ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ Η ΧΡΗΣΗ ΤΟΥΣ ΣΤΑ ΣΟΥΠΕΡΜΑΡΚΕΤ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΠΑΖΙΜΑ ΟΛΓΑ ΑΜ:8286
22/1/2015

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ Η ΧΡΗΣΗ ΤΟΥΣ ΣΤΑ ΣΟΥΠΕΡΜΑΡΚΕΤ

Τ.Ε.Ι Ηπείρου Άρτα

Τμήμα τεχνολογίας πληροφορικής και τηλεπικοινωνιών με κατεύθυνση δίκτυα

Δήλωση πνευματικής ιδιοκτησίας

Η παρούσα εργασία αποτελεί προϊόν αποκλειστικά δικής μου προσπάθειας. Όλες οι πηγές που χρησιμοποιήθηκαν αναφέρονται στην βιβλιογραφία στο τέλος της εργασίας .

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπον καθηγητή μου κύριο Δημόπουλο για την καθοδήγησή και τις πολύτιμες συμβουλές του, που με βοήθησαν σημαντικά στη δημιουργία της εργασίας αυτής.

Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου που όλο αυτό το διάστημα ήτανε δίπλα μου και με στήριζε δίνοντας μου δύναμη και κουράγιο.

Περίληψη

Η παρούσα εργασία αναφέρεται στην εξόρυξη δεδομένων και στη χρήση τους στα σουπερμάρκετ. Αρχικά γίνεται μια εισαγωγική και μια μικρή ιστορική αναφορά και έπειτα αναλύονται έννοιες , ορισμοί και παραδείγματα. Γίνεται αναφορά στον τρόπο διαχείρισης των πελατών με σκοπό την οργάνωση τους. Ακόμα δίνεται ο ορισμός της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων (knowledge detection from data mining) ή αλλιώς KDD και της διαχείρισης πελατειακών σχέσεων (customer relationship management) ή αλλιώς CRM καθώς και παραδείγματα πάνω σε αυτές τις διαδικασίες ώστε να γίνει κατανοητή η λειτουργία τους. Τέλος γίνεται ανάλυση με παραδείγματα ώστε να εξηγήσουμε πως ακριβώς λειτουργεί η εξόρυξη δεδομένων στα σουπερμάρκετ. Αναλύουμε τα παραδείγματα σε ποσοστά και εξηγούμε τη συμπεριφορά των καταναλωτών, του διευθυντή αλλά και τον τρόπο προσαρμογής της επιχείρησης σε σχέση με την προσφορά και τη ζήτηση.

Περιεχόμενα

Εισαγωγή

Η διαδικασία ανακάλυψης της γνώσης από βάσεις δεδομένων

Η ανακάλυψη της γνώσης από βάσεις δεδομένων είναι μια διαδικασία μέσω της οποίας γίνεται προσπάθεια διερευνητικής ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων.

Στόχος της εφαρμογής είναι τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά. Είναι γενικός ορισμός της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων ή αλλιώς KDD που μπορεί να μας δώσει την έννοια της διαδικασίας είναι: « KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων και ενδεχομένως χρήσιμων και κατανοητών προτύπων στα δεδομένα »[7]. Στη συνέχεια θα αναλύσουμε τον ορισμό αυτό, στηριζόμενοι στις λέξεις κλειδιά μιας μικρής φράσης η οποία όμως περιέχει αρκετές πληροφορίες και άλλες έννοιες. Επίσης θα αναλύσουμε το πλαίσιο εφαρμογών και θα περιγράψουμε τα βήματα της διαδικασίας KDD.

Κεφάλαιο 1^{ο}

1.1. Ορισμός

Εξόρυξη δεδομένων είναι η εξεύρεση μιας πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με τη χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστάδων βάσεων δεδομένων.

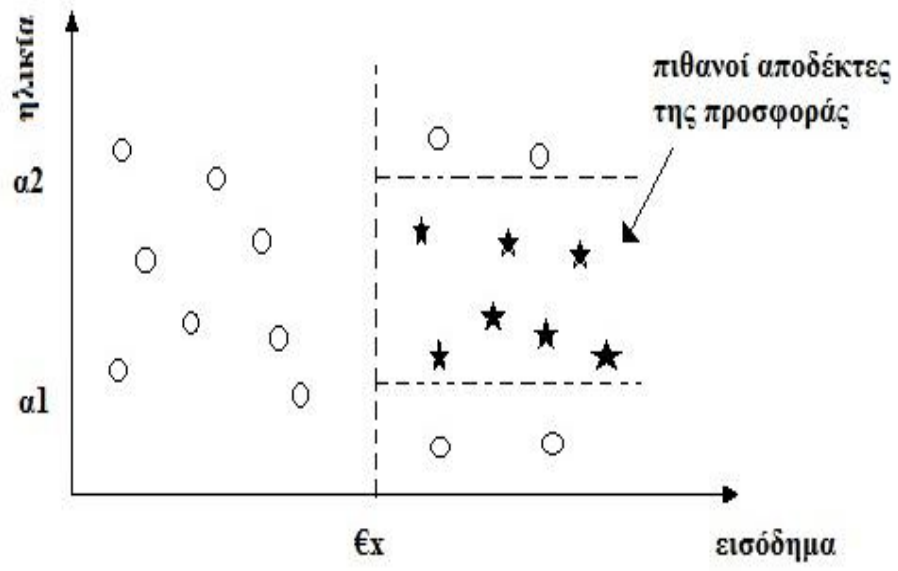
1.2. Ανάλυση ορισμού

Με βάση τον παραπάνω ορισμό αναλύουμε τις παρακάτω έννοιες που αναφέρονται σε αυτόν.

- **Δεδομένα** περιγράφουν οντότητες ή συσχετίσεις του πραγματικού κόσμου. Για παράδειγμα ένα σύνολο εγγραφών που αναφέρονται στην έκδοση δανείων μιας τράπεζας και περιλαμβάνονται ιδιότητες όπως το εισόδημα, η οικογενειακή κατάσταση και το κεφάλαιο που δεν μπορεί να λήξει. Η λέξη δεδομένα σημαίνει: « πραγματικές πληροφορίες που προκύπτουν από την

καταγραφή, την μέτρηση ή την στατιστική και αποτελούν τη βάση υπολογισμών ή επιχειρηματολογίας ».

- **Πρότυπο** είναι μια έκφραση E σε μια γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων $F_E \subseteq F$ εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του. Σε αυτή την περίπτωση το πρότυπο θεωρείται υποσύνολο του Fκαι αφαίρεση του F. Για παράδειγμα, έστω ο κανόνας : « εάν οι δανειολήπτες έχουν εισόδημα > € x^age $[A_1,A_2]$, δηλαδή εισόδημα πάνω από μία τιμή x ευρώ και ηλικία μεταξύ του διαστήματος $[A_1,A_2]$ τότε ανταποκρίνεται στη νέα προσφορά συμπληρωματικού δανείου ». Το σήμα που ακολουθεί απεικονίζει γραφικά αυτή την περιγραφή.



Εικόνα: η περιγραφή της αναφέρεται παραπάνω[7].

- **Διαδικασία** είναι μια αλληλουχία πολλών βημάτων η οποία περιλαμβάνει την προ-επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της εξαγόμενης γνώσης.
- **Εγκυρότητα** είναι ένα από τα πιο βασικά προβλήματα και αντικείμενο στην εξόρυξη δεδομένων. Είναι το εξαγόμενο πρότυπο που θα πρέπει να είναι συνεπές σε νέα δεδομένα με κάποιον βαθμό βεβαιότητας.
- **Πιθανά χρήσιμο** η εξαγωγή των προτύπων θα πρέπει να ακολουθείται από μερικές χρήσιμες διεργασίες όπως η αξιολόγηση από κάποιες συναρτήσεις χρησιμότητας αλλά και η διατήρηση όσο το δυνατόν περισσότερης γνώσης

από τα αρχικά δεδομένα. Η διατήρηση της γνώσης μπορεί να μας βοηθήσει στη λήψη αποφάσεων.

- **Τελικά κατανοητό** στόχος της εξόρυξης δεδομένων είναι να προσδιοριστούν τα πρότυπα και να είναι κατανοητά και από ανθρώπους που δεν είναι ειδικοί, ώστε να μπορέσουν να οδηγηθούν σε σημαντικά συμπεράσματα και αποφάσεις.

Είναι γεγονός ότι στην εξόρυξη δεδομένων έγιναν αρκετές προτάσεις για την ονομασία και για την ανακάλυψη γνώσης. Η εξαγωγή γνώσης, ανακάλυψης της πληροφορίας ή μη επιβλέπουσα αναγνώριση προτύπων είναι μερικές από τις προτάσεις που έγιναν. Πολλές φορές όμως υπάρχει μια σύγχυση στη χρήση των όρων KDD και εξόρυξης δεδομένων.

Για να ξεκαθαρίσουμε τις δύο έννοιες αναφέρουμε το σχολιασμό που γίνεται από τους Fayyad et al (1996-a) με βάση τους οποίους η KDD είναι μια διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα ενώ η εξόρυξη δεδομένων είναι η χρήση αλγορίθμων με στόχο την εξαγωγή πληροφοριών και προτύπων που παράγονται από την διαδικασία KDD.

1.3. Ιστορία και εξέλιξη

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτοι μέθοδοι για τον προσδιορισμό των προτύπων ήταν αυτοί της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η μεγάλη διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Για το λόγο αυτό όσο οι συλλογές δεδομένων αυξάνουν σε όγκο και σε πολυπλοκότητα τόσο η χειρωνακτική ανάλυση των δεδομένων αντικαθίσταται από την επεξεργασία δεδομένων που πλέον γίνεται αυτόματα. Σε αυτό συνέβαλαν και άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενικοί αλγόριθμοι, τα δέντρα απόφασης και η μηχανή υποστήριξης διανυσμάτων. Η εξόρυξη δεδομένων είναι μια διαδικασία εφαρμογής των μεθόδων αυτών στα δεδομένα που έχει σαν σκοπό την ανακάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων.

1.4. Βασικές εργασίες εξόρυξης από δεδομένα

Υπάρχουν επτά βήματα και εργασίες που περιγράφουν τη διαδικασία εξόρυξης από δεδομένα.

- **Κατηγοριοποίηση :** Απεικονίζει τα δεδομένα σε καθορισμένες ομάδες ή κατηγορίες κλάσεις και είναι γνωστή σαν εποπτευόμενη μάθηση επειδή οι κατηγορίες –κλάσεις κατηγοριοποιούνται πριν γίνει η εξέταση στα δεδομένα.
- **Παλινδρόμηση :** Χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μια πραγματική μεταβλητή πρόβλεψης, δηλαδή περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει την απεικόνιση.
- **Ανάλυση χρονοσειρών :** Με την ανάλυση χρονοσειρών γίνεται μελέτη στην τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές αυτές συνήθως λαμβάνονται ανά ίσα χρονικά διαστήματα (ημερήσια, εβδομαδιαία, μηνιαία κοκ), και για να παρασταθούν οπτικά χρησιμοποιείται ένα διάγραμμα χρονοσειρών.
- **Πρόβλεψη :** Η πρόβλεψη μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης .
- **Συσταδοποίηση :** Η συσταδοποίηση είναι παρόμοια με την κατηγοριοποίηση εκτός του ότι οι συστάδες δεν είναι προκαθορισμένες αλλά ορίζονται κυρίως από τα ίδια δεδομένα. Μια ειδική κατηγορία συσταδοποίησης ονομάζεται **κατάτμηση.** Με την κατάτμηση, μια βάση δεδομένων χωρίζεται σε διακριτές ομάδες παρόμοιων εγγράφων που ονομάζονται τμήματα. Η κατάτμηση θεωρείται ίδιου τύπου με την συσταδοποίηση.
- **Παρουσίαση συνόψεων :** Απεικονίζει τα δεδομένα σε υποσύνολα τους με συνοδευτικές απλές περιγραφές. Η σύνοψη των δεδομένων ονομάζεται και χαρακτηρισμός ή γενίκευση. Εξάγει ή παράγει αντιπροσωπευτικά πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό γίνεται ανακτώντας στην πραγματικότητα τμήματα από δεδομένα. Εναλλακτικά μπορούν να εξαχθούν από τα δεδομένα συνοπτικές πληροφορίες.
- **Κανόνες συσχέτισης :** Η ανάλυση συνδέσμων που εναλλακτικά αναφέρεται στη διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. αυτές οι συσχετίσεις χρησιμοποιούνται συχνά στις λιανικές πωλήσεις για να αναγνωριστούν τα προϊόντα που συχνά αγοράζονται μαζί.

Γενικά η ζωή του καθενός μας είναι γεμάτη από πληροφορίες. Ακόμα και χωρίς τη χρήση εξειδικευμένων τεχνικών και μέσων, μπορούμε να αντλήσουμε πολλές πληροφορίες και στοιχεία από το περιβάλλον μας. Για παράδειγμα, καθημερινά παρατηρούμε τους γύρω μας και σημειώνουμε τα χαρακτηριστικά τους, τη γλώσσα που μιλούν, το τι τρώνε, τον τρόπο που εκφράζονται. Αυτή η συνεχής ροή πληροφοριών μας επιτρέπει να γνωρίζουμε πολλά πράγματα για τους άλλους, όπως την ηλικία, τη φυλή, την υπηκοότητα, τις καθημερινές τους προτιμήσεις κτλ. Πάντα

συνειδητά ή ασυνείδητα και χωρίς να το καταλάβουμε διασκορπίζουμε, προς όλες τις κατευθύνσεις, στοιχεία γύρω από την ταυτότητα μας. Κάθε πλευρά του εαυτού μας και της προσωπικότητας μας αποκαλύπτονται κάθε φορά που χρησιμοποιούμε την πιστωτική μας κάρτα, ή πηγαίνουμε στο σουπερμάρκετ και κάνουμε τις αγορές μας χρησιμοποιώντας την κάρτα μέλους. Με την ανάπτυξη της τεχνολογίας, της πληροφορικής και των τηλεπικοινωνιών τα στοιχεία αυτά μπορούν ευκολότερα να συλλεχτούν, να καταγραφούν και να χρησιμοποιηθούν. Σήμερα οι τεχνικές και τα μέσα που χρησιμοποιούνται για την εξόρυξη δεδομένων- πληροφοριών απειλούν το απόρρητο των προσωπικών δεδομένων ειδικότερα στα πλαίσια του διαδικτύου, του ηλεκτρονικού εμπορίου, του άμεσου μάρκετινγκ και της αμφίδρομης διαφήμισης. Αυτή η ευκολία συλλογής και αποθήκευσης των πληροφοριών αντιπροσωπεύει μια σημαντική πρόκληση για το απόρρητο των προσωπικών μας δεδομένων, μιας και ούτε οι εταιρείες που την ασκούν ούτε οι κυβερνήσεις μπορούν να μας διαβεβαιώσουν ότι θα χρησιμοποιηθούν σωστά και χωρίς να δημιουργηθούν προβλήματα.

Η συλλογή και αγοραπωλησία των πληροφοριών και οι πληροφορίες είναι ένα από τα σημαντικότερα περιουσιακά στοιχεία μιας εταιρείας, στις μέρες μας. Ειδικότερα, είναι απαραίτητες για την ανάπτυξη του μάρκετινγκ. Αυτό έχει σαν αποτέλεσμα, οι εταιρείες να έχουν μεγάλο ενδιαφέρον στη συλλογή και στη διαχείριση των στοιχείων των καταναλωτών. Η εξόρυξη δεδομένων αποτελεί ένα από τα σημαντικότερα εργαλεία των εταιρειών. Η εξόρυξη πληροφοριών είναι μια διαδικασία κατά την οποία αυτοματοποιημένες τεχνικές εφαρμόζονται πάνω σε βάσεις δεδομένων με σκοπό την εξαγωγή συμπερασμάτων γύρω από τις αφανείς τάσεις, τις συνήθειες, και τις σχέσεις των στοιχείων μεταξύ τους. Σε αυτή τη διαδικασία συμβάλλου τα νευρωνικά δίκτυα, οι γενετικοί αλγόριθμοι, τα δέντρα αποφάσεων κλπ. Μια εταιρεία χρησιμοποιεί αυτές τις μεθόδους για να αποκτήσει και να διατηρήσει τους πελάτες της, δηλαδή να διατηρήσει το μερίδιο αγοράς, να μειώσει την πιθανότητα απάτης, να προσδιορίσει τα εσωτερικά οργανωτικά κενά της και να μπορέσει να βελτιώσει κατόπιν τις διαδικασίες έτσι ώστε να χαρτογραφήσει τις ανεξερεύνητες περιοχές του διαδικτύου. Η συγκεκριμένη πρακτική θεωρείται απαραίτητη για την επιβίωση κάθε εταιρείας. Αποτελεί την βάση για το σχεδιασμό και την ανάπτυξη νέων προϊόντων. Για παράδειγμα, οι έμποροι λιανικής χρησιμοποιούν τα αποτελέσματα που συλλέγουν από τις ταμιακές μηχανές κάνουν κατευθυνόμενες προωθητικές ενέργειες όπως προσφορές που έχουν σαν βάση την ιστορία αγορών ενός ατόμου. Η εξαγωγή πληροφοριών σχετικών με τα δημογραφικά στοιχεία επιτρέπει στους εμπόρους λιανικής να απευθυνθούν με συγκεκριμένα μηνύματα σε κάθε ομάδα καταναλωτή.

Η επεξεργασία των δεδομένων που συγκεντρώνουν μπορεί να γίνει με τη χρήση συσχετίσεων. Για παράδειγμα, όταν πωλείται μια μπύρα υπάρχει πιθανότητα πώλησης και άλλων προϊόντων όπως είναι τα κράκερ, τα πατατάκια κλπ. Η επεξεργασία των στοιχείων αυτών βοηθάει στον καθορισμό εκείνων των προωθητικών ενεργειών που πιστεύεται ότι θα είναι πιο αποτελεσματικές.

Τα σουπερμάρκετ συλλέγουν τις συναλλαγές από τα σημεία πώλησης και διαβιβάζουν συνεχώς τα στοιχεία των δεδομένων σε μεγάλες βάσεις με όγκο πολλών τεραμπάιτ. Χιλιάδες προμηθευτές χρησιμοποιούν αυτά τα στοιχεία για να αναγνωρίσουν καταναλωτικές συνήθειες έτσι ώστε να τοποθετήσουν τα προϊόντα τους στα κατάλληλα ράφια τους και να διαχειριστούν τα αποθέματα των εμπορευμάτων τους.

1.5. Εξόρυξη δεδομένων ως στάδιο ανακάλυψης της γνώσης σε βάσεις δεδομένων

Οι όροι ανακάλυψης γνώσης σε βάσεις δεδομένων και εξόρυξη γνώσης από δεδομένα χρησιμοποιούνται συχνά και εναλλακτικά για την ίδια έννοια. Έχουν δοθεί πολλές διαφορετικές ονομασίες για την διαδικασία ανακάλυψης κρυμμένων προτύπων από δεδομένα εξαγωγής γνώσης, όπως ανακάλυψη πληροφοριών, εξερευνητική ανάλυση δεδομένων, συγκομιδή πληροφοριών. Η εξόρυξη γνώσης από δεδομένα είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και των προτύπων που παράγονται με την διαδικασία KDD. Η διαδικασία KDD είναι μια διαδικασία που περιλαμβάνει πολλά διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα και οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες είναι η έξοδος.

1.6. Ανακαλύπτοντας την κρυμμένη γνώση

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων, είναι πολύ πιθανό να ανακαλύψουμε την ύπαρξη κρυμμένης γνώσης. Μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξάρτηση ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή. Το είδος αυτό της γνώσης θεωρείται ότι δεν είναι από την αρχή διαθέσιμο, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Διαφορετικά, κρίνεται απαραίτητη η ¨μη προβλεπόμενη¨ ανάκτηση γνώσης που υποστηρίζονται από την εφαρμογή αλγορίθμων. Στόχος είναι η ανακάλυψη γνώσης. Αυτή την ανάγκη έρχεται να καλύψει η εξόρυξη δεδομένων η οποία αποτελεί τον πυρήνα της διαδικασίας ανακάλυψης της γνώσης από βάσεις δεδομένων KDD.

Η διαδικασία KDD αναφέρεται στην διεργασία εξόρυξης γνώσης από μεγάλες αποθήκες δεδομένων και ο όρος εξόρυξη δεδομένων χρησιμοποιείται ως συνώνυμο της διαδικασίας KDD, αλλά αποτελεί παράλληλα και αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και εξαγωγή της γνώσης από διάφορα σύνολα δεδομένων για να γίνει κατανοητή η διαφορά μεταξύ διαδικασίας και εργαλείων ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων, ενώ ο όρος εξόρυξη

δεδομένων αναφέρεται σε τεχνικές που χρησιμοποιούνται για την ανακάλυψη γνώσης. Ένας άλλος όρος που χρησιμοποιείται αντί της εξόρυξης δεδομένων είναι η «εξόρυξη γνώσης». Θεωρείται όμως ότι ο όρος αυτός δεν δίνει έμφαση στην ανάλυση και εξαγωγή προτύπων. Η εξόρυξη δεδομένων αντιπροσωπεύει καλύτερα τη διαδικασία εύρεσης δομών γνώσης που περιγράφουν με ακρίβεια σύνολα πρωτογενών δεδομένων και οι δομές αυτές αναδεικνύουν την κρυμμένη γνώση που δεν είναι άμεσα ορατή και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

Τα τελευταία χρόνια, παρατηρείται μεγάλη ερευνητική και βιομηχανική δραστηριότητα στο χώρο της εξόρυξης δεδομένων και γενικότερα της KDD ο κλάδος αυτός έχει μεγάλη ανάπτυξη, ενώ κάθε χρόνο γίνονται αρκετά διεθνή συνέδρια και εκδίδονται επιστημονικά περιοδικά που προσφέρουν ιδιαίτερα ενδιαφέρον υλικό.

Εικόνα: Το σχήμα παρουσιάζει τα βήματα από τα οποία αποτελείται η διαδικασία KDD[7].

Η διαδικασία KDD αποτελείται από τα επόμενα βήματα:

- **Επιλογή δεδομένων:** Τα δεδομένα που χρησιμοποιούνται για τη διαδικασία της ανακάλυψης γνώσης μπορούν να βρεθούν από πολλές διαφορετικές πηγές δεδομένων. Με την επιλογή αυτή γίνεται συλλογή δεδομένων από διαφορετικές βάσεις δεδομένων με αρχεία και όχι από ηλεκτρονικές πηγές.
- **Προεπεξεργασία :** Τα δεδομένα που πρόκειται να χρησιμοποιηθούν κατά τη διαδικασία, ίσως είναι λανθασμένα. Μπορεί να υπάρχουν δεδομένα από πολλές πηγές που περιλαμβάνουν διαφορετικούς τύπους δεδομένων και διαφορετικές μονάδες μέτρησης. Με την προεπεξεργασία μπορούν να πραγματοποιηθούν πολλές και διαφορετικές δραστηριότητες. Τα λανθασμένα δεδομένα μπορούν να διορθωθούν ή να αφαιρεθούν.
- **Μετασχηματισμός:** Τα δεδομένα που προέρχονται από διαφορετικές πηγές χρειάζονται να μετατραπούν σε ένα κοινό σχήμα για την επεξεργασία τους.

Για μερικά δεδομένα μπορεί να χρειαστεί κωδικοποίηση ή μετασχηματισμός σε χρήσιμα σχήματα.

- **Εξόρυξη γνώσεων από βάσεις δεδομένων:** με βάση το είδος εξόρυξης που είναι να εκτελεστεί σε αυτό το βήμα και εφαρμόζονται αλγόριθμοι στα δεδομένα για να έχουμε τα επιθυμητά αποτελέσματα.
- **Αξιολόγηση :** Είναι σημαντικός ο τρόπος που θα παρουσιαστούν στους χρήστες τα δεδομένα της εξόρυξης επειδή η χρησιμότητα ή όχι των αποτελεσμάτων εξαρτάται από την παρουσίαση αυτή. Για να επιτύχουμε αυτό χρησιμοποιούμε διάφορες τεχνικές οπτικοποίησης όπως είναι οι γραφικές παραστάσεις, τα γεωμετρικά σχήματα κλπ.

1.7. Διαδικασία και απαιτήσεις

Οι διαδικασίες της εξόρυξης δεδομένων έχουν σκοπό τη δημιουργία μοντέλων συναρμολογήσεων ή την εξαγωγή προτύπων των υπό εξέταση δεδομένων. γνωρίζοντας τις παραμέτρους του μοντέλου μέσα από τα δεδομένα που υπάρχουν ή τα πρότυπα που προσδιορίζονται, εφαρμόζουμε τους κατάλληλους αλγορίθμους εξόρυξης δεδομένων.

Οι αλγόριθμοι αυτοί βασίζονται σε τομείς όπως η στατιστική και η μηχανική μάθηση, αλλά η διαφορά τους είναι ότι έχουν τέτοιο σχεδιασμό ώστε να υπάρχει εξελιξιμότητα σε σχέση με το μέγεθος του συνόλου δεδομένων που εισάγονται.

Για να προετοιμαστούν τα δεδομένα για εξόρυξη γνώσης και να παραχθούν αποτελέσματα με περισσότερο νόημα και να γίνουν πιο κατανοητά χρησιμοποιούνται τεχνικές μετασχηματισμού. Για να πραγματοποιηθεί η χρήση αυτών των τεχνικών απαιτούνται ειδικοί τύποι κατανομής δεδομένων. Ακόμα θα πρέπει να συνδυαστούν οι τιμές των γνωρισμάτων ή και να δοθούν νέες τιμές έτσι ώστε να μειωθεί η πολυπλοκότητα τους. Η τροποποίηση των δεδομένων θα πρέπει να γίνεται με προσοχή όπως επίσης με προσοχή θα πρέπει να γίνουν και όλα τα βήματα της διαδικασίας KDD που έχουμε αναφέρει λίγο πιο πάνω. Εάν κατά τη διαδικασία της τροποποίησης γίνει κάποιο λάθος, τότε θα αλλάξουν ριζικά τόσο τα δεδομένα όσο και τα αποτελέσματα από την εξόρυξη γνώσης με αποτέλεσμα τα συμπεράσματα να είναι ανακριβή και πιθανόν λάθος.

1.7.1. Βασικά στάδια εξόρυξης δεδομένων

Υπάρχουν λοιπόν τρία στάδια εξόρυξης δεδομένων και ο διαχωρισμός τους θα μας βοηθήσει να κάνουμε μια αρχική προσέγγιση των διαδικασιών της εξόρυξης δεδομένων αλλά και να εμπεδώσουμε στη συνέχεια τις απαντήσεις της. Έτσι τα τρία στάδια εξόρυξης δεδομένων είναι:

- **Περιγραφή μοντέλου**

Στο στάδιο αυτό της εξόρυξης δεδομένων επιχειρούμε να δηλώσουμε τη λειτουργία του μοντέλου, δηλαδή να δηλώσουμε το στόχο μας, όπως για παράδειγμα την ταξινόμηση, την παλινδρόμηση ή τη συσταδοποίηση.

Επίσης μας ενδιαφέρει η παραστατική μορφή του μοντέλου, δηλαδή η απεικόνιση του έτσι ώστε να ταιριάζει με την απεικόνιση των δεδομένων και να είναι δυνατό να ερμηνευτεί. Χαρακτηριστικά παραδείγματα είναι τα δέντρα απόφασης, τα γραφικά μοντέλα, τα νευρωνικά δίκτυα και τα μοντέλα- συστήματα που βασίζονται σε παραδείγματα ή πιθανότητες.

- **Αξιολόγηση μοντέλου**

Ύστερα από τη δημιουργία του μοντέλου, πρέπει να εξετάσουμε κατά πόσο ταιριάζει με τις συνθήκες της KDD. Προχωράμε, δηλαδή, στην αξιολόγηση του μοντέλου έτσι ώστε να κρίνουμε την εγκυρότητα των προτύπων και την ακρίβεια και χρησιμότητα του μοντέλου. Υπάρχουν διάφορα κριτήρια αξιολόγησης όπως αυτό της μέγιστης πιθανότητας.

- **Αλγόριθμοι ανάκτησης**

Στόχος του σταδίου της ανάκτησης είναι η σύγκριση μοντέλων και παραμέτρων, που δόθηκαν στο σύνολο δεδομένων, της οικογένειας μοντέλων και του κριτηρίου αξιολόγησης. Οι βασικότεροι τύποι αλγορίθμων αναζήτησης είναι αυτοί που αναζητούν παραμέτρους βελτιστοποίησης ενός κριτηρίου αξιολόγησης και αυτοί που αναζητούν μοντέλα αντιπροσώπευσης των δεδομένων.

Υπάρχει ακόμα μία διαδικασία που χρησιμοποιείται για την εξόρυξη δεδομένων και είναι η οπτικοποίηση. Τη διαδικασία αυτή χρησιμοποιούμε για παράδειγμα μια γραφική παράσταση που δείχνει την κατανομή μιας μεταβλητής δεδομένων που είναι πιο κατανοητή και πιο κατατοπιστική για την εξαγωγή συμπερασμάτων. Η χρήση των τεχνικών οπτικοποίησης επιτρέπει στους χρήστες να συνοψίζουν, να εξάγουν και να αντιλαμβάνονται τους μαθηματικούς και πιο περιγραφικούς τρόπους παρουσίασης των αποτελεσμάτων. Μερικές από τις τεχνικές οπτικοποίησης αναφέρονται παρακάτω και είναι οι εξής:

- **Γραφικές παραστάσεις:** Μπορούν να χρησιμοποιηθούν οι παραδοσιακές γραφικές παραστάσεις όπως είναι τα ραβδογράμματα, οι πίτες, τα ιστογράμματα και τα γραμμογράμματα.
- **Γεωμετρικές παραστάσεις:** Οι γεωμετρικές τεχνικές περιλαμβάνουν θηκογράμματα και διαγράμματα διασποράς.

Σχήμα: απεικόνιση ραβδογράμματος, κυκλικού διαγράμματος, εικονογράμματος και χρονογράμματος.[10].

- **Βασισμένες σε εικονίδια:** Χρησιμοποιώντας σχήματα, χρώματα, ή εικονίδια μπορούμε να βελτιώσουμε την παρουσίαση των αποτελεσμάτων.
- **Ιεραρχικές:** Αυτές οι τεχνικές διαιρούν ιεραρχικά το χώρο παρουσίασης σε περιοχές, βασιζόμενες στις τιμές των δεδομένων.
- **Υβριδικές :** Οι προηγούμενες τεχνικές μπορούν να συνδυαστούν σε μια ενιαία παρουσίαση.

Επίσης μπορούν να χρησιμοποιηθούν εργαλεία οπτικοποίησης, για να συνοψίσουν τα δεδομένα. Ακόμα αυτή η διαδικασία μπορεί να χρησιμοποιηθεί για να εμφανίσει πολύπλοκα αποτελέσματα των εργασιών της εξόρυξης γνώσης από τα δεδομένα. Η εξόρυξη των δεδομένων σαν διαδικασία είναι από μόνη της πολύπλοκη και δύσκολα κατανοητή, γι’ αυτό το λόγο τα πρότυπα που ανακαλύπτονται θα πρέπει να ερμηνεύονται και να αξιολογούνται σωστά έτσι ώστε να προκύπτουν πληροφορίες που είναι ακριβείς και να έχουν κάποια ιδιαίτερη σημασία.

1.7.2. Απαιτήσεις της εξόρυξης δεδομένων

Για να έχουμε ένα ολοκληρωμένο αποτέλεσμα από μια διαδικασία εξόρυξης δεδομένων πρέπει αρχικά να ελέγξουμε τα χαρακτηριστικά που περιμένουμε να έχει το σύστημα εξόρυξης δεδομένων, καθώς και τις απαιτήσεις για την εφαρμογή των τεχνικών. Τα κυριότερα ζητήματα που οφείλουμε κάθε φορά να λαμβάνουμε υπόψη είναι:

- **Χειρισμός διαφορετικών τύπων δεδομένων**

Είναι ξεκάθαρο ότι το σύστημα της εξόρυξης δεδομένων μπορεί να εφαρμόζεται σε διαφορετικούς τύπους δεδομένων, καθώς συχνά χρησιμοποιούνται διαφορετικοί τύποι βάσεων δεδομένων σε διαφορετικές εφαρμογές. Ακόμα, παρατηρείται συχνά η ύπαρξη συγγενικών βάσεων δεδομένων. Τέλος ένα τέτοιο σύστημα θα έπρεπε να λειτουργεί ανεξάρτητα από τύπους δεδομένων (δομές δεδομένων και σύνθετα αντικείμενα, υπερκείμενο και στοιχεία πολυμέσων, χωροχρονικά στοιχεία κλπ.)

Η ποικιλία των τύπων δεδομένων και οι διαφορετικοί στόχοι της εξόρυξης δεδομένων κάνουν πιο πιθανή την ύπαρξη ενός συστήματος εξόρυξης δεδομένων που μπορεί να χειριστεί όλα τα είδη δεδομένων. καλό θα ήταν να διαμορφωθούν εξειδικευμένα συστήματα για την εξόρυξη γνώσης πάνω σε συγκεκριμένους τύπους δεδομένων όπως βάσεις δεδομένων, οι χωροχρονικές βάσεις κλπ.

- **Απόδοση και εξελιξιμότητα των αλγορίθμων εξόρυξης δεδομένων**

Για να έχουμε αποτελεσματική εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων, πρέπει να έχουμε αλγορίθμους κατάλληλα προσαρμοσμένους σε αυτά. Επομένως, ο χρόνος εκτέλεσης των αλγορίθμων πρέπει να είναι αποδεκτός και αναμενόμενος για μεγάλες βάσεις δεδομένων. Να σημειώσουμε ότι οι αλγόριθμοι με εκθετική ή πολυωνυμική πολυπλοκότητα δεν θεωρούνται πρακτικοί στη χρήση.

- **Χρησιμότητα, βεβαιότητα, εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων**

Η εξόρυξη γνώσης πρέπει να παρουσιάζει με ακριβή τρόπο τα περιεχόμενα των βάσεων δεομένων και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων θα μπορούσε να εκφραστεί μέσω κάποιων μέτρων βεβαιότητας προσεγγιστικά ή ποσοτικά. Εξαιρέσεις όπως θόρυβος και ακραίες τιμές πρέπει να αντιμετωπιστούν από τα συστήματα εξόρυξης δεδομένων. Το γεγονός αυτό δίνει το κίνητρο για μια συστηματική μελέτη της ποιότητας της εξόρυξης γνώσης κατασκευάζοντας στατιστικά ή αναλυτικά μοντέλα, μοντέλα προσομοίωσης καθώς και εργαλεία αυτών.

- **Εκφράσεις διαφορετικού τύπου για αποτελέσματα**

Όπως μπορούμε να φανταστούμε από μεγάλα σύνολα δεδομένων μπορούν να προκύψουν διαφορετικοί τύποι γνώσεων. Ακόμη, θα ήταν πολύ χρήσιμο να μπορούμε εμείς να ελέγξουμε τη γνώση από διαφορετικές απόψεις και να την εκφράσουμε σε διαφορετικές μορφές. Θεωρείται ότι είναι καλό να μπορούν να εκφραστούν τα ερωτήματα της εξόρυξης δεδομένων και η εξορυγμένη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών διεπαφών των χρηστών. Έτσι η εξόρυξη δεδομένων θα μπορούσε να εφαρμοστεί και από άτομα μη εξειδικευμένα και έτσι η εξορυγμένη γνώση θα χρησιμοποιούταν άμεσα από όλους. Τέλος, απαιτείται το σύστημα να

υιοθετήσει εκφραστικές αναπαραστάσεις της γνώσης, έτσι ώστε να επιτευχθεί η αποτελεσματική αναπαράσταση της γνώσης.

- **Διαλογική ανακάλυψη γνώσης στα πολλαπλά εννοιολογικά επίπεδα**

Είναι δύσκολο να γίνει πρόβλεψη πάνω σε αυτό που θα μπορούσε να ανακαλυφθεί επακριβώς από μια βάση δεδομένων. Γι' αυτό θα μπορούσε να καθοριστεί μια σειρά ερωτήσεων της εξόρυξης δεδομένων προκειμένου διαμορφωθεί η εστίαση στα δεδομένα να δημιουργηθεί ένα λεπτομερέστερο επίπεδο εξόρυξης δεδομένων και να παρατηρηθούν τα αποτελέσματα της εξόρυξης δεδομένων σε πολλαπλά επίπεδα και από διαφορετικές πλευρές. Όλα αυτά μπορούν να επιτευχθούν μέσω μιας διαλογικής ανακάλυψης της γνώσης.

- **Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων**

Σε σχέση με τη σύνδεση των διαφορετικών πηγών δεδομένων, υπάρχει προβάδισμα της ευρέσεως διαθέσιμης σύνδεσης υπολογιστών σε τοπικό και ευρύτερο δίκτυο συμπεριλαμβανομένου του διαδικτύου. Αυτό οδηγεί στη δημιουργία μεγάλων κατανεμημένων και ετερογενών βάσεων δεδομένων.

- **Προστασία δηκτικότητας και ασφάλεια δεδομένων**

Η προστασία και αποτελεσματικότητα των δεδομένων απειλείται στην περίπτωση που αυτά μπορούν να παρατηρηθούν από πολλές διαφορετικές σκοπιές. Είναι σημαντικό να μελετήσουμε πότε μπορούμε να οδηγηθούμε σε μια εισβολή στην ιδιωτικότητα μέσω της KDD και τι μέτρα ασφάλειας μπορούμε να αναπτύξουμε για να εμποδίσουμε την αποκάλυψη των ευαίσθητων πληροφοριών.

Να σημειώσουμε ότι μερικές από τις απαιτήσεις που αναφέραμε παραπάνω μπορούν να φέρουν αντικρουόμενους στόχους. Για παράδειγμα, ο στόχος της προστασίας της ασφάλειας δεδομένων μπορεί να έρχεται σε σύγκρουση με την απαίτηση για διαλογική εξόρυξη πολυεπίπεδης γνώσης από διαφορετικές σκοπιές. Η παρουσίαση των απαιτήσεων αυτών γίνεται στα πλαίσια του ενδιαφέροντος μας για την ανάπτυξη αποτελεσμάτων και εξελίξιμων αλγορίθμων. Για το λόγο αυτό έγιναν συγκεκριμένες ομαδοποιήσεις των απαιτήσεων ώστε να γίνει γενική απεικόνιση.

1.8. Θέματα εξόρυξης από δεδομένα

Υπάρχουν πολλά θέματα υλοποίησης που σχετίζονται με την εξόρυξη γνώσης από δεδομένα και είναι τα εξής:

- **Η Ανθρώπινη αλληλεπίδραση**
- **Η υπερπροσαρμογή :** Η υπερπροσαρμογή εφαρμόζεται όταν το μοντέλο δεν ταιριάζει σε μελλοντικές καταστάσεις. Αυτό μπορεί να

συμβεί είτε από υποθέσεις που γίνονται για τα δεδομένα ή απλά μπορεί να συμβεί επειδή το μέγεθος των δεδομένων και των πληροφοριών είναι μικρό. Επίσης μπορεί να εμφανιστεί ακόμα και σε περιπτώσεις που τα δεδομένα μας είναι σταθερά και αμετάβλητα.

- **Οι ακραίες τιμές:** Υπάρχουν πολλές καταχωρήσεις δεδομένων που δεν ταιριάζουν σωστά στο μοντέλο που έχει αναπτυχθεί. Αυτό συμβαίνει συνήθως στις πολύ μεγάλες βάσεις δεδομένων. Αν το μοντέλο που έχει δημιουργηθεί περιλαμβάνει ακραίες τιμές τότε μπορεί να μην συμπεριφέρεται σωστά για δεδομένα που βρίσκονται μέσα στη βάση.
- **Η ερμηνεία των αποτελεσμάτων:** Με τα σημερινά δεδομένα τα αποτελέσματα από την εξόρυξη γνώσης πρέπει να ερμηνεύονται από ειδικούς του πεδίου, αλλιώς θα είναι χωρίς νόημα για το μέσο χρήστη.
- **Η οπτικοποίηση των αποτελεσμάτων:** Η οπτικοποίηση των αποτελεσμάτων των αλγορίθμων εξόρυξης γνώσης μας βοηθά να δούμε και να κατανοήσουμε ευκολότερα τα αποτελέσματα αυτά.
- **Τα μεγάλα σύνολα δεδομένων:** Στα τεράστια σύνολα δεδομένων δημιουργούνται προβλήματα όταν εφαρμόζονται αλγόριθμοι εξόρυξης γνώσης που έχουν σχεδιαστεί να εφαρμόζονται σε μικρά σύνολα δεδομένων. πολλές εφαρμογές μοντελοποίησης αυξάνονται εκθετικά στον αριθμό των δεδομένων και γι’ αυτό τον λόγο οι εφαρμογές αυτές είναι αποτελεσματικές στα μεγαλύτερα σύνολα δεδομένων. Αποτελεσματικά εργαλεία για να αντιμετωπιστεί το πρόβλημα της κλιμάκωσης είναι η δειγματοληψία και ο παραλληλισμός.
- **Οι υψηλές διαστάσεις:** Το σχήμα μιας συμβατικής βάσης δεδομένων μπορεί να αποτελείται από πολλά διαφορετικά γνωρίσματα. Το πρόβλημα εδώ είναι ότι ίσως δεν χρειάζονται όλα τα γνωρίσματα για να λυθεί ένα συγκεκριμένο πρόβλημα εξόρυξης γνώσης. Αν χρησιμοποιήσουμε κάποια γνωρίσματα μπορεί να εμποδίσουν τη σωστή ολοκλήρωση μιας εργασίας. Η χρήση άλλων γνωρισμάτων μπορεί απλά να αυξήσει τη συνολική πολυπλοκότητα και να μειώσει την απόδοση ενός αλγορίθμου. Αυτό το πρόβλημα μερικές φορές μπορεί να αναφέρεται σαν μια κατάρα των υψηλών διαστάσεων, εννοώντας ότι υπάρχουν πολλά γνωρίσματα που μπλέκονται και είναι δύσκολο να καθοριστεί ποια γνωρίσματα πρέπει να χρησιμοποιηθούν. Μια λύση στο πρόβλημα των υψηλών διαστάσεων είναι να μειωθούν τα γνωρίσματα, κάτι που αναφέρεται ως μείωση των υψηλών διαστάσεων. Δεν είναι όμως πάντα εύκολο να προσδιορίσουμε ποια είναι τα γνωρίσματα που δεν μας είναι χρήσιμα.
- **Τα δεδομένα πολυμέσων:** Οι περισσότεροι από τους αλγόριθμους που έχουν προταθεί κατά καιρούς έχουν σαν στόχο τα παραδοσιακά είδη δεδομένων. η χρήση των δεδομένων πολυμέσων, όμοια με αυτά που

βρίσκουμε στις γεωγραφικές βάσεις δεδομένων, περιπλέκει ή καθιστά ακατάλληλους πολλούς από τους αλγορίθμους αυτούς.

- **Τα ελλιπή δεδομένα:** Κατά τη διάρκεια του βήματος της προεπεξεργασίας στη διαδικασία KDD τα δεδομένα που λείπουν μπορούν να συμπληρωθούν με κατ' εκτίμηση τιμές. Αυτή η προσέγγιση, καθώς και άλλες προσεγγίσεις που αντιμετωπίζουν το πρόβλημα των ελλιπών δεδομένων που ενδεχομένως οδηγούν σε λανθασμένα αποτελέσματα κατά την εξόρυξη από δεδομένα.
- **Τα δεδομένα που αλλάζουν:** Οι βάσεις δεδομένων δεν μπορεί να θεωρηθούν ότι είναι στατικές. Όμως οι περισσότεροι αλγόριθμοι εξόρυξης γνώσης υποθέτουν ότι η βάση δεδομένων είναι στατική. Αυτό απαιτεί ο αλγόριθμος να ξανατρέχει από την αρχή κάθε φορά που αλλάζει η βάση δεδομένων.
- **Ολοκλήρωση:** Η διαδικασία KDD στις μέρες μας δεν αποτελεί μέρος των συνηθισμένων εργασιών επεξεργασίας των δεδομένων. Οι απαιτήσεις της KDD μπορεί να αντιμετωπίζονται σαν ιδιαίτερες, ασυνήθιστες, ή σαν απαιτήσεις της μιας φοράς. Οι απαιτήσεις αυτές γίνονται αποτελεσματικές και όχι αρκετά γενικές για να χρησιμοποιούνται σε συνεχή βάση. Φυσικά ένας επιθυμητός στόχος είναι η ενσωμάτωση των λειτουργιών της εξόρυξης γνώσης σε παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων.
- **Η εφαρμογή:** Αποτελεί μεγάλη πρόκληση στο να προσδιορίσουμε τη σωστή χρήση για μια πληροφορία που προήλθε από τη λειτουργία της εξόρυξης γνώσης. Πράγματι, η αποτελεσματική και σωστή ερμηνεία των αποτελεσμάτων θεωρείται μερικές φορές σαν το πιο δύσκολο έργο από τρέξιμο για τον αλγόριθμο. Επειδή τα δεδομένα είναι πληροφορίες που δεν ήταν γνωστές στο παρελθόν, οι τεχνικές των επιχειρήσεων πρέπει να τροποποιηθούν για να καθορίσουν τον τρόπο με τον οποίο θα χρησιμοποιήσουν τις κρυμμένες πληροφορίες. Αυτά είναι τα θέματα που πρέπει να αντιμετωπιστούν από τους αλγορίθμους και τα προϊόντα εξόρυξης γνώσης από δεδομένα.

1.8.1. Χρησιμότητα και εφαρμογές στον πραγματικό κόσμο

Μερικές από τις εφαρμογές της εξόρυξης δεδομένων, στα πλαίσια ανακάλυψης γνώσης είναι:

- ✓ Ανάλυση οργανικών συνθέσεων
- ✓ Αυτόματη αφαίρεση
- ✓ Προσδιορισμός απειλών στον κλάδο των πιστώσεων
- ✓ Πρόβλεψη κατανάλωσης

- ✓ Οικονομική πρόβλεψη
- ✓ Ιατρική διάγνωση
- ✓ Πρόβλεψη τηλεθέασης
- ✓ Σχεδιασμός παραγωγής
- ✓ Εκτίμηση αυτοκινήτων
- ✓ Πώληση προς συγκεκριμένους στόχους
- ✓ Ανάλυση κινδύνου από τοξικά
- ✓ Βελτιστοποίηση παροχής θερμότητας στα φυτά
- ✓ Πρόβλεψη καιρού

Αυτές είναι μερικές από τις ομάδες εφαρμογών της εξόρυξης δεδομένων ενώ μπορούμε να ανακαλύψουμε και πολλές εφαρμογές αν αναλογιστούμε καθημερινά πρακτικά ζητήματα. Στόχος μας είναι να καταλάβουμε ότι η εξόρυξη δεδομένων αποτελεί το εργαλείο της KDD. Για παράδειγμα αναφέρουμε κάποιες εφαρμογές της KDD στο χώρο των επιχειρήσεων που χρησιμοποιούνται σε κάποιους τομείς όπως:

- ✓ Μάρκετινγκ
- ✓ Επενδύσεις
- ✓ Προσδιορισμός απειλών
- ✓ Βιομηχανική παραγωγή
- ✓ Τηλεπικοινωνίες
- ✓ Καθορισμός δεδομένων

Προφανώς η δράση της KDD σε αυτούς τους τομείς γίνεται μέσω της εξόρυξης δεδομένων. Η ανάπτυξη της διαδικασίας είναι συνεχείς και οφείλεται στη συνεργασία των ερευνητικών πεδίων που αναφέραμε. Το ερώτημα όμως είναι σε τι διαφέρει η KDD από την αναγνώριση προτύπων, τη μηχανική εκμάθηση ή τα άλλα πεδία που συνεισφέρουν στην εξόρυξη δεδομένων;

Η απάντηση δίνεται και είναι η εξής. Τα παραπάνω πεδία προσφέρουν κάποιες από τις μεθόδους της εξόρυξης δεδομένων που χρησιμοποιούνται στο στάδιο της εξόρυξης δεδομένων αν το θεωρήσουμε ως ένα μέρος της διαδικασίας της KDD. Πράγματι η διαδικασία KDD θεωρείται μια σειρά βημάτων μερικά από οποία αποτελούν το μέρος της εξόρυξης δεδομένων. μάλιστα το μέρος αυτό θεωρείται από τα πιο εύκολα κομμάτια της διαδικασίας KDD, ενώ το κομμάτι της προεπεξεργασίας των δεδομένων (καθαρισμός, μετασχηματισμός) θεωρείται πιο περίπλοκο.

Στόχος της KDDείναι η αποκάλυψη γνώσης από δεδομένα, περιλαμβάνοντας όμως και διαδικασίες ελέγχου του τρόπου αποθήκευσης ή πρόσβασης των δεδομένων, αλλά και της δυνατότητας επέκτασης των αλγορίθμων σε πολύ μεγάλα σύνολα δεδομένων χωρίς να μειώνεται η πολυπλοκότητα τους.

Επίσης η KDD ενδιαφέρεται για τον τρόπο με τον οποίο εικονίζουμε τα αποτελέσματα, αλλά και για τη γενικότερη συνεργασία με το ανθρώπινο στοιχείο,

ώστε να συμβάλλει κατά τον καλύτερο δυνατό τρόπο. Άρα οι εφαρμογές του KDD δεν έχουν να κάνουν μόνο με την δράση και την συμπεριφορά των αλγορίθμων. Αυτό αποτελεί καθαρά κομμάτι της εξόρυξης δεδομένων. Όταν όμως έχουμε να ασχοληθούμε με στοιχεία από πραγματικό κόσμο, όλες οι ιδιότητες ενώνονται με τελικό στόχο τη διάκριση χρήσιμων προτύπων από τα δεδομένα.

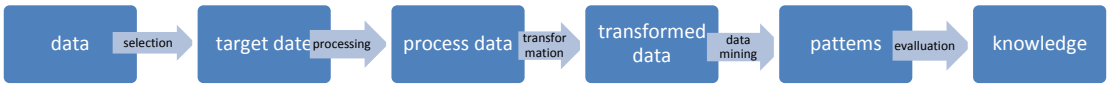
Οι διαδικασίες της KDD διευκολύνονται κατά πολύ μέσω των τεχνολογιών των βάσεων δεδομένων, καθώς και πεδίων όπως η αποθήκευση δεδομένων. Οι τεχνολογίες της αποθήκευσης δεδομένων εξυπηρετούν τη διαδικασία σε δύο πολύ σημαντικά σημεία: τον προορισμό και την πρόσβαση των δεδομένων. Επίσης μία διάσημη προσέγγιση για ανάλυση αποθηκών δεδομένων είναι η OLAP (Online Analytical Processing) ή αλλιώς επισκόπηση της ηλεκτρονικής διαδικασίας ανάλυσης.

Τα εργαλεία της OLAP συμβάλλουν στην πολύ μεταμεταβλητή ανάλυση δεδομένων και στοχεύουν στην υλοποίηση και την υποστήριξη της αλληλεπιδραστικής ανάλυσης δεδομένων. Βέβαια η διαδικασία και τα εργαλεία της είναι ένα βήμα πιο μακριά από αυτά που μπορούν να στηρίξουν τα περισσότερα τυπικά συστήματα βάσεων δεδομένων.

Περισσότερα από τα εργαλεία της OLAP και τις άλλες έννοιες που συμβάλλουν στην εξόρυξη δεδομένων και κατά συνέπεια στην KDD είναι τα συστήματα DBMS, της διαδικασίες της ασαφούς λογικής, την μοντελοποίηση διαστάσεων, την ευρετηριοποίηση κτλ.

1.8.2. Βήματα της διαδικασίας KDD

Η διαδικασία της KDD είναι μια διαλογική και επαναληπτική διαδικασία. Δηλαδή, μπορεί να απαιτηθεί η επιστροφή σε ένα προηγούμενο βήμα όπως φαίνεται στο παρακάτω σχήμα. Ως πρώτο βήμα θεωρείται ο εντοπισμός των στοιχείων της KDD, ενώ στο τέλος αξιοποιούμε την ανακαλυφθείσα γνώση. Στο σχήμα που ακολουθεί, παρατηρούμε τα βήματα της διαδικασίας KDD σε μορφή αλυσίδας.



Εικόνα: διαδικασία ανακάλυψης της γνώσης[7]

Βήμα 1^{ο} ανάπτυξη και κατανόηση της περιοχής της εφαρμογής

Σε αυτό το προκαταρκτικό στάδιο γίνεται προετοιμασία για την κατανόηση του πλαισίου δράσης. Πρέπει να γίνει σαφές ποιες αποφάσεις θα ληφθούν σχετικά με τους μετασχηματισμούς, τους αλγορίθμους, την αναπαράσταση κλπ.

Το βήμα αυτό βοηθά στην κατανόηση των στόχων από τον τελικό χρήστη, καθώς και στην εύρεση του περιβάλλοντος όπου θα δράσει η διαδικασία ανακάλυψης της γνώσης. Στα πλαίσια αυτά περιλαμβάνεται και η προγενέστερη γνώση του τομέα τον οποίο εξετάζουμε. Είναι πιθανό να ζητηθεί επανάληψη του βήματος αυτού στη συνέχεια.

Βήμα 2^{ο} επιλογή και δημιουργία κατάλληλου συνόλου δεδομένων

Έχοντας ορίσει και αναλύσει τους στόχους, θα πρέπει να προσδιορίσουμε και τα δεδομένα που θα χρησιμοποιηθούν. Το βήμα αυτό περιλαμβάνει τον εντοπισμό των δεδομένων που είναι διαθέσιμα, την απόκτηση επιπρόσθετων αν υπάρχει ανάγκη δεδομένων και την ενσωμάτωση όλων αυτών σε ένα σύνολο δεδομένων το οποίο θα περιλαμβάνει τα χαρακτηριστικά που θα ληφθούν υπόψη.

Το βήμα αυτό είναι σημαντικό γιατί η εξόρυξη δεδομένων μαθαίνει και ανακαλύπτει από τα δεδομένα που έχει διαθέσιμα εκείνη τη στιγμή. Σε αυτή τη βάση κατασκευάζονται και τα μοντέλα. Είναι πιθανό, όμως να προκύψουν προβλήματα στην περίπτωση που λείπουν χαρακτηριστικά από κάποιες παρατηρήσεις, καθώς και επίσης μπορεί να δημιουργηθούν και σφάλματα στη μελέτη. Άρα χρειάζεται η μέγιστη δυνατή συλλογή χαρακτηριστικών.

Από την άλλη πλευρά, όμως η ανάγκη αυτή ανεβάζει το κόστος διεξαγωγής της ανάλυσης. Για το λόγο αυτό, η διαδικασία της KDD αναλαμβάνει να αξιοποιήσει αρχικά το βέλτιστο διαθέσιμο σύνολο δεδομένων και στη συνέχεια επεκτείνεται και παρατηρεί τα αποτελέσματα στα πλαίσια της ανακάλυψης γνώσης και μοντελοποίησης.

Βήμα 3^{ο} προεπεξεργασία και καθορισμός δεδομένων

Ένα πολύ σημαντικό σημείο που μας απασχολεί είναι η αξιοπιστία των δεδομένων, η οποία μελετάται μέσα από το απαραίτητο βήμα της διαδικασίας. Στα πλαίσια της αναζήτησης ενός αξιόπιστου συνόλου δεδομένων, οφείλουμε να πραγματοποιήσουμε καθαρισμό δεδομένων.

Με τη χρήση του όρου εννοούμε τη διαχείριση ελλιπών τιμών και την απομάκρυνση του θορύβου ή έκτροπων παρατηρήσεων. Τις διαδικασίες καθαρισμού των δεδομένων μπορούμε να τις πετύχουμε μέσω σύνθετων στατιστικών μεθόδων ή χρησιμοποιώντας έναν αλγόριθμο της εξόρυξης δεδομένων.

Βήμα 4^{ο} μετασχηματισμός δεδομένων

Μέσω της διαδικασίας του μετασχηματισμού δεδομένων τα δεδομένα αλλάζουν σχήμα ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Για το λόγο αυτό εφαρμόζονται μέθοδοι μείωσης διαστάσεων και μετασχηματισμού χαρακτηριστικών. Αποτέλεσμα των εφαρμογών αυτών είναι η μείωση του αριθμού των μεταβλητών που εξετάζονται ή η εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.

Βήμα 5^{ο} επιλογή της κατάλληλης μεθόδου εξόρυξης δεδομένων

Ύστερα από όσα βήματα έχουμε εκτελέσει ήμαστε σε θέση να αποφασίσουμε ποιόν τύπο εξόρυξης δεδομένων θα χρησιμοποιήσουμε. Αυτή η επιλογή βασίζεται περισσότερο στους στόχους της διαδικασίας KDD αλλά και στα βήματα που έχουν ήδη προηγηθεί.

Όμως, όπως έχουμε αναφέρει και θα σχολιάσουμε και στη συνέχεια δύο είναι οι βασικοί στόχοι της εξόρυξης δεδομένων όπως η περιγραφή και η πρόβλεψη. Οι τεχνικές εξόρυξης δεδομένων βασίζονται στην πλειοψηφία τους στην επαγωγική εκμάθηση όπου κατασκευάζονται ένα σαφές ή εννοούμενο μοντέλο μέσω γενίκευσης ενός επαρκούς αριθμού εκπαιδευτικών παραδειγμάτων.

Βασική προϋπόθεση είναι ότι αυτό το μοντέλο εκπαίδευσης θα μπορούσε να εφαρμοστεί σε μελλοντικές περιπτώσεις. Επίσης, η στρατηγική αυτή λαμβάνει υπόψη την περίπτωση μετά-εκμάθησης για το συγκεκριμένο σύνολο των διαθέσιμων δεδομένων.

Βήμα 6^{ο} επιλογή αλγορίθμου εξόρυξης δεδομένων

Έχοντας ορίσει τη στρατηγική, μπορούμε να επιλέξουμε τον τρόπο επίτευξης του στόχου. Σε αυτό το στάδιο εφαρμόζονται ευφυείς μέθοδοι με σκοπό την αναζήτηση προτύπων γνώσης που έχουν πολύ ενδιαφέρον. Για παράδειγμα ένας έλεγχος ακρίβειας θα ήταν καλύτερα να γίνει μέσω νευρωνικών δικτύων, ενώ για την κατανόηση της δομής θα επιλεγόταν τα δέντρα αποφάσεων.

Τα πρότυπα που αναζητούμε θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου όπως οι κανόνες ταξινόμησης, τα δέντρα, η παλινδρόμηση, η συσταδοποίηση, κλπ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα βήματα που αναφέραμε παραπάνω.

Βήμα 7^{ο} εκτέλεση αλγορίθμων

Η κάλυψη των προηγούμενων προϋποθέσεων οδηγεί στο επιθυμητό σημείο που θα εκτελέσουμε τον επιλεγόμενο αλγόριθμο. Είναι πιθανή η επανάληψη του αλγορίθμου αυτού αρκετές φορές μέχρι να προκύψει ικανοποιητικό αποτέλεσμα.

Βήμα 8^{ο} αξιολόγηση

Σε αυτό το βήμα γίνεται εκτίμηση και ερμηνεία των προτύπων που κάνουμε εξόρυξη λαμβάνοντας υπόψη τους στόχους που είχαν τεθεί στο βήμα 1. Επίσης παρατηρούμε την επίδραση των βημάτων 2,3 και 4 (επεξεργασία δεδομένων) στον αλγόριθμο εξόρυξης δεδομένων που έχουν επιλεγεί μέσα από τα βήματα 5,6 και 7 (εξόρυξη δεδομένων). Για παράδειγμα μπορεί να κριθεί αναγκαία η προσθήκη χαρακτηριστικών μεταβλητών στο βήμα 4, ώστε να επαναληφθεί η εφαρμογή της αλυσίδας KDD από εκεί. Το βήμα της αξιολόγησης επικεντρώνεται στην κρίση αν το μοντέλο που προκύπτει είναι κατανοητό και χρήσιμο, καθώς και η επιλογή των πιο ενδιαφερόντων εξαγόμενων προτύπων. Επιπλέον, στο βήμα της αξιολόγησης τεκμηριώνεται η ανακάλυψη της γνώσης και επιπλέον είναι διαθέσιμη για περεταίρω χρήση.

Βήμα 9^{ο} παρουσίαση και χρήση της ανακαλυφθείσας γνώσης

Στο τελευταίο βήμα η εξαγόμενη γνώση ενσωματώνεται στο σύστημα για περισσότερη δράση. Η επιτυχία αυτού του βήματος αποδεικνύει την αποτελεσματικότητα χρήσης της αλυσίδας KDD.

Επιπλέον, μέσα από αυτό το βήμα γίνεται έλεγχος για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση. Είναι πιθανό να αλλάζουν ορισμένες δομές δεδομένων όπως επίσης και κάποιες μεταβλητές μπορεί να μην είναι πλέον διαθέσιμες. Επίσης μπορεί να αλλάζει η περιοχή δράσης των δεδομένων, αφού μπορεί να προκύψει για μια μεταβλητή μια τιμή η οποία να μην έχει αναφερθεί πριν.

Από την απεικόνιση και την καταγραφή των βημάτων της KDD τα βήματα 2,3 και 4 ορίζουν τη διαδικασία προεπεξεργασίας των δεδομένων. η διαδικασία αυτή αποτελεί ένα απαραίτητο στάδιο πριν την εξόρυξη δεδομένων.

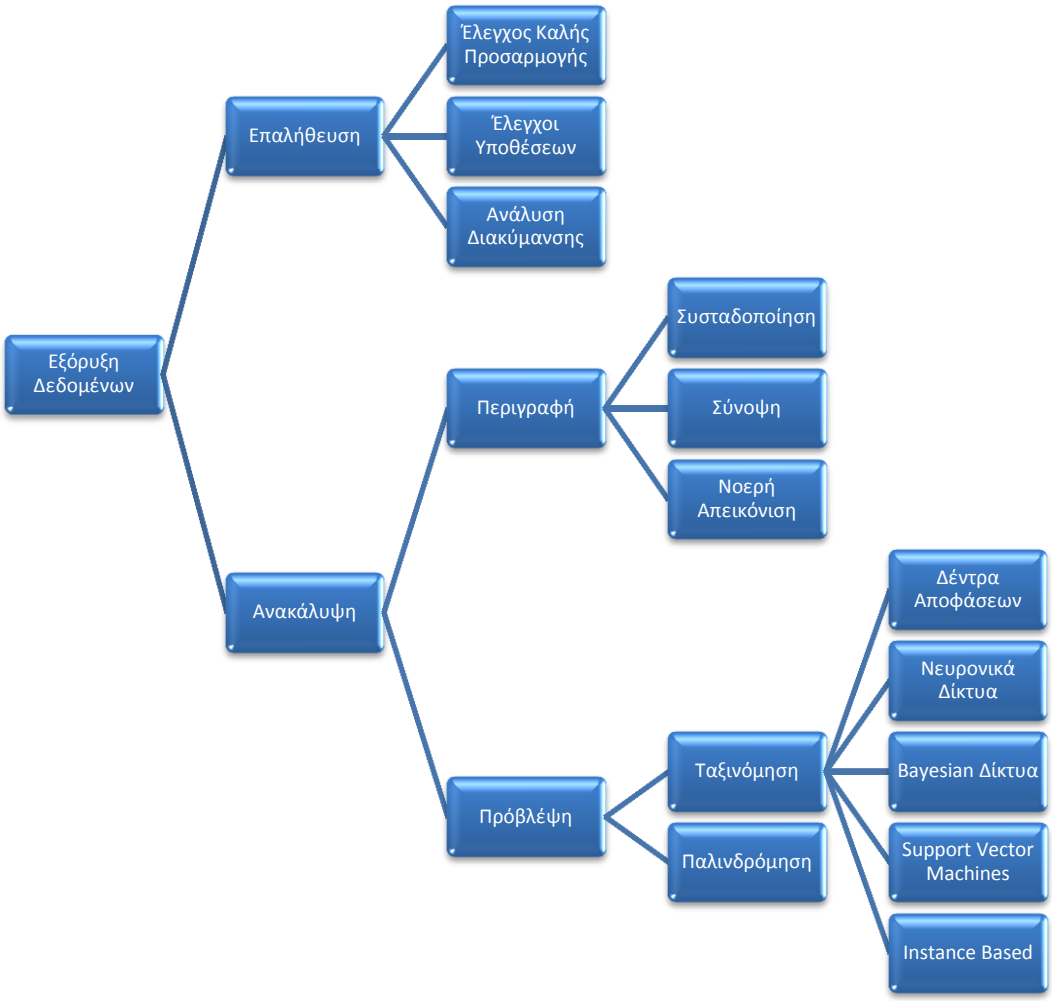
Επίσης τα βήματα 5,6 και 7 αποτελούν στην ουσία τον τομέα εξόρυξης δεδομένων που μας απασχολεί. Έτσι λοιπόν πρέπει να γίνει ταξινόμηση μεταξύ των μεθόδων της εξόρυξης δεδομένων ώστε να ήμαστε σε θέση να κατανοήσουμε τη διαδικασία και τις ανάγκες των βημάτων.

1.9. Διαχωρισμός των μεθόδων εξόρυξης δεδομένων

Έχει ήδη γίνει κατανοητό ότι υπάρχουν αρκετές μέθοδοι εξόρυξης δεδομένων οι οποίες χρησιμοποιούνται για διαφορετικούς σκοπούς να καλύψουν άλλους στόχους. Η ποικιλία των μεθόδων είναι τόσο μεγάλη που είναι αναγκαία η ταξινόμηση της γνώσης σε σχετικές ομάδες ανάλογα με τους στόχους που μπορεί να καλύψει κάθε ομάδα.

Σε αυτή την περίπτωση χωρίζουμε τις μεθόδους σε τέσσερα βασικά μέρη. Αρχικά πρέπει να αναφέρουμε ότι υπάρχουν δύο βασικοί τύποι εξόρυξης δεδομένων: η επαλήθευση και ανακάλυψη. Στον πρώτο γίνεται επαλήθευση των υποθέσεων του χρήστη από το σύστημα, ενώ από το δεύτερο τύπο το σύστημα βρίσκει νέους κανόνες και πρότυπα μέσα από αυτόνομες διαδικασίες.

Οι βασικοί μέθοδοι ανακάλυψης είναι αυτές που εντοπίζουν αυτόματα πρότυπα στα δεδομένα. Το στάδιο αυτό χωρίζεται στην περιγραφή και την πρόβλεψη. Αυτές είναι δύο από τις ομάδες που δημιουργήσαμε. Στο σχήμα είναι εμφανής η ταξινόμηση μεταξύ των εφαρμογών της εξόρυξης δεδομένων ενώ στη συνέχεια δίνουμε τέσσερις ομάδες εργασιών της εξόρυξης δεδομένων.



Εικόνα: Βασικοί τύποι εξόρυξης δεδομένων[7]

1.9.1. Περιγραφική μοντελοποίηση

Στόχος του μοντέλου περιγραφής είναι να γίνει περιγραφή όλου του συνόλου των δεδομένων ή της διαδικασίας που παράγει τα δεδομένα. Ας σκεφτούμε περιγραφές που περιλαμβάνουν μοντέλα για την κατανομή πιθανότητας των δεδομένων πυκνότητας, το διαχωρισμό ενός χώρου ή διαστάσεων σε ομάδες ή την περιγραφή των σχέσεων μεταξύ μεταβλητών.

Οι μέθοδοι περιγραφής έχουν στόχο την ερμηνεία των δεδομένων και επικεντρώνονται στην κατανόηση του τρόπου που σχετίζονται τα δεδομένα. Αυτό γίνεται μέσω μιας καινούριας απεικόνισης ή της σύνοψης με τις οποίες μπορούμε να πούμε ότι αποτελούν μέρος διερευνητικής ανάλυσης δεδομένων. η σημαντικότερη εφαρμογή των περιγραφικών μοντέλων είναι η ανάλυση συστάδων.

Η συσταδοποίηση επιχειρεί να βρει ομάδες παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα χαρακτηριστικά που περιλαμβάνουν. Οι μέθοδοι περιγραφής και

ειδικά η συσταδοποίηση είναι πολύ χρήσιμες σε επαγγέλματα που έχουν σαν βασικό χαρακτηριστικό τους πελάτες και βασίζονται στο CRM γιατί έτσι μπορούν να εντοπίσουν ομάδες πελατών που αναμένεται να έχουν όμοια συμπεριφορά.

1.9.2. Μοντελοποίηση πρόβλεψης

Η κατασκευή ενός μοντέλου πρόβλεψης έχει στόχο τη δυνατότητα πρόγνωσης της τιμής μιας μεταβλητής (απόκριση) μέσα από τις τιμές των άλλων μεταβλητών που είναι γνωστές ως επεξηγηματικές. Αν η μεταβλητή απόκρισης είναι κατηγορική, τότε είμαστε σε θέση να εφαρμόσουμε μια μέθοδο ταξινόμησης. Ένα παράδειγμα είναι η πρόβλεψη αγοράς ενός προϊόντος: ναι ή όχι. Όμως αν έχουμε συνεχή απόκριση, τότε χρησιμοποιούμε τη μέθοδο της παλινδρόμησης. Μια ενδεικτική τιμή είναι η πρόγνωση της μελλοντικής τιμής αποθέματος ενός προϊόντος.

Όπως και να ενεργήσουμε στόχος μας είναι η δημιουργία μοντέλου που ελαχιστοποιεί το σφάλμα στην προβλεφθείσα γνώση και τις πραγματικές τιμές. Ο όρος «πρόβλεψη» χρησιμοποιείται γενικά ως μια έννοια και δε θεωρείται απαραίτητα αυστηρή συνέχεια στο χρόνο. Για παράδειγμα μπορεί να θέλουμε να προβλέψουμε αν ο πελάτης τράπεζας θα πληρώσει το δάνειο του σε συγκεκριμένο μελλοντικό χρονικό διάστημα, αλλά μπορεί και να μας ενδιαφέρει ο προσδιορισμός της διάγνωσης για έναν ασθενή.

Η εξέλιξη της στατιστικής και της μηχανικής μάθησης και η σχετική βιβλιογραφία έχουν δώσει αρκετές μεθόδους πρόβλεψης και μεγάλη πρόοδο στη θεωρία για την καλύτερη και βαθύτερη κατανόηση. Το στοιχείο για να διακρίνουμε τις διαδικασίες πρόβλεψης από αυτές της περιγραφής είναι ο αντικειμενικός σκοπός της πρόβλεψης, δηλαδή μια συγκεκριμένη μεταβλητή, που δεν συμβαίνει στην περιγραφή, αν έχουμε να αναζητήσουμε τεχνικές προώθησης (boosting).

1.9.3. Ανάλυση σαφήνειας

Η ανάλυση αυτού του τύπου χρησιμοποιείται ώστε να ανακαλυφθούν πρότυπα που περιγράφουν ακριβή χαρακτηριστικά μεταξύ των δεδομένων. Τα πρότυπα αυτά απεικονίζονται συνήθως στα πλαίσια κανόνων συνεπαγωγής ή υποομάδων των χαρακτηριστικών.

Η χαρακτηριστικότερη εφαρμογή και αιτία από την οποία ξεκίνησαν οι κανόνες συνάφειας είναι η ανάλυση του «καλαθιού αγοράς». Ας σκεφτούμε τα σουπερμάρκετ όπου οι καταναλωτές γεμίζουν το καλάθι τους με προϊόντα της επιλογής τους. Ο όγκος της πληροφορίας που μπορούμε να συλλέξουμε είναι τεράστιος. Οι κανόνες συνάφειας αξιοποιούν αυτή την πληροφορία. Για παράδειγμα ένας κανόνας μπορεί να

είναι «οι πελάτες που αγοράζουν ψωμί για τοστ, αγοράζουν και αλλαντικά σε ποσοστό 70%». Στον κανόνα αυτό που δώσαμε έχουμε ένα αίτιο (που είναι η αγορά του ψωμιού) το οποίο συνδέεται με ένα αποτέλεσμα(που είναι η αγορά αλλαντικών). Επίσης δίνεται και εκτίμηση για το πόσο πιθανό είναι να συμβεί αυτή η σχέση αιτίας-αποτελέσματος. Οι κανόνες συνάφειας καλούνται και σαν κανόνες «if-then». Άλλες εφαρμογές που πραγματοποιούνται σαν την προώθηση προϊόντων είναι: η τοποθέτηση των προϊόντων στα ράφια καταστημάτων, η διαχείριση αποθεμάτων κλπ.

1.9.4. Ανίχνευση παρεκτροπών

Σε αυτή την ομάδα μεθόδων ανήκουν εργασίες εντοπισμού παρατηρήσεων των οποίων τα χαρακτηριστικά διαφέρουν από αυτά του υπόλοιπου συνόλου δεδομένων. Τέτοιες παρατηρήσεις καλούνται παρέκτροπες ή ακραίες τιμές. Μια σχετική μέθοδος είναι η ανίχνευση αλλαγών και αποκλίσεων.

Ο σκοπός ενός αλγορίθμου ανίχνευσης παρεκτροπών είναι η κάλυψη πραγματικών ανωμαλιών και η αποφυγή λανθασμένου χαρακτηρισμού ενός φυσιολογικού αντικειμένου ως έκτροπο. Δηλαδή, επιθυμούμε ανίχνευση υψηλού επιπέδου όσων αφορά τυχούσες ανωμαλίες, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης.

Ως εφαρμογή της ανίχνευσης παρεκτροπών μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων πιστωτικών καρτών από μια τράπεζα. Αυτό είναι ζήτημα υψίστης σημασίας για τέτοιους οργανισμούς.

1.10. Σύγκριση ταξινόμησης και συσταδοποίησης

Επιθυμώντας να διευκρινίσουμε τη διαφορά μεταξύ ταξινόμησης και συσταδοποίησης πρέπει να σημειώσουμε ότι στην ταξινόμηση επιχειρείται η περιγραφή μιας λειτουργίας που αντιστοιχεί ένα στοιχείο σε μια από τις κατηγορίες οι οποίες είναι ήδη προκαθορισμένες.

Η ταξινόμηση χαρακτηρίζεται από ένα καλά ορισμένο σύνολο κατηγοριών, καθώς και ένα σύνολο προκαθορισμένων παραδειγμάτων. Αντιθέτως η συσταδοποίηση δεν στηρίζεται σε προκαθορισμένες κατηγορίες ή παραδείγματα. Επιπλέον ο στόχος μιας μεθόδου ταξινόμησης είναι από τη μια η εκμάθηση και από την άλλη η κατηγοριοποίηση, δηλαδή η ταξινόμηση. Στην ουσία δημιουργείται ένα μοντέλο που μπορεί να χρησιμοποιηθεί για την ταξινόμηση μελλοντικών δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

Ακόμα αξίζει να αναφέρουμε μια επιπλέον ορολογία για τις μεθόδους πρόβλεψης που είναι η εποπτευόμενη εκμάθηση. Αντιθέτως οι μέθοδοι περιγραφής

χαρακτηρίζονται ως μη εποπτευόμενη εκμάθηση καθώς δεν υπάρχει προηγούμενη πληροφορία πάνω στην οποία μπορούμε να βασιστούμε για την περιγραφή των δεδομένων, και η περιγραφή που κάνουμε στηρίζεται πάνω στη συγκεκριμένη βάση δεδομένων.

Ο όρος της μη εποπτευόμενης μάθησης θεωρείται ότι ταιριάζει καλύτερα κυρίως στη συσταδοποίηση και όχι σε όλες τις μεθόδους περιγραφής, ενώ ο όρος της εποπτευόμενης μάθησης αφορά τις μεθόδους πρόβλεψης γενικά.

1.11 Τύποι δομής μοντέλα και πρότυπα

Με βάση τους ορισμούς της εξόρυξης δεδομένων και της KDD κρίνουμε απαραίτητο να κάνουμε διαχωρισμό μεταξύ των όρων «μοντέλο» και «πρότυπο». Η ουσιαστική διαφορά είναι ο στόχος που έχουμε θέσει και ο τρόπος που θέλουμε να αξιοποιήσουμε και να παρουσιάσουμε τα δεδομένα. Για παράδειγμα ως στόχο μπορεί να έχουμε την περιγραφή των δεδομένων ή τη δημιουργία μιας πρόβλεψης για ένα θέμα που έχει τεθεί.

Σκεφτόμενοι λοιπόν όλα αυτά θεωρούμε ότι μέσα από τις διαδικασίες που θα πραγματοποιήσουμε μπορούμε να οδηγηθούμε σε ένα συνολικό μοντέλο ή σε ένα τοπικό πρότυπο. Όταν μιλάμε για έναν τύπο δομής ενός μοντέλου εννοούμε τη συνολική σύνοψη ενός συνόλου δεδομένων. Με βάση το μοντέλο αυτό εξετάζουμε όλους τους δυνατούς ισχυρισμούς.

Κεφάλαιο 2ο

Εξόρυξη δεδομένων στη διαχείριση πελατειακών σχέσεων

2.1. Εισαγωγή

Η εξόρυξη δεδομένων μπορεί να βοηθήσει τις επιχειρήσεις και να ανταπεξέλθει σε περιόδους που η κατανάλωση είναι χαμηλή, στη διακίνηση των εμπορευμάτων, αλλά και στο να καταφέρουμε να μειώσουμε τη ζημία που οφείλεται σε εσωτερικές απάτες, οι οποίες υπολογίζονται ότι είναι περίπου στο 40%- 50% των αποθεμάτων των επιχειρήσεων λιανικής πώλησης. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στην επίλυση ασυνήθιστων περιπτώσεων που αφορούν επιστροφές εμπορευμάτων, εκπτώσεις, υπερισχύουσες τιμές, πιστωτικές κάρτες, κάρτες καταστημάτων, χρεωστικές κάρτες, εκπτώσεις που οφείλονται σε αποθέματα που αναφέρονται ως κατεστραμμένα ή ελαττωματικά, καθιστώντας με αυτόν τον τρόπο δυνατό τον εντοπισμό της λιανικής απάτης πολύ πιο εύκολη, ακριβής, έγκυρη και οικονομική.

Η ομαλή λειτουργία επιχειρήσεων ή οργανισμών που παράγουν ή διακινούν αγαθά(φυσικά αγαθά ή υπηρεσίες) βασίζεται και στη σωστή διαχείριση των αποθεμάτων των αγαθών αυτών. Αυτό συμβαίνει, στην περίπτωση που είναι απαραίτητη η αποθεματοποίηση των αγαθών ώστε να ικανοποιείται μέσα στο πλαίσιο λειτουργίας του οργανωμένου συστήματος που λέγεται επιχείρηση ή οργανισμός. Κατά μια απλή λογική, τα αποθέματα δεν πρέπει να είναι μεγάλα ώστε το κόστος διαχείρισης να είναι υπερβολικά υψηλό αλλά ούτε και μικρά ώστε να δημιουργούνται άλλα προβλήματα.

Τρεις είναι οι βασικοί λόγοι που υπαγορεύουν την αποθεματοποίηση ενός αγαθού:

- ✓ Η ικανοποίηση της ζήτησης
- ✓ Η ύπαρξη αβεβαιότητας
- ✓ Η αντιμετώπιση της ζήτησης σε έκτακτες καταστάσεις

Η ικανοποίηση της ζήτησης, ακόμα και όταν αυτή είναι επακριβώς γνωστή, πολλές φορές δεν είναι δυνατή λόγω οικονομιών κλίμακας κατά τη διάρκεια της. Έτσι θεωρείται απαραίτητη η αποθεματοποίηση προκειμένου να επιτευχθεί κάποιος συγχρονισμός στην εισροή και την εκροή ενός αγαθού. Ακόμα και σε περιπτώσεις που η ζήτηση παρουσιάζει εποχικότητα ή δεν μπορεί να προβλεφθεί επαρκώς, η αποθεματοποίηση εξασφαλίζει την ομαλή ροή των αγαθών στα σημεία ζήτησης.

Η αβεβαιότητα, που υπεισέρχεται στη ζήτηση ενός αγαθού ή ακόμα και στο χρόνο παραγωγής και διάθεσης, ενδέχεται να οδηγήσει σε έλλειψη που με τη σειρά της δημιουργεί προβλήματα όπως το κόστος. Το γεγονός αυτό οδηγεί στη δημιουργία αποθέματος σαν γεγονός αμυντικού μέτρου.

Η αντιμετώπιση έκτακτων καταστάσεων, όπως η καταστροφή της σοδειάς λόγω θεομηνίας, επιβάλλει την δημιουργία αποθεμάτων ασφαλείας.

Η όλη θεωρεία των αποθεμάτων ξεκινάει από το 1918 με την καθιέρωση του μοντέλου της οικονομικής ποσότητας παραγγελίας και συνεχίστηκε αργότερα κατά τη διάρκεια του Β΄ παγκόσμιου πολέμου με ιδιαίτερη έμφαση στα προβλήματα που προκύπτουν από το στοχαστικό χαρακτήρα της ζήτησης.

Η επιστημονική προσέγγιση στα προβλήματα ελέγχου αποθεμάτων έχει ως στόχο τον προσδιορισμό των σχέσεων μεταξύ εισροής, της αποθεματοποίησης και της εκροής των αγαθών, ώστε να αναπτυχθούν οι σχετικές μέθοδοι ελέγχου. Κατά συνέπεια, το πρόβλημα στο σύνολο του έχει σύνθετη μορφή, γιατί περιέχει την έννοια της πρόβλεψης και στη συνέχεια του προγραμματισμού.

Τα προβλήματα ελέγχου αποθεμάτων, τέλος, απασχολούν σε μεγάλο βαθμό και άλλες θεωρίες της επιχειρησιακής έρευνας, το γραμμικό προγραμματισμό και την προσομοίωση.

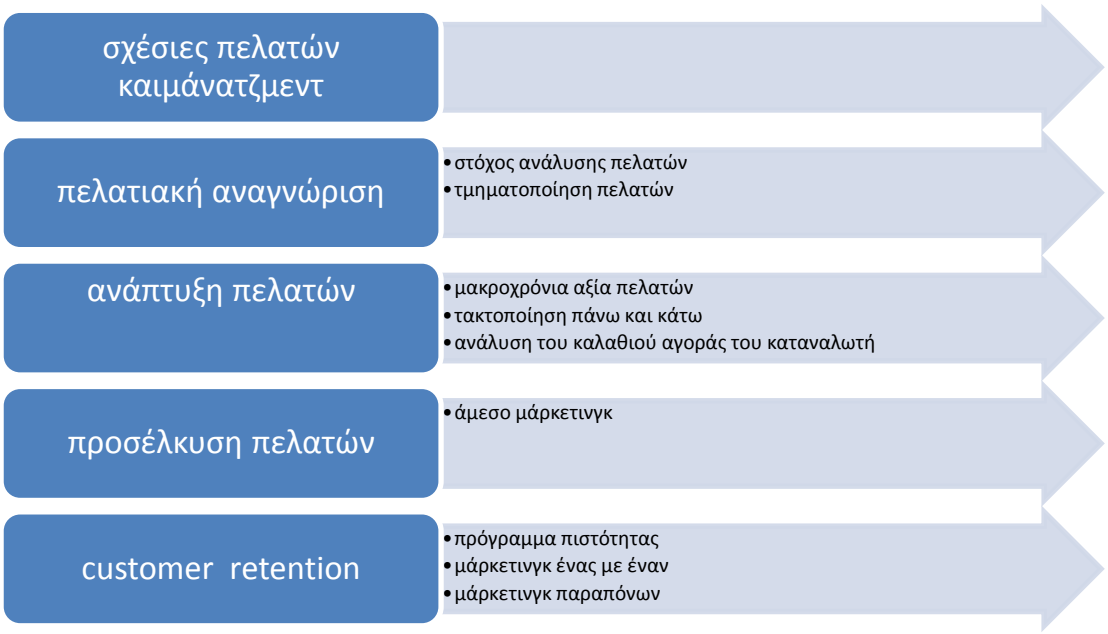
2.2. Έννοιες- Ορισμοί

Τα τελευταία χρόνια έχουν γίνει σημαντικές βελτιώσεις, όσον αφορά την ενσωμάτωση της εξόρυξης δεδομένων στη διαδικασία του CRM ή αλλιώς διαχείριση πελατειακών σχέσεων. Η τάση αυτή αναμένεται να συνεχιστεί, με αποτέλεσμα οι εφαρμογές του CRM να έχουν όλο και περισσότερες δραστηριότητες μάρκετινγκ βασιζόμενες στα αποτελέσματα της εξόρυξης δεδομένων στο κύκλο της ζωής. Ενώ οι μελέτες για την εξόρυξη δεδομένων έχουν επικεντρωθεί κυρίως σε τεχνικές, οι μελέτες για τις πελατειακές σχέσεις έχουν επικεντρωθεί στην αλληλεπίδραση του πελάτη και των στρατηγικών της εταιρείας. Η σωστή διαχείριση πελατειακών σχέσεων μπορεί να επιτευχθεί μόνο μέσω της ενοποίησης της διαδικασίας της ανακάλυψης γνώσης με τη διαχείριση και χρήση γνώσης στις στρατηγικές μάρκετινγκ. Αυτό βοηθάει την επιχείρηση στην αντιμετώπιση των αναγκών των πελατών, με βάση τις πληροφορίες που γνωρίζει για κάθε πελάτη ξεχωριστά, παρά με τη χρήση μιας μαζικής γενίκευσης των χαρακτηριστικών των πελατών.

2.3. Ταξινόμηση των τεχνικών εξόρυξης δεδομένων στις τέσσερεις διαστάσεις του CRM

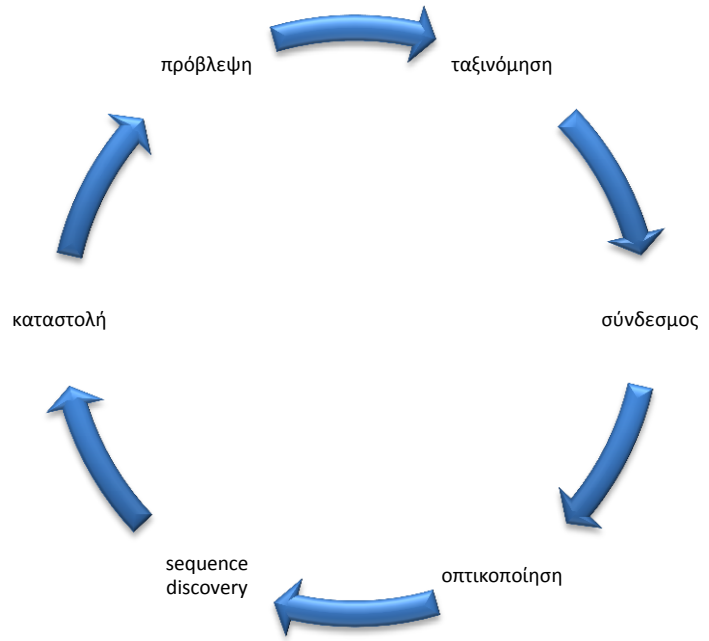
Η ανάλυση και κατανόηση των συμπεριφορών και των χαρακτηριστικών των πελατών αποτελούν τα θεμέλια της ανάπτυξης μιας ανταγωνιστικής στρατηγικής CRM με σκοπό την απόκτηση και διατήρηση των δυνητικών πελατών και τη μεγιστοποίηση της αξίας των πελατών. Τα εργαλεία εξόρυξης δεδομένων και γνώσεις από την τεράστια βάση δεδομένων των πελατών, προσφέρουν μεγάλη υποστήριξη στη λήψη διαφορετικών αποφάσεων CRM.

Στην εικόνα απεικονίζεται η ταξινόμηση ων τεχνικών εξόρυξης δεδομένων που χρησιμοποιούνται στο CRM, ανάλογα με την διάσταση του συστήματος πάνω στην οποία εφαρμόζονται. Οι τεχνικές της εξόρυξης δεδομένων που χρησιμοποιούνται θα πρέπει να βασίζονται στα χαρακτηριστικά των δεδομένων, όπως επίσης και στις απαιτήσεις της επιχείρησης.



Εικόνα: Ταξινόμηση των τεχνικών εξόρυξης δεδομένων στο CRM[6]

Πολύ συχνά, απαιτείται να γίνει συνδυασμός των μοντέλων εξόρυξης δεδομένων για την υποστήριξη ή την πρόβλεψη των επιπτώσεων μιας στρατηγικής CRM. Σε μια τέτοια περίπτωση, η ταξινόμηση των μοντέλων εξόρυξης δεδομένων θα βασίζεται στα σημαντικότερα σημεία της στρατηγικής CRM. Για παράδειγμα, δεδομένου ότι οι σχέσεις μεταξύ των προϊόντων και το κύριο μέλημα της επιχείρησης, στην περίπτωση των προγραμμάτων σταυροειδών πωλήσεων, οι πελάτες μπορούν να κατηγοριοποιηθούν σε ομάδες προτού εφαρμοστεί σε κάθε ομάδα κάποιο μοντέλο συσχετίσεων. Σε τέτοιες περιπτώσεις το πρόγραμμα σταυροειδών πωλήσεων θα πρέπει να ταξινομηθεί με βάση το μοντέλο των συσχετίσεων. Στην περίπτωση του άμεσου μάρκετινγκ, ένα ορισμένο τμήμα των πελατών μπορεί να υποδιαιρεθεί σε ομάδες, έτσι ώστε να σχηματιστούν τα αρχικά τμήματα του μοντέλου τμηματοποίησης. Το πρόγραμμα του άμεσου μάρκετινγκ θα ταξινομηθεί με βάση την τμηματοποίηση, δεδομένου ότι το κύριο μέλημα εδώ είναι η πρόβλεψη της συμπεριφοράς των πελατών.



Εικόνα: Συνέχεια του σχήματος σελίδα34 [6]

2.4. Προετοιμασία των δεδομένων για όλα τα εργαλεία εξόρυξης δεδομένων

Πολύ συχνά χρειάζεται να γίνει προετοιμασία των δεδομένων, προκειμένου να γίνει η είσοδος τους στη διαδικασία της εξόρυξης δεδομένων. Μάλιστα, μπορεί να δαπανηθεί περισσότερος χρόνος κατά την προετοιμασία των δεδομένων, παρά κατά την διαδικασία της εξόρυξης δεομένων. Το πιο σημαντικά βήματα, που χρειάζονται τα δεδομένα να αποκτήσουν την απαιτούμενη μορφή, είναι τα εξής:

- Η διόρθωση ασυνεπούς μορφής δεδομένων και η διόρθωση δεδομένων που η κωδικοποίηση τους δεν είναι συμβατή, οι συντομογραφίες και τα σημεία στίξης.
- Η αφαίρεση ανεπιθύμητων ή περιττών πεδίων. Τα δεδομένα περιέχουν πεδία χωρίς ιδιαίτερη σημασία για την ανάλυση που θέλουμε να κάνουμε. Τα εργαλεία της εξόρυξης δεδομένων μπορεί να ερμηνεύσουν αυτά τα πεδία ως μετρήσεις ή μεγέθη, ειδικά αν πρόκειται για αριθμούς και μπορεί να προκαλέσουν εκτέλεση κύκλων προσπαθώντας να συσχετίσουν αυτά τα πεδία με πραγματικά δεδομένα.
- Οι κωδικοί πρέπει να μεταφραστούν σε κείμενο. Η κλασική μορφή «καθορισμού» δεδομένων περιλαμβάνει βελτίωση ή την αντικατάσταση αινιγματικών κωδικών με ισοδύναμα κειμένου, γραμμένο με αναγνωρίσιμες λέξεις.
- Τα δεδομένα και οι συνδυασμοί τους προέρχονται από διαφορετικές πηγές όπως είναι τα δεδομένα των πελατών που βρίσκονται σε μια κοινή βάση.

- Θα πρέπει να βρεθούν τα πεδία που έχουμε χρησιμοποιήσει για περισσότερους από ένα σκοπό. Ένας τρόπος για να βρούμε τα αρχεία αυτά είναι η καταμέτρηση και η δημιουργία μιας λίστας με όλες τις διαφορετικές τιμές ή χρήσεις που βρίσκονται και υπάρχουν σε ένα πεδίο.
- Ο έλεγχος για στοιχεία που μπορεί να μην είναι φυσιολογικά ή αδύνατα. Κάποια στοιχεία που έχουμε μετρήσει μπορεί να είναι σωστά, αλλά επίσης μπορεί να είναι και ασυνήθιστα. Τέτοια δεδομένα, είναι καλύτερα να μαρκαριστούν με μια ειδική σήμανση, έτσι ώστε να μπορούμε να τα συμπεριλάβουμε ή να τα εξαιρέσουμε από την ανάλυση μας ανάλογα με την περίπτωση.
- Ο έλεγχος για τιμές που μπορεί να λείπουν, ή έχουν αντικατασταθεί από κάποιον προεπιλεγμένο αριθμό.
- Η εφαρμογή ίδιας μεταχείρισης σε τιμές που είναι μηδενικές. Οι τιμές αυτές ενδέχεται να δυσκολέψουν τον τρόπο λειτουργίας του εργαλείου εξόρυξης δεδομένων. σε πολλές περιπτώσεις, η μηδενική αξία αντικαθίσταται από κάποιον προεπιλεγμένο αριθμό.
- Τα ξεχωριστά αρχεία δεδομένων ταξινομούνται σύμφωνα με ένα από τα συγκεντρωτικά του μεγέθη. Κάποιες φορές, μπορεί να είναι επιθυμητός ο εντοπισμός της πώλησης ενός πολύ συγκεκριμένου προϊόντος, όπως ένα τρόφιμο που έχει συγκεκριμένη συσκευασία, μάρκα, τιμή αλλά και από τι υλικά αποτελείται.

2.4.1. Προετοιμασία των δεδομένων ανάλογα με το χρησιμοποιούμενο εργαλείο εξόρυξης δεδομένων

Ανάλογα με το πιο εργαλείο της εξόρυξης δεδομένων χρησιμοποιούμε μπορεί να γίνουν κάποιες επιπλέον μετατροπές στα δεδομένα και αυτές είναι οι ακόλουθες:

- Ο διαχωρισμός των εισερχόμενων πρωτογενών δεδομένων σε τρείς ομάδες. Η πρώτη ομάδα δεδομένων χρησιμοποιείται για την κατάρτιση του εργαλείου εξόρυξης δεδομένων. Ένα εργαλείο συσταδοποίησης, ένα εργαλείο νευρωνικών δικτύων ή ένα εργαλείο δέντρου αποφάσεων που απορροφά την πρώτη σειρά στοιχείων και ορίζει τις παραμέτρους από τις οποίες μπορούν να γίνουν οι μελλοντικές ταξινομήσεις και οι προβλέψεις. Το δεύτερο σύνολο δεδομένων, ελέγχει αυτές τις παραμέτρους για να τσεκάρει πόσο καλά αποδίδει το μοντέλο. Όταν το εργαλείο εξόρυξης δεδομένων έχει ρυθμιστεί σωστά στο πρώτο και στο δεύτερο βήμα αξιολόγησης δεδομένων, εφαρμόζεται στη συνέχεια το τρίτο βήμα αξιολόγησης των δεδομένων, όπου τα συμπλέγματα οι ταξινομήσεις και οι προβλέψεις που προέρχονται από το εργαλείο είναι πλήρως αξιόπιστες και μπορούμε να τις χρησιμοποιήσουμε.

- Η προσθήκη πεδίων που έχουν υπολογιστεί σαν εισροές ή σαν στόχοι. Για παράδειγμα, ένα τέτοιο πεδίο, όπως είναι τα κέρδη ή η ικανοποίηση των πελατών, μπορεί να τεθεί ως στόχος και να επιλέξει το εργαλείο εξόρυξης δεδομένων τους πιο κερδοφόρους πελάτες ή για να επιλέξει τη συμπεριφορά που θέλουμε να ενθαρρύνουμε και αυτό έχει σαν αποτέλεσμα να μπορούμε να γνωρίζουμε και την προτίμηση τους στην αγορά των προϊόντων.
- Η διάταξή τους σε πίνακες με συνεχείς τιμές. Μερικά εργαλεία εξόρυξης δεδομένων ενθαρρύνουν τη διάταξη με διακριτές τιμές σε κλίμακες που είναι τα δέντρα αποφάσεων.
- Η εξομάλυνση των τιμών μεταξύ 0 και 1. Τα εργαλεία νευρωνικών δικτύων συνήθως απαιτούν όλες οι αριθμητικές τιμές να αντιστοιχίζονται σε μια σειρά από το μηδέν και το ένα.
- Η μετατροπή των κειμένων σε αριθμητικές τιμές. Μερικά εργαλεία εξόρυξης δεδομένων μπορούν να λειτουργήσουν μόνο με αριθμητικά δεδομένα εισόδου. Σε αυτές τις περιπτώσεις, οι διακριτές τιμές κειμένου θα πρέπει να αντικατασταθούν από ειδικούς κωδικούς, όπως για παράδειγμα η αντικατάσταση της περιοχής κάθε πελάτη με τον αντίστοιχο ταχυδρομικό του κώδικα.

2.5. Τα βήματα εφαρμογής της εξόρυξης δεδομένων στη διαχείριση πελατειακών σχέσεων

Τα βασικά βήματα της εξόρυξης δεδομένων για ένα αποτελεσματικό σύστημα διαχείρισης πελατών είναι τα ακόλουθα:

1. Καθορισμός του προβλήματος της επιχείρησης. Κάθε εφαρμογή CRM έχει έναν ή περισσότερους επιχειρησιακούς στόχους για τους οποίους θα πρέπει να οικοδομηθεί το κατάλληλο μοντέλο. Ο αποτελεσματικός εντοπισμός του προβλήματος περιλαμβάνει και έναν τρόπο μέτρησης αποτελεσμάτων του έργου του CRM.
2. Δημιουργία βάσης δεδομένων μάρκετινγκ. Είναι απαραίτητη η δημιουργία μιας βάσης δεδομένων, γιατί οι επιχειρησιακές βάσεις δεδομένων συχνά δεν περιέχουν τα δεδομένα που απαιτούνται με τη μορφή που απαιτούνται. Επίσης, εάν υπάρχουν ξεχωριστές βάσεις δεδομένων, όπως για παράδειγμα ξεχωριστή βάση δεδομένων για τους πελάτες, ξεχωριστή βάση δεδομένων για τα προϊόντα και ξεχωριστή βάση δεδομένων για τις συναλλαγές, θα πρέπει να γίνει μια ενσωμάτωση όλων αυτών σε μια ενιαία βάση μάρκετινγκ.
3. Εξερεύνηση των δεδομένων. απαραίτητη κρίνεται η σωστή κατανόηση των δεδομένων προκειμένου να δημιουργηθεί ένα αξιόπιστο μοντέλο προβλέψεων. Χρήσιμη θα μπορούσε να είναι η συγκέντρωση κάποιων αριθμητικών δεικτών, όπως μέσοι όροι και

τυπικές αποκλίσεις, και η εξέταση της διανομής των δεδομένων. Σημαντική βοήθεια προσφέρουν οι γραφικές παραστάσεις και τα εργαλεία απεικόνισης. Η οπτικοποίηση των δεδομένων οδηγεί συχνά σε νέες ιδέες και αυτό με τη σειρά του οδηγεί στην επιτυχία.

4. Η προετοιμασία των δεδομένων για τη μοντελοποίηση. Αρχικά γίνεται η επιλογή στις μεταβλητές πάνω στις οποίες θα κατασκευαστεί το μοντέλο και στη συνέχεια κατασκευάζονται νέες μεταβλητές οι οποίες προέρχονται από τα δεδομένα που ακόμα δεν έχουν επεξεργαστεί. Στη συνέχεια επιλέγονται ένα υποσύνολο ή ένα δείγμα δεδομένων πάνω στο οποίο θα κατασκευαστεί το μοντέλο. Στο τελικό στάδιο γίνεται η μετατροπή των μεταβλητών σύμφωνα με τις απαιτήσεις του αλγόριθμου που έχουμε επιλέξει για να κατασκευάσουμε το μοντέλο.
5. Η αξιολόγηση του μοντέλου. Η κατασκευή του μοντέλου αποτελεί μια επαναληπτική διαδικασία θα πρέπει να διερευνηθούν εναλλακτικά μοντέλα για να βρεθεί το καταλληλότερο στην επίλυση του προβλήματος της επιχείρησης. Κατά τη διαδικασία αυτή της ανίχνευσης του μοντέλου, πολλές φορές μπορεί να οδηγηθούμε σε κάποιο προηγούμενο βήμα και να κάνουμε κάποιες αλλαγές στα δεδομένα, ή ακόμα και να γίνει επαναπροσδιορισμός του αρχικού προβλήματος της επιχείρησης (για εμάς το σουπερμάρκετ).
6. Η αξιολόγηση του μοντέλου. Ίσως είναι το πιο πολύτιμο μέτρο αξιολόγησης του μοντέλου και είναι η ακρίβεια των αποτελεσμάτων. Ένα άλλο μέτρο που χρησιμοποιούμε συχνά είναι η ανύψωση, η οποία μετρά τη βελτίωση που πετυχαίνει το μοντέλο πρόβλεψης. Η μέθοδος αυτή δεν περιλαμβάνει τα κόστη και τα έσοδα, έτσι είναι προτιμότερο να εξετάζουμε τους δείκτες που αναφέρονται στα κέρδη και την απόδοση της επένδυσης.
7. Η ανάπτυξη του μοντέλου και του αποτελέσματος. Στην πραγματικότητα ο τρόπος που ενσωματώνεται η εξόρυξη δεδομένων στην εφαρμογή του CRM καθορίζεται από το είδος της αλληλεπίδρασης της εταιρείας με τον πελάτη. Υπάρχουν δύο βασικοί τρόποι αλληλεπίδρασης: η εισερχόμενη, κατά την οποία οι πελάτες επικοινωνούν με την εταιρεία και η εξερχόμενη, κατά την οποία η εταιρεία επικοινωνεί με τους πελάτες. Στις εισερχόμενες συναλλαγές όπως είναι μια τηλεφωνική παραγγελία ή παραγγελία μέσω ιντερνέτ, η εφαρμογή θα πρέπει να ανταποκριθεί σε πραγματικό χρόνο, κάτι που δεν συμβαίνει με τις εξερχόμενες συναλλαγές.

2.6. Η εξόρυξη δεδομένων στον κύκλο ζωής του πελάτη

Για να είναι αποτελεσματικό ένα σύστημα CRM θα πρέπει να συνδεθούν τα προϊόντα και οι στρατηγικές της εταιρείας με τους στόχους και τους πελάτες της. Ο όρος «κύκλος ζωής του πελάτη» αναφέρεται σε όλα τα στάδια της σχέσης μεταξύ επιχείρησης και του πελάτη. Είναι απόλυτα απαραίτητη η κατανόηση του όρου από την εταιρεία, γιατί σχετίζεται άμεσα με την κερδοφορία της. Ο κύκλος ζωής του πελάτη αποτελείται από τρία στάδια, τα οποία είναι: η απόκτηση του πελάτη, η αύξηση του πελάτη και η διαχείριση του πελάτη.

1. Απόκτηση πελατών: Αποτελεί το πρώτο βήμα του CRM και η εξόρυξη δεδομένων μπορεί να βοηθήσει στην βελτίωση της αποτελεσματικότητας μιας απόπειρας για την απόκτηση πελατών και ελαχιστοποίηση του κόστους.
2. Αύξηση της αξίας των πελατών:
   - Σταυροειδής πωλήσεις: Η εξόρυξη δεδομένων, χρησιμοποιώντας τις πληροφορίες των πελατών στις βάσεις δεδομένων είναι σε θέση να βοηθήσει τον εκπρόσωπο εξυπηρέτησης πελατών να προτείνει τα κατάλληλα επιπρόσθετα προϊόντα στον πελάτη όταν ο πελάτης δεν είναι αποδεκτός σε τέτοιου είδους πωλήσεις. Επίσης μέσω της εξόρυξης δεδομένων μπορούν να ελαχιστοποιηθούν τα παράπονα των πελατών και να αυξηθεί η κερδοφορία της επιχείρησης.
   - Εξατομίκευση πελατών: Μέσω της συσταδοποίησης της εξόρυξης δεδομένων γίνεται εφικτή η ομαδοποίηση των παρόμοιων προϊόντων, έτσι ώστε κάθε φορά που κάποιος πελάτης δείχνει ενδιαφέρον για ένα προϊόν, η επιχείρηση να κάνει τις απαραίτητες συστάσεις για αγορά περισσότερων προϊόντων. Με βάση το προφίλ του πελάτη εντοπίζονται οι πελάτες που μπορεί να έχουν κοινά ενδιαφέροντα για νέα προϊόντα που περιλαμβάνει ο κατάλογος της επιχείρησης.
3. Διατήρηση κερδοφόρων πελατών: Για σχεδόν κάθε εταιρεία, το κόστος απόκτησης ενός νέου πελάτη υπερβαίνει το κόστος διατήρησης κερδοφόρων πελατών. Με τη δημιουργία του προφίλ των επικερδών, αλλά και μη επικερδών πελατών, καθίσταται εφικτότερη η διατήρηση τους ως πελάτες και ο εντοπισμός των πελατών που δεν αποφέρουν αρκετά έσοδα στην επιχείρηση, αλλά μπορούσαν να αποφέρουν στο μέλλον.

Ο κύκλος ζωής του πελάτη αποτελεί ένα καλό πλαίσιο για την εφαρμογή της εξόρυξης δεδομένων στο CRM, αφού είναι σε θέση να προβλέψει την κερδοφορία των δυνητικών πελατών, καθώς αυτοί γίνονται ενεργοί, πόσο καιρό θα είναι ενεργοί πελάτες και ποιες είναι οι πιθανότητες να παύσουν να είναι πελάτες. Βέβαια δε θα είναι ένας ακριβής προγνωστικός δείκτης για το πότε συμβαίνουν οι περισσότερες εκδηλώσεις του κύκλου ζωής, αλλά θα μπορεί να βοηθήσει την επιχείρηση να αναγνωρίζει πρότυπα στα δεδομένα των πελατών της, τα οποία είναι προβλέψιμα.

2.7. Έξυπνα δεδομένα μάρκετινγκ

Το αποτέλεσμα της εφαρμογής της εξόρυξης δεδομένων στον κύκλο της ζωής του πελάτη είναι τα δεδομένα έξυπνου μάρκετινγκ, το οποίο ορίζεται σαν το συνδυασμό του μάρκετινγκ, με βάση τα δεδομένα, και την τεχνολογία με σκοπό τη μεγιστοποίηση της γνώσης και της κατανόησης των πελατών, και τον συνδυασμό των προϊόντων και των δεδομένων συναλλαγής με σκοπό τη βελτίωση της στρατηγικής λήψης των αποφάσεων. Υπάρχουν δύο σημαντικά στοιχεία στα δεδομένα έξυπνου μάρκετινγκ, ο μετασχηματισμός των δεδομένων των πελατών και η ανακάλυψη γνώσης των πελατών.

Τα δεδομένα τα οποία δεν έχουν επεξεργαστεί εξάγονται και μετασχηματίζονται από ένα μεγάλο φάσμα εσωτερικών και εξωτερικών δεδομένων και η συλλογή όλων αυτών των δεδομένων σε μια κεντρική θέση, όπου μπορούν να αναζητηθούν και να διερευνηθούν, αποτελούν το μετασχηματισμό δεδομένων. Η διαδικασία αυτή σχετίζεται μέσω της ανακάλυψης γνώσης των πελατών, κατά την οποία προέρχονται οι πληροφορίες και μπορούν να εξαχθούν χρήσιμα πρότυπα και συμπεράσματα μέσα από τα δεδομένα. Η διαδικασία αυτή θα πρέπει να παρακολουθείται ώστε να διασφαλίζεται ότι τα αποτελέσματα παράγουν αξιοποιήσιμες πληροφορίες.

2.7.1 Τμηματοποίηση πελατών με εργαλεία της εξόρυξης δεδομένων

Οι παραδοσιακές μέθοδοι τμηματοποίησης πελατών στηρίζονται κυρίως σε εμπειρικές ταξινομήσεις ή απλές στατιστικές μεθόδους. Κατηγοριοποιούν τους πελάτες σύμφωνα με κάποιο χαρακτηριστικό συμπεριφοράς ή ιδιότητα και δεν μπορούν να κάνουν πιο πολύπλοκες αναλύσεις. Επίσης καθώς η συσσώρευση των δεδομένων γίνεται ολοένα και μεγαλύτερη, οι παραδοσιακές μέθοδοι δεν μπορούν να ανταπεξέλθουν με το μέγεθος της πολυπλοκότητας.

Πολλές μεγάλες εταιρείες σήμερα κατέχουν τεράστιες βάσεις δεδομένων. Η εξόρυξη δεδομένων δίνει τη δυνατότητα στις επιχειρήσεις να εξάγουν όσο το δυνατόν περισσότερες και αξιοποιήσιμες πληροφορίες μέσα από τον τεράστιο αυτό όγκο δεδομένων. Πρόκειται για μια λύση σε ένα μεγάλο πρόβλημα που αντιμετωπίζουν πολλές επιχειρήσεις και είναι η υπεραφθονία των δεδομένων και μια σχετική έλλειψη κατάλληλου προσωπικού, τεχνολογίας και χρόνου, ώστε να μετατραπούν τα απλά δεδομένα σε ουσιαστικές και αξιοποιήσιμες πληροφορίες σχετικά με τους υπάρχοντες και μελλοντικούς πελάτες. Οι τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται σήμερα στην τμηματοποίηση των πελατών ανήκουν στην κατηγορία της συσταδοποίησης ή των αλγορίθμων των πλησιέστερων γειτόνων.

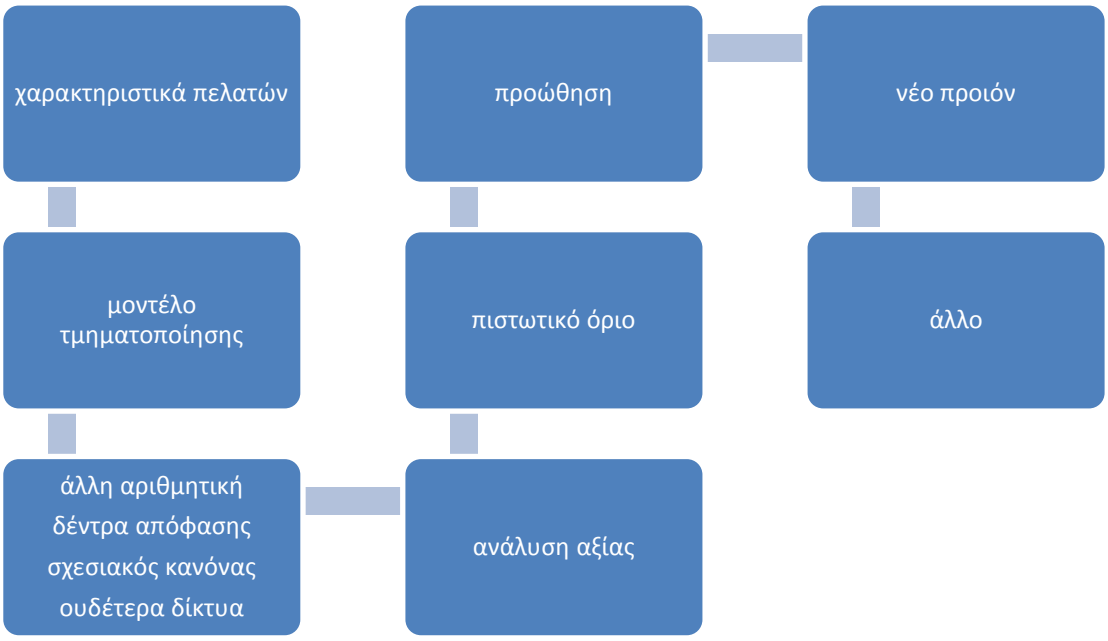
Η τμηματοποίηση πελατών στο άμεσο μάρκετινγκ έχει καταστεί ιδιαίτερα αποτελεσματική, λόγω της ανάπτυξης των εφαρμογών μάρκετινγκ στις βάσεις δεδομένων. Οι προσεγγίσεις της εξόρυξης δεδομένων παρέχουν αποδοτικούς τρόπους για το διαχωρισμό των πελατών σε τμήματα αλλά και την ανάπτυξη στρατηγικών μάρκετινγκ, προσαρμοσμένες στα συγκεκριμένα τμήματα ή άτομα. Οι τεχνικές μάρκετινγκ στις βάσεις δεδομένων έχουν από απλά μοντέλα που περιλαμβάνουν την πρόσφατη πείρα των αγορών του πελάτη, τη συχνότητα των αγορών του, καθώς και το ποσοστό των χρημάτων που έχει δαπανήσει με την επιχείρηση, σε στατιστικές τεχνικές όπως είναι η τεχνική ανίχνευσης περιοχών ενδιαφέροντος και το λογικό μοντέλο παλινδρόμησης.

Η εμφάνιση του στις εφαρμογές του μάρκετινγκ έχει στις βάσεις δεδομένων έχουν δημιουργηθεί και τα ευριζωνικά δίκτυα. Τα εργαλεία της εξόρυξης δεδομένων παρέχουν ολοκληρωμένη υποστήριξη στις διαδικασίες του μάνατζμεντ για την απόκτηση και διατήρηση πελατών, την αύξηση της αξίας που προσφέρουν στην επιχείρηση, την πελατειακή ικανοποίηση, καθώς επίσης και την προώθηση της αφοσίωσης των πελατών.

2.7.2. Μοντέλο τμηματοποίησης πελατών

Η δημιουργία μιας σχεδιασμένης σχέσης ανάμεσα στα αρχικά χαρακτηριστικά του πελάτη αποτελεί το βασικό βήμα για την πελατειακή τμηματοποίηση, κάνοντας χρήση της εξόρυξης δεδομένων. τα πελατειακά δεδομένα περιέχουν διανεμημένα και συνεχή χαρακτηριστικά. Θέτοντας το κάθε χαρακτηριστικό του πελάτη ως ένα σωματίδιο, όλοι οι πελάτες μαζί μιας επιχείρησης δημιουργούν ένα πολυδιάστατο χώρο, ο οποίος έχει οριστεί ως ο χαρακτηριστικός χώρος του πελάτη.

Η πελατειακή τμηματοποίηση, με χρήση τεχνικών εξόρυξης δεδομένων, είναι δυνατή με τη βοήθεια της λειτουργικής ανάλυσης. Η λειτουργική ανάλυση περιλαμβάνει την ανάλυση της αξίας του πελάτη, την ανάλυση της απόδοσης, την ανάλυση της προώθησης κλπ., βασισμένη στα ευρήματα από την σχεδιασμένη σχέση μεταξύ του πελάτη και των βασικών χαρακτηριστικών. Ακόμα περισσότερα ευρήματα θα ανακαλυφθούν με την ανάπτυξη των πρακτικών του μάνατζμεντ στο CRM. Τα νέα αυτά ευρήματα θα προστεθούν στις θεμελιώδης διαστάσεις ενώ η σχεδίαση σχέσεων με τα χαρακτηριστικά των πελατών θα αναδομηθεί στο παρακάτω σχήμα που εικονίζει το μοντέλο τμηματοποίησης πελατών με βάση την εξόρυξη δεδομένων.



Εικόνα : Μοντέλο τμηματοποίησης πελατών με βάση την εξόρυξη δεδομένων[6].

Σύμφωνα με το παραπάνω σχήμα η τμηματοποίηση γίνεται με τη βοήθεια τεχνικών εξόρυξης δεδομένων όπως οι κανόνες συσχετίσεων, τα νευρωνικά δίκτυα και τα δέντρα αποφάσεων. Το μοντέλο αυτό αρχικά τμηματοποιεί τους πελάτες σύμφωνα με την σχεδιασμένη σχέση και συνεχίζει τη διαδικασία με τα διάφορα είδη επιχειρησιακών εφαρμογών.

2.7.3. Πλεονεκτήματα της μεθόδου τμηματοποίησης με εργαλεία εξόρυξης δεδομένων

Οι παραδοσιακές μέθοδοι τμηματοποίησης πελατών κατατάσσουν τους πελάτες σε κατηγορίες σύμφωνα με τα απλά χαρακτηριστικά των πελατών ή των αγοραζόμενων προϊόντων, όπως τα τρόφιμα που αγοράζουμε από τα σουπερμάρκετ. Τα μοντέλα εξόρυξης δεδομένων πλεονεκτούν έναντι των παραδοσιακών μεθόδων τμηματοποίησης, προσφέροντας πολλά πλεονεκτήματα τα οποία είναι:

1) Βελτίωση των προωθητικών ενεργειών: το μοντέλο αυτό μπορεί να βοηθήσει την επιχείρηση να ακολουθήσει τις κατάλληλες προωθητικές στρατηγικές στον κατάλληλο χρόνο και με τα κατάλληλα προϊόντα και υπηρεσίες, με στόχο τους κατάλληλους πελάτες.
2) Ανάλυση της αξίας και της αφοσίωσης των πελατών: οι δυο αυτές μεταβλητές είναι πολύ σημαντικές γιατί επηρεάζουν τη στρατηγική της επιχείρησης. Οι επιχειρήσεις μέσω του μοντέλου αυτού μπορούν να διαχωρίσουν σε βαθμίδες τους πελάτες σύμφωνα με την αναμενόμενη αξία τους και την αφοσίωση τους στην εταιρεία.

3) Ανάλυση του πιστωτικού κινδύνου: η αξιολόγηση του κινδύνου είναι ένας αποτελεσματικός τρόπος για την αξιολόγηση ορισμένων τύπων κινδύνου από πελάτες όπως να μην έχουν αρκετά χρήματα ώστε να πληρώσουν αυτά που θέλουν να αγοράσουν.
4) Δημιουργία νέων προϊόντων έρευνας και ανάπτυξης: οι επιχειρήσεις μπορούν να ανακαλύψουν τις προτιμήσεις των πελατών τους μέσω της ανάλυσης πελατών στη βάση εξόρυξης δεδομένων και να σιγουρευτούν ότι θα υπάρξει συγκεκριμένη ζήτηση για το σχεδιασμένο προϊόν.
5) Η επιβεβαίωση της αγοράς- στόχου: η τμηματοποίηση πελατών με βάση την εξόρυξη δεδομένων είναι σε θέση να εντοπίσει ρητά την αγορά στην οποία απευθύνεται το προϊόν της επιχείρησης και να καταστεί σαφές στις στοχευόμενες πελατειακές ομάδες.

Κεφάλαιο 3ο

3.1 Ανάλυση και παραδείγματα για τον τρόπο λειτουργίας της εξόρυξης δεδομένων στα σουπερμάρκετ

Όπως λοιπόν αναφέραμε στα προηγούμενα κεφάλαια για να καταλάβουμε πως ακριβώς λειτουργεί η εξόρυξη δεδομένων στα σουπερμάρκετ θα δώσουμε παρακάτω κάποια παραδείγματα. Αρχικά το στήσιμο ενός σουπερμάρκετ θα έχει τη δική του λογική με βάση τον τρόπο με τον οποίο είναι τοποθετημένα τα διάφορα προϊόντα στα ράφια και στους διαδρόμους. Ο τρόπος με τον οποίο είναι διαμορφωμένοι οι διάδρομοι έχουν να κάνουν με τις επιλογές των πελατών. Τα μακαρόνια και οι σάλτσες βρίσκονται στον ίδιο διάδρομο, τα καθαριστικά, τα απορρυπαντικά και τα σαμπουάν όπως επίσης και τα αφρόλουτρα βρίσκονται το ένα κοντά στο άλλο στον ίδιο διάδρομο. Τα κρασιά , τα ποτά και τα αναψυκτικά βρίσκονται και αυτά στον ίδιο διάδρομο. Τα χαρτικά και τα παιχνίδια βρίσκονται επίσης σε άλλο διάδρομο. Με αυτό τον τρόπο ταξινόμησης τα προϊόντα έχουν πιο εύκολη πρόσβαση στον πελάτη και στις αγορές του. Έχει παρατηθεί πως τα αγαθά τα οποία συνδέονται μεταξύ τους οι καταστηματάρχες φροντίζουν να τα τοποθετούν το ένα κοντά στο άλλο. Για παράδειγμα, έχει παρατηρηθεί πως όταν κάποιος πελάτης αγοράζει ζάχαρη στο καλάθι του θα βάλει και καφέ, ακόμα όταν κάποιος αγοράζει ζαμπόν θα αγοράσει και κασέρι και ψωμάκια για τοστ. Έτσι λοιπόν όταν ο διευθυντής θελήσει να αυξήσει τις πωλήσεις κάποιων προϊόντων θα προσπαθήσει να κάνει μια αλλαγή στη θέση τους και συγκεκριμένα θα προσπαθήσει να τα τοποθετήσει κοντά σε προϊόντα τα οποία έχουν ζήτηση. Αφού λοιπόν κάναμε μια μικρή εισαγωγή για τον τρόπο με τον οποίο λειτουργεί η οργάνωση του σουπερμάρκετ ας περάσουμε στο πρακτικό μέρος ώστε να εξηγήσουμε όλα αυτά με παράδειγμα.

Το καλάθι του καταναλωτή στο σουπερμάρκετ χρησιμοποιείται σαν μέσω συλλογής πληροφοριών και αυτό μπορούμε να το πετύχουμε μέσω της διαδικασίας των κανόνων συσχέτισης. Οι κανόνες συσχέτισης είναι μια διαδικασία που επιτελείται από μια βάση δεδομένων και εκεί μέσα σε αυτή τη βάση εκτελείται η διαδικασία ανακάλυψης κανόνων και μπορεί να θεωρηθεί ως ένα σύνολο από εγγραφές που περιέχουν ένα σύνολο στοιχείων. Έχοντας λοιπόν υπόψη μας τις συναλλαγές που πραγματοποιούνται από τις ταμειακές μηχανές στο σουπερμάρκετ κάθε στοιχείο αποτελεί ένα προϊόν που αγοράστηκε, ενώ κάθε εγγραφή είναι η λίστα προϊόντων που αγοράστηκαν μια φορά. Έστω ότι το δικό μας σουπερμάρκετ εμπορεύεται πέντε μόνο προϊόντα: ψωμί, γάλα, ζάχαρη, βούτυρο και μπύρα. Άρα στο σουπερμάρκετ μπορούν να πραγματοποιηθούν 31 συνδυασμοί για συναλλαγές. Η υποστήριξη ενός στοιχείου ή ενός συνόλου στοιχείων είναι το ποσοστό των συναλλαγών που εμφανίζεται το στοιχείο αυτό.

Συναλλαγή	Στοιχεία
Σ1	Ψωμί, ζάχαρη, βούτυρο
Σ2	Ψωμί, βούτυρο
Σ3	Ψωμί, γάλα, βούτυρο
Σ4	Μπύρα, ψωμί
Σ5	Μπύρα, γάλα

Εικόνα 1: μας δείχνει τα στοιχεία [7]

Η υποστήριξη του στοιχείου {ψωμί} είναι 80%. Το ίδιο ποσοστό υποστήριξης αντιστοιχεί και στο σύνολο στοιχείων {ψωμί, βούτυρο}. Το στοιχείο {μπύρα} έχει ποσοστό υποστήριξης 40%. Με αυτόν τον τρόπο γίνεται η επιλογή των ποσοστών υποστήριξης.

Κανόνας	Υποστήριξη	Εμπιστοσύνη
Ψωμί → βούτυρο	60%	75%
Βούτυρο → ψωμί	60%	100%
Μπύρα → ψωμί	20%	50%
Βούτυρο → ζάχαρη	20%	33,3%
Ζάχαρη → βούτυρο	20%	100%
Ζάχαρη → γάλα	0%	0%

Εικόνα 2: μέτρα σημαντικότητας υποστήριξης- εμπιστοσύνης στοιχείων[7]

Ένας ορισμός για την έννοια κανόνας συσχέτισης θα μπορούσε να είναι ο εξής: με δεδομένο ένα σύνολο στοιχείων I={I1, I2,… Iμ}και μια βάση δεδομένων από συναλλαγές Σ={ σ1, σ2, σ3,…,σμ} όπου σi={ Ii1, Ii2, Ii3,…Iik }και Iij ανήκει στο I, ένας κανόνας συσχέτισης είναι ένα επαγωγικό συμπέρασμα της μορφής X→Y, όπου X και Y είναι σύνολα στοιχείων (στοιχειοσύνολα) και X U Y = ø.

Οι αλγόριθμοι κανόνων συσχέτισης προσπαθούν να ανακαλύψουν τις σημαντικές συσχετίσεις. Για να κρίνουν το πόσο σημαντικός είναι ένας κανόνας υπολογίζονται

δυο μέτρα, η υποστήριξη του κανόνα και η εμπιστοσύνη. Η υποστήριξη για έναν κανόνα Χ→ Υ είναι το ποσοστό των συναλλαγών που περιέχουν το Χ U Υ. Η εμπιστοσύνη του κανόνα, η οποία υποδεικνύει την ισχύ του, είναι το κλάσμα του πλήθους των συναλλαγών που περιέχουν το Χ U Υ προς το πλήθος των συναλλαγών που περιέχουν το Χ. Τα μέτρα αυτά, για κάποιους κανόνες που προκύπτουν από τις συναλλαγές της εικόνας 1 και παρουσιάζονται στην εικόνα 2. Για παράδειγμα, ο κανόνας βούτυρο → ψωμί έχει εμπιστοσύνη 100%. Αυτό σημαίνει ότι ο κανόνας είναι πολύ ισχυρός αφού κάθε πελάτης που αγοράζει βούτυρο αγοράζει και ψωμί. Επίσης, ο κανόνας αυτός έχει υποστήριξη 60% και αυτό σημαίνει ότι τα προϊόντα αυτά που συνθέτουν τον κανόνα, εμφανίζονται στο 60% των συναλλαγών. Η εξόρυξη τέτοιου είδους κανόνων μπορεί να βοηθήσει τη διεύθυνση του σουπερμάρκετ στην σωστή χωροθέτηση των προϊόντων, στον αποτελεσματικό σχεδιασμό μιας διαφημιστικής καμπάνιας κοκ.

Αλγόριθμος Παραγωγή_Κανόνων_Συσχέτισης
Είσοδος:
ΒΔ !Βάση συναλλαγών
Ι !Σύνολο στοιχείων
Σ !Σύνολο συχνών στοιχειοσυνόλων
ε !Εμπιστοσύνη
Έξοδος:
ΚΣ !Κανόνες Συσχέτισης
Αρχή
1. Βρες τα συχνά στοιχεία
2. *ΚΣ* ← Ø
3. Για κάθε στοιχειοσύνολο *σ* που ανήκει στο *Σ*
4. Για κάθε στοιχείο *χ* του *σ*
5. Αν (υποστήριξη(σ)/υποστήριξη(χ))≥ ε τότε
6. *ΚΣ*← *ΚΣ* + {*χ* →(*σ*-*χ*)}

Τέλος
Σχήμα 1: Αλγόριθμος παραγωγής κανόνων

Οι αλγόριθμοι ανακάλυψης συσχετίσεων επιλύουν Τα προβλήματα της εύρεσης των κανόνων διασπώντας τα σε δυο μέρη: α) εύρεση των συχνών στοιχειοσυνόλων, β) δημιουργία κανόνων από τα συχνά στοιχειοσύνολα. Ένα συχνό στοιχειοσύνολο είναι ένα σύνολο στοιχείων του οποίου το πλήθος των εμφανίσεων ξεπερνά ένα προκαθορισμένο κατώφλι. Η διαδικασία εύρεσης των συχνών στοιχειοσυνόλων θεωρείται αρκετά απλή αλλά πολύ δαπανηρή (για 5 στοιχεία έχουμε 31 πιθανά στοιχειοσύνολα, για 30 προϊόντα έχουμε 1073741823). Οι αλγόριθμοι κανόνων συσχέτισης διαφέρουν μεταξύ τους στον τρόπο που ανακαλύπτουν τα συχνά στοιχειοσύνολα. Όταν έχουν βρεθεί όλα τα στοιχειοσύνολα, η δημιουργία κανόνων είναι μια απλή διαδικασία και παρουσιάζεται στο Σχήμα 1.

Ορίζοντας ως κατώφλι υποστήριξης υ=30% και εμπιστοσύνης ε=50% και χρησιμοποιώντας τις συναλλαγές της εικόνας 2, λαμβάνουμε τα ακόλουθα συχνά

στοιχειοσύνολα που ικανοποιούν τα υ (δηλ., ο αριθμός των εμφανίσεων στη βάση δεδομένων είναι πάνω από 30%).

Σ={ μπύρα, 40%}, {ψωμί, 60%}, {γάλα, 40%}, {βούτυρο, 60%}, {ψωμί, βούτυρο,60%} εκτελώντας λοιπόν τον αλγόριθμο του παραπάνω Σχήματος 1, προκύπτουν δυο κανόνες ψωμί → βούτυρο και βούτυρο → ψωμί που ικανοποιούν το κατώφλι της εμπιστοσύνης ε. ο πρώτος έχει εμπιστοσύνη 75% ενώ ο δεύτερος έχει εμπιστοσύνη 100%. Δηλαδή (υποστήριξη {ψωμί, βούτυρο}/υποστήριξη {ψωμί}) =60/80 = 0,75 και (υποστήριξη {βούτυρο, ψωμί}/υποστήριξη{βούτυρο}) = 60/60 =1.

Βλέποντας λοιπόν πως λειτουργεί η αλληλουχία μεταξύ των προϊόντων ο διευθυντής της επιχείρησης του σουπερμάρκετ μπορεί να προσθέσει σιγά σιγά κάποια νέα προϊόντα στα ράφια. Στο παραπάνω παράδειγμα είδαμε να έχουμε αυξημένη ζήτηση στο ψωμί και το βούτυρο, επομένως το νέο προϊόν που ίσως θα ήθελε η επιχείρηση να πουλήσει να είναι η μαρμελάδα, η μερέντα ή το μέλι. Και αυτό γιατί μπορούν να χρησιμοποιηθούν μαζί στο πρωινό. Αν τώρα στο παράδειγμα μας είχε αυξημένη ζήτηση η μπύρα τότε θα ο διευθυντής του σουπερμάρκετ θα μπορούσε να προσθέσει σαν νέο προϊόν προς πώληση τα πατατάκια ή τους ξηρούς καρπούς που συνοδεύουν συνήθως τη μπύρα. Όλα τα ζευγάρια που αναφέραμε είναι προϊόντα που συνηθίζουμε να αγοράζουμε σε συνδυασμό μεταξύ τους στην καθημερινότητά μας. Αυτά τα προϊόντα μπορούν να τοποθετηθούν το ένα δίπλα στο άλλο, προκειμένου ο πελάτης να μην δυσκολευτεί να τα εντοπίσει. Αν τώρα ο διευθυντής θελήσει να δελεάσει τον πελάτη να αγοράσει και κάτι άλλο θα τοποθετήσει αυτά τα προϊόντα σε διαφορετικούς διαδρόμους και ράφια. Με τον τρόπο αυτό αναγκάζει τον καταναλωτή-πελάτη να ψάξει για τα προϊόντα που χρειάζεται και θέλει να αγοράσει και επειδή περνούν ανάμεσα από άλλα πολλά προϊόντα υπάρχει μεγάλη πιθανότητα να μπουν στον πειρασμό να πάρουν μαζί τους και προϊόντα τα οποία δεν συνοδεύονται με αυτά τα οποία αναφέραμε παραπάνω. Έτσι η επιχείρηση καταφέρνει να αυξήσει τις πωλήσεις της και μαζί και το κέρδος της. Με τον ίδιο τρόπο λειτουργούν όλοι οι πιθανοί συνδυασμοί που μπορούμε να κάνουμε στις αγορές μας.

Πέρα όμως από τις αλυσιδωτές αγορές που αναφέραμε σημαντικό ρόλο στην πώληση ενός προϊόντος επίσης παίζει και η διαφήμιση. Η συσκευασία, τα χρώματα που χρησιμοποιούνται σε αυτή και το σημείο το οποίο είναι τοποθετημένο το αγαθό προς πώληση συμβάλουν και αυτά με τη σειρά τους. Δεν είναι τυχαίο ότι τα αγαθά που είναι τοποθετημένα στα ράφια τα οποία βρίσκονται στο ύψος των ματιών μας έχουν και την ανάλογη ζήτηση σε σχέση με αυτά τα αγαθά τα οποία βρίσκονται στο τελευταίο ράφι κάτω χαμηλά.. Ακόμα και η σειρά με την οποία είναι τοποθετημένα παίζει και αυτό με τη σειρά του κάποιο ρόλο. Οι προσφορές και οι εκπτώσεις έχουν και αυτές ένα μέρος ευθύνης στην πώληση του προϊόντος. Όλα αυτά λοιπόν μια επιχείρηση τα μελετά καλά πριν διαθέσει στην αγορά κάποιο νέο προϊόν. Η εταιρεία –επιχείρηση για να καταφέρει να υλοποιήσει όλα τα παραπάνω στάδια προκειμένου

το αγαθό να είναι διαθέσιμο στην αγορά και κατά συνέπεια στον καταναλωτή θέλει να έχει και το ανάλογο κέρδος.

Κεφάλαιο 4ο

4.1 Πειραματική διαδικασία και επαλήθευση της θεωρίας

Ποιο αναλυτικά για να δώσουμε μια εικόνα των ελλήνων καταναλωτών και των σουπερμάρκετ ακολουθεί η παρακάτω περιγραφή.

Παρακολουθήσαμε τη συμπεριφορά των καταναλωτών στο σουπερμάρκετ και καταγράψαμε τις αγορές τους. Έχουμε σαν στοιχεία τις αποδείξεις από τις ταμειακές μηχανές αλλά και πίνακες με τα αποτελέσματα που θα μας βοηθήσουν να κατανοήσουμε τη συσχέτιση μεταξύ κάποιων προϊόντων αλλά και τη συχνότητα με την οποία αγοράζονται από τους καταναλωτές όπως έχουμε αναφέρει στα παραπάνω κεφάλαια.

Ακολουθούν πίνακες με την περιγραφή από τις αγορές των καταναλωτών.

Προϊόντα 1ου καλαθιού
Πενιρλί
Ρύζι
Σοκολάτα
Οινόπνευμα
Τροφή σκύλου
Βαφή μαλλιών
Σακούλα

Προϊόντα 2ου καλαθιού
τυρί
μπανάνες
Μήλα
Φυτά
Κρεμμύδια
Αγγούρια
Παγωτό
Σφολιάτα
Φιλαδέλφια
Ανθότυρο
Γάλα
Φέτα
Γιαούρτι

Σοκολάτες
Κοτόπουλο
Σακούλες

Προϊόντα 3ου καλαθιού
Σοκολάτα

Προϊόντα 4ου καλαθιού
Αλεύρι
Κρουασάν
Υγρό πιάτων
Απιονισμένο νερό

Προϊόντα 5ου καλαθιού
ξηροί καρποί

Προϊόντα 6ου καλαθιού
Κοτόπουλο
Πάριζα
Τυρί
Γάλα
Ζύμη
Μουστάρδα
Κριθαράκι
Κέτσαπ
Ρύζι

Προϊόντα 7ου καλαθιού
Γιαούρτι
Γάλα
Ανθότυρο
Ψωμί
Ψωμί για τοστ
Σοκολάτα
Κρεμοσάπουνο
Σακούλα

Προϊόντα 8ου καλαθιού
Τυροπιτάκια
Λουκανοπιτάκια
Τυρί

Πάριζα
Σαλάμι αέρος
Αλεύρι
Νερό
Σοκολάτα
Κρουασάν
Γαριδάκια
Μεμβράνη
Μαλακτικό

Προϊόντα 9ου καλαθιού
Μπανάνες
Αχλάδια
Κρεμμύδια
Μπουγάτσα
Γιαούρτι
Ξύδι
Μανιτάρια
Χυμός τομάτας
Φρυγανιές
Παστέλι
Ψωμί
Παξιμάδια
Σακούλα
Ψωμί για τοστ

Προϊόντα 10ου καλαθιού
Τυρί
Γιαούρτι
Ζάχαρη
Ρύζι
Ψωμί
Σακούλα

Προϊόντα 11ου καλαθιού
Μπισκότα
Τραυμαπλάστ
Αθλητικά παπούτσια
Μπατονέτες

Προϊόντα 12ου καλαθιού
Ντομάτες
Μήλα

Αγγούρια
Κρεμμύδια
Πατάτες
Χυμός τομάτας
Πιπεριές
Μπύρα
Σακούλα
Σερβιέτες
Χαρτοπετσέτες
Χαρτί υγείας
Υγρό για τα τζάμια

Προϊόντα 13ου καλαθιού
Γιαούρτι
Νερό
Ψωμί

Προϊόντα 14ου καλαθιού
Λάδι
Ρύζι
Αλάτι
Σοκολάτες
Παστέλι
Σακούλα
Υγρό για τα πιάτα
Χαρτοπετσέτες

Προϊόντα 15ου καλαθιού
Φυτά
Ελληνικός καφές
Φρυγανιές
Κέτσαπ
Σακούλα
Ταμπλέτες για τα άλατα
Σκόνη πλυντηρίου

Προϊόντα 16ου καλαθιού
Μπανάνες
Γιαούρτι
Γάλα
Πιπέρι
Μακαρόνια
Χυμός τομάτας

Γίγαντες
Πορτοκαλάδα
Νερό
Χαρτί υγείας
Οδοντόβουρτσα

Προϊόντα 17ου καλαθιού
Γάλα
Ψωμί
Κρουασάν
Σφουγγαρίστρα
Σακούλες

Προϊόντα 18ου καλαθιού
Γαλοπούλα
Γκούντα
Αλεύρι
Λάδι
Φρυγανιές
Τσίχλες
Ξηροί καρποί
Παστέλι
Χαρτί κουζίνας
Μαλακτικό
Σακούλα
Απορρυπαντικά
Οδοντόκρεμα

Προϊόντα 19ου καλαθιού
Νερό
Ζάχαρη
Αλάτι
Μανταλάκια
Γαλοπούλα
Γκούντα
Χυμός
Γάλα

Προϊόντα 20ου καλαθιού
Δημητριακά
Νερό
Ψωμί για τοστ
Κοτόπουλο

Παρακολουθήσαμε 20 καταναλωτές και για το παράδειγμά μας έχουμε επιλέξει 10 τυχαίους πελάτες. Θα χρησιμοποιήσουμε σύμβολα για τους πελάτες (π).

Πελάτες	Περιεχόμενο καλαθιού
Π1	Μπανάνες, αχλάδια, κρεμμύδια, μπουγάτσα, γιαούρτι, ξύδι, μανιτάρια, χυμός ντομάτας, ψωμί, φρυγανιές, παξιμάδια
Π2	Κασέρι, γιαούρτι, ζάχαρη, ρύζι, ψωμί
Π3	Γιαούρτι, ψωμί, νερό
Π4	Γαλοπούλα, γιαούρτι, γάλα, χυμό λεμόνι, μπισκότα, σοκολάτες, ψωμάκια για τοστ
Π5	Γάλα, κρουασάν, ψωμί, σακούλες κατάψυξης, σφουγγαρίστρα
Π6	Μπανάνες, γιαούρτι, γάλα, κέτσαπ, πιπέρι, μακαρόνια, χυμός ντομάτας, γίγαντες, πορτοκαλάδα, νερό
Π7	Ξηροί καρποί
Π8	Τυρί, μπανάνες, μήλα, φυτά, κρεμμύδια, αγγούρια, παγωτό, σφολιάτα, τυρί κρέμα, ανθότυρο, γάλα, γιαούρτι, σοκολάτες, κοτόπουλο
Π9	Κοτόπουλο, πάριζα, τυρί, μακαρόνια, μουστάρδα, κριθαράκι, κέτσαπ, ρύζι
Π10	Γιαούρτι, γάλα, ανθότυρο, ψωμί, για τοστ, σοκοφρέτες, ξηροί καρποί, κρεμοσάπουνο

Εικόνα : Πίνακας που μας περιγράφει το καλάθι του καταναλωτή

Παρατηρώντας τον παραπάνω πίνακα συμπεραίνουμε ότι:

Οι δέκα πελάτες έχουν αγοράσει κοινά προϊόντα. Αυτά είναι το γιαούρτι, το ψωμί, το γάλα και το νερό. Για να αποτυπώσουμε τη συχνότητα τους και τα ποσοστά θα χρησιμοποιήσουμε πίνακες στατιστικής.

Αρχικά θα κάνουμε έναν επιπλέον πίνακα όπου θα φαίνονται τα προϊόντα που είναι κοινά στους καταναλωτές.

Καταναλωτές	Κοινά προϊόντα
Π1	Ψωμί , γιαούρτι
Π2	Γιαούρτι, ψωμί
Π3	Γιαούρτι, ψωμί, νερό
Π4	Γιαούρτι, γάλα
Π5	Γάλα, ψωμί
Π6	Γιαούρτι, γάλα, νερό

Π7	Ξηροί καρποί
Π8	Γάλα, γιαούρτι, τυρί
Π9	Πάριζα, τυρί
Π10	Γιαούρτι, γάλα, ψωμί, ξηροί καρποί

Εικόνα : πίνακας που μας δείχνει τα αγαθά που είναι κοινά στο καλάθι των καταναλωτών

Από τους πίνακες παρατηρούμε ότι το αγαθό με την μεγαλύτερη συχνότητα και κατά συνέπεια ζήτηση είναι το γιαούρτι, ακολουθεί το γάλα και το ψωμί και έπειτα ακολουθούν και τα υπόλοιπα αγαθά. Το συμπέρασμα όμως αυτό το επαληθεύουμε και το κατανοούμε καλύτερα και από τον πίνακα των συχνοτήτων σχετικών και αθροιστικών που ακολουθεί:

Προϊόντα	xi	Ni	fi	Fi	fi%	Fi%
Γιαούρτι	7	7	0,304	0,304	30,4	30,4
Γάλα	5	12	0,217	0,521	21,7	52,1
Ψωμί	5	17	0,217	0,739	21,7	73,9
Νερό	2	19	0,087	0,826	8,7	82,6
Τυρί	1	20	0,044	0,869	4,4	86,9
Ξηροί καρποί	2	22	0,087	0,956	8,7	95,6
Πάριζα	1	23	0,044	1	4,4	100
Σύνολο (ν)	23		1		100	

Υπολογίζω την αθροιστική συχνότητα Ni:

Ni=x1+x2+x3+x4+x5+x6+x7=7+5+5+2+1+2+1=23

Υπολογίζω τις σχετικές συχνότητες των μεταβλητών (fi):

f1=x1/ν → f1=7/23 → f1=0,304

f2=x2/ν → f2=5/23 → f2=0,217

f3=x3/ν → f3=5/23 → f3=0,217

f4=x4/ν → f4=2/23 → f4=0,087

f5=x5/ν → f5=1/23 → f5=0,044

f6=x6/ν → f6=2/23 → f6=0,087

f7=x7/ν → f7=1/23 → f7=0,044

Υπολογίζω τις σχετικές αθροιστικές συχνότητες (Fi)

F1=N1/ν → F1=7/23 → F1=0,304

F2=N2/ν → F2=12/23 → F2=0,521

F3=N3/ν → F3=17/23 →F3 =0,738

F4=N4/ν → F4=19/23 → F4=0,826

F5=N5/ν → F5=20/23 → F5=0,869

F6=N6/ν → F6=22/23 → F6=0,956

F7=N7/ν → F7= 23/23 → F7=1

Υπολογίζω την σχετική συχνότητα επί τοις εκατό % (fi%)

f1%=f1*100 → f1%=30,4

f2%=f2*100 → f2%=21,7

f3%=f3*100 → f3%=21,7

f4%=f4*100 → f4%=8,7

f5%=f5*100 → f5%=4,4

f6%=f6*100 → f6%=8,7

f7%=f7*100 → f7%=4,4

Υπολογίζω σχετική αθροιστική συχνότητα επί τοις εκατό % (Fi%)

F1%=F1*100 → F1%=30,4

F2%=F2*100 → F2%=52,1

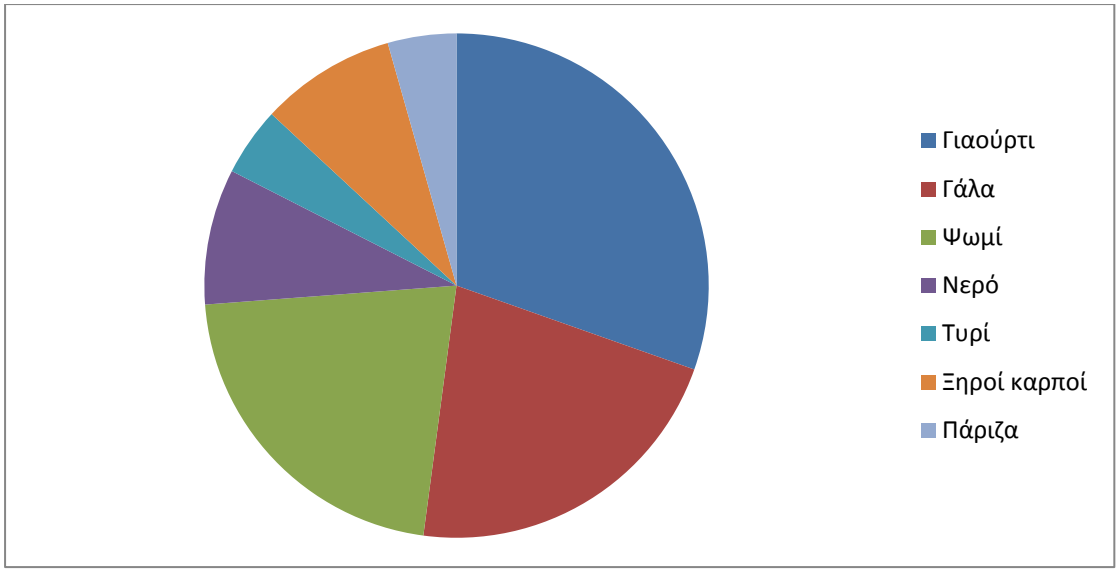
F3%=F3*100 → F3%=73,9

F4%=F4*100 → F4%=82,6

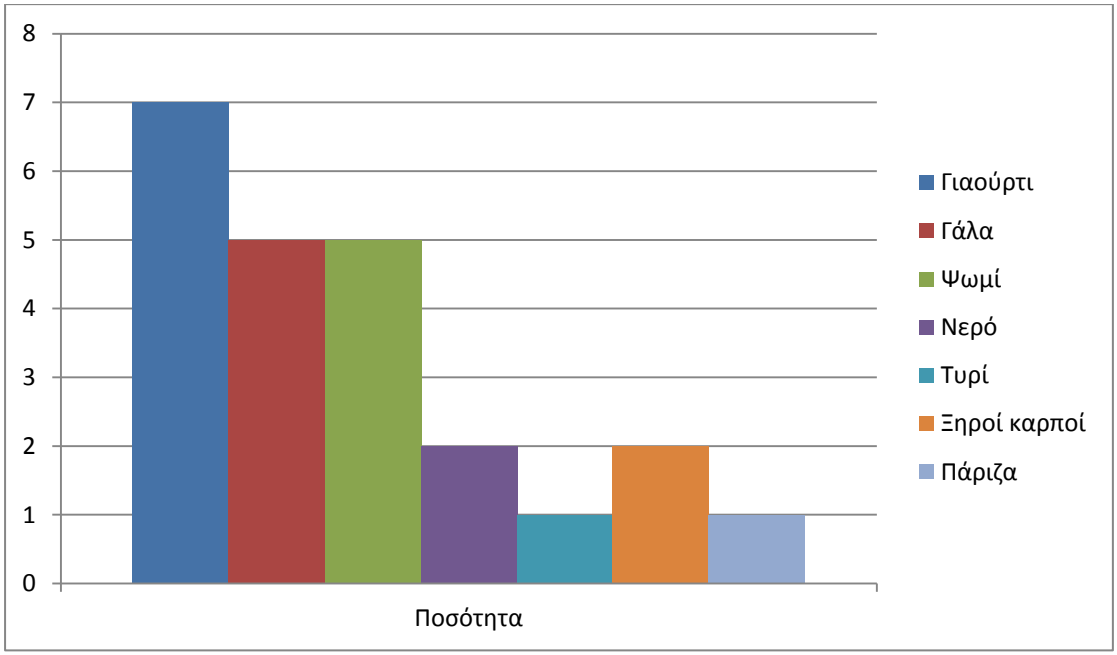
F5%=F5*100 → F5%=86,9

F6%=F6*100 → F6%=95,6

F7%=F7*100 → F7%=100



Εικόνα 1: Μας δείχνει πως απεικονίζεται γραφικά οι σχετικές συχνότητες fi% των προϊόντων του πίνακα στη σελίδα 53.



Εικόνα 2: Μας δείχνει την γραφική απεικόνιση των συχνοτήτων xi του πίνακα στην σελίδα 53

Συμπεράσματα

Με το πέρας της εργασίας αυτής βγάζουμε κάποια συμπεράσματα τα οποία είναι τα εξής:

Αρχικά η εξόρυξη δεδομένων χρησιμοποιείται με διάφορους τρόπους στην καθημερινότητά μας. Άλλοτε βγάζοντας συμπεράσματα από μεγάλες βάσεις δεδομένων και άλλοτε από απλές. Στην παρούσα εργασία η βάση δεδομένων με την οποία έχει να κάνει η εργασία είναι σχετική με κάποιες καταναλωτικές συνήθειες πελατών που με τη σειρά τους διαμορφώνουν το στήσιμο των προϊόντων στο σουπερμάρκετ.

Ανάλογα λοιπόν με τον όγκο της πληροφορίας που έχουμε να επεξεργαστούμε έχουμε και το ανάλογο κόστος. Αν οι πληροφορίες χρησιμοποιούνται με σωστό τρόπο μπορούν να βοηθήσουν στη βελτίωση της τεχνολογίας, του τρόπου οργάνωσης και λειτουργίας κάποιων επιχειρήσεων αλλά και στον τρόπο αντιμετώπισης των όποιων προβλημάτων που μπορεί να προκύψουν.

Ακόμα μπορούμε να βγάλουμε συμπεράσματα και να καταγράψουμε τις καταναλωτικές συνήθειες των πελατών έτσι ώστε να δημιουργήσουμε καινοτόμα προϊόντα ή ακόμα και προσφορές στις τιμές κάποιων προϊόντων που θα μπορούσαν να αυξήσουν το κέρδος της επιχείρησης αλλά και να φέρουν σε αυτή νέους πελάτες. Τέλος μπορούμε να καταλάβουμε πως κάθε τι που χρησιμοποιούμε στην καθημερινότητά μας συνδέεται στενά με την εξόρυξη δεδομένων χωρίς εμείς να το καταλαβαίνουμε.

Αναφορές – βιβλιογραφία

1) http//el.Wikipedia.org/wiki/εξόρυξη δεδομένων

2) Εγκυκλοπαίδεια πληροφορικής και τεχνολογίας υπολογιστών. Εκδόσεις νέων τεχνολογιών

3) http//digilib.unip.gr/dspace/bitstream/unipi/48981/athanasopoulou

4) Data Mining, εισαγωγικά και προηγμένα θέματα εξόρυξης γνώσης από δεδομένα, Margaret H Dunham,επιμέλεια ελληνικής έκδοσης: Βασίλης Βερύκιος & Γιάννης Θεοδωρίδης, εκδόσεις νέων τεχνολογιών

5) http//www.diapluw.org/index.php.?id=19

6) http//dspace.lib.uom.gr/bitstream/259/140816/goulouzoiMc2012.pdf

7) Πανεπιστήμιο Πειραιά, τμήμα στατιστικής και ασφαλιστικής, πτυχιακό πρόγραμμα σπουδών στην εφαρμοσμένη στατιστική, εξόρυξη δεδομένων (data mining) και κατηγορικά δεδομένα, Γεράσιμος Ε. Σταυλιώτης (Ιούνιος 2008)

8) Ανακάλυψη κανόνων συσχέτισης από εκπαιδευτικά δεδομένα, σημειώσεις κ. Στέφανος Ουγγιάρογλου

9) Fayyad, UM, Piatesky- Shapiro, G, Smyth, P. and uthurusamy, R.(1996-a), Advances in knowledge Discovery and Data mining , AAAI Press

10) www.youtube.com/watch?v=RS796QY708E