της και τα κέρδη που αποκομίζονται, αλλά και του πελάτη. Έτσι μπορεί να αναπτυχθεί μια μακροπρόθεσμη σχέση μεταξύ αυτού και της επιχείρησης. Έτσι η προώθηση αγαθών διαθέτει κάποιες μεθόδους, με τις οποίες η επιχείρηση μπορεί να ικανοποιήσει τους πελάτες της, να πάρει ένα μερίδιο αγοράς και να προσπαθήσει να μείνει σε αυτήν [4] .

Μέσω του μάρκετινγκ η κάθε επιχείρηση θα πρέπει να πείσει τους πελάτες της να την εμπιστευθούν. Αυτό γίνεται όταν η επιχείρηση προσφέρει αξιόπιστα προϊόντα. Οι πελάτες θα το καταλάβουν όταν αφού τα χρησιμοποιήσουν, διαπιστώσουν ότι έχουν καλύψει τις ανάγκες τους. Νέοι πελάτες να αγοράζουν τα προϊόντα της, αλλά και οι προηγούμενοι να μείνουν σε αυτήν πιστοί. Για να μπορέσει η επιχείρηση να διατηρήσει τους πελάτες της θα πρέπει να βρει τι ακριβώς θέλουν και πως θα τους ικανοποιήσει. Αυτό μπορεί να πραγματοποιηθεί, με καλύτερες τιμές από τους ανταγωνιστές, καλύτερη ποιότητα των προϊόντων ή υπηρεσιών και ευχάριστο περιβάλλον εξυπηρέτησης. Έτσι, η επιχείρηση θα καταφέρει να έχει ένα ανταγωνιστικό πλεονέκτημα, με αποτέλεσμα οι πελάτες να στηρίζουν την επιχείρηση, να μένουν σε αυτήν και να της έχουν εμπιστοσύνη. Όλα αυτά θα οδηγήσουν την επιχείρηση σε μία ηγετική θέση μέσα στην αγορά, στην οποία η επιχείρηση θα υπερτερεί από τους ανταγωνιστές της. Επίσης για να προσελκύσει νέους, μπορεί να εφαρμόσει την τεχνική της διαφήμισης, η οποία αναλύεται πιο κάτω.

1.2 ΜΕΙΓΜΑ ΜΑΡΚΕΤΙΝΓΚ

Το μείγμα Μάρκετινγκ (marketing mix) προτάθηκε από τον Jerome McCarthy [5]. Αποτελεί ένα από τα συνηθισμένα εργαλεία τα οποία η κάθε επιχείρηση χρησιμοποιεί, ώστε να επιτύχει τους στόχους της, όπως την προσέλκυση νέων καταναλωτών με την μείωση των τιμών. Είναι γνωστό, και ως τα 4p (σύμφωνα με το αρχικό γράμμα της κάθε λέξης στα

αγγλικά). Αυτό γιατί τέσσερα στοιχεία λαμβάνονται υπόψη από την επιχείρηση: το προϊόν (product), η τιμή (price), η τοποθεσία (place), και η προώθηση (promotion).

Όσον αφορά το προϊόν είναι το αποτέλεσμα παραγωγής της επιχείρησης μετά τη συλλογή, επεξεργασία και σύνθεση των πρώτων υλών, όπου πρώτη ύλη είναι ένα συστατικό μέρος που χρησιμοποιείται για να παραχθεί το προϊόν. Για παράδειγμα για να παραχθεί το ψωμί, ως πρώτη ύλη χρησιμοποιείται το αλεύρι. Η τιμή είναι το πόσο κοστίζει το προϊόν για να το αγοράσει ο πελάτης. Η τοποθεσία είναι η θέση που είναι τοποθετημένα τα προϊόντα στα ράφια των καταστημάτων. Είναι χαρακτηριστικό ότι κάποια προϊόντα, είναι τοποθετημένα σε τέτοια σημεία, ώστε να προσελκύσουν τους καταναλωτές να τα αγοράσουν, όπως για παράδειγμα οι τσίχλες είναι τοποθετημένες στα ταμεία του σούπερ μάρκετ. Τέλος η προώθηση αναφέρεται κυρίως στην περαιτέρω προβολή των προϊόντων προς τον πελάτη προκειμένου να τους ελκύσει να τα αγοράσουν. Ένας τέτοιος τρόπος είναι αυτό της διαφήμισης. Αναφορά σε αυτή γίνεται στην επόμενη ενότητα.

Στον κλάδο του μάρκετινγκ είναι αρκετά δύσκολο να διαχωρίσουμε τα προϊόντα από τις υπηρεσίες. Για παράδειγμα, ένα εστιατόριο παρέχει στον πελάτη του όχι μόνο προϊόν (π.χ. καλής ποιότητας φαγητό), αλλά και το περιβάλλον (π.χ. μουσική). Αυτό που χαρακτηρίζει μια υπηρεσία είναι ότι είναι άυλη, δηλαδή δεν έχει κάποια διάσταση, γιατί παράγεται την στιγμή που ο καταναλωτής ζητά να του παραχθεί η υπηρεσία. Έτσι στο παραπάνω παράδειγμα το φαγητό είναι το προϊόν, ενώ η μουσική η υπηρεσία που το συνοδεύει. Το 1981 έγινε μία επέκταση των τεσσάρων στοιχείων του μείγματος μάρκετινγκ σε τρία επιπλέον, τα οποία καθορίζουν κάποια επιπλέον στοιχεία των υπηρεσιών. Έτσι για τις υπηρεσίες, το μείγμα μάρκετινγκ είναι τα 7p. Τα τρία επιπλέον στοιχεία αποτελούνται από τους ανθρώπους (people), την διαδικασία (process) και τις φυσικές αποδείξεις (physical evidence). Συγκεντρωτικά και τα 7p απεικονίζονται στην εικόνα 1[6].

Εικόνα 1 Τα 7p του μείγματος Μάρκετινγκ

Οι άνθρωποι είναι τα άτομα που λαμβάνουν μέρος στην πραγματοποίηση της υπηρεσίας. Είναι οι εργαζόμενοι οι οποίοι έρχονται σε επαφή με τον πελάτη. Επίσης οι πελάτες κάθε επιχείρησης. Αποτελούν το πιο σημαντικό τμήμα κερδοφορίας της επιχείρησης, αλλά και διατήρησης καλών σχέσεων με εκείνη και αξιοπιστίας των πελατών προς αυτή. Η διαδικασίες είναι οι δραστηριότητες που γίνονται για να παραδοθεί μια υπηρεσία στον πελάτη. Τέλος οι φυσικές αποδείξεις συνδέουν τον υπάλληλο και τον πελάτη μέσω ενός αγαθού. Για παράδειγμα ας πάρουμε το ξενοδοχείο. Ο σχεδιασμός, ο εφοδιασμός, ο φωτισμός και η διακόσμηση του αποτελεί μία καλή οπτική στα μάτια του πελάτη όσον αφορά την ποιότητα της υπηρεσίας που προσφέρεται [7].

1.3 Η ΣΥΜΒΟΛΗ ΤΗΣ ΔΙΑΦΗΜΙΣΗΣ ΣΤΟ ΜΑΡΚΕΤΙΝΓΚ

Ένα αναπόσπαστο κομμάτι του μάρκετινγκ είναι αυτό της διαφήμισης το οποίο ανήκει στο μείγμα προβολής και επικοινωνίας της επιχείρησης. Μέσω της διαφήμισης γίνεται αφενός γνωστό ένα προϊόν ή μία υπηρεσία στον καταναλωτή και αφετέρου τον δελεάζει να το αγοράσει, ενώ πριν δεν τον ενδιέφερε, αναφερόμενο στα θετικά στοιχεία του. Με αυτό τον τρόπο αυξάνονται οι πωλήσεις του. Παράλληλα μπορεί να ενημερώνει το κοινό για την τιμή, τη διαθεσιμότητα, τους τυχόν κινδύνους κ.λπ. Μπορεί να γίνει με πολλά μέσα ενημέρωσης όπως: τηλεόραση, αφίσες, περιοδικά,, ραδιόφωνο, μέσα κοινωνικής δικτύωσης και άλλα [8] .

Εφόσον η επιχείρηση καταλάβει τι ακριβώς περιμένουν οι καταναλωτές, θα προσαρμόσουν την παραγωγή των προϊόντων, οι καταναλωτές θα μείνουν ευχαριστημένοι με τη χρήση τους, με αποτέλεσμα να επιζητούν περισσότερα προϊόντα από την συγκεκριμένη εταιρεία. Έτσι δεν θα υπάρξει μόνο αύξηση τον πωλήσεων, αλλά και προσέλκυση νέων καταναλωτών αφού η επιχείρηση θα αποκτήσει μια καλή φήμη στην αγορά. Χρησιμοποιούνται διάφοροι τρόποι ώστε να παραχθούν νέα προϊόντα. Τέτοιοι τρόποι είναι το market basket analysis και η τμηματοποίηση της αγοράς. Αναφορά σε αυτά θα γίνει στα κεφάλαια 4 και 5 αντίστοιχα.

1.4 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΑΡΚΕΤΙΝΓΚ

Η εξόρυξη γνώσης αποτελεί σημαντική τεχνική για την επεξεργασία των δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων σε σχέση με περιπτώσεις μάρκετινγκ. Αυτό μπορούμε να το συμπεράνουμε από το γεγονός ότι τα αποτελέσματα που βγαίνουν από την διαδικασία της εξόρυξης γνώσης είναι πολύ πιο κατανοητά και μπορούν να εξαχθούν αρκετά πιο γρήγορα. Σαν περιπτώσεις μάρκετινγκ εννοούμε τεχνικές που χρησιμοποιεί η κάθε επιχείρηση ώστε να καταλάβει το προφίλ των πελατών της. Για παράδειγμα, οι επιχειρήσεις χρησιμοποιούν την τμηματοποίηση της αγοράς, ώστε να χωρίσουν σε κομμάτια την αγορά, να καταλάβουν τι αγοράζουν οι πελάτες της και να στοχεύσουν σε ένα συγκεκριμένο κομμάτι αυτής. Για να γίνει όμως η επεξεργασία των δεδομένων και να έχουμε τις απαντήσεις που χρειάζονται, θα πρέπει να εφαρμοστούν τεχνικές εξόρυξης γνώσης. Παρακάτω παραθέτονται κάποια προβλήματα μάρκετινγκ τα οποία μπορούν να επιλυθούν εφαρμόζοντας τεχνικές εξόρυξης δεδομένων.

Η πρόβλεψη για μια επιχείρηση είναι πολύ σημαντική. Με αυτήν μπορεί να προβλεφθεί η συμπεριφορά των καταναλωτών της και να δράσει αναλόγως. Αυτό επιλύεται με μια από τις τεχνικές εξόρυξης γνώσης, την ταξινόμηση (Classification). Στόχος της ταξινόμησης είναι να χωρίσει τους πελάτες σε κατηγορίες με κάποια βασικά κριτήρια, ώστε να προβλέψει μια μελλοντική συμπεριφορά τους. Για παράδειγμα, μια επιχείρηση, βλέπει σε μια κατηγορία καταναλωτών ότι στρέφεται σε συγκεκριμένα προϊόντα, με αποτέλεσμα να παράγει περισσότερα, ώστε να υπάρχει επαρκής ποσότητα προϊόντος στην αγορά. Οι επιχειρήσεις για να αυξήσουν τις πωλήσεις τους πρέπει αρχικά να καταλάβουν τι αγοράζουν οι καταναλωτές, πότε το αγοράζουν και πως. Αυτό επιτυγχάνεται με την ανάλυση καλαθιού αγοράς (market basket analysis). Το πρόβλημα αυτής της ανάλυσης επιλύεται με την τεχνική των κανόνες συσχέτισης (Association rules). Αυτή η τεχνική εφαρμόζει κανόνες, οι οποίοι δείχνουν τι αγοράζουν οι καταναλωτές και μαζί με ποια άλλα προϊόντα το αγοράζουν. Τέλος, μια άλλη μέθοδος που χρησιμοποιούν οι επιχειρήσεις είναι η τμηματοποίηση της αγοράς, η οποία χωρίζει την αγορά σε τμήματα τα οποία μπορούν να αντιμετωπιστούν με μια συγκεκριμένη στρατηγική.. Αυτό υλοποιείται μέσω της τεχνικής της συσταδοποίησης ή ομαδοποίησης (Clustering).

Οι τεχνικές εξόρυξης γνώσης επιλύουν τις παραπάνω τεχνικές μάρκετινγκ. Για να μιλήσουμε όμως για τεχνικές εξόρυξης γνώσης θα πρέπει πρώτα να αναφέρουμε την θεωρία της. Η θεωρία της εξόρυξης γνώσης αναφέρεται στο κεφάλαιο που ακολουθεί.

2 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

2.1 ΟΡΙΣΜΟΣ

Η έννοια της εξόρυξης δεδομένων (data mining) πρωτοεμφανίζεται στο [1] την δεκαετία του 1980. Υπάρχουν αρκετοί ορισμοί οι οποίοι την χαρακτηρίζουν. Εμείς θα αναφέρουμε κάποιους από αυτούς.

- Ορισμός 1: Εξόρυξη δεδομένων ονομάζουμε την έρευνα και ανάλυση με αυτόματα ή ημιαυτόματα μέσα μεγάλων ποσοτήτων δεδομένων με σκοπό να ανακαλύψουμε πρότυπα και κανόνες [9].
- Ορισμός 2: Μέθοδος ανακάλυψης ενδιαφέρουσας δομής σε μεγάλες βάσεις δεδομένων (μοτίβα, πρόβλεψη κανόνων καθώς και μη συνήθεις περιπτώσεις).
- Ορισμός 3: Εξελιγμένη δυνατότητα αναζήτησης δεδομένων που χρησιμοποιεί στατιστικούς αλγόριθμους για να ανακαλύψει μοτίβα και συσχετίσεις στα δεδομένα [10].

Με βάσει τα αποτελέσματα και τα συμπεράσματα που βγαίνουν από την διαδικασία της εξόρυξης δεδομένων, η κάθε επιχείρηση αποφασίζει πως θα διαθέσει τα προϊόντα της στην αγορά και στους τελικούς καταναλωτές. Τα δεδομένα προς επεξεργασία μπορεί να είναι από καλά οργανωμένες αποθήκες δεδομένων ή από διάφορες μη δομημένες πηγές. Επίσης μπορεί να προέρχονται από το γραμμωτό κώδικα των προϊόντων, δηλαδή από τον έλεγχο στα ταμεία των καταστημάτων ακόμα και από μια απλή λίστα συναλλαγής όταν μιλάμε για μικρό αριθμό συναλλαγών.

Συνήθως τα δεδομένα συναλλαγών παρουσιάζονται σε έναν πίνακα δύο διαστάσεων ο οποίος ονομάζεται σύνολο δεδομένων. Οι γραμμές περιέχουν παρατηρήσεις με προηγούμενες πωλήσεις και ονομάζονται συναλλαγές. Οι στήλες περιέχουν πληροφορίες για την κάθε παρατήρηση και ονομάζονται ως χαρακτηριστικά (attributes), ή μεταβλητές (variables). Τα χαρακτηριστικά αυτά μπορεί να είναι κατηγορηματικά ή αριθμητικά, ανάλογα με τον τύπο τιμών που παίρνουν. Παράδειγμα κατηγορηματικού χαρακτηριστικού είναι η επαρχία της κατοικίας ενός ατόμου: παίρνει ένα σύνολο ονομάτων, η οποία μπορεί να αντιστοιχηθεί με ακέραιους αριθμούς. Παράδειγμα αριθμητικού χαρακτηριστικού είναι το ποσό εξερχόμενων τηλεφωνημάτων δύο πελατών Α και Β. Κάνουν τηλεφωνήματα σε μία εβδομάδα για €27 και €36, η διαφορά μεταξύ των ποσών που ξοδεύονται είναι ίση με €9 και ότι ο Α έχει καταναλώσει τα ¾ που ποσού που ξοδεύεται από τον Β [11] .

2.2 ΤΕΧΝΙΚΕΣ

Για την επεξεργασία των δεδομένων και την εξαγωγή χρήσιμων συμπερασμάτων χρησιμοποιούνται τεχνικές εξόρυξης γνώσης. Οι τεχνικές αυτές είναι πολλές. Κάποιες συνδυάζονται με υπολογιστικά προγράμματα, άλλες με την βοήθεια της στατιστικής ή σε συνδυασμό και των δυο. Οι πιο συχνές, και αυτές που θα αναφερθούν στα επόμενα κεφάλαια είναι:

- η ταξινόμηση (Classification)
- οι κανόνες συσχέτισης (Association Rules)
- η συσταδοποίηση (Clustering).

Η ταξινόμηση χωρίζει τα δεδομένα σε κατηγορίες και χρησιμοποιεί τον ταξινομητή Bayes, ο οποίος περιγράφεται μέσα από το θεώρημα Bayes, δέντρα αποφάσεων τα οποία δημιουργούνται μέσω των αλγορίθμων ID3 και C4.5, τεχνητά νευρωνικά δίκτυα, και τον αλγόριθμο κοντινότερου γείτονα (k nearest neighbor - k-nn). Οι κανόνες συσχέτισης

χρησιμοποιούν κανόνες, ώστε σύμφωνα με αυτούς να βρεθεί μια λύση, όπου εφαρμόζονται οι όροι εμπιστοσύνη και υποστήριξη. Για παράδειγμα, {καφές}→{ζάχαρη}, που σημαίνει ότι κάποιος αν αγοράσει καφέ, τότε θα αγοράσει και ζάχαρη. Ένας αλγόριθμος που χρησιμοποιείται στους κανόνες συσχέτισης είναι ο apriori. Η συσταδοποίηση οργανώνει τα δεδομένα σε συστάδες-ομάδες,. Χρησιμοποιεί τον αλγόριθμο κ-μέσων (k-means), ο οποίος ομαδοποιεί τα δεδομένα σύμφωνα με το κέντρο της κάθε συστάδας.

Στα επόμενα κεφάλαια, αφού ορίσουμε την κάθε τεχνική, στη συνέχεια θα αναλύσουμε και από ένα παράδειγμα, ώστε να γίνει πιο κατανοητό στους αναγνώστες. Θα δούμε κάθε μία πώς εφαρμόζεται κυρίως στο μάρκετινγκ, αλλά και σε άλλους τομείς, καθώς και κάποια εργαλεία που χρησιμοποιεί η καθεμιά.

2.3 ΓΙΑΤΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Το ανταγωνιστικό περιβάλλον που υπάρχει στους κλάδους των επιχειρήσεων καθιστά αναγκαία την εφαρμογή νέων τρόπων ώστε μια επιχείρηση να κερδίσει ένα ανταγωνιστικό πλεονέκτημα. Αυτό θα επιτευχθεί με την ικανοποίηση των πελατών, με την επαρκή ποσότητα των προϊόντων στα ράφια, με τις χαμηλότερες τιμές σε σχέση με τους ανταγωνιστές και άλλα. Όλα αυτά επιτυγχάνονται σε ένα βαθμό με την βοήθεια της εξόρυξη γνώσης σε σχέση με τα προϊόντα, τις τιμές τους και γενικά για το τι αγοράζει και τι δεν αγοράζει ο καταναλωτής και πως το αγοράζει. Έτσι βελτιώνεται το μάρκετινγκ της επιχείρησης και οι πωλήσεις της.

«Οι επιχειρήσεις όλων των μεγεθών μιμούνται τις μικρές, προσανατολισμένες στις υπηρεσίες επιχειρήσεων, δημιουργώντας σχέσεις με τους πελάτες τους» [9]. Για παράδειγμα ένα παντοπωλείο σε ένα χωριό, γνωρίζει καλύτερα τι θέλουν οι πελάτες γιατί έχει λιγότερες συναλλαγές. Επειδή σε μία μεγάλη επιχείρηση δεν είναι εύκολο να αναγνωριστεί τι ακριβώς θέλει ο πελάτης, η επιχείρηση εφαρμόζει την εξόρυξη δεδομένων.

2.4 ΔΙΑΔΙΚΑΣΙΑ

Στην αναφορά μας στον κλάδο της εξόρυξης δεδομένων, θεωρούμε πως είναι χρήσιμο, να δώσουμε τα βήματα της λειτουργίας της εξόρυξης δεδομένων, ώστε να γίνει πιο κατανοητή. Η διαδικασία της εξόρυξης δεδομένων είναι επαναληπτική, κατά την οποία, τα μοντέλα και οι τεχνικές διαδραματίζουν καθοριστικό ρόλο.

Το πρώτο βήμα είναι ο ορισμός των στόχων. Σε αυτό το βήμα λαμβάνονται αποφάσεις και έτσι γνωρίζονται κάποια πράγματα. Αναφέρεται το πρόβλημα και οι στόχοι. Το δεύτερο βήμα είναι η συλλογή δεδομένων από διάφορες πηγές και η ενσωμάτωση. Επειδή, τα δεδομένα δεν θα είναι μόνο από μια πηγή, είναι απαραίτητο να συνδυαστούν και να ενσωματωθούν. Το τρίτο βήμα είναι η διερευνητική ανάλυση. Σε αυτό το στάδιο γίνεται μια διερεύνηση των δεδομένων για πιθανά λάθη, όπως οι ημερομηνίες γέννησης να είναι αποδεκτές ή οι πωλήσεις να μην έχουν αρνητικά ποσά. Το τέταρτο βήμα είναι η επιλογή του χαρακτηριστικού. Αφαιρούνται εκείνα που χρησιμοποιούνται λιγότερο, ώστε να μην υπάρχουν περιττές πληροφορίες. Επίσης τα νέα χαρακτηριστικά των αρχικών μεταβλητών υπάρχουν και αυτά. Το πέμπτο βήμα είναι η ανάπτυξη του μοντέλου και η επαλήθευση. Τέλος η πρόβλεψη και ερμηνεία. Το μοντέλο που επιλέγεται πρέπει να πραγματοποιηθεί και να χρησιμοποιηθεί για να επιτύχει το πρώτο βήμα των στόχων [11].

Τα κυριότερα στοιχεία της διαδικασίας εξόρυξης δεδομένων είναι: Η εύρεση των δεδομένων από διαφορές πηγές, η μετατροπή τους σε χρήσιμες πληροφορίες και η φόρτωση δεδομένων συναλλαγών από και προς τις αποθήκες δεδομένων, η αποθήκευση και διαχείρισή τους σε μια πολυδιάστατη βάση δεδομένων, η πρόσβαση στα δεδομένα σε αναλυτές και

επαγγελματίες IT, η ανάλυση δεδομένων μέσω κατάλληλου λογισμικού εφαρμογών και τέλος η παρουσίαση αποτελεσμάτων (γραφικές απεικονίσεις, πίνακες, κτλ) [12] .

Τα εργαλεία της εξόρυξης δεδομένων επεξεργάζονται πληροφορίες ώστε να φτιάξουν συγκεκριμένα μοτίβα μέσω κάποιων δραστηριοτήτων. Οι δράσεις αυτές χωρίζονται σε τρείς κατηγορίες. Στην κατηγορία εξερεύνησης (Discovery), κατά την οποία γίνεται μια εύρεση κρυμμένων μοτίβων σε βάσεις δεδομένων χωρίς να υπάρχει μια καθορισμένη ιδέα ή υπόθεση, σχετικά με το τι μοτίβα μπορεί να είναι. Η δεύτερη κατηγορία είναι η πρόβλεψη μοντέλου (Predictive Modeling) όπου αναπτύσσονται μοτίβα από τις βάσεις και σύμφωνα με αυτά γίνεται μια μελλοντική πρόβλεψη. Τέλος η τρίτη κατηγορία είναι η Forensic ανάλυση στην οποία εφαρμόζονται τα μοτίβα ώστε να βρεθούν ανώμαλα ή ασυνήθιστα μοτίβα [10] .

2.5 ΕΦΑΡΜΟΓΕΣ

Οι εφαρμογές της εξόρυξης γνώσης θα μπορούσαμε να πούμε ότι είναι πολλές, αφού τις συναντάμε σχεδόν σε όλους τους τομείς. Ένας τομέας που εφαρμόζεται η εξόρυξη δεδομένων είναι το Σχεσιακό Μάρκετινγκ. Παραδείγματα αυτού είναι η ταυτοποίηση τμημάτων του πελάτη στις εκστρατείες μάρκετινγκ, η πρόβλεψη του ποσοστού των θετικών απαντήσεων αυτών των εκστρατειών καθώς και η ανάλυση των προϊόντων που αγοράζονται μαζί από τους πελάτες, η λεγόμενη ανάλυση καλαθιού αγοράς (market basket analysis) η οποία παρουσιάζεται εκτενέστερα στο τέταρτο κεφάλαιο [11] .

Η εξόρυξη δεδομένων δεν εφαρμόζεται μόνο στο τομέα των επιχειρήσεων, αλλά και σε άλλους τομείς. Άλλες εφαρμογές, είναι η ιατρική διάγνωση, στην οποία οι θεράποντες ιατροί χρησιμοποιούν την εξόρυξη δεδομένων, για την αξιολόγηση της κατάστασης των ασθενών, την πορεία της θεραπείας και για νέες πρακτικές. Χαρακτηριστικό παράδειγμα είναι η Mayo κλινική, η οποία χρησιμοποιεί το IBM ώστε να αντιληφθεί την πορεία της θεραπείας σε ασθενείς με ίδια ηλικία, ίδια χαρακτηριστικά ασθενείας και ιατρικό ιστορικό. Επίσης, χρησιμοποιείται στην γενετική, ώστε να διαγνωστεί κατά πόσο η αλλαγή του DNA ενός ανθρώπου μπορεί να δημιουργήσει την εμφάνιση ασθενειών, αλλά και να ταυτοποιηθούν συγκεκριμένες σειρές DNA. Ένα αντιπροσωπευτικό δείγμα εφαρμογής, της εξόρυξης γνώσης στην γενετική είναι το [13] , στο οποίο δίνεται μια λεπτομερής περιγραφή.

Ένα άλλο χαρακτηριστικό παράδειγμα είναι η εφαρμογή της εξόρυξης δεδομένων σε ποδοσφαιρικούς και άλλων αθλημάτων αγώνες. Οι προπονητές σύμφωνα με στατιστική ανάλυση και εικόνες αναγνωρίζουν προβλήματα σύμφωνα με την σύνθεση των παιχτών και την επίδοσή τους ανάλογα με την θέση στον αγωνιστικό χώρο (άμυνα ή επίθεση). Αυτό μπορεί να το δει κάποιος πιο λεπτομερώς στο [14], στο οποίο χρησιμοποιούνται νευρωνικά δίκτυα για την ανάλυση ενός αγώνα μπάσκετ.

Ακόμη, χρησιμοποιείται στην ηλεκτρολογία για την επίβλεψη σε κυκλώματα όσον αφορά την μόνωσή τους. Σε έρευνες εκπαίδευσης σχετικά με το τι μπορεί να αποσυντονίσει τον μαθητή/σπουδαστή. Σε ανίχνευση απάτης πιστωτικής κάρτας ή σε τηλεφωνικές επικοινωνίες, κατά την οποία σύμφωνα με στατιστική ανάλυση των δεδομένων φαίνεται μια κλοπή που έχει γίνει. Στην αξιολόγηση κινδύνων, δηλαδή κατά πόσο μια απόφαση έχει μεγάλο ρίσκο, σε αυτό βοηθάει και η πρόβλεψη για το τι μπορεί να συμβεί αν παρθεί μια συγκεκριμένη απόφαση. Τέλος, εφαρμόζεται στην εξόρυξη γνώσης από κείμενα, αναγνώριση εικόνας ή εξόρυξη στο παγκόσμιο ιστό [11] .

Τέλος, θα πρέπει να πούμε ότι μια λεπτομερής αναφορά των εφαρμογών της εξόρυξης δεδομένων, είναι το [15] στο οποίο για κάθε τεχνική δίνεται η εφαρμογή της και βιβλιογραφικές αναφορές τα οποία θα μπορούσε να δει κάποιος.

3 ΤΑΞΙΝΟΜΗΣΗ (CLASSIFICATION)

3.1 ΟΡΙΣΜΟΣ ΤΑΞΙΝΟΜΗΣΗΣ

Μία από τις τεχνικές εξόρυξης δεδομένων είναι η ταξινόμηση ή κατηγοριοποίηση (classification). Ταξινόμηση είναι η διαδικασία εκμάθησης μιας συνάρτησης στόχου (target function) f (μοντέλο) που απεικονίζει κάθε σύνολο γνωρισμάτων x σε μια από τις προκαθορισμένες ετικέτες κλάσεις [16]. Η ταξινόμηση αναφέρεται στην κατηγοριοποίηση των δεδομένων σε μία ή περισσότερες κατηγορίες και βασίζεται στην ήδη υπάρχουσα γνώση για να κατηγοριοποιήσει νέα δεδομένα. Είναι μια τεχνική μάθησης με επίβλεψη ή αλλιώς ένα μοντέλο πρόβλεψης, το οποίο σαν στόχο έχει να κατηγοριοποιήσει δεδομένα, για τη πρόβλεψη ενός γεγονότος. Εφαρμόζοντας την ταξινόμηση στην διαδικασία του μάρκετινγκ η επιχείρηση στοχεύει στην δημιουργία ενός μοντέλου το οποίο θα προβλέπει την μελλοντική συμπεριφορά των καταναλωτών της, ταξινομώντας αναφορές πωλήσεων σε έναν αριθμό από προκαθορισμένες κατηγορίες, με βάση ορισμένα κριτήρια [17].

3.2 ΤΑΞΙΝΟΜΗΣΗ ΠΡΟΪΟΝΤΩΝ ΣΤΗΝ ΠΡΟΩΘΗΣΗ ΑΓΑΘΩΝ

Η ταξινόμηση είναι πολύ σημαντική για την διαδικασία του μάρκετινγκ. Μέσω αυτής είναι αρκετά πιο εύκολο να βγουν συμπεράσματα για την αγοραστική συμπεριφορά των καταναλωτών, αλλά και για μελλοντικές αγορές τους. Αυτό γίνεται με την κατηγοριοποίηση των καταναλωτών, αλλά και των προϊόντων. Έτσι οι στρατηγικές μπορεί να γίνουν πιο αποτελεσματικές. Αυτό συμβαίνει γιατί υπάρχουν μικρότερα κομμάτια (κλάσεις) να αντιμετωπιστούν, αλλά και γιατί η στρατηγική που θα επιλεχθεί θα στοχεύει σε μια συγκεκριμένη κατηγορία. Τέλος, με την κατηγοριοποίηση επιλέγεται η κατάλληλη στρατηγική, η οποία θα είναι η καλύτερη για μια συγκεκριμένη κλάση. Με αυτόν τον τρόπο οδηγούμαστε στην καλύτερη εφαρμογή του μάρκετινγκ.

Η ταξινόμηση προϊόντων σε κατηγορίες βοηθά τους προμηθευτές, να αποφασίζουν στρατηγικές και μεθόδους που θα προωθήσουν έτσι το προϊόν ή την υπηρεσία της επιχείρησης. Υπάρχουν διάφοροι τύποι ταξινόμησης. Οι επιχειρήσεις επικεντρώνουν τις προσπάθειές τους, με βάση τις αγορές των πελατών. Στη συνέχεια με βάση τις αγορές, η επιχείρηση σχεδιάζει το μάρκετινγκ, για να καθορίσει την ομάδα-στόχο. Κατά τον Scott Dacko, υπάρχουν κάποιες ταξινομήσεις προϊόντων : Η πρώτη είναι η ευκολία εμπορευμάτων, η οποία αναφέρεται στα προϊόντα, τα οποία οι πελάτες αγοράζουν συχνά. Τέτοια παραδείγματα είναι το γάλα, τα μακαρόνια, τα δημητριακά, δηλαδή προϊόντα τα οποία είναι απαραίτητα. Μία δεύτερη ταξινόμηση, είναι οι αγορές εμπορευμάτων. Αυτή η ταξινόμηση γίνεται με βάση τις παρατηρήσεις των πελατών στην τιμή, την ποιότητα και φυσικά στην αξία που προσφέρει σε αυτούς το αντίστοιχο προϊόν. Δηλαδή, στο πόσο ακριβό θεωρούν οι καταναλωτές ότι είναι το κάθε προϊόν, τι ποιότητα προσφέρεται σε αυτούς και συνολικά στην εικόνα του προϊόντος. Επίσης τα εξειδικευμένα προϊόντα τα οποία αναφέρονται στις μάρκες προϊόντων (brandnames). Τέτοια παραδείγματα είναι τα ρολόγια, οι υπολογιστές ή τα αυτοκίνητα Ferrari. Τέλος τα αζήτητα εμπορεύματα. Σε αυτή την περίπτωση οι καταναλωτές δεν έχουν την τάση να αγοράζουν κάποια προϊόντα, όπως οι μπαταρίες [18].

3.3 ΠΡΟΒΛΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

Ο σκοπός των μοντέλων ταξινόμησης είναι να προσδιορίσουν τις επαναλαμβανόμενες σχέσεις μεταξύ των επεξηγηματικών μεταβλητών που περιγράφουν τα παραδείγματα που ανήκουν στην ίδια κατηγορία. Μετά έρχονται οι κανόνες κατάταξης που χρησιμοποιούνται για να προβλέψουν την κλάση των παραδειγμάτων για τα οποία είναι γνωστές μόνο οι τιμές των επεξηγηματικών χαρακτηριστικών. Υπάρχουν 3 συνιστώσες του προβλήματος

ταξινόμησης: μία γεννήτρια παρατηρήσεων (Generator), ένας επιβλέπων (Supervisor) της τάξης στόχου και ένας αλγόριθμος ταξινόμησης (Algorithm). Όσον αφορά τη γεννήτρια, εξάγει τυχαία διανύσματα x των παραδειγμάτων σύμφωνα με μία άγνωστη πιθανότητα Px(x). Ο επόπτης, επιστρέφει για κάθε διάνυσμα x των παραδειγμάτων την αξία της κλάσης στόχου σύμφωνα με μία κατανομή $P_Y|X(y|x)$ η οποία είναι άγνωστη. Ο αλγόριθμος ταξινόμησης AF, λέγεται επίσης και ταξινομητής (Classifier) (μία τάξη) επιλέγει μία συνάρτηση $f^* \in F$, ώστε να ελαχιστοποιηθεί η κατάλληλη καθορισμένη απώλεια συνάρτησης [11] .

Για την εφαρμογή της ταξινόμησης και την εξαγωγή χρήσιμων συμπερασμάτων χρησιμοποιούνται κάποια εργαλεία-αλγόριθμοι. Τα πιο συχνά εργαλεία που χρησιμοποιούντα σε αυτήν την τεχνική είναι, τα νευρωνικά δίκτυα τα οποία αποτελούνται από νευρώνες όπως οι νευρώνες του ανθρώπινου εγκεφάλου. Τα νευρωνικά δίκτυα έχουν σαν είσοδο δεδομένα, τα οποία επεξεργάζονται και εξάγουν χρήσιμες πληροφορίες. Επίσης χρήσιμα, τα δέντρα αποφάσεων τα οποία θέτουν μια σειρά ερωτήσεων με σκοπό την επίλυση τους. Η επίλυση ενός δέντρου αποφάσεων μπορεί να είναι είτε σε μορφή γραφικής αναπαράστασης είτε με την μορφή κανόνων. Τέλος εργαλεία της ταξινόμησης είναι και ο αλγόριθμος κοντινότερου γείτονα, ο οποίος κατατάσσει ένα σημείο ανάλογα με τις συντεταγμένες που έχουν τα γειτονικά του σημεία , ο Naives Bayes ταξινομητής, τα Bayesian Belief δίκτυα και οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines). Στις επόμενες υποενότητες, θα αναφέρουμε κάποια από αυτά.

3.3.1 ΤΑΞΙΝΟΜΗΤΗΣ BAYES

Ο ταξινομητής Bayes βασίζεται σε ένα συγκεκριμένο θεώρημα. Το Θεώρημα Bayes, μας λέει ότι: Έχουμε ένα πρότυπο x=[x(1), x(2), … , x(d)] το οποίο είναι γνωστό. Θα πρέπει να ταξινομηθεί σε μία από τις k κατηγορίες C_1, C_2 ,…, C_K, έχοντας την υψηλότερη εκ των υστέρων δεσμευμένη πιθανότητα. Ισχύει ότι

$(C_i|x)\ p(x) = p(x|C_i) \cdot P(C_i)$, όπου:

$P(C_i)$ είναι η εκ των προτέρων πιθανότητα της κλάσης C_i

$p(x)$ είναι η συνάρτηση πυκνότητας πιθανότητας του γεγονότος x

$p(x|C_i)$ είναι η δεσμευμένη πυκνότητα πιθανότητας του γεγονότος x δοθέντος της της κλάσης C_i και

$P(C_i \mid x)$: η εκ των υστέρων πιθανότητα της κλάσης C_i δοθέντος του x.

Προβλέπεται η κλάση του x αν και μόνο αν $P(C_i \mid x) > P(C_j \mid x)$

Αλλά $P(C_i|x) = P(x|C_i)\ P(C_i) / p(x)$. Επειδή p(x) είναι σταθερό για όλες τις κλάσεις, μόνο το $P(x|C_i)\ P(C_i)$ πρέπει να μεγιστοποιηθεί.

Τα βήματα του αλγορίθμου είναι τα εξής:

(α) Υπολογισμός της εκ των προτέρων πιθανότητας της κλάσης C_i , $P(C_i)$:

Εάν είναι άγνωστες συχνά υποθέτουμε πώς όλες οι κλάσεις είναι ισοπίθανες , δηλαδή $P(C_1) = P(C_2) = \ldots = P(C_m)$ και έτσι μόνο η $P(x|C_i)$ πρέπει να μεγιστοποιηθεί.

Αλλιώς υπολογίζεται από τον τύπο $P(C_i) = |C_{i,D}| / D$, όπου $|C_{i,D}|$ το πλήθος των προτύπων του D που ανήκουν στην κλάση C_i και |D| το πλήθος των προτύπων του συνόλου δεδομένων.

(β) Υπολογισμός της $P(x|C_i) = P(x_1|C_i) \cdot P(x_2|C_i) \cdot \ldots \cdot P(x_d|C_i)$ [19] .

Ακολουθεί ένα παράδειγμα με τη βοήθεια του πίνακα [20].

Α/Α	**Ηλικία**	**Εισόδημα**	**Μαθητής**	**Πιστωτική εκτίμηση**	**Κλάση: αγορά υπολογιστή**
1	Νεολαία	Υψηλό	Όχι	Δίκαιη	Όχι
2	Νεολαία	Υψηλό	Όχι	Άριστη	Όχι
3	Μεσήλικας	Υψηλό	Όχι	Δίκαιη	Ναι
4	Γηραιότερος	Μεσαίο	Όχι	Δίκαιη	Ναι
5	Γηραιότερος	Χαμηλό	Ναι	Δίκαιη	Ναι
6	Γηραιότερος	Χαμηλό	Ναι	Άριστη	Όχι
7	Μεσήλικας	Χαμηλό	Ναι	Άριστη	Ναι
8	Νεολαία	Μεσαίο	Όχι	Δίκαιη	Όχι
9	Νεολαία	Χαμηλό	Ναι	Δίκαιη	Ναι
10	Γηραιότερος	Μεσαίο	Ναι	Δίκαιη	Ναι
11	Νεολαία	Μεσαίο	Ναι	Άριστη	Ναι
12	Μεσήλικας	Μεσαίο	Όχι	Άριστη	Ναι
13	Μεσήλικας	Υψηλό	Ναι	Δίκαιη	Ναι
14	Γηραιότερος	Μεσαίο	Όχι	Άριστη	Όχι

Πίνακας 1 Παράδειγμα Ταξινομητή Bayes

Στο πίνακα εμφανίζονται τα εξής χαρακτηριστικά:

- Η ηλικία, που αφορά τα άτομα και μπορεί να λάβει τιμές «Νεολαία», «Μεσήλικες» και «Γηραιότερα»,
- Το εισόδημα με τιμές «Χαμηλό», «Μεσαίο» και «Υψηλό»,
- Ο μαθητής, που λαμβάνει τιμή «Ναι» ή «Όχι»,
- Η πιστωτική εκτίμηση, που μπορεί να είναι είτε «Δίκαιη» είτε «Άριστη», καθώς και
- Η «κλάση αγορά υπολογιστή» με τιμές «Ναι» ή «Όχι» ανάλογα με το αν θα αγοράσει υπολογιστή ή όχι, με τις αντίστοιχες τιμές τους κάθε φορά.

Θα θέλαμε να βρούμε την κλάση του προτύπου X = (Ηλικία = Νεολαία , Εισόδημα = Μεσαίο, Μαθητής = Ναι, Πιστωτική Εκτίμηση = Δίκαιη).
Οι κλάσεις C_i είναι δύο: $C_1 \rightarrow$Ναι και $C_2 \rightarrow$Όχι. Η κλάση C_1 είναι η υπόθεση ένας πελάτης να αγοράσει έναν υπολογιστή (Αγορά υπολογιστή = Ναι). Η κλάση C_2 είναι η υπόθεση ένας πελάτης να μην αγοράσει έναν υπολογιστή (Αγορά υπολογιστή = Όχι). Θα πρέπει να μεγιστοποιήσουμε το γινόμενο $P(X|C_i)P(C_i)$, για i= 1,2. Ξεκινάμε υπολογίζοντας τις πιθανότητες:
Η $P(C_i)$ είναι η εκ των προτέρων πιθανότητα της κάθε κλάσης: Η εκ των προτέρων πιθανότητα ένας πελάτης να αγοράσει έναν υπολογιστή (Αγορά υπολογιστή = Ναι) είναι :
$P(C_1) = 9 /14 = 0.643$. Η εκ των προτέρων πιθανότητα ένας πελάτης να μην αγοράσει έναν υπολογιστή (Αγορά υπολογιστή = Όχι) είναι : $P(C_2) = 5 /14 = 0.357$.
Στη συνέχεια υπολογίζουμε τις δεσμευμένες πιθανότητες $P(X|C_i)$:
P (Ηλικία = Νεολαία | C_1) = 2 / 9 = 0.222.
P (Ηλικία = Νεολαία | C_2) = 3 / 5 = 0.6.
P (Εισόδημα = Μεσαίο | C_1) = 4 / 9 = 0.444.
P (Εισόδημα = Μεσαίο | C_2) = 2 / 5 = 0.4.
P (Μαθητής = Ναι | C_1) = 6 / 9 = 0.667.
P (Μαθητής = Ναι | C_2) = 1 / 5 = 0.2.
P (Πιστωτική Εκτίμηση = Δίκαιη | C_1) = 6 / 9 = 0.667.

P (Πιστωτική Εκτίμηση = Δίκαιη | C_2) = 2 / 5 = 0.4.
Επίσης: P(X|C_1) = P (Ηλικία = Νεολαία | C_1) · P (Εισόδημα = Μεσαίο | C_1) ·
P (Μαθητής = Ναι | C_1) · P (Πιστωτική Εκτίμηση = Δίκαιη | C_1) = 0.222 · 0.444 · 0.667 · 0.667 = 0.044
P(X|C_2) = P (Ηλικία = Νεολαία | C_2) · P ((Εισόδημα = Μεσαίο | C_2) · P (Μαθητής = Ναι | C_2) · P (Πιστωτική Εκτίμηση = Δίκαιη | C_2) = 0.6 · 0.4 · 0.2 · 0.4 = 0.019.
Τέλος βρίσκουμε την κλάση που μεγιστοποιεί το γινόμενο P(X|C_i)P(C_i):
P(X|C_1)P(C_1) = 0.044 · 0.643 = 0.028.
P(X|C_2)P(C_2) = 0.019 · 0.357 = 0.007.
Επειδή P(X|C_1)P(C_1) > P(X|C_2)P(C_2), το πρότυπο X θα ταξινομηθεί στην κατηγορία C_1, δηλαδή: "Αγορά υπολογιστή" = "Ναι" . Που σημαίνει ότι ένας καταναλωτής ο οποίος είναι νεαρός, έχει εισόδημα μεσαίο, είναι μαθητής και η πιστωτική του κάρτα είναι δίκαιη, είναι πιθανόν να αγοράσει έναν ηλεκτρονικό υπολογιστή.

3.3.2 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ

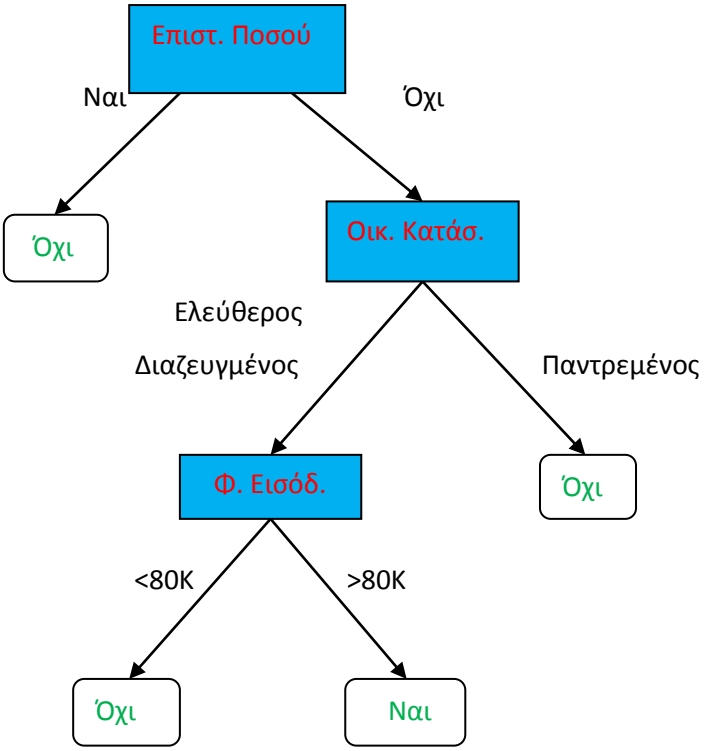
Τα δέντρα αποφάσεων (decision trees) είναι μία απλή και διαδεδομένη τεχνική ταξινόμησης, επιλέγοντας κάθε φορά τη διαδρομή που χρειάζεται κάθε φορά. Ακολουθεί ένα παράδειγμα αυτού: Θέλουμε να προβλέψουμε εάν ένα άτομο είναι απατεώνας ή όχι. Έχουμε έναν πίνακα με τα δεδομένα τα οποία συγκεντρώνονται στον πίνακα 2.

Α/Α	**Επιστροφή Ποσού**	**Οικογενειακή Κατάσταση**	**Φορολογητέο Εισόδημα**	**Κλάση:Απάτη**
1	Ναι	Ανύπανδρος	125K	Όχι
2	Όχι	Παντρεμένος	100K	Όχι
3	Όχι	Ανύπανδρος	70K	Όχι
4	Ναι	Παντρεμένος	120K	Όχι
5	Όχι	Διαζευγμένος	95K	Ναι
6	Όχι	Παντρεμένος	60K	Όχι
7	Ναι	Διαζευγμένος	220K	Όχι
8	Όχι	Ανύπανδρος	85K	Ναι
9	Όχι	Παντρεμένος	75K	Όχι
10	Όχι	Ανύπανδρος	90K	Ναι

Πίνακας 2 Παράδειγμα Δέντρου Αποφάσεων

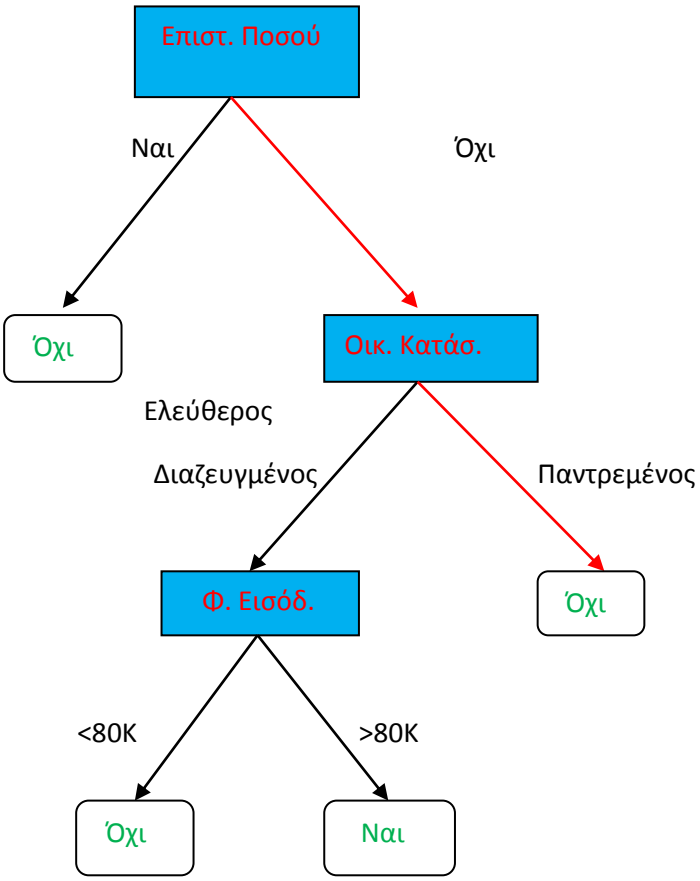
Στον πίνακα εμφανίζονται τα χαρακτηριστικά:

- Η επιστροφή ποσού, που λαμβάνει τις τιμές «Ναι» και «Όχι»,
- Η οικογενειακή κατάσταση, η οποία χωρίζεται σε «Διαζευγμένος», «Ανύπανδρος» και «Παντρεμένος»,
- Το φορολογητέο εισόδημα με τις ανάλογες τιμές και τέλος
- Η κλάση «απάτη», λαμβάνει τις τιμές «Ναι» και «Όχι», ανάλογα με τις ανάλογες τιμές που παίρνει κάθε φορά. Η μετατροπή του παραπάνω πίνακα σε δέντρο απόφασης παρουσιάζεται στην εικόνα 2.



Εικόνα 2 Μετατροπή πίνακα σε δέντρο απόφασης

Το χαρακτηριστικό Επιστροφή Ποσού είναι ο αρχικός κόμβος προς διαχωρισμό, ή αλλιώς η λεγόμενη ρίζα του δέντρου. Τα χαρακτηριστικά οικογενειακή κατάσταση και φορολογητέο εισόδημα ονομάζονται εσωτερικοί κόμβοι, όπου και εκεί γίνεται διαχωρισμός. Οι τιμές ναι και όχι είναι το τελευταίο επίπεδο του δέντρου, ή αλλιώς τα λεγόμενα φύλλα του. Θέλουμε τώρα να εξετάσουμε την περίπτωση όπου «Επιστροφή Ποσού = Όχι» , «Οικογενειακή Κατάσταση = Παντρεμένος» , «Φορολογητέο Εισόδημα = 80K" και ψάχνουμε να βρούμε εάν έχει εξαπατήσει ή όχι. Στην εικόνα 3 φαίνεται με τα κόκκινα βέλη η διαδρομή που έχει ακολουθηθεί μέχρι να απαντήσουμε στο ερώτημα.



Εικόνα 3 Διαδρομή και λύση

Επομένως βλέπουμε ότι το χαρακτηριστικό απάτη παίρνει την τιμή όχι , δηλαδή «Απάτη = Όχι». Δηλαδή κάποιος με «Επιστροφή Ποσού = Όχι» , «Οικογενειακή Κατάσταση = Παντρεμένος» , «Φορολογητέο Εισόδημα = 80K», δεν έχει απατήσει. [21] .

Ο πιο διαδεδομένος αλγόριθμος που χρησιμοποιείται για την δημιουργία δέντρων αποφάσεων είναι ο ID3 (Iterative Dichotomiser 3), ο οποίος βασίζεται σε μια επιστημονική αρχή, στο ξυράφι ή Λεπίδα του Όκαμ, αναφέρεται στο ότι: "Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από όσες είναι απαραίτητες και αποτελείται από έναν μηχανισμό διαχωρισμού, την εντροπία της πληροφορίας, ο οποίος επιλέγει εκείνη την ανεξάρτητη μεταβλητή που οδηγεί σε περισσότερο συμπαγές δέντρο". Ο ID3 περιγράφεται ως εξής:

```
1.  Βρες την ανεξάρτητη μεταβλητή η οποία αν χρησιμοποιηθεί
ως κριτήριο διαχωρισμού των δεδομένων εκπαίδευσης θα οδηγήσει
σε κόμβους κατά το δυνατό διαφορετικούς σε σχέση με την
εξαρτημένη μεταβλητή .
2.  Κάνε τον διαχωρισμό.
3.  Επανέλαβε την διαδικασία για κάθε έναν από τους κόμβους
που προέκυψαν μέχρι να μην είναι δυνατός περαιτέρω
διαχωρισμός.
```

Αλγόριθμος 1 ID3

Πιο αναλυτικά ο ID3 είναι ένας άπληστος αλγόριθμος ο οποίος κατασκευάζει ένα δέντρο από πάνω προς τα κάτω. Χρησιμοποιεί το λεγόμενο Κέρδος πληροφορίας (Information Gain) ή σε συντομία Gain για να επιλέξει το καλύτερο χαρακτηριστικό σαν αρχικό κόμβο ή κόμβος αποφάσεων, όπου το μέγιστο κέρδος είναι η υψηλότερη πληροφορία κέρδους για να κάνουμε τον διαχωρισμό. Χρησιμοποιεί την λεγόμενη Εντροπία (Entropy) για

να μετρήσει την πληροφορία και συγκεκριμένα την τυχαιότητα σε ένα σύνολο δεδομένων, δηλαδή την ομοιογένειά τους. Υπολογίζεται ως εξής:

Entropy(S) = - $P_{\theta\varepsilon\tau\iota\kappa\acute{o}}$ $\log_2 P_{\theta\varepsilon\tau\iota\kappa\acute{o}}$ – $P_{\alpha\rho\nu\eta\tau\iota\kappa\acute{o}}$ $\log_2 P_{\alpha\rho\nu\eta\tau\iota\kappa\acute{o}}$, όπου:

S = σύνολο δεδομένων

$P_{\theta\varepsilon\tau\iota\kappa\acute{o}}$ = ποσοστό θετικών δεδομένων

$P_{\alpha\rho\nu\eta\tau\iota\kappa\acute{o}}$ = ποσοστό αρνητικών δεδομένων.

Για παράδειγμα εάν το S είναι ένα σύνολο 14 δεδομένων με 9 "Ναι" και 5 "Όχι" , τότε Entropy(S) = $-\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log 2\left(\frac{5}{14}\right) = 0.940$.

Η απάντηση θα είναι 0 ή 1. Εάν είναι 0, σημαίνει ότι όλα τα δεδομένα ανήκουν σε μία μόνο μία κατηγορία. Εάν είναι 1, τότε τα δεδομένα ανήκουν σε περισσότερες κατηγορίες.

Ακολουθεί ένα πιο ολοκληρωμένο παράδειγμα εφαρμογής του αλγορίθμου, σύμφωνα με τον πίνακα 3:

Α / Α	**Ύψος**	**Τύπος μαλλιών**	**Χρώμα ματιών**	**Κλάση**
1	Κοντός	Ξανθά	Μπλε	+
2	Ψηλός	Ξανθά	Καφέ	–
3	Ψηλός	Κόκκινα	Μπλε	+
4	Ψηλός	Σκούρα	Καφέ	–
5	Κοντός	Σκούρα	Μπλε	–
6	Ψηλός	Σκούρα	Μπλε	–
7	Ψηλός	Ξανθά	Μπλε	+
8	Κοντός	Ξανθά	Καφέ	–

Πίνακας 3 Δεδομένα αλγορίθμου ID3

Σύμφωνα με το παραπάνω σύνολο δεδομένων θέλουμε να δημιουργήσουμε το αντίστοιχο δέντρο απόφασης. Στην αρχή υπολογίζουμε την εντροπία Ε(S) όλων των δεδομένων. Έτσι: Ε (3+ , 5-) = $-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.954$. Στη συνέχεια εξετάζουμε το χαρακτηριστικό ύψος, με τη βοήθεια του πίνακα 4:

Ύψος	**+**	**–**	**Σύνολο**
Κοντός	1	2	3
Ψηλός	2	3	5
			8

Πίνακας 4 Δεδομένα χαρακτηριστικού Ύψους

Μετά υπολογίζουμε το κέρδος του: από την συνολική εντροπία αφαιρούμε την εντροπία του Ύψους: Gain (Υψος) = E(S) − E (Υψος)

E (Υψος) = $\frac{3}{8}$ E(Κοντός) + $\frac{5}{8}$ E (Ψηλός)

E (Κοντός) = $-\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = 0.918$

E (Ψηλός) = $\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.97$

Άρα E (Υψος) = $\frac{3}{8}$ E (Κοντός) + $\frac{5}{8}$ E (Ψηλός) = $\frac{3}{8} \times 0.918 + \frac{5}{8} \times 0.97 = 0.9505$, και

Gain (Υψος) = E(S) − E (΄Υψος) = 0.954 – 0.9505 = <u>0.0045</u>

Συνεχίζουμε με το χαρακτηριστικό Τύπο μαλλιών, με τη βοήθεια του πίνακα 5:

Τύπος μαλλιών	+	−	**Σύνολο**
Ξανθά	2	2	4
Κόκκινα	1	0	1
Σκούρα	0	3	3
			8

Πίνακας 5 Δεδομένα χαρακτηριστικού Τύπος μαλλιών

Gain (Τύπος μαλλιών) = E (S) – E (Τύπος μαλλιών)

E (Τύπος μαλλιών) = $\frac{4}{8}$ E (Ξανθά) + $\frac{1}{8}$ E (Κόκκινα) + $\frac{3}{8}$ E (Σκούρα)

E (Ξανθά) = − ($\frac{2}{4}$) $\log_2$ ($\frac{2}{4}$) − ($\frac{2}{4}$) $\log_2$ ($\frac{2}{4}$) = 1

E (Κόκκινα) = − 1 $\log_2$ 1− 0 $\log_2$ 0 = 0

E (Σκούρα) = − 0 $\log_2$ 0− $\frac{3}{3}$ $\log_2$ ($\frac{3}{3}$) = 0

Άρα: E (Τύπος μαλλιών) = $\frac{4}{8}$ E (Ξανθά) + $\frac{1}{8}$ E (Κόκκινα) + $\frac{3}{8}$ E (Σκούρα)

= $\frac{4}{8}$ × 1 + $\frac{1}{8}$ × 0 + $\frac{3}{8}$ × 0 = $\frac{4}{8}$ και

Gain (Τύπος μαλλιών) = E (S) – E (Τύπος μαλλιών) = 0.954 − $\frac{4}{8}$ = <u>0.454</u>

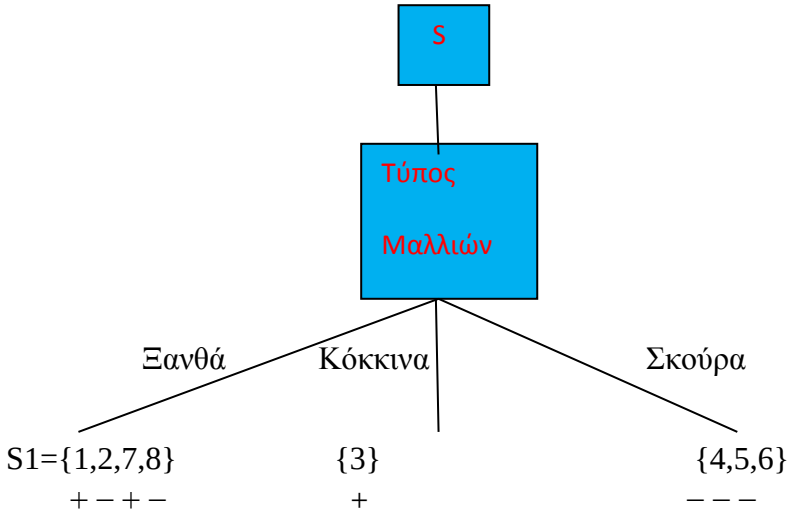
Τέλος υπολογίζουμε το χαρακτηριστικό Χρώμα ματιών, με τη βοήθεια του πίνακα 6:

Χρώμα ματιών	+	−	**Σύνολο**
Μπλε	3	2	5
Καφέ	0	3	3
			8

Πίνακας 6 Δεδομένα χαρακτηριστικού Χρώμα ματιών

Gain (Χρώμα ματιών) = E(S) – E (Χρώμα ματιών)
E(Χρώμα ματιών) = $\frac{5}{8}$ E (Μπλε) + $\frac{3}{8}$ E (Καφέ)
E (Μπλε) = − $\frac{3}{5}$ $\log_2$ ($\frac{3}{5}$) − ($\frac{2}{5}$) $\log_2$ ($\frac{2}{5}$) = 0.970
E (Καφέ) = − $\frac{0}{3}$ $\log_2$ ($\frac{0}{3}$) − ($\frac{3}{3}$) $\log_2$ ($\frac{3}{3}$) = 0 – 0 = 0
Άρα: E(Χρώμα ματιών) = $\frac{5}{8}$ E (Μπλε) + $\frac{3}{8}$ E (Καφέ) = $\frac{5}{8}$ × 0.970 + $\frac{3}{8}$ × 0 = 0.606 και
Gain (Χρώμα ματιών) = E(S) – E (Χρώμα ματιών) = 0.954 – 0.606 = <u>0.348</u>

Τώρα θα συγκρίνουμε και τα τρία Gain και έπειτα διαχωρίζουμε το χαρακτηριστικό με το μεγαλύτερο από τα 3. Έτσι παρατηρούμε ότι το μεγαλύτερο είναι το Gain(Τύπος μαλλιών). Άρα διαχωρίζουμε το χαρακτηριστικό Τύπο μαλλιών, σύμφωνα με την παρακάτω βοηθητική εικόνα 4:



Εικόνα 4: Διαχωρισμός "Τύπος μαλλιών"

$E(S1) = -\frac{2}{4}\log_2(\frac{2}{4}) - (\frac{2}{4})\log_2(\frac{2}{4}) = 1$

Μετά υπολογίζουμε το κέρδος του χαρακτηριστικού Ύψους: από την συνολική εντροπία αφαιρούμε την εντροπία του: Gain (Υψος) = E(S1) − E (Υψος).

Έχουμε τον πίνακα 7:

Ύψος	+	−	**Σύνολο**
Κοντός	1	1	2
Ψηλός	1	1	2
			4

Πίνακας 7 Δεδομένα Ύψους μετά τον διαχωρισμό " Τύπος μαλλιών "

$E(Υψος) = \frac{2}{4}E(Κοντός) + \frac{2}{4}E(Ψηλός)$

$E(Κοντός) = -\frac{1}{2}\log_2(\frac{1}{2}) - (\frac{1}{2})\log_2(\frac{1}{2}) = 1$

$E(Ψηλός) = -\frac{1}{2}\log_2(\frac{1}{2}) - (\frac{1}{2})\log_2(\frac{1}{2}) = 1$ και

$E(Υψος) = \frac{2}{4}E(Κοντός) + \frac{2}{4}E(Ψηλός) = \frac{2}{4}\times 1 + \frac{2}{4}\times 1 = 1$

Άρα: Gain (Υψος) = E(S1) − E (Υψος) = 1 −1 = $\underline{0}$

Συνεχίζουμε με τον υπολογισμό του χαρακτηριστικού Χρώμα ματιών σύμφωνα με τον παρακάτω πίνακα 8:

Χρώμα ματιών	+	−	**Σύνολο**
Μπλε	2	0	2
Καφέ	0	2	2
			4

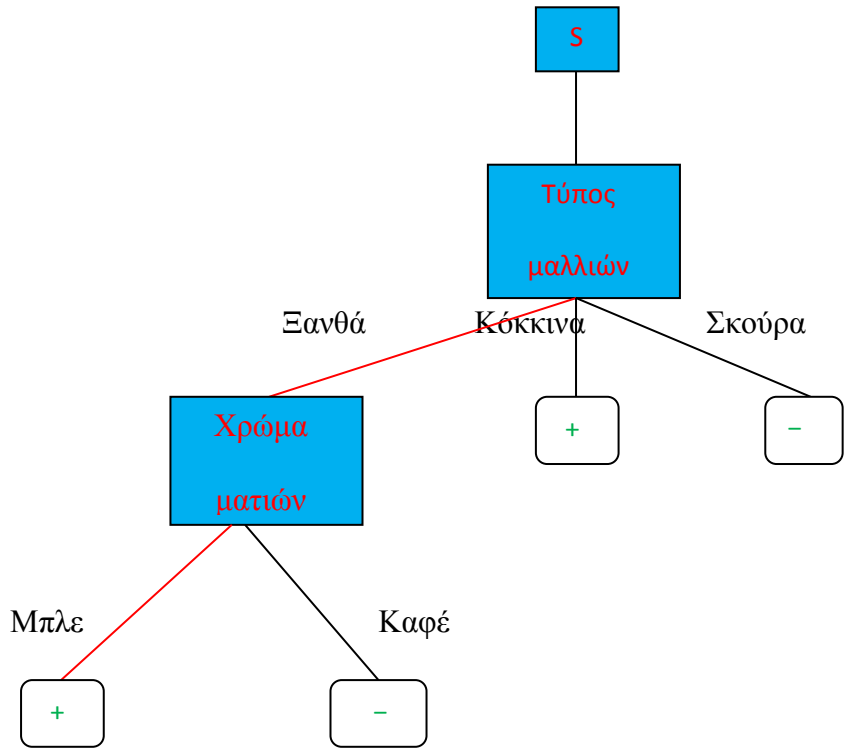
Πίνακας 8 Δεδομένα Χρώμα ματιών μετά τον διαχωρισμό " Τύπος μαλλιών "

$E(Χρώμα\ ματιών) = \frac{2}{4}E(Μπλε) + \frac{2}{4}E(Καφέ)$

$E(Μπλε) = -\frac{2}{2}\log_2(\frac{2}{2}) - (\frac{0}{2})\log_2(\frac{0}{2}) = 0$

$E(Καφέ) = -\frac{0}{2}\log_2(\frac{0}{2}) - (\frac{2}{2})\log_2(\frac{2}{2}) = 0$ και

E (Χρώμα ματιών) = $\frac{2}{4}$ E (Μπλε) + $\frac{2}{4}$ E (Καφέ) = $\frac{2}{4}$ × 0 + $\frac{2}{4}$ × 0 = 0
Άρα Gain(Χρώμα ματιών) = E(S1) − E (Χρώμα ματιών) = 1 – 0 = 1
Καταλήγουμε βλέποντας ότι το μεγαλύτερο κέρδος είναι το Gain (Χρώμα ματιών). Άρα διαχωρίζουμε με βάση το χρώμα ματιών σύμφωνα με την εικόνα 5:



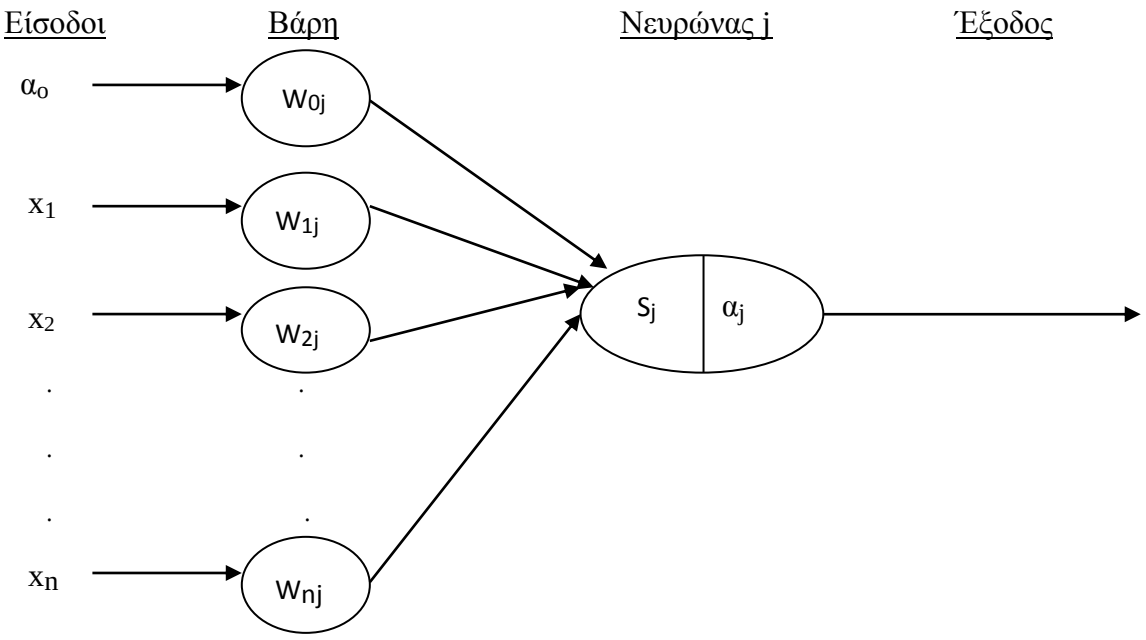
Εικόνα 5 Διαχωρισμός χαρακτηριστικού Χρώμα ματιών

Η διαδικασία ολοκληρώθηκε. Τέλος θέλουμε να κατηγοριοποιήσουμε το σύνολο δεδομένων: X = {Κοντός, Ξανθός, Μπλε,?}. Ακολουθώντας το μονοπάτι που φαίνεται με τις κόκκινες γραμμές, βλέπουμε ότι το X κατηγοριοποιείτε στην κλάση + [22] .

Η επέκταση του ID3 αλγορίθμου είναι ο C4.5 ο οποίος παράγει δέντρα αποφάσεων που μπορούν να χρησιμοποιηθούν για την ταξινόμηση των δεδομένων για αυτό και αναφέρεται ως στατιστικός ταξινομητής (statistical classifier) [23] .

3.3.3 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Τεχνητό νευρωνικό δίκτυο (artificial neural networks) ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων. Είναι μια εύκολη μέθοδος για την εκμάθηση αριθμητικών και διανυσματικών συναρτήσεων, ώστε να προβλεφθούν κάποια γεγονότα. Έχουν το πλεονέκτημα σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή δεδομένα που παρουσιάζουν λάθη, όπως για παράδειγμα λανθασμένη καταχώριση. Όμως δεν είναι εύκολο να εξηγηθεί η γνώση που εξάγουν. Ένα τεχνητό νευρωνικό δίκτυο αποτελείται από νευρώνες όπως και οι νευρώνες του ανθρώπινου εγκεφάλου [23] . Ένα από τα πιο απλά νευρωνικά δίκτυα είναι τα Perceptrons, τα οποία αποτελούνται από μία ή περισσότερες εισόδους (inputs), δηλαδή τα δεδομένα εισόδου (x_1, x_2, …, x_n), συναρτήσεις και εξόδους σε επόμενα επίπεδα ή στην τελική έξοδο, η οποία σε αυτή την περίπτωση θα είναι μία, όπως φαίνεται παρακάτω στην εικόνα 6:



Εικόνα 6 Αναπαράσταση τεχνητού νευρώνα

Οι είσοδοι x_1, x_2, …, x_n έχουν και τα δικά τους βάρη (weight) w_{1j}, w_{2j},…,w_{nj}. Κάθε είσοδος πολλαπλασιάζεται με το βάρος της, δηλαδή $x_1 \times w_{1j}$, $x_2 \times w_{2j,\ldots,}$ $x_n \times w_{nj}$ Επίσης υπάρχει μία ακόμα είσοδος, η α_o η οποία ονομάζεται τάση πόλωσης και έχει πάντα σταθερή τιμή, συνήθως −1 ή 1, η οποία και αυτή πολλαπλασιάζεται με το βάρος της, δηλαδή $\alpha_0 \times w_{0j,}$ Η συνολική είσοδος στον νευρώνα j είναι το άθροισμα όλων των εισόδων, δηλαδή $S_j = \sum_1^n x_n \times w_{nj}$. Η έξοδος του προκύπτει από την εφαρμογή της συνάρτησης ενεργοποίησης στη συνολική του είσοδο S_j σύμφωνα με τον τύπο $a_j=\Phi(Sj)$. Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης. Εμείς θα παρουσιάσουμε την πιο συνηθισμένη, την λεγόμενη βηματική συνάρτηση ενεργοποίησης. Όλους του τύπους κανείς μπορεί να τους μελετήσει στο [24]. Η βηματική συνάρτηση ενεργοποίησης ορίζεται από το τύπο: $\Phi(S) = \{1,$ αν $S > 0$ και 0 αν $S \leq 0\}$. Υπάρχουν οι λεγόμενες λογικές συναρτήσεις όπως η AND, OR, NOT, κα. Εμείς θα αναλύσουμε τη λειτουργία του νευρώνα για τη συνάρτηση AND, σύμφωνα με τη βηματική συνάρτηση ενεργοποίησης, στο παρακάτω παράδειγμα, σύμφωνα με τον πίνακα 9. Πιο αναλυτικά μπορεί να διαβάσει κανείς για αυτές στο [25] .

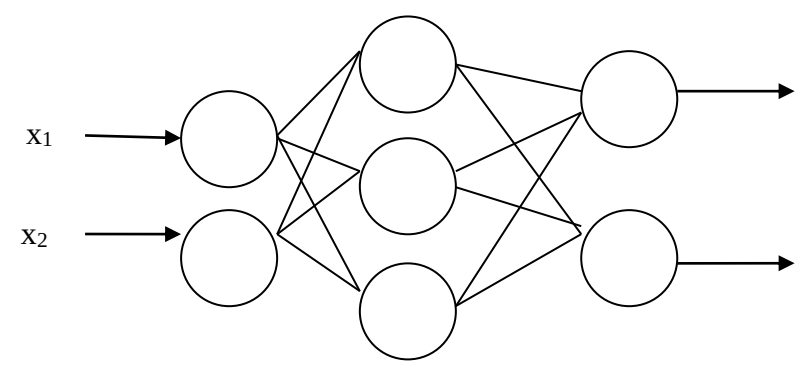
X	**Y**	**Z**	$\mathbf{w_x}$	$\mathbf{w_y}$	$\mathbf{w_z}$	**S**	**Φ(S)**
0	0	1	1	1	−1,5	−1,5	0
0	1	1	1	1	−1,5	−0,5	0
1	0	1	1	1	−1,5	−0,5	0
1	1	1	1	1	−1,5	+0,5	1

Πίνακας 9 Δεδομένα συνάρτησης AND

Πιο συγκεκριμένα το αποτέλεσμα της συνάρτησης AND επιστρέφει 1 αν όλες οι είσοδοι είναι 1. Αν μία είσοδος είναι 0,τότε η συνάρτηση επιστρέφει την τιμή 0. Στο παράδειγμα μας μόνο στην τελευταία περίπτωση το αποτέλεσμα είναι 1, αφού όλες οι είσοδοι έχουν τιμή 1. Έπειτα θα δείξουμε πώς βγήκε το αποτέλεσμα S και το $\Phi(S)$ σε κάθε περίπτωση. Στην αρχή υπολογίζουμε το $x \times w_{x,}$ $y \times w_{y,}$ $z \times w_z$. Στην πρώτη περίπτωση έχουμε: $x \times w_x = 0 \times 1 = 0$, $y \times w_y = 0 \times 1 = 0$, $z \times w_z = 1 \times (-1{,}5) = -1{,}5$. Σύμφωνα με τον

τύπο: $S_j = \sum_1^n x_n \times w_{nj,}$ $S = 0 + 0 + (-1,5) = -1,5$ Επίσης σύμφωνα με τον τύπο: $\Phi(S) = \{1,$ αν $S > 0$ και 0 αν $S \leq 0\}$ και εφόσον $\underline{S = -1,5} < 0$, τότε $\underline{\Phi(S) = 0}$. Στην δεύτερη περίπτωση έχουμε: $x \times w_x = 0 \times 1 = 0$, $y \times w_y = 1 \times 1 = 1$, $z \times w_z = 1 \times (-1,5) = -1,5$. Σύμφωνα με τον τύπο: $S_j = \sum_1^n x_n \times w_{nj,}$ $S = 0 + 1 + (-1,5) = -0,5$, Επίσης σύμφωνα με τον τύπο: $\Phi(S) = \{1,$ αν $S > 0$ και 0 αν $S \leq 0\}$ και εφόσον $\underline{S = -0,5} < 0$, τότε $\underline{\Phi(S) = 0}$. Στην τρίτη περίπτωση έχουμε: $x \times w_x = 1 \times 1 = 1,$ $y \times w_y = 0 \times 1 = 0,$ $z \times w_z = 1 \times (-1,5) = -1,5$. Σύμφωνα με τον τύπο: $S_j = \sum_1^n x_n \times w_{nj,}$ $S = 1 + 0 + (-1,5) = -0,5$, Επίσης σύμφωνα με τον τύπο: $\Phi(S) = \{1,$ αν $S > 0$ και 0 αν $S \leq 0\}$ και εφόσον $\underline{S = -0,5} < 0$, τότε $\underline{\Phi(S) = 0}$. Στην τέταρτη περίπτωση έχουμε: $x \times w_x = 1 \times 1 = 1,$ $y \times w_y = 1 \times 1 = 1,$ $z \times w_z = 1 \times (-1,5) = -1,5$. Σύμφωνα με τον τύπο: $S_j = \sum_1^n x_n \times w_{nj,}$ $S = 1 + 1 + (-1,5) = +0,5$ Επίσης σύμφωνα με τον τύπο: $\Phi(S) = \{1,$ αν $S > 0$ και 0 αν $S \leq 0\}$ και εφόσον $\underline{S = +0,5} < 0$, τότε $\underline{\Phi(S) = 1}$.

Επίσης υπάρχουν τεχνητά νευρωνικά δίκτυα με περισσότερα από ένα επίπεδα όπως φαίνεται παρακάτω στην εικόνα 7. Αρχικά έχουμε τις εισόδους x_1, x_2, …, x_n, στο επόμενο επίπεδο έχουμε το κρυμμένο επίπεδο και τέλος την έξοδο [26] .



Εικόνα 7 Τεχνητό νευρωνικό Δίκτυο με ένα κρυφό επίπεδο

Στο παράδειγμα που ακολουθεί θα δούμε πως εφαρμόζεται ένα νευρωνικό δίκτυο στο μάρκετινγκ, από το [27] . Ένα σουπερμάρκετ στην Κίνα είχε χωρίσει τον τομέα των φρούτων σε δύο τμήματα, στα εγχώρια και στα εισαγόμενα, με σκοπό την αποφυγή του συνωστισμού. Ήταν ένα τμήμα που είχε κόσμο καθημερινά, και η διεύθυνση ήθελε να το αναπτύξει ακόμα περισσότερο, με σκοπό την μείωση των πελατών τις ημέρες γιορτής όπως είναι τα Χριστούγεννα και την επέκταση τους τις καθημερινές. Για να συλλέξει η επιχείρηση τα δεδομένα που θα χρησιμοποιούνταν, ερωτήθηκαν πεντακόσιοι είκοσι καταναλωτές, και τα δεδομένα που χρησιμοποιήθηκαν ήταν τετρακόσια ενενήντα πέντε. Το 30% χρησιμοποιήθηκε για την εκπαίδευση του δικτύου, το 20% για την διακοπή της εκπαίδευσης και το υπόλοιπο 50% χρησιμοποιήθηκε για τον τελικό έλεγχο. Αυτή η διαδικασία ονομάζεται Hold-out. Οι ερωτήσεις ήταν σχετικές με τις πεποιθήσεις και στάσεις των καταναλωτών, με κριτήρια όπως την τιμή των φρούτων, την εμφάνιση και το αν είναι φρέσκα, αλλά και την καταναλωτική συμπεριφορά τους με βάση την κοινωνική θέση, την καριέρα και γενικά για την προσωπικότητά τους. Τα χαρακτηριστικά (attributes) του συγκεκριμένου προβλήματος φαίνονται στον παρακάτω πίνακα 10 όπου τα χ_1—χ_{11} είναι εξαρτημένες μεταβλητές και η y είναι ανεξάρτητη.

Χαρακτηριστικά	Τύπος Μεταβλητής
Όψη φρούτου (χ_1)	Συνεχής
Κατάσταση Συσκευασίας (χ_2)	Συνεχής
Λιγότερα χημικά (χ_3)	Συνεχής
Καλή Γεύση (χ_4)	Συνεχές
Διαφορετική Γεύση (χ_5)	Συνεχής
Αν είναι φρέσκο(χ_6)	Συνεχής
Κατορθώματα ατόμων(χ_7)	Συνεχής
Ποσότητα Φρούτων(χ_8)	Συνεχής
Προσωπικότητα (χ_9)	Συνεχής
Κοινωνική Θέση(χ_{10})	Συνεχής
Τιμή(χ_{11})	Κατηγορηματική
Αγοραστική Πρόθεση (y)	Κατηγορηματική

Πίνακας 10 Τύπος μεταβλητών

Οι μέθοδοι που χρησιμοποιήθηκαν για το συγκεκριμένο παράδειγμα ήταν τα τεχνητά νευρωνικά δίκτυα, η μέθοδος Cart, και η παλινδρόμηση. Οι δύο επιπλέον μέθοδοι χρησιμοποιήθηκαν σε σύγκριση με τα τεχνητά νευρωνικά δίκτυα για το ποια από τις τρείς είναι η αποτελεσματικότερη μέθοδος. Το τεχνητό νευρωνικό δίκτυο που επιλέχθηκε ήταν οπισθοδιάδοσης με ένα κρυφό επίπεδο. Ο αλγόριθμος που χρησιμοποιήθηκε ήταν ο Lavenberg-Marguard. Για τον αριθμό των κρυφών κόμβων αναφέρεται χαρακτηριστικά είναι ότι "το πλήθος των κρυφών κόμβων θα πρέπει να είναι ένα σχετικά μικρό κλάσμα της εισόδου ή θα πρέπει να υπολογιστεί ως δύο φορές η ρίζα των κόμβων εισόδου συν τους κόμβους εξόδου¨. Αν το δίκτυο αποτυγχάνει να συγκλίνει μια λύση, τότε απαιτούνται περισσότεροι κρυφοί κόμβοι. Αντίθετα αν συγκλίνει σε μια λύση, τότε απαιτούνται λιγότεροι κρυφοί κόμβοι. Στο συγκεκριμένο παράδειγμα χρησιμοποιήθηκαν τέσσερεις κρυφοί κόμβοι. Τα αποτελέσματα των μεθόδων που χρησιμοποιήθηκαν φαίνονται στον πίνακα 11.

Χαρακτηριστικά	Παλινδρόμηση	Νευρωνικά Δίκτυα	CART
X_1	1.062	0.015	3.71
X_2	0.928	0.072	0.00
X_3	1.060	0.266	27.55
X_4	1.107	0.227	100.0(1)
X_5	0.996	0.066	25.22
X_6	0.926	0.028	8.21
X_7	1.030	0.111	22.35
X_8	1.019	0.093	22.14
X_9	0.991	0.042	11.34
X_{10}	1.005	0.087	24.5
X_{11}	2.073(1)	0.269(1)	57.38

Πίνακας 11 ΤΝΔ Περιγραφή των χαρακτηριστικών

Όπως βλέπουμε για την παλινδρόμηση και τα νευρωνικά δίκτυα η πιο σημαντική μεταβλητή είναι η χ_{11} δηλαδή η τιμή των φρούτων, ενώ για την μέθοδο CART η πιο σημαντική μεταβλητή είναι η χ_4 δηλαδή το πόσο γευστικά είναι. Που σημαίνει ότι για τις δυο πρώτες μεθόδους, για να αγοράσει ένας καταναλωτής ένα συγκεκριμένο φρούτο παρακολουθεί αρχικά τις τιμές, ενώ για την μέθοδο της παλινδρόμησης σημαντικό ρόλο παίζει η καλή γεύση. Στον πίνακα 12 που ακολουθεί μπορούμε να συμπεράνουμε πια είναι η καλύτερη μέθοδος για την επίλυση του προβλήματος αφού τα νευρωνικά δίκτυα μπορούν να "εκπαιδευτούν" πιο εύκολα από τις άλλες δύο τεχνικές και το τελικό αποτέλεσμα θα είναι πιο ακριβές.

Μέθοδος	**Εκπαίδευση**	**Hold-out**
Παλινδρόμηση	67.7%	63.56%
Νευρωνικά Δίκτυα	89.59%	65.18%
CART	82.66%	61.13%

Πίνακας 12 ΤΝΔ αποδοτικότητα των μεθόδων

3.3.4 ΑΛΓΟΡΙΘΜΟΣ ΚΟΝΤΙΝΟΤΕΡΟΥ ΓΕΙΤΟΝΑ

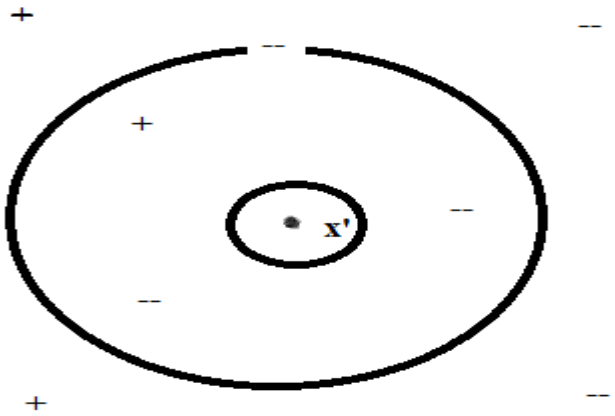
Ο αλγόριθμος κοντινότερου γείτονα είναι ένα εργαλείο εξόρυξης γνώσης το οποίο ταξινομεί κοντινά σημεία τα οποία έχουν ίδια χαρακτηριστικά. Δηλαδή, αυτό που θέλουμε να επιτύχουμε εφαρμόζοντας τον αλγόριθμο κοντινότερου γείτονα είναι να εντάξουμε ένα άγνωστο σημείο σε μια ομάδα με βάση τα γειτονικά του σημεία. Για αυτό χρησιμοποιείται η κοντινότερη απόσταση, η οποία μπορεί να μετρηθεί με την ευκλείδεια απόσταση ή την Manhattan ή την απόσταση Mahalanobis. Για παράδειγμα, έχουμε ένα σημείο Α το οποίο θέλουμε να το ταξινομήσουμε σε μια ομάδα, κοιτάμε τα γειτονικά σημεία Β, Γ και Δ σε ποια ομάδα ανήκουν. Αν τα περισσότερα από αυτά ανήκουν σε μια συγκεκριμένη ομάδα, τότε το σημείο Α εντάσσεται σε αυτήν την ομάδα. Παρακάτω ακολουθεί ο αλγόριθμος k-nn από το [19]:

1. Αναζήτηση στα N πρότυπα εκπαίδευσης x_i των k κοντινότερων γειτόνων στο πρότυπο ψ. Η παράμετρος k ορίζεται από τον χρήστη
2. Από τους k κοντινότερους γείτονες, προσδιορίζεται το πλήθος k_i των προτύπων που ανήκουν στην κλάση $C_{i..}$ Επίσης $\sum_{i=1}^{p} ki = k$.
3. Το πρότυπο ψ ανατίθεται στην κλάση C_i για την οποία ισχύει $\kappa_i > k_j$, , $i \neq j$. Δηλαδή το πρότυπο ψ ανατίθεται σε εκείνη την κλάση στην οποία ανήκουν οι περισσότεροι k κοντινότεροι γείτονες του ψ.

Αλγόριθμος 2 K-NN

Παρακάτω δίνεται ένα παράδειγμα του αλγορίθμου. Στην παρακάτω εικόνα 8 βλέπουμε δύο κύκλους, καθώς και δύο χαρακτηρισμούς (+ , --) και ένα σημείο x' που περιγράφεται από το σύνολο χαρακτηριστικών $<\alpha_1(\chi'),\alpha_2(\chi').....\alpha_n(\chi')$, όπου $\alpha_\rho(\chi')$είναι το r χαρακτηριστικό της και ένός αποθηκευμένου παραδείγματος χ, που από το σύνολο χαρακτηριστικών $<\alpha_1(\chi), \alpha_2(\chi)....\alpha_n(\chi),y(\chi)>$ είναι το αθροισμα των Ευκλείδιων απόστασεων όλων των

χαρακτηριστικών των 2 σημείων.όπου d(χ,χ)=$\sqrt{\sum_{r=1}^{n}(a_r(x) - a_r(x''))^2}$. Θέλουμε το χ' να ταξινομηθεί σε έναν από τους δύο χαρακτηρισμούς, (+, --).



Εικόνα 8 Κατηγοριοποίηση σημείου με βάση τον k-NN

Το σημείο χ' χαρακτηρίζεται ως + , διότι υπάρχει μόνο 1 πλησιέστερος γείτονας, και ως -- , διότι υπάρχουν πέντε πλησιέστεροι γείτονες [23] .

Ας δούμε στη συνέχεια που εφαρμόζεται ο αλγόριθμος knn. Ο αλγόριθμος knn εφαρμόζεται στο να προβλεφθεί το χρηματιστήριο. Αναφέρεται στην αποκάλυψη των τάσεων της αγοράς, προγραμματίζοντας τις στρατηγικές επένδυσης, που προσδιορίζουν τον καλύτερο χρόνο να αγοραστούν τα αποθέματα και ποια αποθέματα στην αγορά. Επίσης προβλέπεται η τιμή του αποθέματος βάση των οικονομικών στοιχείων. [28]

Ένα άλλο παράδειγμα που αναφέρεται στα εισοδήματα είναι: εάν παρατηρήσουμε τους ανθρώπους στη γειτονιά μας, όλοι έχουν περίπου τα ίδια εισοδήματα με τα δικά μας. Εάν ο γείτονας μας έχει ένα εισόδημα μεγαλύτερο από $100,000, οι πιθανότητες είναι καλές ότι έχουμε και εμείς επίσης ένα υψηλό εισόδημα. Αυτές οι πιθανότητες είναι μεγαλύτερες όταν όλοι οι γείτονες μας έχουν εισοδήματα πάνω από $100,000 από ότι αν έχουν εισοδήματα των $20,000 [29] .

Τέλος υπάρχουν κάποια πλεονεκτήματα και μειονεκτήματα αυτού. Τα πλεονεκτήματα είναι τα εξής: απλότητα, αποτελεσματικότητα, καθώς και ανταγωνιστική απόδοση ταξινόμησης. Τα μειονεκτήματα είναι τα εξής: μπορεί να έχει κακή απόδοση του χρόνου εκτέλεσης όταν το σύνολο κατάρτισης είναι μεγάλο. Είναι ευαίσθητο στα ανόμοια ή περιττά χαρακτηριστικά, επειδή όλα τα χαρακτηριστικά αναφέρονται στην ομοιότητα και συνεπώς στην ταξινόμηση. Επίσης το κόστος υπολογισμού είναι αρκετά υψηλό, επειδή πρέπει να υπολογίσουμε όλες τις αποστάσεις σε όλα τα δείγματα κατάρτισης [28] .

3.4 ΕΦΑΡΜΟΓΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Η ταξινόμηση εφαρμόζεται σε πολλούς και διαφορετικούς τομείς, όπως είναι ο τομέας της ιατρικής, των οικονομικών, του μάρκετινγκ και άλλοι. Στον τομέα του μάρκετινγκ, η ταξινόμηση εφαρμόζεται όταν τα στελέχη μάρκετινγκ, θέλουν να μάθουν αν ένας πελάτης αγοράσει κάποιο προϊόν, με βάση το προφίλ του. Αυτό είναι χρήσιμο για τις επιχειρήσεις, ώστε να προβλέπουν τι χρειάζονται οι πελάτες της και να εφαρμόζουν την κατάλληλη στρατηγική, ταξινομώντας πελάτες με ίδια χαρακτηριστικά. Ας δούμε στη συνέχεια ένα πιο συγκεκριμένο παράδειγμα: Τι τύποι δεδομένων χρειαζόμαστε για το προφίλ του πελάτη;

Μπορείς να χρησιμοποιήσεις την βάση δεδομένων των υπαρχόντων πελατών ή τα αποτελέσματα από έρευνες/συνεντεύξεις. Αυτό μπορεί να είναι πληροφορίες για τις αγορές τους, τα ενδιαφέροντά τους, καθημερινές ανάγκες και πιο προσωπικές πληροφορίες, όπως για παράδειγμα την ηλικία τους, το φύλο τους, οικογενειακή κατάσταση, εισόδημα, κτλ. Πως μπορούν οι πελάτες να ταξινομηθούν; Οι κατηγορίες θα πρέπει να δημιουργηθούν ανάλογα με το αντικείμενο της ανάλυσης της βάσης δεδομένων των πελατών. Ο πελάτης, ο οποίος ενδιαφέρεται για ένα συγκεκριμένο προϊόν, αντιστοιχούν τέσσερις κατηγορίες: "Ναι", "Όχι", "Δεν είμαι σίγουρος", "Δεν γνωρίζω για το προϊόν". Συνδυάζοντας τα δεδομένα που λαμβάνονται με την ανάλυση των πελατών σε είδη διαφημίσεων, δημιουργείται μία στοχευόμενη εκστρατεία μάρκετινγκ, βάση των αποτελεσμάτων της εξόρυξης δεδομένων. . Με αυτόν τον τρόπο η επιχείρηση θα δει που μπορεί να βελτιστοποιήσει τα προϊόντα της ή τις στρατηγικές της [30] .

Μία άλλη εφαρμογή της ταξινόμησης είναι η ανακάλυψη απάτης. Σε αυτή την περίπτωση, εάν ανιχνευθεί μία απάτη τραπεζικής συναλλαγής από την βάση δεδομένων συναλλαγών της τράπεζας, εντάσσεται σε δύο κατηγορίες: δόλια και μη δόλια. Κατασκευάζεται ένα μοντέλο από τα στοιχεία δεδομένων του δείγματος με γνωστές ετικέτες κλάσης και χρησιμοποιείται αυτό το μοντέλο για την πρόβλεψη της κατηγορίας αντικειμένων, στο πληθυσμό του οποίου οι κατηγορίες δεν είναι γνωστές. Κάθε γραμμή από τη βάση δεδομένων περιέχει ένα ή περισσότερα χαρακτηριστικά πρόβλεψης, η οποία καθορίζει την προβλεπόμενη ετικέτα κατηγορίας της γραμμής σύμφωνα με το κατασκευασμένο μοντέλο. Αυτά τα μοντέλα είναι κατασκευασμένα συχνά χρησιμοποιώντας δέντρα αποφάσεων ή νευρωνικά δίκτυα [31] .

Επίσης, η ταξινόμηση μπορεί να χρησιμοποιηθεί για την έγκαιρη διάγνωση ασθενειών, και συγκεκριμένα για ασθένειες καρδιάς. Στο [32] χρησιμοποιώντας τον αλγόριθμο κοντινότερου γείτονα τα δεδομένα χωρίζονται, σε δύο ομάδες, στην ομάδα υγιής και στην ομάδα άρρωστος. Η εφαρμογή του συγκεκριμένου αλγορίθμου δείχνει ότι ταξινομεί τα δεδομένα αποδοτικότερα από άλλα εργαλεία όπως τα νευρωνικά δίκτυα, σύμφωνα με την ακρίβεια, δηλαδή με το πόσο σωστά έχουν ταξινομηθεί τα δεδομένα. Η ακρίβεια μετράται σαν τον αριθμό των δειγμάτων που έχουν ταξινομηθεί σωστά, προς τον συνολικό αριθμό των δειγμάτων για τον τελικό έλεγχο.

Μια εφαρμογή των τεχνικών νευρωνικών δικτύων στον αθλητισμό και ειδικότερα στο μπάσκετ, αναφέρεται στο [14]. Τα δεδομένα που συλλέγονται στο τέλος κάθε παιχνιδιού χρησιμοποιούνται για την ανάλυση της απόδοσης της ομάδας και όπως καταλαβαίνουμε αυτό βοηθά κάθε προπονητή που θέλει να αναγνωρίσει λάθη που κάνουν οι παίκτες του, να δει ποιος παίκτης είναι καλός σε μια συγκεκριμένη θέση του παρκέ και αρκετά ακόμα που έχουν να κάνουν με την γενική, αλλά και πρακτική εικόνα της ομάδας του. Στην συγκεκριμένη εφαρμογή, αναλύονται δεδομένα από ομάδες μπάσκετ της Σερβίας, από το 2006 έως το 2011. Ανάλογα με τα ποσοστά των βολών επιτυχίας ή αποτυχίας της κάθε ομάδας μπορούμε να διαπιστώσουμε αν η ομάδα αυτή μπορεί να κερδίσει τους αντιπάλους της. Για παράδειγμα, αν το ποσοστό των τρίποντων,, σε μία ομάδα είναι κοντά στο ογδόντα τις εκατό τότε η συγκεκριμένη ομάδα έχει αρκετές πιθανότητες να προκριθεί.

Μια άλλη εφαρμογή της ταξινόμησης θα μπορούσε να είναι όταν ένας υπάλληλος μιας τράπεζας ο οποίος εγκρίνει δάνεια χρειάζεται να μάθει ποιες από τις αιτήσεις δανείων είναι «ασφαλείς» ή «υψηλού κινδύνου» για την τράπεζα.

4 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ (ASSOCIATION RULES)

Πολλές επιχειρήσεις συλλέγουν μεγάλες ποσότητες δεδομένων οι οποίες αναφέρονται στις αγορές των πελατών που μαζεύονται στα ταμεία των καταστημάτων. Ο κάθε πωλητής ενδιαφέρεται για την ανάλυση των δεδομένων για να μάθει για το τι αγοράζουν οι πελάτες, ποια προϊόντα αγοράζονται περισσότερο, κτλ. Με αυτά ασχολείται η ανάλυση καλαθιού αγοράς (market basket analysis), στο οποίο αναφορά γίνεται πιο κάτω. Αυτή η τεχνική χρησιμοποιείται κυρίως για την επεξεργασία πολλών δεδομένων. Επίσης, η εξόρυξη κανόνων συσχέτισης επιδιώκει να βρει κανόνες συσχέτισης οι οποίοι ικανοποιούν τις προκαθορισμένες απαιτήσεις υποστήριξης (support) και εμπιστοσύνης (confidence). Ως κανόνας συσχέτισης ορίζεται η μέθοδος ανακάλυψης σχέσεων μεταξύ των διαφόρων μεταβλητών σε μεγάλες βάσεις δεδομένων [13] . Ένας άλλος ορισμός είναι μια έκφραση της μορφής $X \rightarrow Y$ όπου X και Y είναι στοιχεία συνόλου $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \emptyset$ [16] . Τέλος οι κανόνες συσχέτισης χρησιμοποιούνται αρκετά συχνά επειδή είναι αποτελεσματικοί και γρήγοροι στην επεξεργασία μεγάλου πλήθος δεδομένων.

4.1 ΑΝΑΛΥΣΗ ΚΑΛΑΘΙΟΥ ΑΓΟΡΑΣ

Ένα πρόβλημα που επιλύεται με του κανόνες συσχέτισης είναι η ανάλυση καλαθιού αγοράς (market basket analysis). Η ανάλυση καλαθιού αγοράς αναφέρεται στις πληροφορίες που αντλούνται για το τι αγοράζουν οι πελάτες, για να πάρουν μία πρώτη άποψη για το ποιοι είναι και γιατί κάνουν ορισμένες αγορές. Δείχνει ποια προϊόντα πρόκειται να αγοραστούν μαζί καθώς και ποια προωθούνται περισσότερο [9] .

Αυτή η τεχνική έχει ευρέως εφαρμοστεί σε σουπερμάρκετ. Τα δεδομένα από την ανάλυση καλαθιού αγοράς στην πιο πρωτόγονη μορφή τους θα μπορούσαν να είναι μια λίστα συναλλαγής από τις αγορές των καταναλωτών αναφέροντας μόνο τα αντικείμενα που είναι μαζί και αναφέροντας τις τιμές τους [33] .

Στόχος της ανάλυσης καλαθιού αγοράς είναι να βρει ποια αγαθά αγοράζονται μαζί, ώστε να τοποθετηθούν σε ένα συγκεκριμένο σημείο. Αυτό γίνεται, για να προωθηθούν περισσότερο τα προϊόντα σε σχέση με άλλα αγαθά. Για παράδειγμα ένας καταναλωτής θα αγοράσει καφέ, όμως μαζί με τον καφέ μπορεί να αγοράσει ζάχαρη, νερό ή και φίλτρα του καφέ για την καφετιέρα. Η ζάχαρη θα βρίσκεται δίπλα ακριβώς από τον καφέ, για λόγους ευκολίας, αλλά και για λόγους τακτικής μάρκετινγκ. Τα δεδομένα για το κάθε προϊόν θα μπορούσαν να είναι μια λίστα για το τι έχουν αγοράσει οι πελάτες ή μια βάση δεδομένων με όλα τα χαρακτηριστικά της κάθε συναλλαγής. Δηλαδή, ημερομηνία, ώρα κλπ.

Η ανάλυση καλαθιού αγοράς στοχεύει στην παροχή εικόνας για συγγένειες προϊόντων. Φαίνονται σημαντικές πληροφορίες για τη διαφήμιση, για την αλλαγή στα ράφια ή τους καταλόγους, καθώς και για την ανάπτυξη ειδικών προσφορών προϊόντων ή υπηρεσιών. Επίσης στην παροχή προτάσεων του προϊόντος σύμφωνα με την συμπεριφορά του πελάτη [34] . Με βάση λοιπόν τους κανόνες συσχέτισης μπορούμε να βρούμε τι περιέχει ένα καλάθι αγοράς και ανάλογα με το τι θέλει να επιτύχει η κάθε επιχείρηση να καταστρώσει μια νέα στρατηγική. Στην επόμενη ενότητα θα αναφερθούμε σε παραδείγματα κανόνων συσχέτισης.

4.2 ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ

Το κάθε πρόβλημα, το οποίο θέλουμε να επιλύσουμε με την τεχνική των κανόνων συσχέτισης αντιμετωπίζεται σαν δύο υπό-προβλήματα. Το ένα υπό-πρόβλημα είναι η εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Item sets) και το δεύτερο υπό-πρόβλημα είναι η δημιουργία κανόνων συσχέτισης. Ένας κανόνας συσχέτισης έχει δύο μέρη, την συνθήκη η οποία καθορίζει το αποτέλεσμα και το αποτέλεσμα. Ένα απλό παράδειγμα είναι: {μολύβι}→{γόμα}. Αυτό ερμηνεύεται ως εξής: εάν ένας πελάτης αγοράσει μολύβι, τότε θα αγοράσει και γόμα. Ένα πιο ολοκληρωμένο παράδειγμα είναι αυτό που παρουσιάζεται παρακάτω στον πίνακα 13.

A/A	ΑΝΤΙΚΕΙΜΕΝΑ
1	Ψωμί, Γάλα
2	Ψωμί, Πάνα, Μπύρα, Αυγά
3	Γάλα, Πάνα, Μπύρα, Κόκα Κόλα
4	Ψωμί, Γάλα, Πάνα, Μπύρα
5	Ψωμί, Γάλα, Πάνα, Κόκα Κόλα

Πίνακας 13 Παράδειγμα Κανόνων Συσχέτισης

Κάθε γραμμή είναι μία συναλλαγή και στη δεύτερη στήλη βλέπουμε τα αντικείμενα-στοιχεία (items) ή αλλιώς προϊόντα. Η αναπαράσταση είναι σε μορφή 0 και 1. Το 1 δείχνει αν το προϊόν εμφανίζεται στη συναλλαγή και 0 αν όχι σύμφωνα με τον πίνακα 14:

A/A	Ψωμί	Γάλα	Πάνα	Μπύρα	Αυγά	Κόκα Κόλα
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Πίνακας 14 Παράδειγμα κανόνων συσχέτισης σε μορφή 0 και 1

I= $\{i_1,i_2,\ldots i_d\}$ όλα τα αντικείμενα , T= $\{t_1,t_2,\ldots,t_n\}$ όλες οι συναλλαγές. Κάθε συναλλαγή ti περιέχει ένα υποσύνολο αντικειμένων του I. Το στοιχειοσύνολο περιέχει 0 ή περισσότερα αντικείμενα του I. Για παράδειγμα {Milk,Bread}. Γενικά: k-στοιχειοσύνολο είναι ένα στοιχειοσύνολο με k στοιχεία. Το πλάτος συναλλαγής είναι ο αριθμός στοιχείων της συναλλαγής. Μία συναλλαγή t_i περιέχει ένα στοιχειοσύνολο X, με την προϋπόθεση ότι το X είναι ένα υποσύνολο της t_i. Για παράδειγμα η δεύτερη συναλλαγή περιέχει το στοιχειοσύνολο {Ψωμί, Πάνα}, αλλά όχι το {Ψωμί, Γάλα}.

Στη συνέχεια μετράμε την υποστήριξη (support count) ενός στοιχειοσυνόλου. Δείχνει το πόσο συχνά εμφανίζεται το στοιχειοσύνολο στο σύνολο των συναλλαγών T. Ο τύπος είναι: σ(X) = | {ti | X⊆ ti , ti∈T} | , όπου | . | δείχνει το πλήθος των στοιχείων του συνόλου. Παράδειγμα: σ({Γάλα, Ψωμί, Πάνα}) = 2 επειδή υπάρχουν δύο συναλλαγές που περιέχουν και τα τρία αντικείμενα του στοιχειοσυνόλου. Η υποστήριξη (support) ενός στοιχειοσυνόλου είναι, το ποσοστό συναλλαγών που περιέχει ένα στοιχειοσύνολο. Υπολογίζεται ως εξής: s(X) = σ(X) / N .Για παράδειγμα s({Γάλα ,Ψωμί, Πάνα}) = 2/5 = 0.4 Στη συνέχεια μετράμε την υποστήριξη του κανόνα, καθώς και την εμπιστοσύνη αυτού. Η υποστήριξη δείχνει πόσο συχνά είναι εφαρμόσιμος ο κανόνας. Υπολογίζεται ως εξής: $\underline{s (X\rightarrow Y) = \sigma(X\cup Y) / N}$

Η εμπιστοσύνη (confidence) δείχνει πόσο συχνά τα αντικείμενα στο στοιχειοσύνολο Y εμφανίζονται σε συναλλαγές που περιέχουν το X. Υπολογίζεται ως εξής:

$\underline{c (X\rightarrow Y) =\sigma(X\cup Y) / \sigma(X)}$

Ακολουθεί ένα παράδειγμα: Έστω ο κανόνας συσχέτισης {Ψωμί, Μπύρα}→{Πάνα}.

Το X∪Y = {Ψωμί, Μπύρα, Πάνα} εμφανίζεται δύο φορές σε σύνολο πέντε συναλλαγών, τότε η υποστήριξη του κανόνα είναι: s = σ{Ψωμί, Μπύρα, Πάνα} / N = 2 / 5 = 0.4 = 40%

Η εμπιστοσύνη, επειδή υπάρχουν δύο συναλλαγές που περιέχουν τα αντικείμενα του στοιχειοσυνόλου {Ψωμί, Μπύρα} θα είναι:

c = σ (**{**Ψωμί, Μπύρα, Πάνα}) /σ{Ψωμί, Μπύρα} = 2 / 2 = 1 = 100% [16] .

Με άλλα λόγια ένας κανόνας με μικρή υποστήριξη μπορεί να εμφανίζετε τυχαία. Δίνεται λιγότερη σημασία-χρησιμότητα, γιατί αφορά μικρό αριθμό συναλλαγών. Η εμπιστοσύνη μετρά την αξιοπιστία-βεβαιότητα του συμπεράσματος του κανόνα. Όσο μεγαλύτερη είναι, τόσο μεγαλύτερη η πιθανότητα εμφάνισης του Y σε κανόνες που περιέχουν το X.

4.3 Ο ΑΛΓΟΡΙΘΜΟΣ APRIORI

Υπάρχουν αρκετοί αλγόριθμοι για την εύρεση κανόνων συσχέτισης σε μεγάλες βάσεις δεδομένων. Ένας από τους πιο σημαντικούς είναι ο αλγόριθμος apriori, ο οποίος χρησιμοποιώντας δυαδικές τιμές εντοπίζει συσχετίσεις στις συναλλαγές από κατηγορηματικά χαρακτηριστικά [33] .

Περιλαμβάνει δυο βασικά βήματα: τη δημιουργία των συχνών συνόλων αντικειμένων και τη δημιουργία των κανόνων συσχέτισης. Η διαδικασία της δημιουργίας συχνών συνόλων αντικειμένων περιλαμβάνει δύο στάδια: αρχικά δημιουργείται ένα σύνολο υποψήφιων συχνών αντικειμένων και στη συνέχεια χρησιμοποιώντας το όριο υποστήριξης (support), δημιουργείται το σύνολο των συχνών συνόλων αντικειμένων. Η διαδικασία επαναλαμβάνεται πραγματοποιώντας διαδοχικά περάσματα στα δεδομένα μέχρι να βρεθούν είτε τα συχνά σύνολα αντικειμένων ενός προκαθορισμένου επιπέδου ή τα μέγιστα συχνά σύνολα αντικειμένων. Επιπλέον αποτελείται από ένα βήμα συνένωσης και ένα βήμα κλαδέματος. Για τη δημιουργία των κανόνων συσχέτισης ελέγχεται η εμπιστοσύνη (confidence) όλων των κανόνων που προκύπτουν από τα μέγιστα συχνά σύνολα αντικειμένων και στο τέλος μένουν εκείνοι των οποίων η εμπιστοσύνη ξεπερνά το όριο που τέθηκε. Παρακάτω θα δώσουμε τον κώδικα αυτού του αλγορίθμου:

1. $K=1$
2. $F_k=\{i/i\in I \wedge \sigma(\{i\})\geq N*\min sup\}$(βρες όλα τα συχνά 1-item sets)
3. Repeat
4. $K=k+1$
5. $Ck=apriori\text{-}gen(F_k\text{-}1)$ (δημιούργησε candidate item sets)
6. For each transaction $t\in T$ do
7. $Ct=subset(C_k,t)$ (ταυτοποιεί όλα τα candidates τα οποία ανήκουν στο t)
8. For each candidate item set $C\in ct$ do
9. $\sigma(c)=\sigma(c)+1$
10. End for
11. End for
12. $F_k=\{c/c\in C_k \wedge \sigma(c)\geq N*\min sup\}$ (εξάγει τα συχνά k-item sets)
13. Until $F_K=\emptyset$
14. Result=UF_K. [16]

Αλγόριθμος 3 APRIORI

Ακολουθεί ένα παράδειγμα εφαρμογής του αλγορίθμου APRIORI με τη βοήθεια του πίνακα 15:

Α/Α	Ψωμί	Καφές	Γάλα	Ζάχαρη
1	1	0	1	0
2	0	1	0	0
3	1	0	1	1
4	0	1	0	1
5	1	0	1	1
6	1	1	1	0
7	1	0	0	1
8	1	1	1	1
9	0	0	1	1
10	1	1	0	1

Πίνακας 15 Παράδειγμα APRIORI (dataset)

Στο παραπάνω πίνακα φαίνονται 10 διαφορετικές συναλλαγές από ένα supermarket. Στην πρώτη βλέπουμε για παράδειγμα ότι ένα άτομο αγόρασε Ψωμί και Γάλα (αναγράφεται ως 1), αλλά όχι Καφέ και Ζάχαρη (αναγράφεται ως 0).
Η υποστήριξη που ζητάμε είναι sup = 40% και η εμπιστοσύνη conf = 80%. Στο πρώτο βήμα υπολογίζουμε την υποστήριξη όλων των αντικειμένων, δηλαδή δημιουργούμε το πρώτο σύνολο:
S {Ψωμί} = 7 / 10 = 70%
S {Καφές} = 5/10 = 50%
S {Γάλα} = 6/10 = 60%
S {Ζάχαρη} = 7/10 = 70%
Άρα στο πρώτο σύνολο βλέπουμε ότι όλα είναι μεγαλύτερα από την υποστήριξη που ζητάμε. Άρα το πρώτο σύνολο είναι:{ Ψωμί , Καφές , Γάλα , Ζάχαρη}.
Μετά παίρνουμε τα ζεύγη αντικειμένων για να δημιουργηθεί το σύνολο των ζευγών αντικειμένων:{Ψωμί,Καφές},{Ψωμί,Γάλα},{Ψωμί,Ζάχαρη},{Καφές,Γάλα},{Καφές, Ζάχαρη},{Γάλα, Ζάχαρη}.
Στη συνέχεια υπολογίζουμε την υποστήριξη και απορρίπτονται εκείνα που δεν ξεπερνούν το όριο ελάχιστης υποστήριξης, ώστε να δημιουργηθεί το δεύτερο σύνολο συχνών ζευγών:
S {Ψωμί, Καφές} = 3/10 = 30% < sup (απορρίπτεται)
S {Ψωμί, Γάλα} = 5/10 = 50% ≥ sup
S {Ψωμί, Ζάχαρη} =5/10=50% ≥ sup
S {Καφές, Γάλα} =2/10=20% ≤sup (απορρίπτεται)
S {Καφές, Ζάχαρη} =3/10=30% ≤ sup (απορρίπτεται)
S {Γάλα, Ζάχαρη} = 4/10=40% ≥ sup
Τελικά το δεύτερο σύνολο συχνών ζευγών είναι: {{Ψωμί, Γάλα},{Ψωμί, Ζάχαρη},{Γάλα, Ζάχαρη}}.
Στη συνέχεια βρίσκουμε τις τριάδες αντικειμένων: Υπάρχει το βήμα της συνένωσης: {Ψωμί, Γάλα}∪{Ψωμί, Ζάχαρη}= {Ψωμί, Γάλα, Ζάχαρη}.
Στη συνέχεια εκτελούμε το βήμα του κλαδέματος: Οι επιμέρους δυάδες ανήκουν στο δεύτερο σύνολο συχνών ζευγών. Άρα προκύπτει το σύνολο ζευγών αντικειμένων: {{Ψωμί, Γάλα, Ζάχαρη}}. Η υποστήριξη του είναι: S({Ψωμί, Γάλα, Ζάχαρη}) = 3/10 = 30%

(απορρίπτεται). Η διαδικασία τελειώνει εδώ και επομένως το μέγιστο συχνό σύνολο αντικειμένων είναι το δεύτερο.
Το επόμενο βήμα είναι να βρούμε τους κανόνες από τα συχνά σύνολα(δηλαδή μόνο το δεύτερο) βάση της εμπιστοσύνης τους. ‰
{Ψωμί, Γάλα}:
Ψωμί → Γάλα : εμπιστοσύνη = 5/7 = 71% < conf (απορρίπτεται)
Γάλα → Ψωμί : εμπιστοσύνη = 5/6 = 83% > conf
{Ψωμί, Ζάχαρη}
Ψωμί → Ζάχαρη: εμπιστοσύνη = 5/7 = 71% < conf (απορρίπτεται)
Ζάχαρη → Ψωμί: εμπιστοσύνη = 5/7 = 71% < conf (απορρίπτεται)
{Γάλα, Ζάχαρη}
Γάλα → Ζάχαρη: εμπιστοσύνη = 4/6 = 66% < conf (απορρίπτεται)
Ζάχαρη → Γάλα : εμπιστοσύνη = 4/7 = 57% < conf (απορρίπτεται)
Τελικά δημιουργείται ο κανόνας Γάλα → Ψωμί, δηλαδή όποιος αγοράζει Γάλα, τότε αγοράζει και Ψωμί [23] .

Ο Apriori είναι αποτελεσματικός κατά την διάρκεια παραγωγής της διαδικασίας. Χρησιμοποιεί τεχνικές κλαδέματος για να αποφευχθεί η μέτρηση ορισμένων στοιχειοσυνόλων, ενώ εγγυάται την πληρότητα. Ωστόσο, υπάρχουν δύο σημεία συμφόρησης του αλγορίθμου Apriori. Το ένα είναι η υποψήφια (candidate) διαδικασία παραγωγής που χρησιμοποιεί το μεγαλύτερο μέρος του χρόνου, του χώρου και της μνήμης. Ένα άλλο εμπόδιο είναι η πολλαπλή σάρωση της βάσης δεδομένων. Με βάση τον Apriori, πολλοί νέοι αλγόριθμοι έχουν σχεδιαστεί με ορισμένες τροποποιήσεις ή βελτιώσεις [11] .

Ένας άλλος αλγόριθμος είναι ο AIS , ο οποίος ήταν ο πρώτος αλγόριθμος που προτάθηκε για την εξόρυξη κανόνων συσχέτισης. Σε αυτόν τον αλγόριθμο μόνο τα στοιχεία ακολουθούμενων ενώσεων παράγονται, δηλαδή η συνέπεια αυτών των κανόνων περιέχει μόνο ένα στοιχείο, για παράδειγμα, μπορούν να δημιουργηθούν κανόνες όπως το $X \cap Y \Rightarrow Z$, αλλά όχι τους κανόνες αυτούς ως $X \Rightarrow Y \cap Z$. Το κύριο μειονέκτημα του αλγορίθμου AIS είναι ότι πολλά υποψήφια στοιχειοσύνολα που τελικά αποδείχθηκε ότι είναι μικρά έχουν παραχθεί, το οποίο σημαίνει ότι απαιτείται περισσότερος χώρος και απόβλητα με πολύ προσπάθεια που αποδείχθηκε, ότι ήταν άχρηστη. Ταυτόχρονα αυτός ο αλγόριθμος, απαιτεί πάρα πολλά περάσματα σε ολόκληρη την βάση δεδομένων. Επίσης υπάρχουν οι αλγόριθμοι SetM , F-P Growth, και οι Apriorihybrid και AprioriTID [35] .

4.4 ΕΦΑΡΜΟΓΕΣ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

Μία από τις κύριες εφαρμογές, μέσω των οποίων υλοποιούνται οι κανόνες συσχέτισης είναι η ανάλυση καλαθιού αγοράς. Όταν ένας πελάτης αγοράζει κάποια προϊόντα σε ένα κατάστημα και παίρνει μία απόδειξη πληρωμής, αυτή η συναλλαγή καταγράφεται στο σύστημα. Μαζί με τη λίστα των αγορασμένων προϊόντων αποθηκεύεται η τιμή, η ημερομηνία, η ώρα και ο τόπος της συναλλαγής. Πιο συγκεκριμένα, ένας κανόνας που μπορεί να υπάρξει είναι: Εάν {αυγά, γάλα, φρέσκα φρούτα} → {λαχανικά}. Αυτός ο απλός κανόνας αναφέρεται σε παρόμοια προϊόντα που αγοράζονται μαζί. Ο παραπάνω κανόνας ερμηνεύεται ως εξής: Όταν τα αυγά, το γάλα και τα φρέσκα φρούτα αγοράζονται, τότε υπάρχει πιθανότητα να αγοραστούν και τα λαχανικά [34] .

Μία άλλη εφαρμογή είναι η εξόρυξη παγκόσμιου ιστού. Αναφέρεται στο ποιες σελίδες επισκέφτηκε ο χρήστης σε μία περίοδο η οποία είναι και αυτή μία συναλλαγή καθώς φαίνεται η ώρα της επίσκεψης. Ένας κανόνας συσχέτισης μπορεί να είναι: εάν ένας χρήστης επισκέφτηκε τη σελίδα www.facebook.com, θα επισκεφθεί μέσα σε πέντε μέρες και την σελίδα www.hotmail.com με πιθανότητα 0.50.

Τέλος στο [36] παρουσιάζεται η εφαρμογή των κανόνων συσχέτισης στην εκπαίδευση. Σύμφωνα με τα εκπαιδευτικά δεδομένα της βάσης δεδομένων ESOG (Electra School Occupational Guidance) γίνεται εξόρυξη γνώσης, με βάση τον αλγόριθμο apriori και εξάγονται πληροφορίες για την επιλογή των υποψηφίων στα ανώτατα εκπαιδευτικά ιδρύματα, τον διαχωρισμό των μαθητών σε ομάδες με περίπου ίδια ενδιαφέροντα, την αξιολόγηση του ερωτηματολογίου του ESOG, και γενικά των ενδιαφερόντων ανάλογα με το φύλο. Οι υποψήφιοι δηλώνουν μέσω ενός ερωτηματολογίου, τον βαθμό ενδιαφέροντος για διάφορα μαθήματα και το ESOG παρουσιάζει μια ταξινομημένη λίστα των ιδρυμάτων που θα τους ενδιέφεραν. Για την εξαγωγή των κανόνων συσχέτισης δημιουργήθηκε ένα αρχείο για επεξεργασία που περιείχε 511 γραμμές, δηλαδή πεντακόσιοι έντεκα μαθητές και 24 γνωρίσματα δηλαδή κάθε γραμμή είχε 24 πεδία σχετικά με τα ενδιαφέροντα του μαθητή και χρησιμοποιήθηκε η συγγραφή εντολών σε SQL για την ανάκτηση των δεδομένων από την βάση του Πανελλήνιου Σχολικού δικτύου καθώς και μια εφαρμογή C που συνένωσε τα αποτελέσματα των ερωτημάτων SQL. Για την εξαγωγή των κανόνων συσχέτισης χρησιμοποιήθηκε το ανοιχτό λογισμικό WEKA εφαρμόζοντας τον αλγόριθμο apriori. Οι κανόνες είχαν την μορφή Μαθηματικά = Πάρα πολύ, Λατινικά = λίγο→ Πληροφορική = Πάρα πολύ. Αυτός ο κανόνας σημαίνει ότι αν ενός μαθητή του αρέσουν τα μαθηματικά πάρα πολύ αλλά δεν του αρέσουν τα λατινικά τότε θα του αρέσει πάρα πολύ η πληροφορική. Κάποιοι από τους κανόνες συσχέτισης δείχνουν τον διαχωρισμό των μαθητών σε τεχνολογικούς και σε θετικούς, όπως Φυσική = καθόλου, Στατιστική = καθόλου → Μαθηματικά = καθόλου ή Ιστορία = καθόλου, Αρχαία ελληνικά = καθόλου → Λατινικά = καθόλου.

5 ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

Η τεχνική της Συσταδοποίησης ή Ομαδοποίησης είναι μία από τις παλαιότερες τεχνικές που χρησιμοποιείται στην εξόρυξη δεδομένων. Βασίζεται στην μάθηση χωρίς επίβλεψη (Unsupervised Method). Χρησιμοποιείται στην τμηματοποίηση της αγοράς, δηλαδή στην διαδικασία μάρκετινγκ της επιχείρησης να χωρίσουν σε τμήματα την αγορά, ανάλογα με τα ενδιαφέροντα και τα χαρακτηριστικά των πελατών (ενότητα 5.2) . Με αυτόν τον τρόπο η επιχείρηση μπορεί να έχει καλύτερα αποτελέσματα στο συνολικό της κέρδος, αφού θα έχει διαφορετικές στρατηγικές μάρκετινγκ σε διαφορετικά κομμάτια της αγοράς.

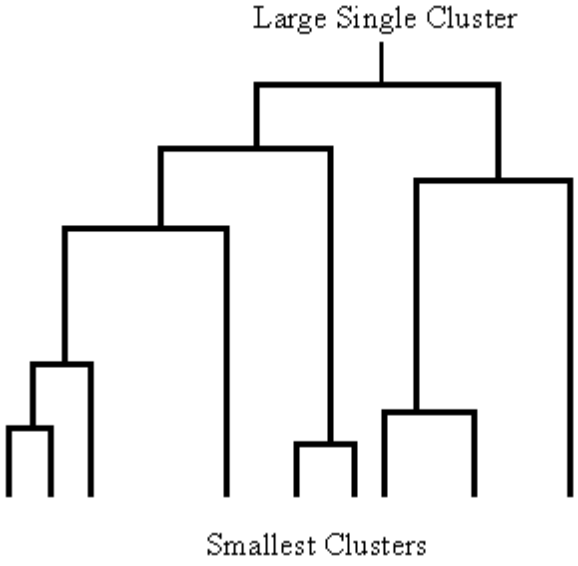
Η διαφορά που υπάρχει στην μέθοδο ομαδοποίησης (Clustering) με την μέθοδο της ταξινόμησης (Classification), είναι ότι στην ομαδοποίηση, αναλύονται δεδομένα τα οποία έχουν τακτοποιηθεί σε κλάσεις, ενώ αντίθετα στην ταξινόμηση τα δεδομένα αναλύονται χωρίς να υπάρχει η γνώση σε ποια κατηγορία ανήκουν. Για να γίνει η ομαδοποίηση των δεδομένων χρησιμοποιούνται γεωμετρικοί και στατιστικοί μέθοδοι, αλλά και νευρωνικά δίκτυα.

5.1 ΟΡΙΣΜΟΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Η συστάδοποίηση είναι η μέθοδος με την οποία οι εγγραφές είναι ομαδοποιημένες μαζί. Με αυτό τον τρόπο δίνεται στον τελικό χρήστη η άποψη του τι συμβαίνει στη βάση δεδομένων [10] .

Τα μοντέλα ομαδοποίησης στοχεύουν στο να διαιρούν τις εγγραφές ενός συνόλου δεδομένων σε ομοιογενείς ομάδες παρατηρήσεων, οι οποίες ονομάζονται συστάδες, έτσι ώστε οι παρατηρήσεις που ανήκουν σε μία ομάδα να είναι παρόμοιες με άλλες, και διαφορετικές από αυτές που ανήκουν σε άλλες ομάδες [11] .

Υπάρχουν δύο τύποι ομαδοποίησης: εκείνοι που δημιουργούν μία ιεραρχία ομαδοποιήσεων και εκείνοι που δεν διαθέτουν. Η ιεραρχική ομαδοποίηση δημιουργεί μία ιεραρχία συστάδων από μικρές σε μεγάλες. Έτσι μπορούμε να επιλέξουμε τον αριθμό των ομάδων που είναι επιθυμητό. Το πλεονέκτημα αυτής, είναι ότι υπάρχει η δυνατότητα να επιλεχθούν όλες, οι μισές ή λίγες συστάδες. Παρουσιάζονται σε δέντρο στο οποίο οι μικρότερες συστάδες εμφανίζονται μαζί για να σχηματίσουν το επόμενο επίπεδο όπως φαίνεται στην εικόνα 9.



Εικόνα 9 Ιεράρχηση Συστάδων

Οι μη ιεραρχικές τεχνικές γενικά είναι γρηγορότερες για να δημιουργηθούν από την ιστορική βάση δεδομένων, αλλά έχουν κάποια μειονεκτήματα. Χωρίζονται σε δύο μεθόδους: στην μια γίνεται ένα πέρασμα των στοιχείων ώστε να δημιουργηθούν οι συστάδες (Single pass method) και στην άλλη πραγματοποιούνται πολλά περάσματα από την μια συστάδα στην άλλη μέχρι να πραγματοποιηθεί η καλύτερη ομαδοποίηση (Reallocation). Επιπλέον υπάρχουν δύο ειδών αλγόριθμοι ιεραρχικής ομαδοποίησης. Οι Συσσωρευτικοί αλγόριθμοι (Agglomerative) και οι Διαιρετικοί (Divisive). Οι Agglomerative αρχίζουν με πολλές συστάδες οι οποίες περιέχουν μόνο μια αναφορά. Οι συστάδες που είναι πιο κοντά συγκεντρώνονται και φτιάχνουν το επόμενο επίπεδο, μέχρι να δημιουργηθεί το υψηλότερο επίπεδο. Οι Divisive κάνουν ακριβώς το αντίθετο, δηλαδή ξεκινούν από μια αναφορά με όλες τις συστάδες και ομαδοποιούνται σε μικρότερα κομμάτια μέχρι να δημιουργηθεί και το χαμηλότερο επίπεδο [29].

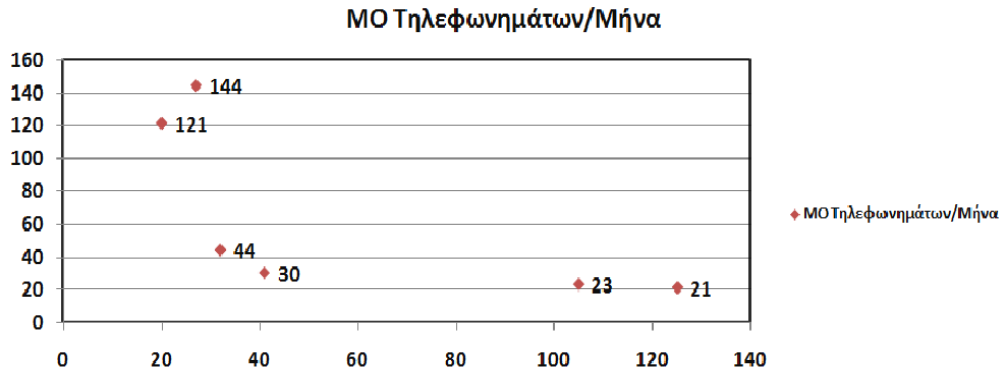
Ένα απλό παράδειγμα για την διαδικασία της συσταδοποίησης φαίνεται παρακάτω στον πίνακα 16: έχουμε έναν πίνακα με αναφορές από τηλεφωνικές επικοινωνίες, ώστε να ομαδοποιηθούν οι χρήστες σε κατηγορίες.

ID ΠΕΛΑΤΗ	**ΜΟ ΜΗΝΥΜΑΤΩΝ /ΜΗΝΑ**	**ΜΟ ΤΗΛΕΦΩΝΗΜΑΤΩΝ/ΜΗΝΑ**
1	27	144
2	32	44
3	41	30
4	125	21
5	105	23
6	20	121

Πίνακας 16 Παράδειγμα Συσταδοποίησης

Όπως μπορούμε να δούμε στο παρακάτω σχήμα ο πελάτης με ID 1 και ο 6 έχουν μικρή κατανάλωση σε αποστολή μηνυμάτων και μεγάλη κατανάλωση σε τηλεφωνήματα. Για αυτό τον λόγο έχουν συγκεντρωθεί μαζί και έχουν φτιάξει μια ομάδα πελατών η οποία μπορεί να

λέγεται “Υψηλή κατανάλωση τηλεφωνημάτων”. Επίσης, οι 3 και 2 έχουν συγκεντρωθεί μαζί σχηματίζοντας την ομάδα "Τυπικοί καταναλωτές". Τέλος, οι 4 και 5 έχουν σχηματίσει την ομάδα "Χρήστες μηνυμάτων" [34]. Όλα τα παραπάνω απεικονίζονται στην εικόνα 10:



Εικόνα 10 Παράδειγμα Συσταδοποίησης

5.2 ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΤΗΣ ΑΓΟΡΑΣ

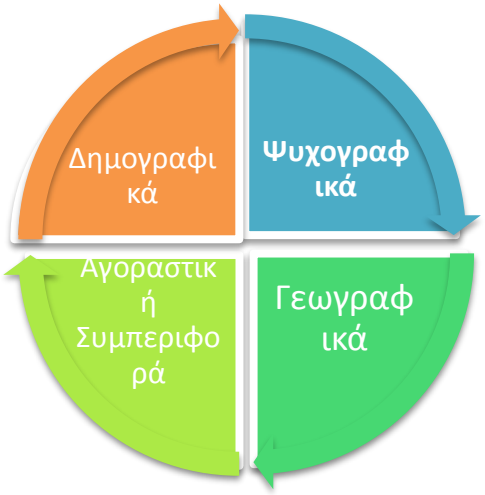
5.2.1 ΟΡΙΣΜΟΣ ΤΜΗΜΑΤΟΠΟΙΗΣΗΣ

Αποτελεί μία από τις σημαντικότερες και συνήθεις τακτικές κάθε επιχείρησης. Η τμηματοποίηση (segmentation) αναφέρεται στην υποδιαίρεση της αγοράς σε ίδια τμήματα πελατών, διαλέγεται εύκολα και μπορεί να αντιμετωπιστεί ως μία μικρότερη εξειδικευμένη αγορά. Αποτελεί μία διαδικασία την οποία εκτελεί η κάθε επιχείρηση σε ομάδες πελατών της, ώστε να μπορεί να τους διαχειριστεί καλύτερα. Αυτό γίνεται διότι αυτοί διαφέρουν μεταξύ τους, έχουν ιδιαίτερες απαιτήσεις και η επιχείρηση προσαρμόζει τα προϊόντα της ανάλογα με αυτές. Για παράδειγμα όπως ηλικιακά, δηλαδή άλλα προϊόντα θα αγοράσει ένας καταναλωτής 20 ετών και άλλα ένας 50 ετών. Ανάλογα με την ιδιότητά τους μαμά ή κοπέλα, ανάλογα με το εισόδημά τους, αν είναι νέοι πελάτες ή έχουν ξαναγοράσει κάποιο από το προϊόν της επιχείρησης, ή αν αγοράζουν συχνά. Όταν μιλάμε για μεγάλες αγορές όπου τα δεδομένα είναι πολλά και τα χαρακτηριστικά των πελατών δεν είναι ευδιάκριτα, τότε χρειάζεται η μέθοδος της εξόρυξης δεδομένων, η ομαδοποίηση.

5.2.2 ΚΡΙΤΗΡΙΑ ΤΜΗΜΑΤΟΠΟΙΗΣΗΣ ΑΓΟΡΑΣ

Μια επιχείρηση μπορεί να φτιάξει σε ομάδες τους πελάτες της σύμφωνα με διάφορα κριτήρια. Τα κύρια κριτήρια τμηματοποίησης της αγοράς ταξινομούνται σε:

- Γεωγραφικά
- Δημογραφικά
- Ψυχογραφικά
- Αγοραστικής Συμπεριφοράς



Εικόνα 11 Κριτήρια Τμηματοποίησης της αγοράς

Στα γεωγραφικά κριτήρια χωρίζουμε την αγορά σε διάφορες περιοχές. Έτσι η επιχείρηση μπορεί να διαφοροποιήσει τους πελάτες της σε αυτούς των μεγάλων αστικών κέντρων, μεγάλων πόλεων, κωμοπόλεων και χωριών. Αυτό προκύπτει γιατί οι καταναλωτές των διαφόρων γεωγραφικών περιοχών έχουν διαφορετικές ανάγκες. Για παράδειγμα θέλουμε οι διαστάσεις των πόλεων να έχουν πληθυσμό κάτω των πέντε χιλιάδων.

Όσον αφορά τα δημογραφικά κριτήρια, αυτά αναφέρονται στην ηλικία, το φύλο, το εισόδημα, το επάγγελμα, την κοινωνική τάξη, την θρησκεία, κτλ.

Τα ψυχογραφικά κριτήρια μπορεί να είναι προσωπικότητα ή ο τρόπος ζωής, κα. Για παράδειγμα υπάρχουν άνθρωποι που χαίρονται έτσι όπως ζουν και θέλουν τα τελευταία προϊόντα της τεχνολογίας. Άλλοι αγοράζουν προϊόντα που θα δείχνουν την υψηλή κοινωνική τους θέση. Υπάρχουν και οι απλοί άνθρωποι που ψάχνουν τα συνηθισμένα προϊόντα.

Όσον αφορά την αγοραστική συμπεριφορά οι καταναλωτές χωρίζονται με βάση τις ανάγκες που έχουν ούτως ώστε να τις ικανοποιήσουν. Μπορεί να είναι με βάση τη χρησιμότητα του προϊόντος, τη ποσότητα που αγοράζουν, κα.

Όπως έχει αναφερθεί η τμηματοποίηση της αγοράς βοηθά τις επιχειρήσεις να εφαρμόσει συγκεκριμένες στρατηγικές μάρκετινγκ σε συγκεκριμένα τμήματα αγοράς Αυτός ο διαχωρισμός έχει κάποια πλεονεκτήματα. Ένα από τα πλεονεκτήματα της είναι να βλέπει τις ευκαιρίες της αγοράς και να σχεδιάζει στρατηγικές μάρκετινγκ, έτσι ώστε να εξασφαλίζει το μεγαλύτερο δυνατό αποτέλεσμα. Δηλαδή, χωρίζοντας την αγορά σε ομάδες ανάλογα με την τάση της αγοράς η επιχείρηση μπορεί να εφαρμόσει μια στρατηγική σε ένα συγκεκριμένο τμήμα το οποίο πλεονεκτεί. Για παράδειγμα, τους τελευταίους μήνες οι πωλήσεις ενός συγκεκριμένου προϊόντος έχουν αυξηθεί στις γυναίκες από είκοσι έως τριάντα ετών. Η επιχείρηση θα επικεντρωθεί σε αυτή την ομάδα πελατών, ώστε να αυξήσει τις πωλήσεις τις ακόμα περισσότερο. Επίσης γνωρίζοντας τα ιδιαίτερα χαρακτηριστικά του κάθε τμήματος, τα στελέχη μάρκετινγκ της επιχείρησης μπορούν να παράγουν το κατάλληλο προϊόν, να χρησιμοποιήσουν κερδοφόρες στρατηγικές τιμολόγησης, να επιλέξουν τα σωστά δίκτυα διανομής, και την διαφήμιση αυτού του τμήματος. Τέλος έχουν τη δυνατότητα να συντονίζουν και να κατευθύνουν τον προϋπολογισμό του μάρκετινγκ σε αυτό το τμήμα της αγοράς που θεωρείται πιο κερδοφόρο [37] .

5.3 ΑΛΓΟΡΙΘΜΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

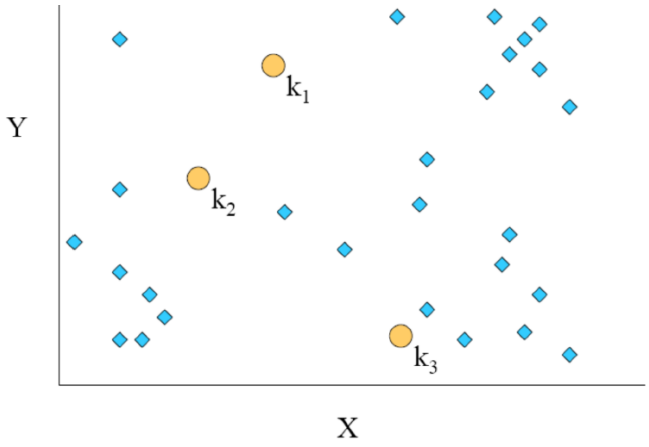
Υπάρχουν αρκετοί αλγόριθμοι ομαδοποίησης, που χρησιμοποιούνται για την επεξεργασία των δεδομένων και την τελική παρουσίασή τους σε ομάδες. Όπως έχουμε αναφέρει χωρίζονται στους Agglomerative και στους Divisive.

Θα ξεκινήσουμε να αναφέρουμε τον αλγόριθμο k-μέσων (k means): Είναι ο πιο διαδεδομένος αλγόριθμος για την συσταδοποίηση των δεδομένων, ο οποίος ανήκει στους Διαχωριστικούς αλγόριθμους (Divisive). Είναι ο πιο αποτελεσματικός αλλά και αρκετά γρήγορος. Αυτό γιατί μπορεί να διαχειριστεί πολλές αναφορές και σε μεγάλες βάσεις δεδομένων, σε πίνακες πολλών διαστάσεων. Επίσης, δεν χρειάζεται να διαβάσει όλες τις αναφορές. αλλά να πάρει μόνο μερικές ώστε να γίνει η ομαδοποίηση. Η είσοδος του k-means περιέχει το πλήθος των συστάδων κ που θέλουμε να δημιουργήσουμε, και ένα σύνολο δεδομένων. Η ομαδοποίηση των δεδομένων γίνεται με την τυχαία επιλογή κ στοιχείων κάθε ένα από τα οποία αρχικά αναπαριστά το κέντρο της συστάδας. Ανάλογα με την απόσταση που έχει το συγκεκριμένο στοιχείο με το κέντρο της συστάδας μπαίνει σε μια ομάδα και το κέντρο ενημερώνεται. Η έξοδός του περιέχει τις συστάδες με την μορφή ενός διαγράμματος. Η ομαδοποίηση συνεχίζεται μέχρις ότου δεν υπάρξει καμία αλλαγή στο αποτέλεσμα. Η διαδικασία του k-means χρησιμοποιείται σε ένα ολοκληρωμένο μαθηματικό λογισμικό το matlab. Αναφορά σε αυτό γίνεται στο επόμενο κεφάλαιο. Παρακάτω θα δώσουμε τον αλγόριθμο του Κ-Μέσων από το [23] .

Αλγόριθμος Κ-μέσων
Είσοδος:
Σύνολο δεδομένων $D=\{x_1,\ldots\ldots, x_n\}$
Αριθμός Ομάδων k
Έξοδος:
Ομάδες C_i
1.// ανάθεση τυχαίων κέντρων
για i=1,….,k κάνε:
θεώρησε m_i ως ένα τυχαίο στοιχείο από το D;
2.// Συσταδοποίηση
Όσο υπάρχουν αλλαγές στις ομάδες C_i κάνε:
2α. //δημιουργία ομάδων
για i=1,…,k κάνε
$C_i=\{x\epsilon\ D/\ d(m_i, x) \leq d(m_j, x)$ για όλα τα $j=1,\ldots,k, j\neq k\}$;
2β.// υπολογισμός νέων κέντρων
για i=1,….,k κάνε
m_i= το μέσο διάνυσμα των σημείων που ανήκουν στην ομάδα C_i ;

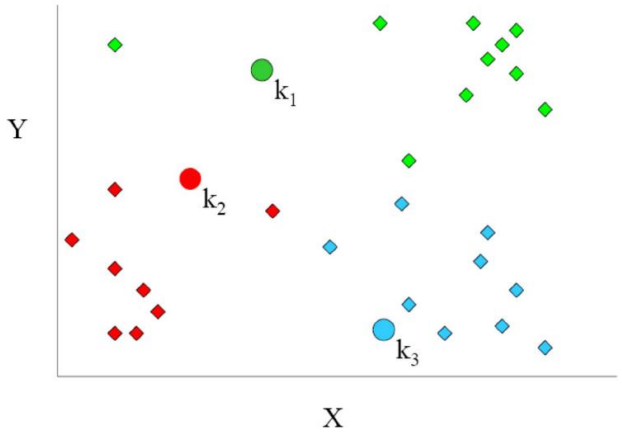
Αλγόριθμος 4 Κ-μέσων

Ακολουθεί μία γραφική αναπαράσταση των βημάτων του αλγορίθμου: Η αρχική κατάσταση είναι το πλήθος των επιθυμητών συστάδων κ = 3 , αρχικά τυχαία επιλεγμένα κέντρα κ_1 , κ_2, κ_3 , όπως φαίνονται στην εικόνα 12:



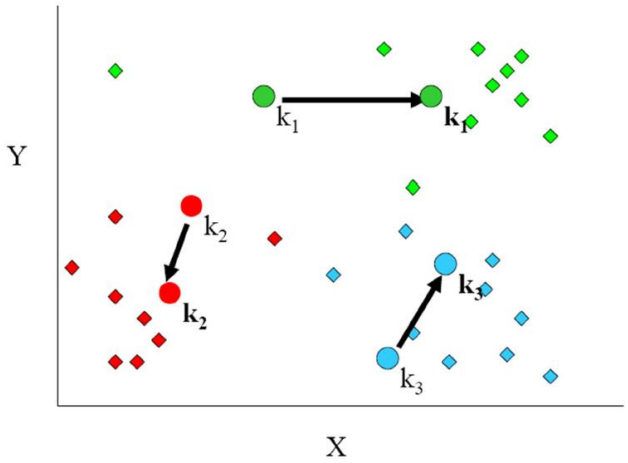
Εικόνα 12 Αρχική κατάσταση

Στη συνέχεια τα σημεία ανατίθενται στο πιο κοντινό τους αρχικό κέντρο όπως φαίνεται στην εικόνα 13:



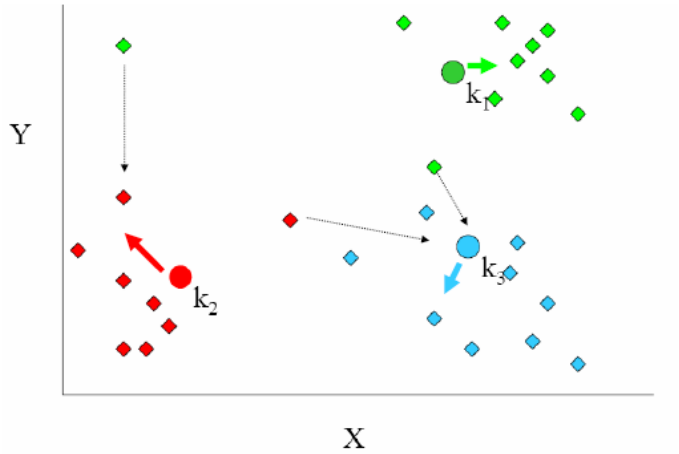
Εικόνα 13 Ανάθεση Σημείων στο πιο κοντινό τους κέντρο

Μετά υπολογίζεται ξανά το κέντρο βάρους κάθε συστάδας όπως απεικονίζεται στην εικόνα 14:



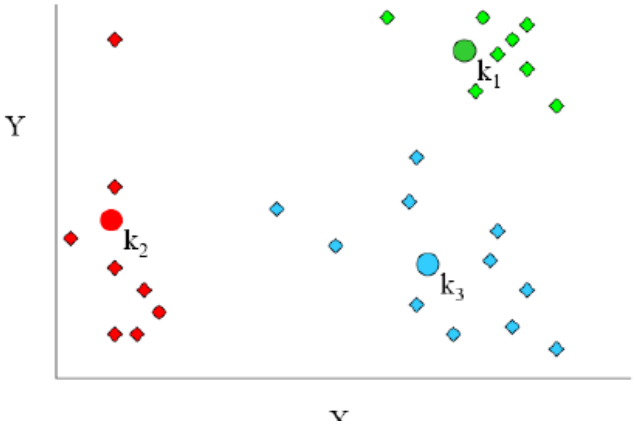
Εικόνα 14 Νέο κέντρο βάρους συστάδων

Ακολουθεί η νέα ανάθεση των σημείων και υπολογίζονται τα νέα κέντρα όπως φαίνονται στην εικόνα 15:



Εικόνα 15 Νέα ανάθεση σημείων και υπολογισμός νέων κέντρων

Σε αυτό το σημείο που βρισκόμαστε βλέπουμε ότι δεν υπάρχει καμία αλλαγή. Η διαδικασία τελειώνει όπως παρουσιάζεται στην εικόνα 16:



Εικόνα 16 Πλήρης ομαδοποίηση σημείων σε συστάδες

Συμπερασματικά ο αλγόριθμος k-means προσπαθεί επαναληπτικά να μειώσει την απόσταση όλων των σημείων από ένα σημείο της συστάδας.

Ακολουθεί ένα πιο συγκεκριμένο παράδειγμα εφαρμογής του αλγορίθμου κ–means: Έχουμε τα σημεία Α = (1,1), Β = (2,2), Γ = (2,3), Δ = (4,4), Ε = (4,6), Ζ = (5,5). Χρησιμοποιώντας την απόσταση Manhattan , η οποία ορίζεται ως

$$d(x,y) = |x_1 - x_2| + |y_1 - y_2|$$

όπου $x= (x_1,x_2)$ και $y= (y_1,y_2)$, θέλουμε να ομαδοποιήσουμε το σύνολο σημείων σε δύο ομάδες. Ως αρχικά κέντρα βάρους χρησιμοποιούμε τα σημεία $C_1 = (0,0)$ και $C_2 = (6,6)$. Επειδή υπάρχουν δύο ομάδες σημαίνει πως k = 2. Υπολογίζουμε τις αποστάσεις όλων των σημείων από τα δύο κέντρα:

Σημείο	Απόσταση από κέντρο C_1	Απόσταση από κέντρο C_2
Α	$d(A, C_1) = \|1-0\| + \|1-0\| = 2$	$d(A, C_2) = \|1-6\| + \|1-6\| = 10$
Β	$d(B, C_1) = \|2-0\| + \|2-0\| = 4$	$d(B, C_2) = \|2-6\| + \|2-6\| = 8$
Γ	$d(\Gamma,C_1) = \|2-0\| + \|3-0\| = 5$	$d(\Gamma,C_2) = \|2-6\| + \|3-6\| = 7$
Δ	$d(\Delta,C_1) = \|4-0\| + \|4-0\| = 8$	$d(\Delta,C_2) = \|4-6\| + \|4-6\| = 4$
Ε	$d(E,C_1) = \|4-0\| + \|6-0\| = 10$	$d(E,C_2) = \|4-6\| + \|6-6\| = 2$
Ζ	$d(Z,C_1) = \|5-0\| + \|5-0\| = 10$	$d(Z,C_2) = \|5-6\| + \|5-6\| = 2$

Άρα το σημείο Α ανήκει στην πρώτη ομάδα γιατί $d(A,C_1) < d(A,C_2)$. Το σημείο Β ανήκει στην πρώτη ομάδα γιατί $d(B,C_1) < d(B,C_2)$. Το σημείο Γ ανήκει στην πρώτη ομάδα γιατί $d(Γ,C_1) < d(Γ,C_2)$. Το σημείο Δ ανήκει στην δεύτερη ομάδα γιατί $d(Δ,C_1) > d(Δ,C_2)$. Το σημείο Ε ανήκει στην δεύτερη ομάδα γιατί $d(E,C_1) > d(E,C_2)$. Τέλος το σημείο Z ανήκει στην δεύτερη ομάδα γιατί $d(Z,C_1) < d(Z,C_2)$. Άρα η πρώτη ομάδα O_1 σχηματίζεται από τα σημεία $O_1=\{A,B,Γ\}$, ενώ η δεύτερη $O_2=\{Δ,E,Z\}$.

Το επόμενο βήμα είναι ο υπολογισμός των νέων κέντρων των δύο ομάδων: ο μέσος της πρώτης ομάδας είναι: $C_1 = (1+2+2 / 3 , 1+2+3 / 3) = (5/3 , 6/3) = (5/3 , 2)$. Ο μέσος της δεύτερης ομάδας είναι: $C_2 = (4+4+5 / 3 , 4+6+5 / 3) = (13/3 , 15/3) = (13/3 , 5)$.

Επειδή τα δύο κέντρα των ομάδων έχουν αλλάξει, εκτελούμε εκ νέου τα παραπάνω βήματα. Υπολογίζουμε τις αποστάσεις των σημείων από τα νέα κέντρα:

Σημείο	Απόσταση από κέντρο C_1	Απόσταση από κέντρο C_2
A	$d(A,C_1) = \|1-5/3\| + \|1-2\| = 1,667$	$d(A,C_2) = \|1-13/3\| + \|1-5\| = 7,33$
B	$d(B,C_1) = \|2-5/3\| + \|2-2\| = 0,333$	$d(B,C_2) = \|2-13/3\| + \|2-5\| = 5.33$
Γ	$d(Γ,C_1) = \|2-5/3\| + \|3-2\| = 1,333$	$d(Γ,C_2) = \|2-13/3\| + \|3-6\| = 4,33$
Δ	$d(Δ,C_1) = \|4-5/3\| + \|4-2\| = 4,333$	$d(Δ,C_2) = \|4-13/3\| + \|4-5\| = 1,33$
E	$d(E,C_1) = \|4-5/3\| + \|6-2\| = 6,333$	$d(E,C_2) = \|4-13/3\| + \|6-5\| = 1,33$
Z	$d(Z,C_1) = \|5-5/3\| + \|5-2\| = 6,333$	$d(Z,C_2) = \|5-13/3\| + \|5-5\| = 0,66$

Βλέπουμε ότι οι ομάδες δεν μεταβλήθηκαν και τα κέντρα τους δεν άλλαξαν. Άρα οι ομάδες είναι: $O_1=\{A,B,Γ\}$ με κέντρο $C_1=(5/3,2)$ και η ομάδα $O_2=\{Δ,E,Z\}$ με κέντρο $C_2=(13/3,5)$ [16].

Στη συνέχεια κάνουμε μια μικρή αναφορά στον αλγόριθμο k- medoids: Είναι μια παραλλαγή του k-means αλγόριθμου. Τα medoids είναι αντικείμενα από ένα σύνολο δεδομένων τα οποία μοιάζουν μεταξύ τους. Χρησιμοποιείται κυρίως όταν ένα κέντρο δεν έχει προσδιοριστεί. Απαιτεί αρκετά μεγάλες επαναλήψεις στο πέρασμα των δεδομένων και δεν ενδύκνυται στην διαίρεση των συστάδων σε μεγάλα datasets [38] .

5.4 ΕΦΑΡΜΟΓΕΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Όπως αναφέραμε σε αυτό το κεφάλαιο, μία από τις εφαρμογές της συσταδοποίησης είναι στο χώρο του μάρκετινγκ. Ομαδοποιώντας τους πελάτες μιας επιχείρησης με βάση τις συμπεριφορές τους στο τι αγοράζουν, μπορεί να δημιουργηθεί μία ομαδοποίηση εξειδικευμένης αγοράς η οποία θα βοηθήσει τις προωθητικές ενέργειες για διαφημιστικούς λόγους. Όλα αυτά έχουν σαν αποτέλεσμα, η επιχείρηση να στοχεύσει σε συγκεκριμένα τμήματα αγοράς και να εφαρμόσει μια συγκεκριμένη στρατηγική για κάθε τμήμα. Όπως καταλαβαίνουμε, είναι αδύνατον να εφαρμοστεί μια στρατηγική για ολόκληρη την αγορά αφού οι πελάτες έχουν διαφορετική αγοραστική συμπεριφορά. Έτσι με την συγκεκριμένη τεχνική εξόρυξης γνώσης οι πελάτες με ίδια χαρακτηριστικά χωρίζονται σε τμήματα που μπορούν να αντιμετωπιστούν πιο εύκολα [34] .

Μία άλλη εφαρμογή είναι στον χώρο των ασφαλίσεων. Γίνεται εντοπισμός των κατόχων ασφάλισης αυτοκινήτου με ένα υψηλό μέσο όρο απαιτεί κόστος. Επίσης στην πολεοδομία. Βρίσκονται τα συγκροτήματα κατοικιών σύμφωνα με τον τύπο του σπιτιού, την αξία και την γεωγραφική θέση [39] .

Στο [40] περιγράφεται μια εφαρμογή του αλγορίθμου Κ-μέσων για την ομαδοποίηση ομοιογενώς κατανεμημένων δεδομένων σε ένα περιβάλλον όπως τα δίκτυα αισθητήρων. Η επικοινωνία με τους γειτονικούς κόμβους είναι ασύγχρονη και δίνεται μια θεωρητική ανάλυση του αλγορίθμου, σε συνδυασμό με την συγκεντρωτική προσέγγιση που απαιτεί την λήψη όλων των παρατηρούμενων δεδομένων σε ένα μόνο χώρο. Δηλαδή, οι κόμβοι του

δικτύου δεν επικοινωνούν μεταξύ τους και θέλουμε να ομαδοποιήσουμε τα δεδομένα των κόμβων εφαρμόζοντας των αλγόριθμο κ-μέσων, όταν τα δεδομένα όλου του δικτύου είναι σε ένα συγκεκριμένο σημείο. Τα αποτελέσματα δείχνουν ότι, όταν όλα τα δεδομένα μεταδίδονται σε μία κεντρική τοποθεσία για την εφαρμογή του αλγορίθμου συσταδοποίησης, το κόστος επικοινωνίας (ένας σημαντικός παράγοντας σε δίκτυα αισθητήρων τα οποία είναι συνήθως εξοπλισμένα με περιορισμένη ισχύ μπαταρίας) είναι σημαντικά μικρότερο. Όπως επίσης η ακρίβεια των λαμβανόμενων κέντρων είναι υψηλή και ο αριθμός των δειγμάτων τα οποία έχουν λανθασμένη επισήμανση είναι επίσης μικρή.

Επίσης στο [41] εφαρμόζεται μια παραλλαγή του αλγορίθμου κ-μέσων σε κείμενα συνεχούς ροής (streaming text) η οποία αναφέρεται ως OSKM (online spherical k-means). Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί την τεχνική « ο νικητής-τα παίρνει-όλα.» και σαν τελικά αποτελέσματα έχουμε, ότι όταν ο παράγοντας μνήμης έχει μικρή παράμετρο , δηλαδή το να αφαιρεί γρήγορα χαρακτηριστικά που δεν χρησιμοποιούνται, οδηγούμαστε σε καλύτερα αποτελέσματα συσταδοποίησης.

6 ΧΡΗΣΗ ΛΟΓΙΣΜΙΚΟΥ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Για να εξάγουμε χρήσιμες πληροφορίες από τα διαθέσιμα δεδομένα χρησιμοποιούνται υπολογιστικές μέθοδοι όπως στατιστική ανάλυση, δέντρα αποφάσεων, νευρωνικά δίκτυα, κανόνες συσχέτισης και γραφική αναπαράσταση, ώστε να μετατραπούν τα δεδομένα σε πληροφορίες. Κάποια λογισμικά που χρησιμοποιούντα για την διαδικασία του data mining είναι: το SAS Enterprise Miner, SPSS Clementine, Statistica Data Miner, MS SQL Server, Polyanalyst, Knowledge STUDIO, αρκετά αποδοτικοί αλγόριθμοι, αλλά απαιτούν άριστες εγκαταστάσεις σε σχέση με τον χειρισμό των δεδομένων και μετά δεδομένων. Επίσης, χρησιμοποιείται το ελεύθερο λογισμικό Weka (Waikato Environment for Knowledge Analysis) Free (GPLed) Java package with GUI διαθέσιμο στην http://www.cs.waikato.ac.nz/ml/weka/

Τέλος χρησιμοποιούνται R packages E.g. r part, class, tree, n net, c clust, deal, GeneSOM, knn Tree, mlbench, randomForest, subselect. Άλλα ελεύθερα λογισμικά είναι το Karypis Lab και το Illimine (Illinois Data Mining System). Τέλος, ένα εμπορικό λογισμικό που χρησιμοποιείται είναι το Oracle Data mining and miner και το IBM DB2 Intelligent Miner.

6.1 ΛΟΓΙΣΜΙΚΟ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΕΙΤΑΙ ΣΤΗΝ ΤΑΞΙΝΟΜΗΣΗ

Τα λογισμικά που χρησιμοποιούντα στην μέθοδο της ταξινόμησης για την επεξεργασία των δεδομένων και την εξαγωγή τους, είναι αρκετά. Όπως έχουμε αναφέρει, τα εργαλεία που χρησιμοποιούνται είναι τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων, αλγόριθμοι κοντινότερου γείτονα, κανόνες (Rule-based methods), memory based reasoning, Naives Bayes, Bayesian Belief δίκτυα, και support vector machines (SVM). Έτσι θα αναφέρουμε λογισμικά για κάποια από αυτά τα εργαλεία.

Για τα δέντρα αποφάσεων, υπάρχουν επαγγελματικά λογισμικά όπως το AC2, το οποίο παρέχει γραφικά εργαλεία για την προετοιμασία των δεδομένων και την δημιουργία δέντρων. Alice d'Isoft 6.0, μια εκδοχή της ISoft's decision-tree-based AC2 data-mining product, είναι σχεδιασμένο από mainstream business users. Το AngossKnowledgeSEEKER, παρέχει την ρίσκου, την προετοιμασία των δεδομένων και την εξερεύνηση γνώσης για καλύτερο διαχωρισμό και για την κατανόηση καταναλωτικής συμπεριφοράς. Επίσης, υπάρχουν και τα :Angoss StrategyBUILDER, BigML, C5.0/See5, CART 5.0 decision-tree software, Citrus Technology Replay Professional.

Για τα νευρωνικά δίκτυα στην Neural Network FAQ liST υπάρχουν δωρεάν και επαγγελματικά λογισμικά από την WarrenSarleofSAS. Επίσης στο Portal for forecasting with neural networks υπάρχουν λογισμικά, δεδομένα και άλλα. Ενδεικτικά παραδείγματα είναι τα: Alyuda Neuro Intelligence, Bio Compi Model (tm) ,BrainMaker. Δωρεάν λογισμικά για τα νευρωνικά δίκτυα είναι το: Nuclass 7, το Scengy RPF και το Sharky Neural Network.

Για τα SVM επαγγελματικά :KXEN ,Tiberius, και το Treparel KMXBig Data Text Analytics &Visualization. Τα δωρεάν λογισμικά είναι :BSVM για μεγάλα προβλήματα ταξινόμησης, e1071 Rpackage, KernelMachines , LS-SVMlab, LeastSquares – Support Vector Machines Matlab/CToolbox ,LIBSVM, SVM-light από τον ThorstenJoachims.

Για τα Rule-based methods είναι : Compumine Rule Discovery System, Datamite, DMT Nuggets, PolyAnalyst, SuperQuery, WizWhy, XpertRule Miner Δωρεάν είναι: AutoUniv, CBA, DM-II ,KINOsuite-PR, PNC2 Rule Induction System. Περισσότερα λογισμικά διατίθεται στην http://www.kdnuggets.com/software/classification.html

6.2 ΛΟΓΙΣΜΙΚΟ ΓΙΑ ΕΥΡΕΣΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

Τα λογισμικά που χρησιμοποιούνται είναι το AzmySuperQuery , το IBMSPSS Modeler Suiteτο οποίο περιλαμβάνει και την ανάλυση καλαθιού αγοράς. Επίσης το LPA Data mining Toolkit υποστηρίζει την ανακάλυψη κανόνων από σχετικές βάσεις δεδομένων.

Τα δωρεάν λογισμικά είναι το Apriori, βρίσκει κανόνες συσχέτισης μέσω του αλγόριθμου apriori. Επίσης τοFP-growth, Eclatand DIC implementations από Bart Goethals και τέλος το ARtool το οποίο είναι μια συλλογή από αλγόριθμους σε δυαδικές βάσεις δεδομένων. Περισσότερες πληροφορίες και σύνδεσμοι στην http://www.kdnuggets.com/software/associations.html.

6.3 ΛΟΓΙΣΜΙΚΟ ΓΙΑ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Ένα λογισμικό που χρησιμοποιείται στην ομαδοποίηση είναι το Clustan Graphics3, με ιεραρχική ανάλυση από πάνω προς τα κάτω. Επίσης το CMSR Data Minerτο οποίο δημιουργήθηκε για επιχειρήσεις. Το IBM SPSS Modeler, εμπεριέχει τα Kohonen δίκτυα , τον Two-step cluster και τον k-means, αλγόριθμο. Τέλος, το MATLAB εμπεριέχει αλγόριθμους όπως ο k-means. Ελεύθερα και ανοιχτού τύπου είναι τα CLUTO, το Autoclass Cκαι το Databionic ESOM Tools. Περισσότερα λογισμικά και σελίδες στην http://www.kdnuggets.com/software/clustering.html και στην http://www.classification-society.org/csna/mda-sw / .

Τέλος, δύο από τα συστήματα ομαδοποίησης τα οποία βοηθούν την επιχείρηση στην τμηματοποίηση της αγοράς είναι: το PRIZM^{TM} από την εταιρεία Claritas και το MicroVision^{TM}από την εταιρεία Equifax. Αυτές οι εταιρείες έχουν ομαδοποιήσει τον πληθυσμό από δημογραφικές πληροφορίες σε τμήματα που πιστεύουν ότι είναι χρήσιμα για άμεσο μάρκετινγκ και πωλήσεις. Για να γίνει η ομαδοποίηση χρησιμοποίησαν πληροφορίες, όπως το εισόδημα, την ηλικία, το επάγγελμα, κτλ.

6.4 ΛΟΓΙΣΜΙΚΟ MATLAB

Σε προηγούμενα κεφάλαια μας αναφερθήκαμε στο λογισμικό Matlab. Σε αυτό το κεφάλαιο θα μιλήσουμε πιο αναλυτικά για αυτό. Το όνομά του είναι σύνθετο δύο λέξεων: MATrix και LABoratory. Είναι μια προγραμματιστική γλώσσα. Χρησιμοποιείται για την επίλυση μαθηματικών προβλημάτων και τον προγραμματισμό, καθώς περιέχει εντολές από την γλώσσα προγραμματισμού C++, όπως την while, την switch και την if. Όσον αφορά τα μαθηματικά μπορεί να εκτελέσει συναρτήσεις πραγματικές και άλλες. Τέλος στην στατιστική μπορεί να απεικονίσει ιστογράμματα, ραβδοδιαγράμματα και άλλα [42].

6.4.1 ΛΕΙΤΟΥΡΓΙΑ ΛΟΓΙΣΜΙΚΟΥ

Το συγκεκριμένο λογισμικό λαμβάνει τα δεδομένα από το περιβάλλον εργασίας και δίνει στο χρήστη τα αποτελέσματα των υπολογισμών. Επίσης μπορεί να εισάγει και δεδομένα από εξωτερικά αρχεία, καθώς επίσης και να εξάγει αποτελέσματα σε διαφορετικά αρχεία [43] . Για να εκτελέσουμε κάποια εντολή σε κάποιο πρόγραμμα χρησιμοποιούμε συναρτήσεις της μορφής function(x),όπου function το όνομα της συνάρτησης και x αυτό που θέλουμε να ταξινομήσουμε. Στις επόμενες υποενότητες θα δούμε πώς εφαρμόζονται οι τεχνικές που αναφέραμε και αναλύσαμε στα προηγούμενα κεφάλαια σε αυτό.

6.4.2 ΕΦΑΡΜΟΓΗ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ MATLAB

Ας ξεκινήσουμε να δούμε πως δημιουργούμε μία ταξινόμηση κατά Bayes. Η εντολή είναι η εξής: nb=NaiveBayes.fit(training,class). Με αυτό τον τρόπο δημιουργεί ένα

αντικείμενο Naive Bayes ταξινομητή (nb.training) το οποίο είναι μία μήτρα από δεδομένα εκπαίδευσης. Οι γραμμές αντιστοιχούν σε παρατηρήσεις και οι στήλες σε χαρακτηριστικά. Κάθε στοιχείο της κατηγορίας δείχνει σε ποια κατηγορία ανήκει η αντίστοιχη γραμμή.

Ας δούμε μετά πώς εφαρμόζεται ο αλγόριθμος ταξινόμησης knn στο Matlab. Συντάσσεται ως εξής :

IDX = knnsearch(X,Y)
[IDX,D] = knnsearch(X,Y)
[IDX,D] = knnsearch(X,Y,'*Name*',*Value*) , όπου:

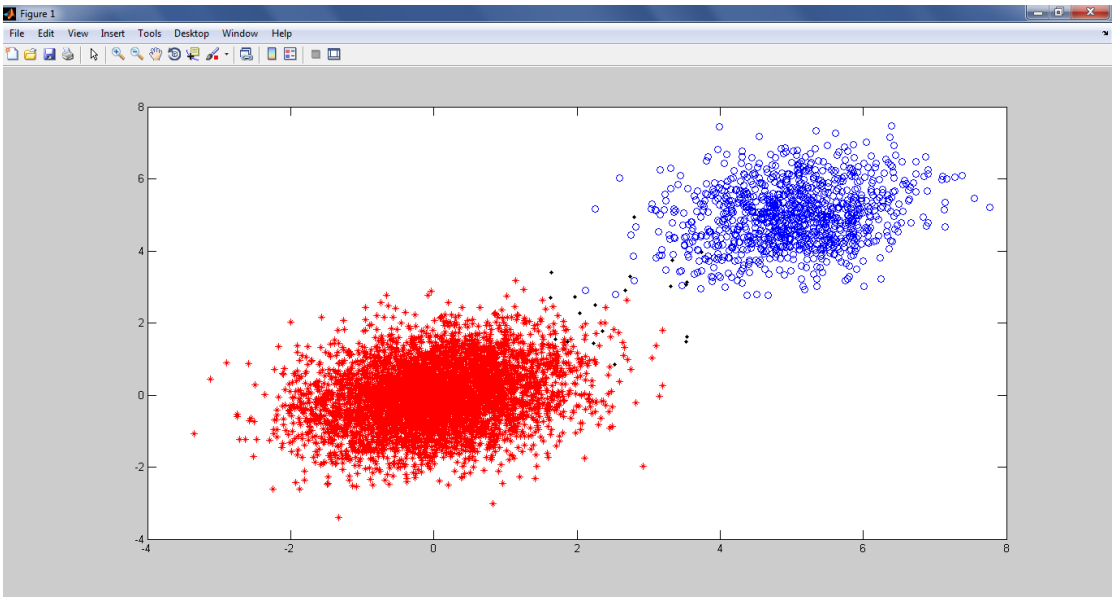
IDX = knnsearch(X,Y) βρίσκει τον κοντινότερο γείτονα στο X για κάθε σημείο στο Y. Το IDX είναι, ένα διάνυσμα στηλών με τις σειρές. Κάθε σειρά περιέχει τον δείκτη του κοντινότερου γείτονα στο X για την αντίστοιχη σειρά στο Y.
[IDX,D] = knnsearch(X,Y) επιστρέφει ένα από 1 διανυσματικό D περιέχοντας τις αποστάσεις μεταξύ κάθε παρατήρησης στο Y και της αντίστοιχης πιο στενής παρατήρησης στο X. Δηλαδή το D(i) είναι η απόσταση μεταξύ του X(IDX(i),:) και Y(i,:).
[IDX,D] = knnsearch(X,Y,'*Name*',*Value*) δέχεται ένα ή περισσότερα προαιρετικά όνομα και αξία [44] .

Ας δούμε στη συνέχεια ένα παράδειγμα κατηγοριοποίησης απόστασης knn στο Matlab. Τα διανύσματα των δεδομένων ταξινομούνται σε δύο ισοπίθανες κλάσεις C_1 και C_2. Οι κλάσεις μοντελοποιούνται μέσω της κανονικής κατανομής με μέσες τιμές $m_1 = [0\ 0]^T$ και $m_2 = [5.0\ 5.0]^T$. Ο πίνακας συνδιασπορών και των δύο κατανομών είναι: S1 = S2 = [0.8 0.2 0.2 0.8]. Θέλουμε να δημιουργήσουμε δύο σύνολα δεδομένων X_1 και X_2, τα οποία αποτελούνται από 5000 και 1000 σημεία αντίστοιχα. Το σύνολο εκπαίδευσης θα είναι: X = [X1 X2].
Στη συνέχεια θέλουμε να δημιουργήσουμε ένα διάνυσμα γραμμής, η διάσταση του οποίου θα είναι αντίστοιχη με το άθροισμα των διαστάσεων των συνόλων X_1 και X_2. Σε κάθε θέση του διανύσματος θα είναι ένας αριθμός που θα αντιστοιχεί στην κλάση που ανήκει το κάθε δεδομένο εκπαίδευσης. Για την κλάση C_1 θα είναι 1, ενώ για την κλάση C_2 θα είναι 2.
Στη συνέχεια θέλουμε να δημιουργήσουμε ένα νέο σύνολο δεδομένων ds με 20 στοιχεία με μέση τιμή $m = [2.5\ 2.5]^T$ με τον ίδιο πίνακα συνδιασπορών.
Οι αντίστοιχες εντολές στο πρόγραμμα είναι:

```
m1 = [0 0]';
S1 = [0.8 0.2; 0.2 0.8];
N1 = 5000;
X1 = mvnrnd(m1, S1, N1);

m2 = [5.0 5.0]';
S2 = S1;
N2 = 1000;
X2 = mvnrnd(m2, S2, N2);

X = [X1 ; X2];
[r,c] = size(X);
classX = [ ones(1,N1)  2*ones(1,N2) ];

figure(1)
plot(X(classX(:)==1,1), X(classX(:)==1,2), 'r*', X(classX(:)==2,1), X(classX(:)==2,2), 'bo');

n = 20;
m = [2.5 2.5]';
ds = mvnrnd(m, S1, n);
```

```
hold on
plot(ds(:,1), ds(:,2), 'k.', 'MarkerSize', 10);
```

Μετά την εκτέλεση των εντολών παίρνουμε την εικόνα 17:

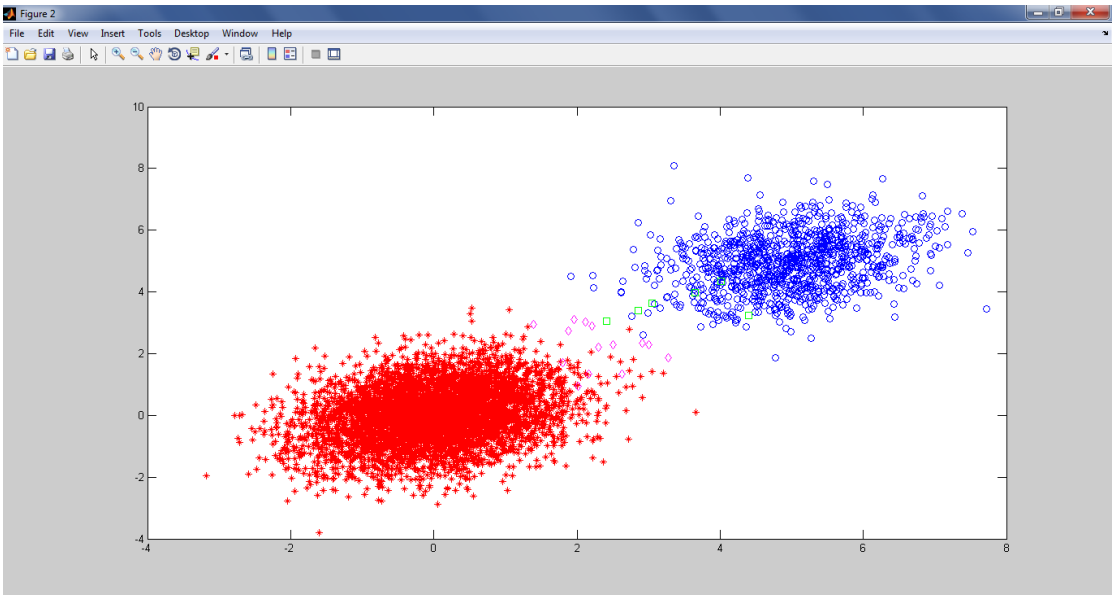


Εικόνα 17 Παράδειγμα k-means στο MATLAB

Τέλος θέλουμε να ταξινομήσουμε τα σημεία του συνόλου ds με k = 3 και με χρήση της Ευκλείδειας απόστασης. Κάνουμε κλήση της συνάρτησης knnclassify(). Οι εντολές είναι:

```
k=3;
class_ds = knnclassify(ds, X, classX, k, 'Euclidean', 'nearest');
figure(2)
plot(X(classX(:)==1,1), X(classX(:)==1,2), 'r*', X(classX(:)==2,1), X(classX(:)==2,2), 'bo');
hold on
plot(ds(class_ds(:)==1,1), ds(class_ds(:)==1,2), 'md', ds(class_ds(:)==2,1),
ds(class_ds(:)==2,2), 'gs', 'MarkerSize', 7);
```

Μετά την εκτέλεση των παραπάνω εντολών παίρνουμε την εικόνα 18:



Εικόνα 18 Παράδειγμα 2 k-means στο MATLAB

Έτσι τώρα τα δεδομένα έχουν ταξινομηθεί [45] .

6.4.2.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ ΣΤΟ MATLAB

Τώρα θα δούμε πως εφαρμόζονται τα δέντρα αποφάσεων στο MATLAB. Αρχικά, γράφουμε τον κώδικα. Θα χρησιμοποιήσουμε ένα σύνολο δεδομένων το Iris, το οποίο είναι ενσωματωμένο στο πρόγραμμα και απλώς θα το "καλέσουμε". Θα χρησιμοποιηθεί ο αλγόριθμος ID3 για να κατασκευάσουμε το δέντρο αποφάσεων. Αρχικά θα φορτώσουμε το σύνολο δεδομένων(fisheriris) στο MATLAB.

```
; data= load('fisheriris');
ds= data.meas;
class= data.species
```

Στη συνέχεια θα δημιουργήσουμε και θα τρέξουμε τα δεδομένα όπως φαίνεται παρακάτω:

```
ds_tr = [ds(1:40,:) ; ds(51:90,:) ; ds(101:140,:)];
class_ds_tr = [class(1:40) ; class(51:90) ; class(101:140)];
ds_ts = [ds(41:50,:) ; ds(91:100,:) ; ds(141:end,:)];
class_ds_ts = [class(41:50) ; class(91:100) ; class(141:150)];
```

Μετά θα κατασκευάσουμε το δέντρο αποφάσεων όπως παρακάτω:

```
DT = classregtree(ds_tr,class_ds_tr,'names',{'SL' 'SW' 'PL' 'PW'})
view(DT)
fit_ts = eval(DT,ds_ts);
```

Υπολογίζουμε ότι η αναλογία είναι σωστά ταξινομημένη:

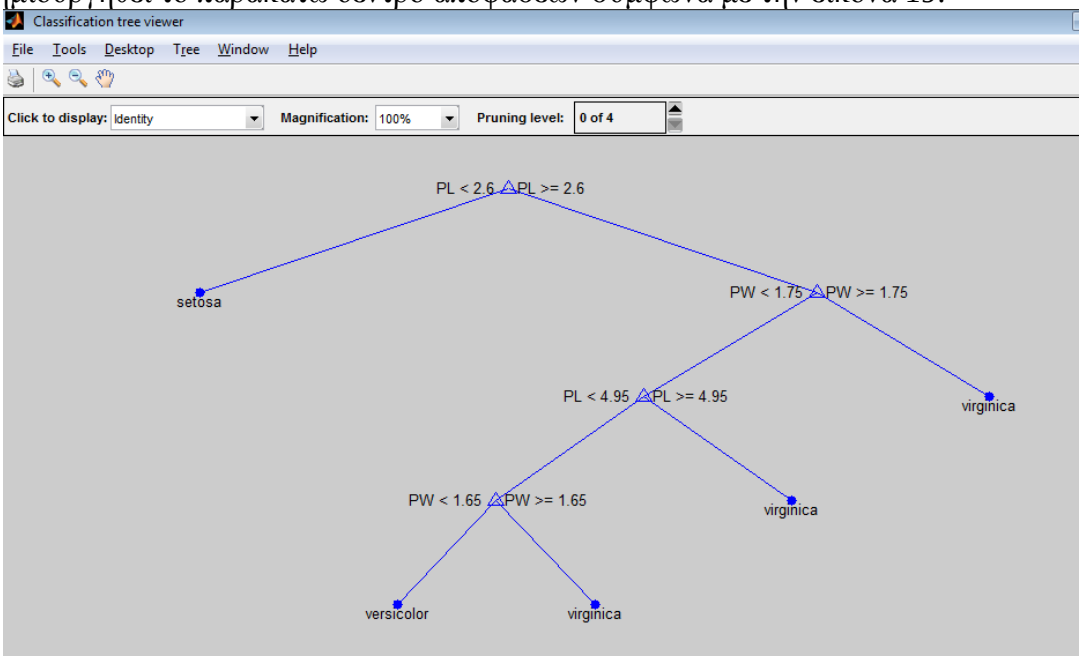
```
perf_dec_tree = sum(strcmp(fit_ts,class_ds_ts))/length(class_ds_ts);
fprintf(' The performance of Decision Tree to classify the objects of the ds_test is:  %5.2f ',
perf_dec_tree);
```

Στη συνέχεια το πρόγραμμα θα τυπώσει τα παρακάτω:

```
DT =
```

Decision tree for classification
1 if PL<2.6 then node 2 elseif PL>=2.6 then node 3 else setosa
2 class = setosa
3 if PW<1.75 then node 4 elseif PW>=1.75 then node 5 else versicolor
4 if PL<4.95 then node 6 elseif PL>=4.95 then node 7 else versicolor
5 class = virginica
6 if PW<1.65 then node 8 elseif PW>=1.65 then node 9 else versicolor
7 class = virginica
8 class = versicolor
9 class = virginica

The performance of Decision Tree to classify the objects of the ds_test is: 1.00 >> και θα δημιουργηθεί το παρακάτω δέντρο αποφάσεων σύμφωνα με την εικόνα 19:



Εικόνα 19 Παράδειγμα δέντρων αποφάσεων στο MATLAB

6.4.3 ΕΦΑΡΜΟΓΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΣΤΟ MATLAB

Ας δούμε πώς συντάσσεται ο αλγόριθμος ομαδοποίησης k-means. Οι εντολές είναι:
idx = kmeans(X,k)
idx = kmeans(X,k,Name,Value)
[idx,C] = kmeans(___)
[idx,C,sumd] = kmeans(___)
[idx,C,sumd,D] = kmeans(___), όπου

idx = kmeans(X,k) χωρίζει τις παρατηρήσεις X στοιχείων στις συστάδες και επιστρέφει το idx που περιέχει τους δείκτες συστάδων κάθε παρατήρησης. Οι σειρές του X αντιστοιχούν στα σημεία και οι στήλες αντιστοιχούν στις μεταβλητές.
idx = kmeans(X,k,Name,Value) επιστρέφει τους δείκτες συστάδων από ένα ή περισσότερα ονόματα και αξίες κάθε φορά.
[idx,C] = kmeans(___) επιστρέφει τις συστάδες θέσεις από την μήτρα C.

[idx,C,sumd] = kmeans(___)επιστρέφει τις αποστάσεις σημείων στο διάνυσμα sumd.
[idx,C,sumd,D] = kmeans(___)επιστρέφει τις αποστάσεις από κάθε σημείο σε κάθε μήτρα D [46] .

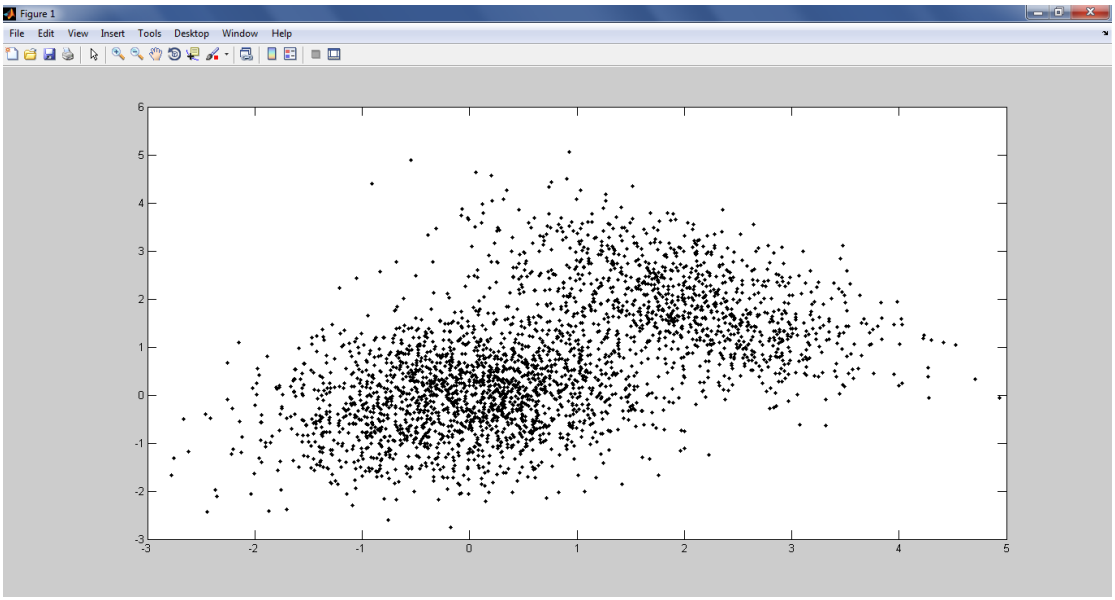
Ας δούμε στη συνέχεια ένα πιο συγκεκριμένο παράδειγμα αυτού του αλγορίθμου. Το παρακάτω παράδειγμα αναφέρεται στο χώρο των δύο διαστάσεων. Έχουμε ένα σύνολο δεδομένων X, και αυτά θέλουμε να ομαδοποιηθούν σε δύο ομάδες C_1 και C_2. Η πρώτη ομάδα αποτελείται από 2000 δεδομένα και έχουν μέση τιμή $m_1 = [0\ 0]^T$ και πίνακα συνδιασπορών S_1 = [0.8 0.2 0.2 0.8].

Η δεύτερη ομάδα αποτελείται από 1000 δεδομένα και έχουν μέση τιμή $m_2 = [2\ 2]^T$ και πίνακα συνδιασπορών S_2 = [0.9 -0.5 0.5 0.9].

Το σύνολο δεδομένων είναι: X=[X1 ; X2]. Οι αντίστοιχες εντολές στο πρόγραμμα Matlab είναι:

```
m1 = [0 0]';
S1 = [0.8 0.2; 0.2 0.8];
N1 = 2000;
X1 = mvnrnd(m1, S1, N1);

m2 = [2.0 2.0]';
S2 = [0.9 -0.5; -0.5 0.9];
N2 = 1000;
X2 = mvnrnd(m2, S2, N2);

X = [X1 ; X2];
[r,c] = size(X);

figure(1)
plot(X(:,1), X(:,2), 'k.');
```

Έτσι θέλουμε να δημιουργηθεί ένα σχήμα με όνομα figure(1).Με την εντολή plot δημιουργούμε το σχήμα. Μετά την εκτέλεση των εντολών δημιουργείται η εικόνα 20:

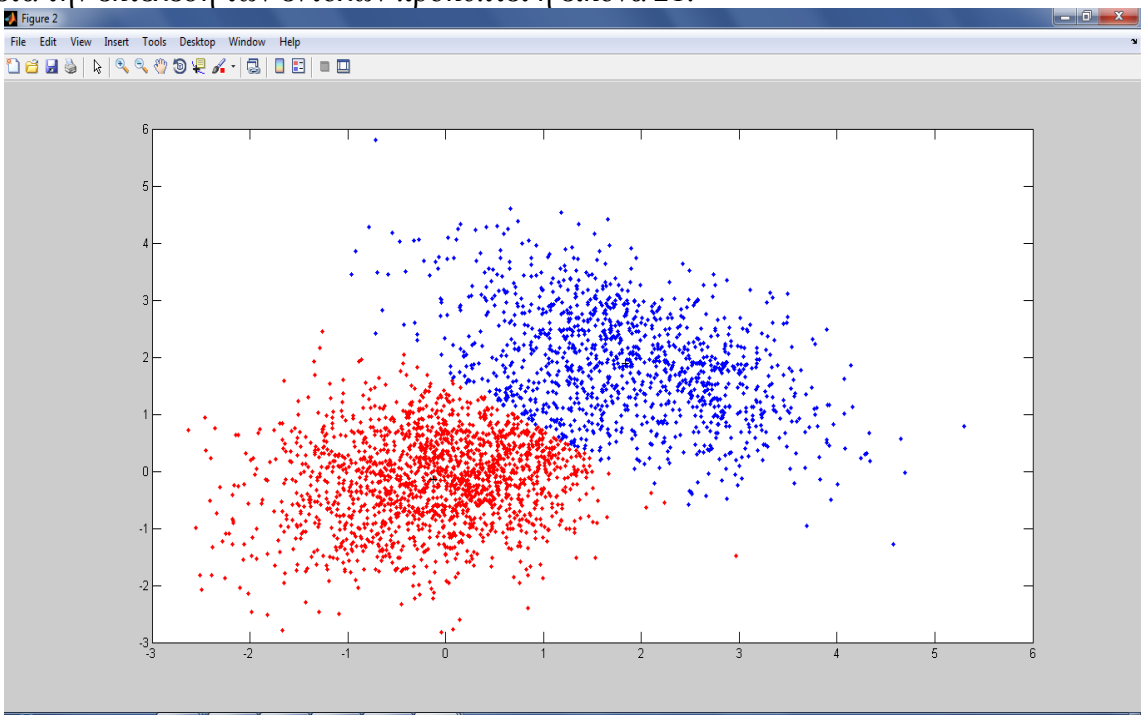


Εικόνα 20 Παράδειγμα k-means, Δεδομένα μη ομαδοποιημένα

Όπως βλέπουμε στην παραπάνω εικόνα 20 τα δεδομένα δεν είναι ομαδοποιημένα. Συνεχίζουμε στην ομαδοποίησή τους. Οι ομάδες είναι δύο. Άρα k = 2. Οι επιπρόσθετες εντολές είναι:

```
k=2;
opts = statset('Display','final');
[cidx, ctrs, sumd] = kmeans(X, k, 'Distance','sqEuclidean', 'Replicates',1, 'Options',opts);
sum2 = sum(sumd);

figure(2)
plot(X(cidx==1,1),X(cidx==1,2),'r.',X(cidx==2,1),X(cidx==2,2),'b.', ctrs(:,1),ctrs(:,2),'k+');
```

Η εντολή opts = statset('Display','final'); μας δείχνει ότι επιλέγουμε(opts) τα στατιστικά στοιχεία(statset) σε ποιο επίπεδο του αποτελέσματος('Display') θα εμφανιστούν. Θέλουμε το τελικό αποτέλεσμα('final').
Μετά γίνεται η ομαδοποίηση του X με βάση το k. Χρησιμοποιούμε την Ευκλείδεια απόσταση('Distance','sqEuclidean').Το 'Replicates' είναι το πόσες φορές να επαναλάβουμε την ομαδοποίηση χρησιμοποιώντας νέες θέσεις κέντρου βάρους της αρχικής ομαδοποίησης. Εδώ γίνεται μία φορά.
Μετά την εκτέλεση των εντολών προκύπτει η εικόνα 21:



Εικόνα 21 k-means, Δεδομένα ομαδοποιημένα

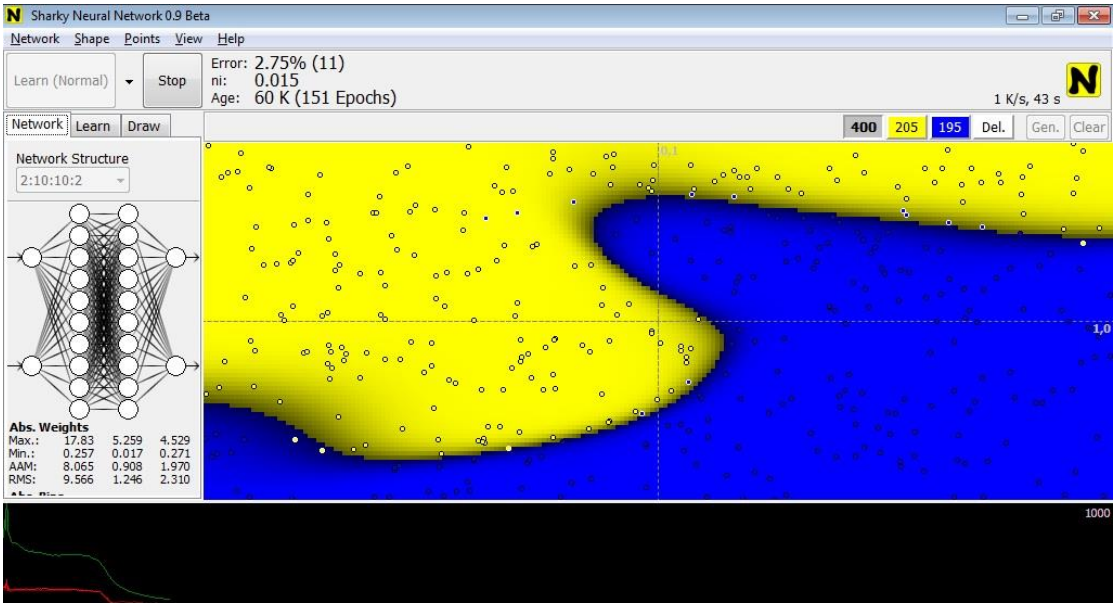
Παρατηρούμε ότι τα δεδομένα έχουν ομαδοποιηθεί στις δύο ομάδες [46] .

6.5 SHARKY NEURAL NETWORK

Υπάρχουν λογισμικά που χρησιμοποιούνται για την εκπαίδευση νευρωνικών δικτύων με σκοπό την ταξινόμηση. Ένα δωρεάν λογισμικό είναι το sharky neural network το οποίο είναι εύκολο και χρησιμοποιείται για διασκέδαση, και για εκπαιδευτικούς σκοπούς. Μπορούμε να το κατεβάσουμε από την Shark time Software http://www.sharktime.com/

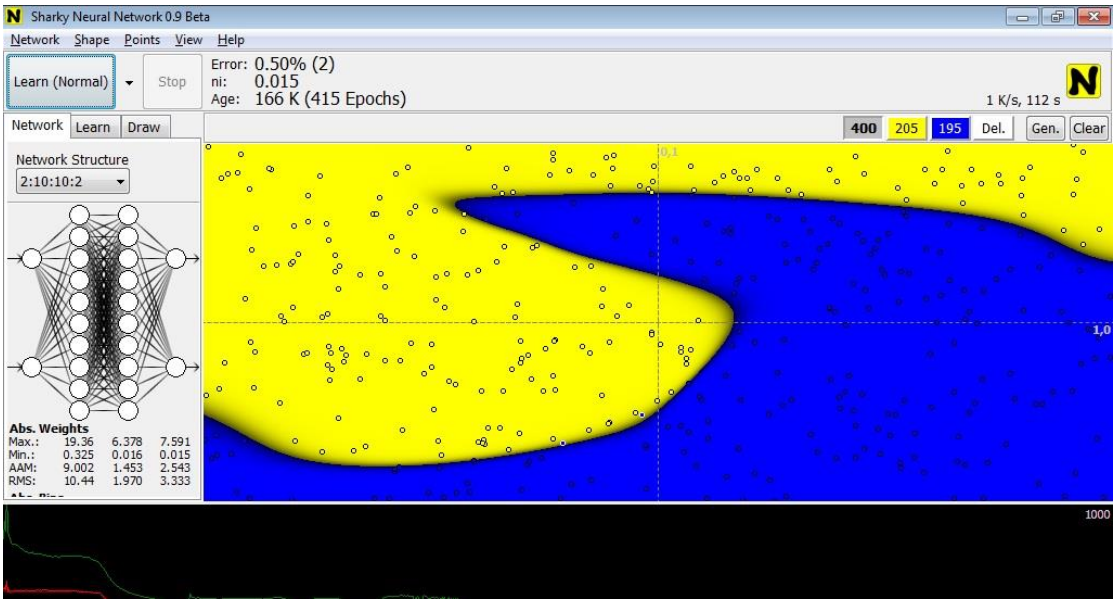
Ο σκοπός είναι να ταξινομήσει σημεία σε επίπεδο δύο διαστάσεων σε δύο κλάσεις(μπλε και κίτρινο). Παρακάτω θα δώσουμε ένα παράδειγμα ταξινόμησης σε αυτό το λογισμικό. Οι είσοδοι είναι σημεία που έχουν δύο τιμές (–X,Y) και η συνάρτηση που χρησιμοποιείται είναι η ($f(x)=2/(1+e^{-\beta x})-1$).

Υπάρχουν 400 σημεία, 205 κίτρινα και 195 μπλε. Θα χρησιμοποιήσουμε την δομή 2: 10:10:2, δηλαδή 2 αρχικούς κόμβους, 10 στο 1ο κρυφό επίπεδο, 10 στο 2ο κρυφό επίπεδο και 2 εξόδους. Το σχήμα δεν έχει σχέση με την ταξινόμηση των σημείων, αλλά αντιπροσωπεύει, αν ένα σημείο είναι μπλε ή κίτρινο. Αν θέλουμε να εισάγουμε σημεία που υπάρχουν στην ιστοσελίδα επιλέγουμε points→open points. Αφού έχουμε διαμορφώσει το δίκτυο πατάμε το κουμπί Learn για να ξεκινήσει η ταξινόμηση. Κατά την διάρκεια της ταξινόμησης(δεν έχει τελειώσει η ταξινόμηση των σημείων) έχουμε ότι έντεκα σημεία δεν έχουν ταξινομηθεί, όπως φαίνεται στην παρακάτω εικόνα 22:



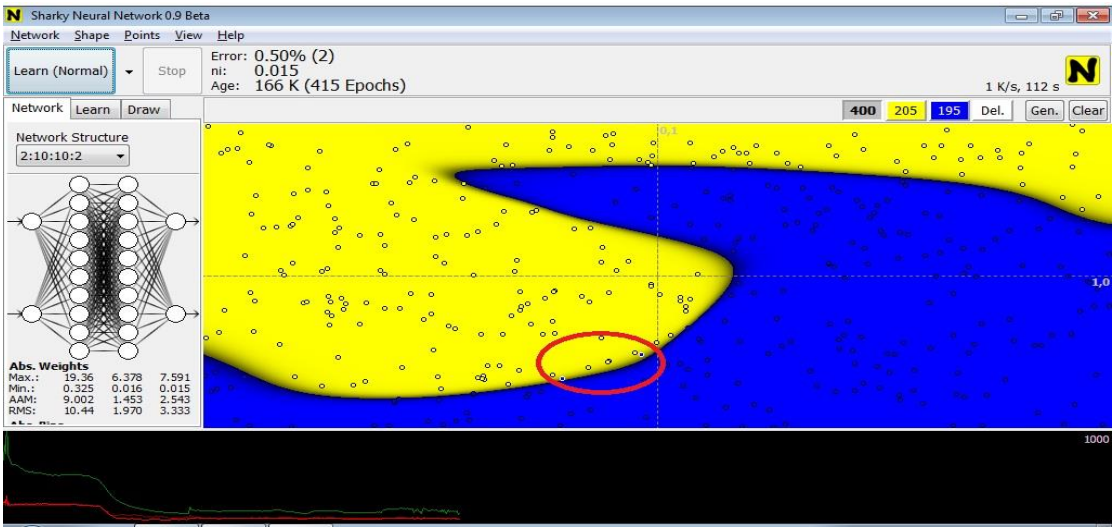
Εικόνα 22 Διαδικασία Sharky Neural Network

Το τελικό αποτέλεσμα είναι η παρακάτω εικόνα 23. Όπως βλέπουμε τα σημεία έχουν ταξινομηθεί, όμως δύο από αυτά δεν ξέρουμε αν είναι μπλε ή κίτρινο. Αυτό έχει σαν αποτέλεσμα το νευρωνικό δίκτυο που φτιάξαμε να έχει error0,50%



Εικόνα 23 Διαδικασία Sharky Neural Network

Τα σημεία που δεν ταξινομήθηκαν φαίνονται παρακάτω στην εικόνα 24 και είναι αυτά που έχουν χρώμα μπλε με άσπρο περίγραμμα.



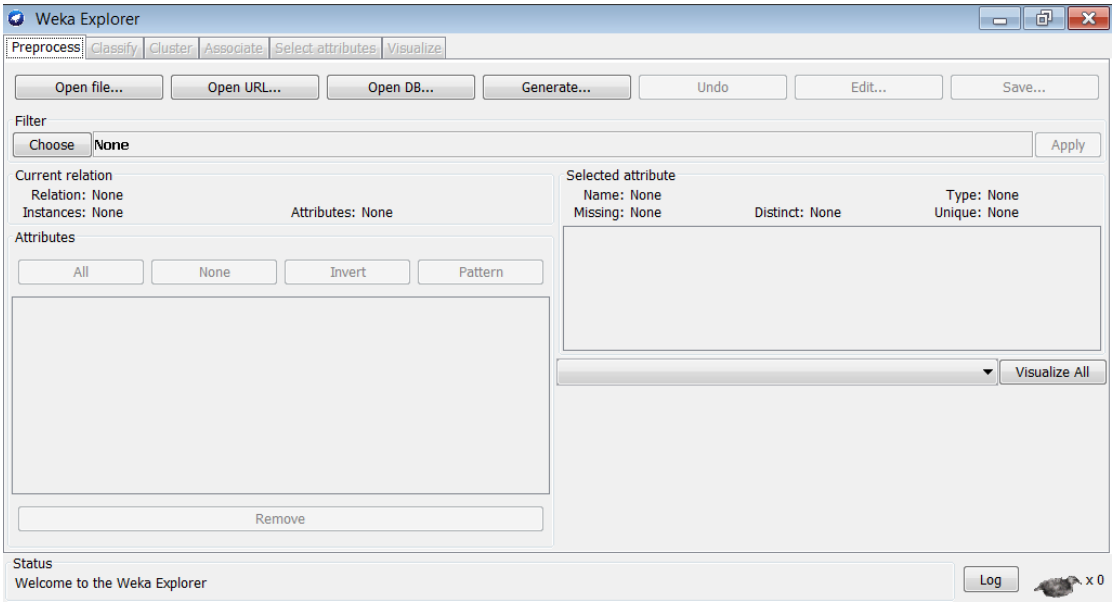
Εικόνα 24 Διαδικασία Sharky Neural Network

Η διαδικασία που ακολουθήσαμε στην συγκεκριμένη περίπτωση είναι απλή, όμως σε πιο περίπλοκα προβλήματα η εκπαίδευση των νευρωνικών δικτύων είναι πολύπλοκη και χρονοβόρα. Δηλαδή για να ολοκληρωθεί η διαδικασία μπορεί να περάσουν ώρες, αλλά και μέρες. Υπάρχει η δυνατότητα ο χρήστης να δημιουργήσει δικά του σημεία στο επίπεδο ή να εισάγει σημεία από ένα αρχείο μέσω του points→open points.

6.6 ΛΟΓΙΣΜΙΚΟ WEKA

Το WEKA (Waikato Environment for Knowledge Analysis), είναι ένα ανοιχτό λογισμικό που χρησιμοποιείται κυρίως για προβλήματα εξόρυξης δεδομένων. Η έκδοση που θα χρησιμοποιήσουμε είναι η 3.7.12 και είναι διαθέσιμη στην http://www.cs.waikato.ac.nz/ml/weka καθώς και αρκετές άλλες προηγούμενες εκδοχές του. Χρειάζεται η εγκατάσταση της γλώσσας προγραμματισμού JAVA 1.4.0 και άνω. Επίσης τα

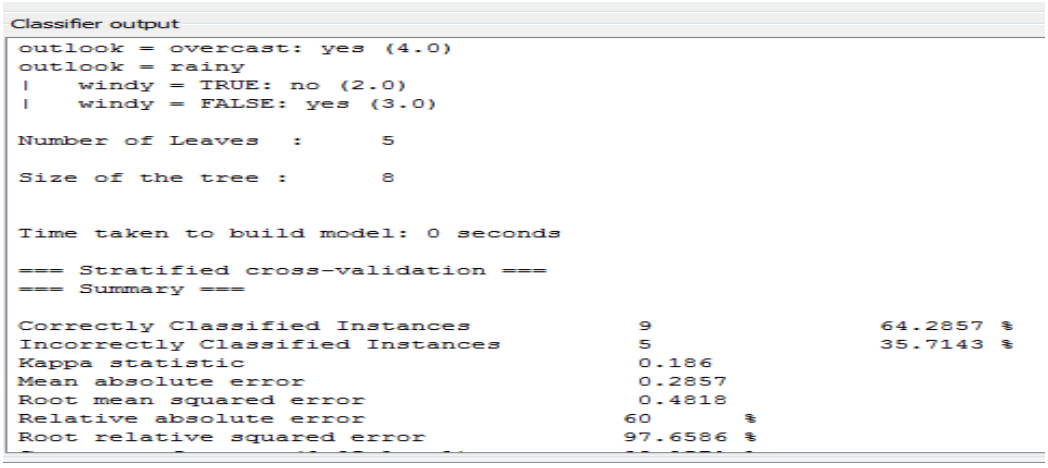
αρχεία που χρησιμοποιούνται σαν είσοδο από το πρόγραμμα είναι αρχεία arff τα οποία περιέχουν δεδομένα. Στον παγκόσμιο ιστό είναι διαθέσιμα αρκετά έτοιμα αρχεία arff, όμως υπάρχει η δυνατότητα να δημιουργήσουμε δικά μας αρχεία από ένα excel. Εισάγουμε τα δεδομένα σε ένα αρχείο excel ,στην συνέχεια το μετατρέπουμε σε CSV και τέλος σε ένα arff format. Στη συνέχεια θα δούμε μερικές εφαρμογές του, για την επίλυση προβλημάτων, όπως τα δέντρα αποφάσεων, ο αλγόριθμος k-means, και ο apriori. Να πούμε επίσης ότι μπορούν να χρησιμοποιηθούν κι άλλες εφαρμογές όπως νευρωνικά δίκτυα, οι αλγόριθμοι id3 και ο knn. Στην εικόνα 25, βλέπουμε αυτό που εμφανίζεται όταν ανοίγουμε το weka.



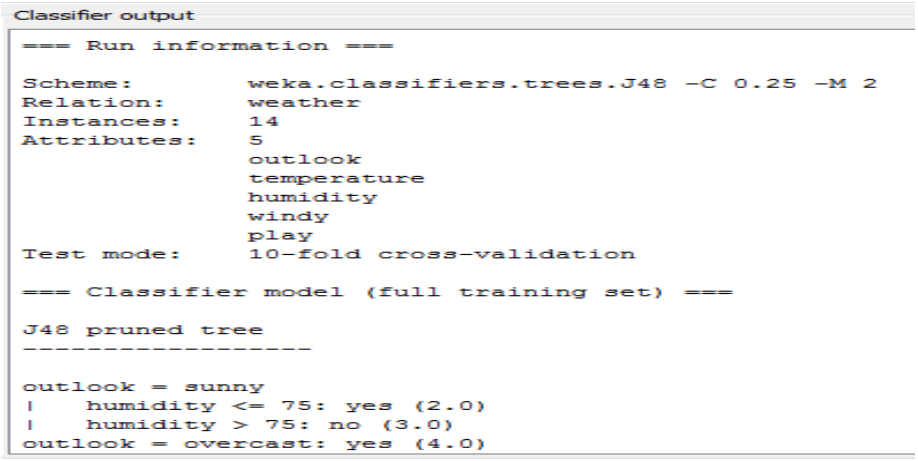
Εικόνα 25 Παρουσίαση WEKA

6.6.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ ΣΤΟ WEKA

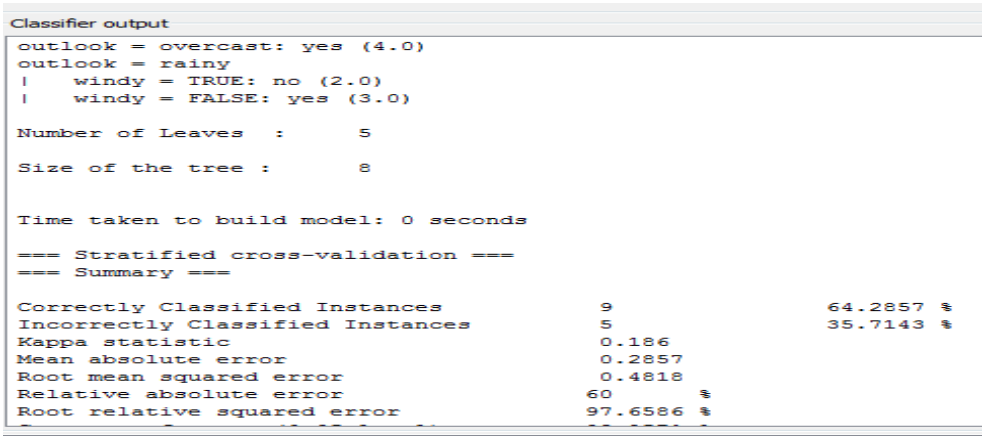
Αρχικά ανοίγουμε το πρόγραμμα και πατάμε, δεξιά την επιλογή Explorer. Θα εμφανιστεί ένα άλλο παράθυρο, στο οποίο θα εισάγουμε τα δεδομένα, στην επιλογή open file. Για την παρουσίαση του παραδείγματος, θα χρησιμοποιήσουμε το αρχείο weather.arff το οποίο δημιουργείται κατά την εγκατάσταση του προγράμματος και βρίσκεται στον φάκελο WEKA στον δίσκο C του υπολογιστή. Για την εκτέλεση θα χρησιμοποιήσουμε τον αλγόριθμο j48. Για να επιλέξουμε τον συγκεκριμένο αλγόριθμο θα πάμε επάνω στο κουμπί classify και θα επιλέξουμε στο κουμπί choose→ Classifiers→ trees→J48. Τέλος θα επιλέξουμε το κουμπί Start και κάτω αριστερά, θα εμφανιστούν τα αποτελέσματα όπως φαίνονται στην εικόνα 26, εικόνα 27 και εικόνα 28. Αρχικά δίνεται μια περιγραφή των δεδομένων, δηλαδή ότι υπάρχουν 14 στιγμιότυπα και 5 χαρακτηριστικά τα οποία είναι τα (outlook, temperature, humidity, windy, play) και ότι χρησιμοποιήθηκε η μέθοδος 10-fold cross-validation για να γίνει μια εκτίμηση του μοντέλου που παράχθηκε. Στη συνέχεια, παρουσιάζονται τα στοιχεία του δέντρου που παράχθηκε, ότι δηλαδή, είναι ένα «ψαλιδισμένο» δέντρο, έχει 5 «φύλλα» και μέγεθος 9.



Όπως βλέπουμε στην παραπάνω εικόνα, ο χρόνος που χρειάστηκε για να παραχθεί το δέντρο είναι 0 δευτερόλεπτα, και 9 από τα 14 στιγμιότυπα ταξινομήθηκαν σωστά, ενώ 5 από τα 14 δεν ταξινομήθηκαν σωστά. Στην συνέχεια παρουσιάζονται κάποιες μεταβλητές σφάλματος και μεταβλητές απόδοσης της ταξινόμησης στην εικόνα 26.



Εικόνα 26 Αποτελέσματα j48 WEKA



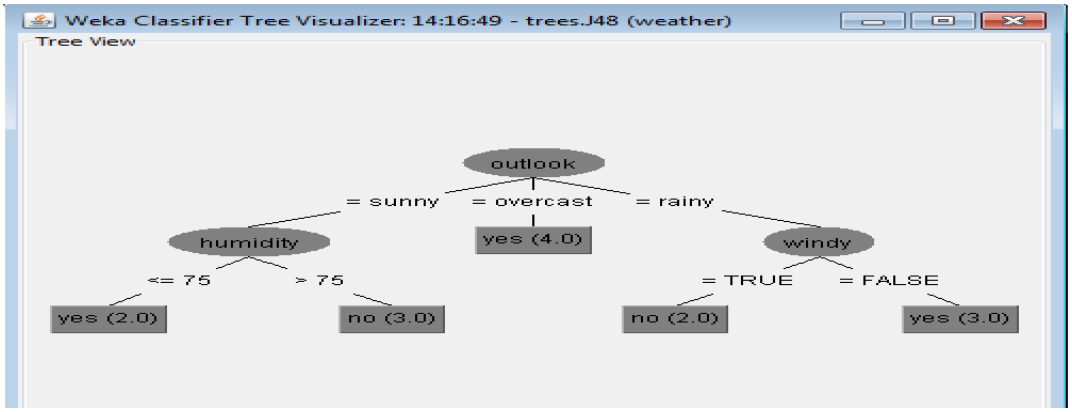
Εικόνα 27 αποτέλεσμα j48 weka

Classifier output

```
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 60      %
Root relative squared error             97.6586 %
Coverage of cases (0.95 level)          92.8571 %
Mean rel. region size (0.95 level)      64.2857 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,778    0,600    0,700      0,778    0,737      0,189    0,789     0,847     yes
                 0,400    0,222    0,500      0,400    0,444      0,189    0,789     0,738     no
Weighted Avg.    0,643    0,465    0,629      0,643    0,632      0,189    0,789     0,808

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

Εικόνα 28 Αποτελέσματα j48 weka

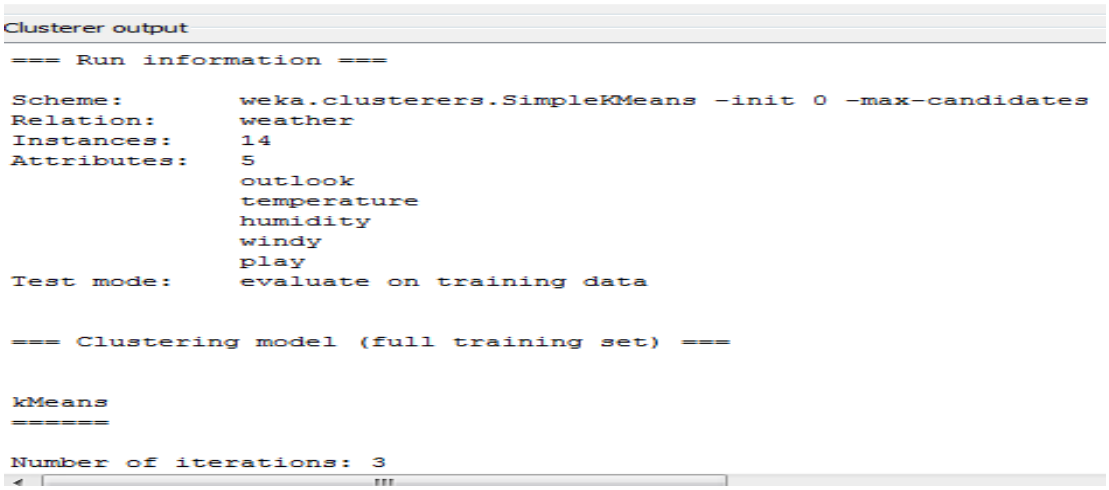
Για να δούμε το δέντρο αποφάσεων, που δημιουργήθηκε, θα πατήσουμε δεξί κλικ κάτω αριστερά, και θα επιλέξουμε Visual tree. Θα εμφανιστεί, το δέντρο αποφάσεων όπως παρακάτω εικόνα 29.



Εικόνα 29 Παρουσίαση δέντρου αποφάσεων WEKA

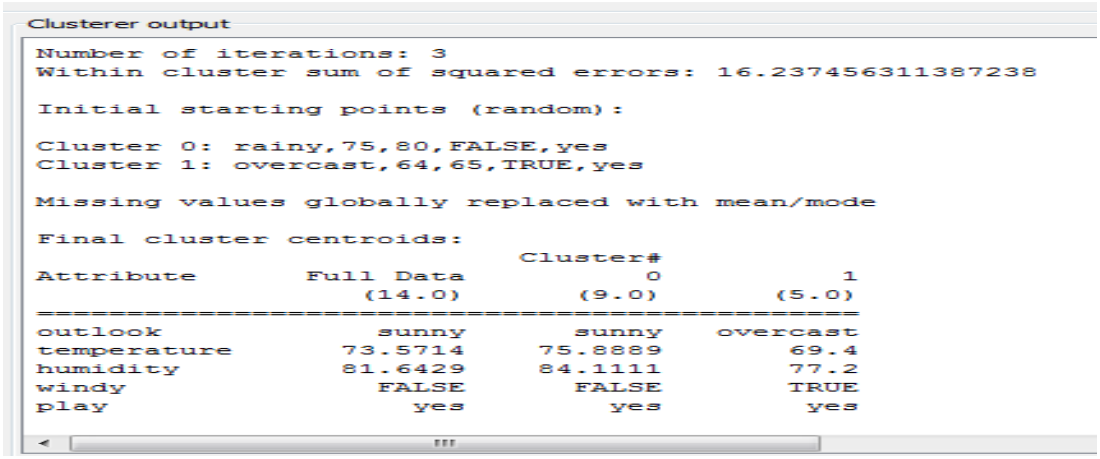
6.6.2 Ο ΑΛΓΟΡΙΘΜΟΣ Κ ΜΕΣΩΝ ΣΤΟ WEKA

Όπως έχουμε αναφέρει, στην ομαδοποίηση χρησιμοποιείται ο αλγόριθμος K-means. Έτσι παρακάτω θα δούμε πως μπορούμε να δούμε τα αποτελέσματα του στο weka. Αρχικά ανοίγουμε το πρόγραμμα και εισάγουμε τα δεδομένα όπως έχουμε αναφέρει σε προηγούμενη ενότητα. Σε αυτήν την περίπτωση θα χρησιμοποιήσουμε το αρχείο weather.arff. Τα αποτελέσματα φαίνονται παρακάτω στην εικόνα 30, εικόνα 31, εικόνα 32 και εικόνα 33.

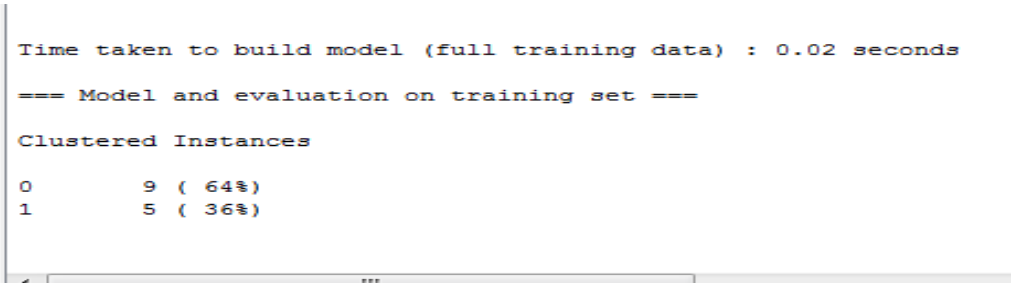


Εικόνα 30 k-means weka

Σαν αποτέλεσμα δείχνει ότι, χρειάστηκαν 5 επαναλήψεις, κάποιες μεταβλητές σφαλμάτων, όπως το συνολικό σφάλμα των τετραγώνων των αποστάσεων όλων των παραδειγμάτων από τα αντίστοιχα κέντρα των συστάδων, το κέντρο κάθε συστάδας, την τυπική απόκλιση για τα παραδείγματα κάθε συστάδας, και τελικά τον αριθμό που περιλαμβάνει κάθε συστάδα όπως φαίνονται στις παρακάτω εικόνες.

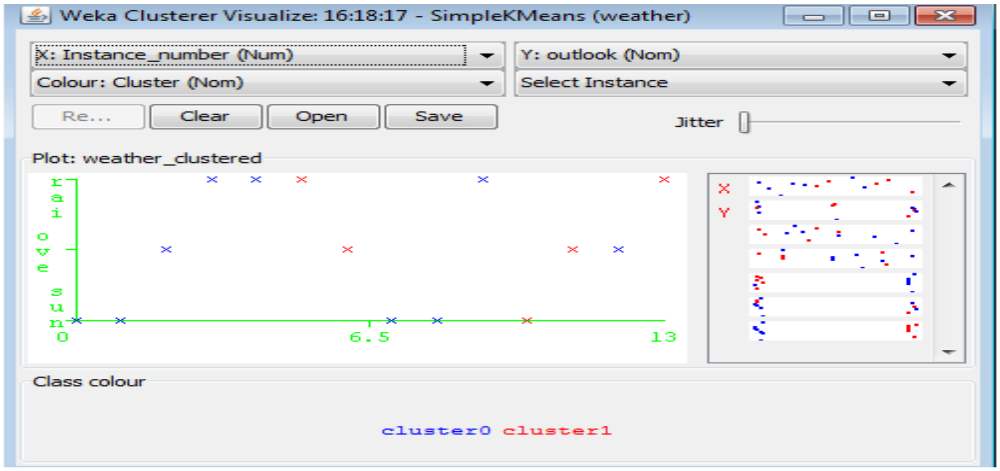


Εικόνα 31 Αποτελέσματα k-means WEKA



Εικόνα 32 Αποτελέσματα k-means WEKA

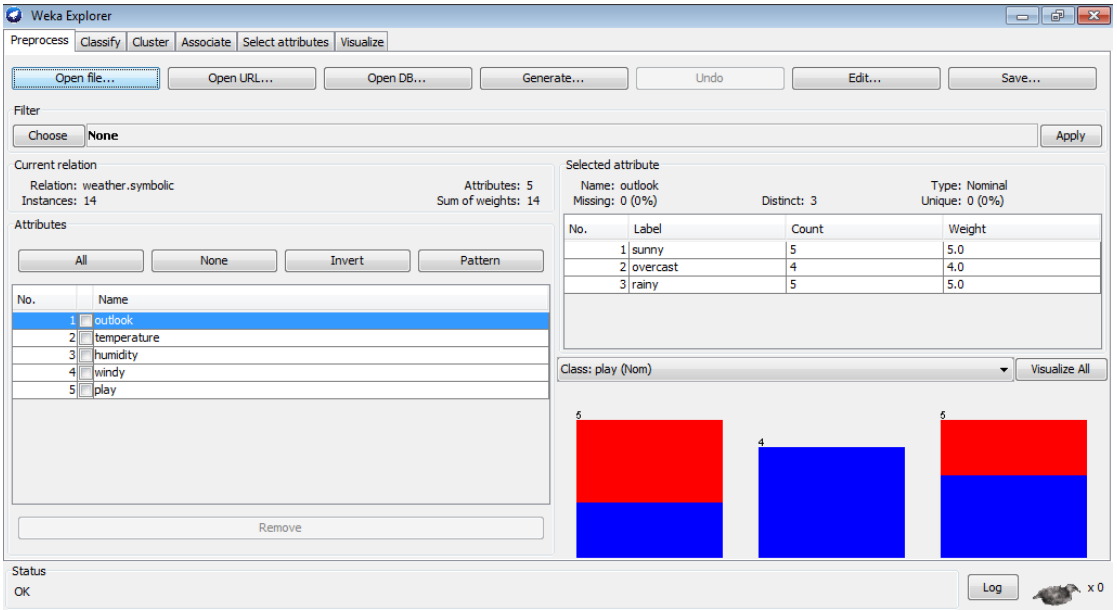
Για να δούμε μια από τις γραφικές αναπαραστάσεις των αποτελεσμάτων θα πρέπει να πατήσουμε δεξί κλικ κάτω αριστερά, και να επιλέξουμε visualize cluster assignment. Θα εμφανιστεί όπως στην παρακάτω εικόνα.



Εικόνα 33 Γραφική απεικόνιση Clustering

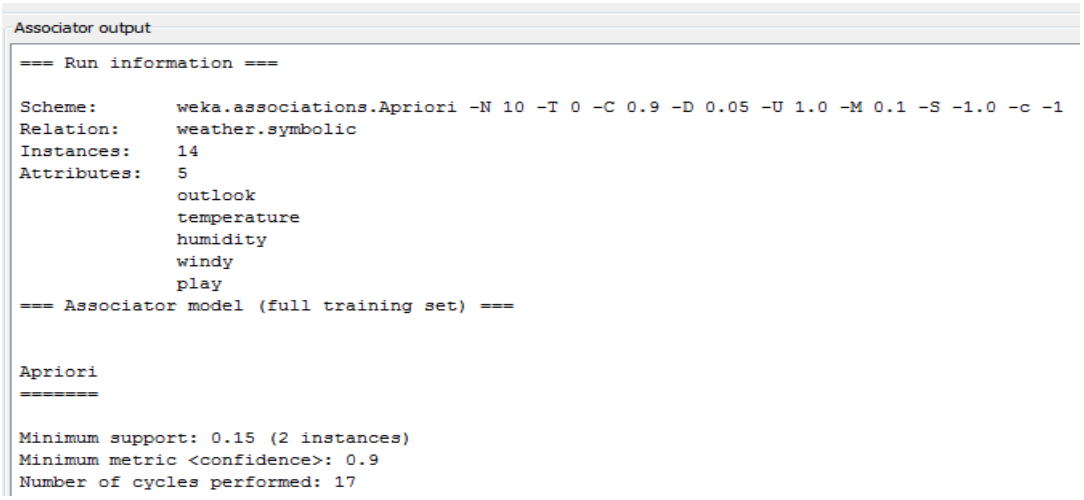
6.6.3 Ο APRIORI ΣΤΟ WEKA

Τέλος, θα αναφέρουμε πως εφαρμόζεται ο αλγόριθμος apriori στο weka και τι μπορεί να δούμε στα αποτελέσματά του. Αρχικά, εισάγουμε τα δεδομένα τα οποία θα είναι σε μορφή nominal, αυτή την φορά, και θα είναι το αρχείο weather.nominal.arff το οποίο θα το βρούμε στον δίσκο C→προγράμματα→WEKA. Την διαδικασία την βλέπουμε παρακάτω στην εικόνα 34.

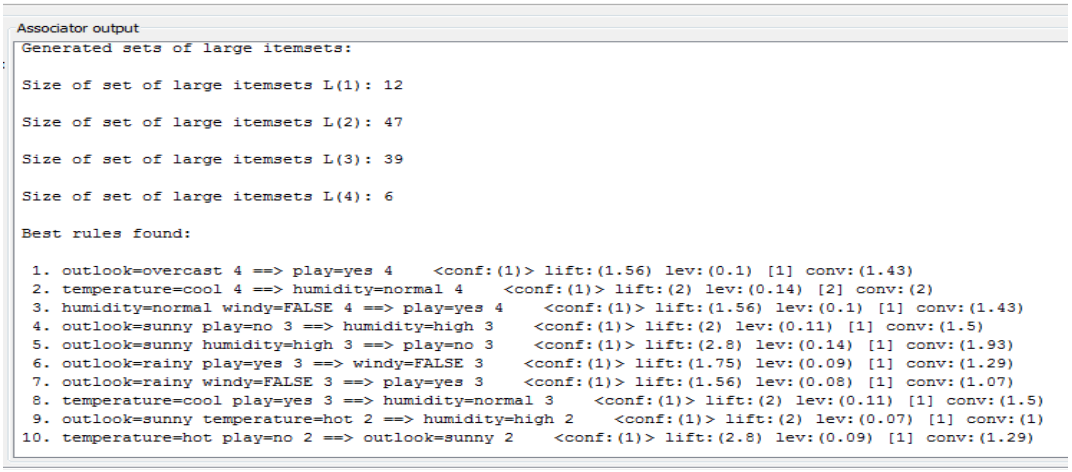


Εικόνα 34 Apriori WEKA-Εισαγωγή δεδομένων

Θα επιλέξουμε στο πάνω μέρος το Assotiate, start και θα εμφανιστούν τα παρακάτω αποτελέσματα στην εικόνα 35 και εικόνα 36.



Εικόνα 35 Αποτελέσματα apriori- weka



Εικόνα 36 Αποτελέσματα apriori- weka

Έτσι σαν αποτέλεσμα έχουμε ότι ο αλγόριθμος έψαξε για κανόνες συσχέτισης για τους οποίους η ελάχιστη τιμή είναι confidence=0.9, τα δεδομένα περάστηκαν 17 φορές ώστε να βρεθούν οι 10 καλύτεροι κανόνες οι οποίοι φαίνονται στην Εικόνα 36, και ότι ελέχθησαν κανόνες με support μέχρι 0.1Π1. Τέλος, αναφέρεται ο αριθμός των large item-sets και οι κανόνες.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Lovell, Michael C.** The review of economics and statistics . 1983.

2. **Wikipedia.** Philip Kotler. [Ηλεκτρονικό] 2 Μαιος 2015. http://en.wikipedia.org/wiki/Philip_Kotler.

3. **Wikimarkt.** Η έννοια του marketing. [Ηλεκτρονικό] 2015. http://wikimarkt.wikispaces.com//Η+έννοια+του+μάρκετινγκ.

4. **Σ.Δημητριάδη, ΑΜ Τζωρτζάκη.** *Marketing: Αρχές, Στρατιγηκές Εφαρμογές.* s.l. : ROSILI, 2010.

5. **Wikipedia.** E. Jerome McCarthy. [Ηλεκτρονικό] 30 Απρίλιος 2015. http://en.wikipedia.org/wiki/E._Jerome_McCarthy.

6. **Marketing Made Simple.** Welcome to the most sensible guide to marketing on the web. [Ηλεκτρονικό] http://www.marketing-made-simple.com.

7. **Vliet, Vincent van.** Service Marketing mix – 7 P's. [Ηλεκτρονικό] 2015. http://www.toolshero.com/service-marketing-mix-7ps/.

8. **Βικιπαίδεια.** Διαφήμιση. [Ηλεκτρονικό] 24 Ιανουάριος 2015. http://el.wikipedia.org/wiki/Διαφήμιση.

9. **Michael J.A.Berry, Gordon Linoff.** *Data Mining Techniques For Marketing , Sales and Customer Support.* s.l. : John wiley & Sons, Inc, 1997. σ. 5.

10. **Chris Rygielski, Jyun-Cheng Wang b, David C.Yen.** *Data mining techniques for customer relationship management.* 2002. σ. 5.

11. **Vercellis, Carlos.** *Business Intelligence:Data Mining and Optimization for Decision Making.* s.l. : John Willey & sons, 2009.

12. **Stan Mack, UCLA.** Data Mining: What is Data Mining? [Ηλεκτρονικό] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

13. **George Tzanis, Christos Berberidis,Ioannis Vlahavas.** *Biological Data Mining.* Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης : s.n.

14. **Branko Markoski, Zdravko Ivankovic and Miodrag Ivkovic.** Using Neural Networks in Preparing and Analysis of Basketball Scouting. [Ηλεκτρονικό] 2012. http://dx.doi.org/10.5772/48178.

15. **Adem Karahoca.** Data Mining Applications in Engineering and Medicine. [Ηλεκτρονικό] 29 Αύγουστος 2012. http://www.intechopen.com/books/data-mining-applications-in-engineering-and-medicine.

16. **P.Tan, M.Steinbach,V.Kumar.** *Introduction to data mining-Association Analysis:Basic Concepts and Algorithms.* s.l. : Wesley.

17. **E.W.T. Ngai, Li Xiu etc.** *Application of data mining techniques in customer relationship management: A literature review and classification.* s.l. : ELSEVIER, 2009. σ. 2595.

18. **Chuck Dye, Demand Media.** The Classification of Products in Marketing. [Ηλεκτρονικό] http://yourbusiness.azcentral.com/classification-products-marketing-16806.html.

19. **Sergios Theodoridis, Konstantinos Koutroumbas.** *An Introduction to Pattern Rcognition:A MATLAB Approach.* 2010.

20. **Jiawei Han, Micheline Kamber.** *Data mining: Concepts and techniques.* B'. s.l. : Εκδόσεις Morgan Kaufman, 2006. σ. 286.

21. **Πιτουρά, Ευαγγελία.** [Ηλεκτρονικό] 2007-2008. http://www.cs.uoi.gr/%20pitoura/courses/dm/index.html.

22. **Neymark, Allan.** ID3 Algorithm. [Ηλεκτρονικό] 2007.

23. **ΒΛΑΧΑΒΑΣ, ΚΕΦΑΛΑΣ,ΒΑΣΙΛΕΙΑΔΗΣ.ΚΟΚΚΟΡΑΣ,ΣΑΚΕΛΛΑΡΙΟΥ.** *Τεχνητή Νοημοσύνη Γ' ΕΚΔΟΣΗ.* Β'. ΘΕΣΣΑΛΟΝΙΚΗ : Πανεπιστήμιο Μακεδονίας, 2006.

24. **Βικιπαίδεια.** Νευρωνικό Δίκτυο. [Ηλεκτρονικό] 2015. http://el.wikipedia.org/wiki/Νευρωνικό_δίκτυο.

25. **Γιώργος Στεφανίδης, Αλέξανδρος Χατζηγεωργίου.** *Λογικές συναρτήσιες.* 2η. s.l. : Κλειδάριθμος, 2010.

26. **Νίκος Σίμου, Βαγγέλης Σπύρου.** Νευρωνικά Δίκτυα 2007-2008. [Ηλεκτρονικό] 3 Δεκέμβριος 2007.

27. **AHN, JAESOO KIM & HEEJUNE.** *A New Perspective for Neural Networks: Application to a Marketing Management Problem.* Seoul, Korea : JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 25, 1605-1616 (2009), 2009.

28. **Sadegh Bafandeh Imandoust, Mohammad Bolandraftar.** *Application of K-Nearest Neighbor (KNN). Approach fo rPredicting Economic Events: Theoretical Background.* 2013.

29. **Alex Berson, Stephen Smith, Kurt Thearling.** *Building Data Mining Applications for CRM.*

30. **Smirnov-M, H.** Data Mining and Marketing. [Ηλεκτρονικό] Ιούλιος 2007.

31. **Sreekumar Pulakkazhy, R.V.S. Balan.** *Data Mining in Banking and its Applications-A Review.* 2013.

32. **M.Akhil jabbar, B.L Deekshatulu, Priti Chandra.** Science Direct. *Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.* [Ηλεκτρονικό] 2013.

33. **Delen, David L.Olson & Dursun.** *Advanced Data mining Techniques.* s.l. : Springer-Verlag, 2008. σσ. 63-75.

34. **Konstantinos Tsiptsis, Antonios Chorianopoulos.** *Data Mining Techniques in CRM: Inside Customer Segmentation.* 2009.

35. **V.Chobe, Trupti A. Kumbhare & Prof. Santosh.** *An overview of association Rule Mining Algorithms.* s.l. : International Journal of computer Science and Information Technologies, 2014.

36. **Στέφανος Ουγιάρογλου.** *Ανακάλυψη δεδομένων συσχέτισης από εκπαιδευτικά δεδομένα .* s.l. : 6ο πανελλήνιο συνέδριο των εκπαιδευτικών για τις ΤΠΕ .

37. **Γ. Πετρώφ, Κ. Τζωρτζάκης,Α. Τζωρτζάκη.** *Μάρκετνγκ Μάνατζμεντ Η Ελληνική Προσέγγιση.* Β'. s.l. : Rosili, 2002.

38. **Wikipedia.** Wikipedia. *Medoid.* [Ηλεκτρονικό] 10 Οκτώβριος 2014. http://en.wikipedia.org/wiki/Medoid.

39. **Stefanowski, Jerzy.** Data Mining - Clustering. [Ηλεκτρονικό] 2008/2009. http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf.

40. **Sanghamitra Bandyopadhyay, Chris Giannella,Ujjwal Maulik, Hillol Kargupta.** *Peer to peer data streams .* 2005.

41. **Zhong, Shi.** *Efficient Streaming Text Clustering .* Florida : Department of Computer Science and Engineering.

42. **Wikipedia.** MATLAB. [Ηλεκτρονικό] 25 Απρίλιος 2015. http://el.wikipedia.org/wiki/MATLAB.

43. **Ευάγγελος, Κατσάνος.** Βασικά στοιχεία για τη χρήση του MATLAB & Εφαρμογή σε προβλήματα κατασκευών. [Ηλεκτρονικό] edusoft.civil.auth.gr/TE4800/Matlab%20Notes%20and%20Codes/Matlab%20Notes.pdf.

44. **Mathworks.** knnsearch. [Ηλεκτρονικό] 1994-2015. http://www.mathworks.com/help/stats/knnsearch.html;jsessionid=ac376fb81d9b3591464fcd830e44.

45. **MathWorks.** knnclassify. [Ηλεκτρονικό] 1994-2015. http://www.mathworks.com/help/bioinfo/ref/knnclassify.html.

46. **MarthWorks.** kmeans. [Ηλεκτρονικό] 1994-2015. http://www.mathworks.com/help/stats/kmeans.html.

ΓΛΩΣΣΑΡΙ

Ταξινομητής Bayes: Classifier Bayes
Neural Networks : Νευρωνικά Δίκτυα
Knn (K Nearest Neighbor) : αλγόριθμος κοντινότερου γείτονα
Aassociation Rules: Κανόνες Συσχέτισης
Support Vector Machines (SVM): Μηχανές Διανυσμάτων Υποστήριξης
Support: Υποστήριξη
Segmentation: Τμηματοποίηση
Marketing: Προώθηση Αγαθών
Marketing Mix: Μείγμα Μάρκετινγκ
Market basket analysis: Ανάλυση Καλαθιού Αγοράς
K-means: Αλγόριθμος κ μέσων
Information Gain: Κέρδος πληροφορίας
Frequent Item Sets: Στοιχειοσύνολα
Entropy: Εντροπία
Decision trees: Δέντρα Αποφάσεων
Data Mining: Εξόρυξη Δεδομένων
Cross selling: Διασταυρούμενες πωλήσεις
Confidence: Εμπιστοσύνη
Clustering: Ομαδοποίηση - Συσταδοποίηση
Classification: Ταξινόμηση