ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΗΠΕΙΡΟΥ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΤΕΧΝΟΛΟΓΙΚΟ
ΕΚΠΑΙΔΕΥΤΙΚΟ
ΙΔΡΥΜΑ
ΤΕΙ ΗΠΕΙΡΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας: Εξεύρεση συσχετίσεων με χρήση μεθόδων εξόρυξης δεδομένων σε δεδομένων καλαθιού αγορών

Παπαδόπουλος Νικόλαος - Α.Μ. 7614

Επιβλέπων καθηγητής
Νικόλαος Γιαννακέας

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας: Εξεύρεση συσχετίσεων με χρήση μεθόδων εξόρυξης γνώσης σε δεδομένα καλαθιού αγορών

Παπαδόπουλος Νικόλαος - Α.Μ. 7614

Επιβλέπων Καθηγητής
Νικόλαος Γιαννακέας

- Άρτα 2015 -

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θελα να ευχαριστήσω θερμά τον καθηγητή κο. Γιαννακέα Νικόλαο κυρίως για την εμπιστοσύνη που μου έδειξε, και την υπομονή που έκανε κατά τη διάρκεια υλοποίησης της πτυχιακής εργασίας. Όπως επίσης και για την πολύτιμη βοήθεια και καθοδήγηση του, για την επίλυση διάφορων θεμάτων.

Θα ήθελα επίσης να απευθύνω τις ευχαριστίες μου στους γονείς μου, οι οποίοι στήριξαν τις σπουδές μου με διάφορους τρόπους, φροντίζοντας για την καλύτερη δυνατή μόρφωση μου.

Τέλος ευχαριστώ το φιλαράκι μου από το στρατό που με βοήθησε σε μεγάλο βαθμό στην υλοποίηση αυτής της εργασίας.

Περίληψη

Στα πλαίσια της συγκεκριμένης πτυχιακής εργασίας, μελετήθηκε η περιοχή της εξόρυξης γνώσης από δεδομένα, τόσο ως προς το θεωρητικό κομμάτι όσο και ως προς το πρακτικό, εστιάζοντας κυρίως στην περίπτωση της εξόρυξης από δεδομένα καλαθιού. Οι μέθοδοι που χρησιμοποιήθηκαν για αυτό το σκοπό είναι τα δένδρα απόφασης και ο αλγόριθμος apriorigια την εξαγωγή των κανόνων συσχέτισης. Τα δεδομένα που χρησιμοποιήθηκαν συλλέχθηκαν σε διάστημα ενός μηνός από ένα κοινό παντοπωλείο γειτονίας, έχοντας χωρίσει το σύνολο των προϊόντων σε οκτώ διαφορετικές γενικότερες κατηγορίες. Τα ποσοστά που προέκυψαν δεν αποδεικνύουν κάποια πολύ ισχυρή συσχέτιση των καταναλωτών συνηθειών, αλλά πιθανότατα αυτό οφείλεται στο μικρό όγκο δεδομένων που είχαμε στη διάθεση μας.

Abstract

In this work, the scientific domain of data mining through data was studied in depth. The two aspects that we examined was both the theoretical and the practical, and we mainly focused on the market basket task of data mining. For this purpose, the methods that were used are decision trees and apriori algorithm for extracting association rules. The data that we used, were collected during a period of 30 days in a common grocery shop and were classified in eight different categories. The results did not show any strong correlation of consumer habits, but the most probable reason why this happened is the small size of the available collected data.

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Ο σκοπός της παρούσας εργασίας είναι η μελέτη του αντικειμένου της εξόρυξης γνώσης και πιο συγκεκριμένα η εφαρμογή αυτής σε δεδομένα καλαθιού, δηλαδή σε δεδομένα που καταγράφονται είτε από ηλεκτρονικά μαγαζιά είτε από τα συνήθη καταστήματα με σκοπό την καταγραφή, τη μελέτη και την επεξεργασία των καταναλωτικών συνηθειών των πελατών του εκάστοτε καταστήματος.

Οι μέθοδοι που χρησιμοποιήθηκαν ανήκουν στο γενικότερο πλαίσιο της μηχανικής μάθησης. Πιο συγκεκριμένα, χρησιμοποιήθηκε το περιβάλλον του WEKA, στο οποίο έχουν υλοποιηθεί οι περισσότεροι και οι ποιο δημοφιλείς αλγόριθμοι που χρησιμοποιούνται σε αυτό το επιστημονικό πεδίο. Όσον αφορά το κομμάτι της ταξινόμησης, χρησιμοποιήθηκε ο ταξινομητής C4.5 που ανήκει στα δένδρα απόφασης ενώ για την εξαγωγή των κανόνων συσχέτισης χρησιμοποιήθηκε ο apriori αλγόριθμος. Επίσης χρησιμοποιήθηκε το προγραμματιστικό πακέτο της Rγια την εξαγωγή του πίνακα συσχέτισης καθώς και γράφθηκαν ορισμένα scriptsστη Pythonπροκειμένου να πραγματοποιηθούν αποδοτικά και ευέλικτα οι διάφορες παραμετροποιήσεις στο δικό μας σύνολο δεδομένων για να μπορεί να εισαχθεί στο περιβάλλον του WEKAμε ορθό τρόπο.

Στο παρόν σημείο, κρίνεται σκόπιμο να δοθεί μία σύντομη περιγραφή του κάθε κεφαλαίου που περιέχεται στην εργασία. Συνεπώς, στο πρώτο κεφάλαιο πραγματοποιείται μία ιστορική αναδρομή του πεδίου της εξόρυξης γνώσης καθώς και αναφέρονται χρήσιμοι ορισμοί και λεπτομέρειες σχετικά με το συγκεκριμένο πεδίο, όπως οι κυριότερες εφαρμογές που βρίσκει αυτό στη σύγχρονη ζωή καθώς και οι σημαντικότερες διαδικασίες που υλοποιούνται στα πλαίσια αυτού. Στη συνέχεια, ακολουθεί το δεύτερο κεφάλαιο όπου αναφέρονται περισσότερες λεπτομέρειες για την εξόρυξη γνώσης δεδομένων καλαθιού και γίνεται αναφορά σε παλιότερες αντίστοιχες εργασίες. Στο τρίτο κεφάλαιο, περιγράφονται εκτενώς οι αλγόριθμοι που χρησιμοποιήθηκαν για την ταξινόμηση και την εξαγωγή των κανόνων συσχέτισης που διέπουν το πρωτότυπο σύνολο δεδομένων που συλλέχθηκε από ένα κοινό παντοπωλείο. Επίσης, σε αυτό το κομμάτι περιγράφονται και κάποιοι στατιστικοί δείκτες που είναι αρκετά χρήσιμοι για την κατανόηση της επίδοσης και της συμπεριφοράς των αλγορίθμων που χρησιμοποιήθηκαν, καθώς και για την ορθή

ανάγνωση τω αποτελεσμάτων. Στο τέταρτο κεφάλαιο εναποτίθενται τα αποτελέσματα που προέκυψαν από τη συγκεκριμένη έρευνα. Περιλαμβάνονται πίνακες με συνοπτικά αποτελέσματα και κάποια ενδεικτικά δένδρα απόφασης που αποσαφηνίζουν τις συσχετίσεις μεταξύ των δεδομένων μας. Τέλος, στο πέμπτο κεφάλαιο έχουμε τα συμπεράσματα που προέκυψαν και ένα σύντομο σχολιασμό αυτών.

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή στη Μηχανική μάθηση και Εξόρυξη Δεδομένων

1.1 Ιστορική αναδρομή

Το επιστημονικό πεδίο της Εξόρυξης Γνώσης (DataMining) είναι άμεσα συνυφασμένο με τα αντικείμενο της Μηχανικής Μάθησης (MachineLearning) και της Επιστήμης Των Υπολογιστών (ComputerScience).Τα τελευταία χρόνια έχει μάλιστα παρατηρηθεί μία δραματική στροφή προς αυτόν τον χώρο, τόσο από τη σκοπιά της εκπαίδευσης - που αφορά τη διδασκαλία του αντικειμένου αυτού σε προπτυχιακούς και μεταπτυχιακούς φοιτητές, αλλά και σε ερευνητικό επίπεδο με δεκάδες συνέδρια και περιοδικά να διεξάγονται και να δημοσιεύονται αντίστοιχα σε ετήσια βάση – όσο και στο εμπορικό και άκρως πιο πρακτικό κομμάτι. Χαρακτηριστικά παραδείγματα είναι οι δεκάδες ιστότοποι που προφέρουν onlineυπηρεσίες στους υποψήφιους πελάτες τους με σκοπό τη βελτίωση των παρεχόμενων υπηρεσιών των τελευταίων ή την πρόβλεψη πιθανών καταστάσεων, όπως σε συστήματα μελέτης καιρικών φαινομένων ή ακόμη και πρόβλεψη μελλοντικών οικονομικών καταστάσεων, όπως η πτώχευση, σε οποιουδήποτε είδους επιχειρήσεις. Παρά το γεγονός λοιπόν πως τα τελευταία χρόνια, η εξόρυξη γνώσης αποτελεί ένα ιδιαίτερα ενδιαφέρον αντικείμενο και έχει καταφέρει να τραβήξει το ενδιαφέρον πολλών επιστημόνων και επιχειρηματιών, η πρώτη ουσιαστική εμφάνιση αυτού του τομέα, χρονολογείται περίπου στο 1700 και στηρίχθηκε αρχικά στο θεώρημα των πιθανοτήτων του Bayes. Στη συνέχεια, τα επόμενα αξιοσημείωτα επιστημονικά επιτεύγματα που αξιοποιούνται πλέον από τις εφαρμογές εξόρυξης γνώσης από δεδομένα, ήταν η τεχνική της Παλινδρόμησης (Regression)(1920), τα Νευρωνικά Δίκτυα (NeuralNetworks) και ο ταξινομητής του πλησιέστερου ή γενικότερα των K πλησιέστερων γειτόνων (KNearestNeighbors – KNNclassifier) που έκαναν την εμφάνιση τους στα μέσα του 20^{ου} αιώνα. Τα επόμενα χρόνια και πιο συγκεκριμένα στις αρχές του 1960 ξεκινάει και πρακτικά η άνθηση της Μηχανικής Μάθησης, όπου

εξαιτίας της τεράστιας αύξησης του μεγέθους των δεδομένων, η συνάφεια των δύο αυτών πεδίων με τα αποτελέσματα της επιστήμης των υπολογιστών και τις νέες αλγοριθμικές ιδέες και τακτικές που εμφανίζονται εκείνη την εποχή μπορούμε να πούμε πως έφτασε στο αποκορύφωμα της. Ως αποτέλεσμα, οδηγηθήκαμε στην εμφάνιση εξαιρετικά χρήσιμων εργαλείων και τεχνικών που αποτελούν ακόμη και σήμερα σε πολλά προβλήματα οι αποδοτικότερες προσεγγίσεις με σκοπό την εξόρυξη γνώσης. ΟιπιοσημαντικέςαπόαυτέςήτανταΔένδραΑποφάσεων (DecisionTrees), η Συσταδοποίηση (Clustering) και οι Γενετικοί Αλγόριθμοι (GeneticAlgorithms) έως τέλη του 1960, καθώς και οι Μηχανές Διανυσματικής Στήριξης(SupportVectorMachines)και οι Αποθήκες Δεδομένων (DataWarehouses) γύρω στο 1990 [1].

Παρατηρώντας κανείς την προηγούμενη αναδρομή παρατηρεί πως ουσιαστικά το πεδίο της Εξόρυξης Γνώσης χρησιμοποιεί και επηρεάζεται πρακτικά από μία σειρά από άλλα επιστημονικά πεδία, ευρύτερα και προγενέστερα από το ίδιο. Με μία αυστηρή παρατήρηση αυτών, θα μπορούσαμε να κατονομάσουμε τα εξής:

- Μηχανική Μάθηση
- Στατιστική
- Αλγόριθμοι
- Βάσεις Δεδομένων
- Τεχνητή Νοημοσύνη
- Ανάκτηση Πληροφοριών

Η συσχέτιση μάλιστα των παραπάνω πεδίων συνηθίζεται να αποδίδεται με κάποιο blockδιάγραμμα, ανάλογο της Εικόνας 1 που ακολουθεί. Σημαντικό επίσης ιστορικό στοιχείο το οποίο τονίζει τη συνεργασία και την ενοποίηση των παραπάνω επιστημονικών πεδίων είναι το γεγονός πως στις αρχές του 1960, γνωστοί στατιστικολόγοι χρησιμοποιούσαν τον όρο “DataFishing” ή και τον “DataDredging” για να υπογραμμίσουν τις λανθασμένες προεκτάσεις που μπορεί να αποκτήσει το πεδίο της Εξόρυξης Γνώσης αν κανείς προσπαθήσει να αναλύσει δεδομένα χωρίς να έχει ορίσει μία εκ των προτέρων γνωστή υπόθεση. Για ένα σύντομο μάλιστα χρονικό διάστημα, χρησιμοποιήθηκε και ο όρος “DatabaseMining” μεταξύ άλλων, όπως οι

“DataArchaeology”, “InformationHarvesting”, “InformationDiscovery” και “KnowledgeExtraction”.Αργότερα, καθιερώθηκε ο ευρύτατα χρησιμοποιούμενα πλέον όρος “DataMining” από τους ειδικούς του χώρου των βάσεων δεδομένων. Τέλος, αξίζει να αναφερθεί πως ο GregoryPiatetsky-Shapiro επινόησε τον όρο «Ανακάλυψη γνώσης σε βάσεις δεδομένων» (“KnowledgeDiscoveryinDatabases”) για την πρώτη ημερίδα αφιερωμένη σε αυτό το αντικείμενο και κατέληξε να καθιερωθεί πλέον στους περισσότερου επιστημονικούς κλάδους με τα αρχικά KDD, σε αντίθεση με τη βιομηχανία όπου το “DataMining” φαίνεται ακόμη και σήμερα πιο οικείο [2].

Εικόνα 1: Επικοινωνία της Εξόρυξης Γνώσης με τα υπόλοιπα επιστημονικά πεδία.

1.2 Εισαγωγή στην Εξόρυξη Γνώσης

Έχοντας αναφερθεί στην ιστορική αναδρομή σχετικά με τη εξόρυξη γνώσης και την επιρροή που έχει δεχθεί αυτός ο κλάδος από άλλους συναφείς επιστημονικούς κλάδους, κρίνεται κατάλληλο να δοθεί ένας κατατοπιστικός ορισμός για αυτό το πεδίο ευθύς αμέσως:

*Ορισμός:*η εξόρυξη γνώσης αποτελεί την εξερεύνηση και την ανάλυση μεγάλου όγκου δεδομένων με σκοπό την ανακάλυψη ουσιωδών και μη τετριμμένων προτύπων ή/και κανόνων που να διέπουν είτε το σύνολο των δεδομένων είτε κάποια επιμέρους υποσύνολα αυτού, τα οποία συνδέονται με κάποια ουσιαστική σχέση μεταξύ τους.

Ο παραπάνω ορισμός συγκεντρώνει τα περισσότερα στοιχεία που αναφέρονται στους διάφορους ορισμούς που δίνουν οι περισσότερες βιβλιογραφικές πηγές [1], [2], [3], [4], [6]. Εξαιτίας της γενικής ερμηνείας που μπορεί να αποδοθεί στον παραπάνω ή σε όμοιους με αυτόν ορισμό από άτομα που προέρχονται από διαφορετικό κάθε φορά χώρο, δημιουργούνται κατά καιρούς συγχύσεις για το αν τελικά μια διαδικασία ή μία εφαρμογή άπτεται του αντικειμένου της εξόρυξης γνώσης. Πολύ συνοπτικά μπορούμε να αναφέρουμε πως ζητήματα που αφορούν την επεξεργασία ερωτημάτων ή άλλες αφαιρετικές διεργασίες, καθώς και οι περιπτώσεις των έμπειρων συστημάτων και των στατιστικών προγραμμάτων μικρής κλίμακας δεν αποτελούν αντικείμενο έρευνας της εξόρυξης γνώσης και δεν συνδέονται άμεσα το συγκεκριμένο επιστημονικό κλάδο.

Αντιθέτως, μιλώντας γενικά για το αντικείμενο της εξόρυξη γνώσης, θα μπορούσε να ειπωθεί πως κάνει τηνεμφάνιση της με δύο διαφορετικούς και αρκετά γενικούς τρόπους: την κατευθυνόμενη (directed) και τη μη-κατευθυνόμενη (undirected) μορφή της [6]. Πιο συγκεκριμένα, στην πρώτη μορφή ανήκουν οι προσπάθειες να ερμηνευθεί ή και να κατηγοριοποιηθεί κάποιο συγκεκριμένο χαρακτηριστικό των δεδομένων, όπως το εισόδημα των πολιτών ή το είδος μουσικής. Από την άλλη πλευρά, η μη-κατευθυντική μορφή, αναζητά και εντείνει τις προσπάθειες της στην εύρεση προτύπων ή ομοιοτήτων μεταξύ υποσυνόλων των συνολικών εγγραφών που δίνονται σαν είσοδο χωρίς τη χρήση κάποιου αυστηρώς ορισμένου κριτηρίου ή χρησιμοποιώντας προκαθορισμένες εντολές.

Όσον αφορά τη βασική λειτουργία που εκτελείται κατά τη διαδικασία των αλγορίθμων της εξόρυξη γνώσης, μπορεί να διακριθεί σε τρία στάδια [1]:

1. Εκτίμηση ενός μοντέλου: απαιτείται το βέλτιστο ταίριασμα του μοντέλου στα διαθέσιμα δεδομένα,
2. Προτίμηση: ορισμός κριτηρίων με βάση τα οποία επιλέγεται κάποιο μοντέλο έναντι των υπολοίπων και
3. Αναζήτηση: περιγραφή της τεχνικής με βάση την οποία θα εκτελεστεί η αναζήτηση στα διαθέσιμα δεδομένα.

Από την παραπάνω περιγραφή διαπιστώνει κανείς πως η εξόρυξη γνώσης είναι άμεσα συσχετισμένη με τη δημιουργία μοντέλων. Η έννοια της λέξης μοντέλο αφορά στην παρούσα περίπτωση απλώς ένα αλγόριθμο ή μία ομάδα κανόνων, τα οποία συνδέουν ένα πλήθος εισόδων με κάποια συγκεκριμένη έξοδο, χρησιμοποιώντας τις έννοιες που περιγράφουν μία κοινή διαδικασία στο πεδίο των βάσεων δεδομένων. Τεχνικές όπως η παλινδρόμηση, τα νευρωνικά δίκτυα, τα δένδρα απόφασης, οι ensembleταξινομητές και άλλες εξίσου σημαντικές τεχνικές που άπτονται στην εξόρυξη γνώσης, αποτελούν βασικό και αναπόσπαστο στοιχείο της δημιουργίας μοντέλων. Κάτω από τις κατάλληλες συνθήκες άλλωστε, ένα υποψήφιο μοντέλο μπορεί να ευνοήσει την αντίληψη και την κατανόηση του τρόπου διάδρασης και επιρροής των εκάστοτε χαρακτηριστικών που περιλαμβάνει η προς εξέταση συλλογή δεδομένων, είτε για την πρόβλεψη κάποιων μελλοντικών γεγονότων είτε για την ουσιαστική επέμβαση στα διάφορα κρίσιμα χαρακτηριστικά των δεδομένων που περιγράφουν κάποια συμπεριφορά, με σκοπό τη βελτίωση ή την αύξηση κάποιων δεικτών προς όφελος της επιχείρησης ή του φυσικού προσώπου που διαχειρίζεται αυτά. Χαρακτηριστικό παράδειγμα αποτελούν οι τράπεζες, όπου συλλέγοντας τα δεδομένα από τους πελάτες τους αλλά και από πελάτες άλλων τραπεζών, έχουν στη διάθεση τους ένα αντιπροσωπευτικό μοντέλο που περιγράφει τη συμπεριφορά του καλού/ουδέτερου/κακού πελάτη και μπορούν ανά πάσα στιγμή να αποκτήσουν μία πρόβλεψη για τη πιστωτική ικανότητα ή συμπεριφορά ενός νέου πελάτη ή ενός υπάρχοντος πελάτη για μελλοντική συναλλαγή ή συνεργασία με αυτές.

Μια επιπλέον χρησιμότητα που προκύπτει από τη δημιουργία μοντέλων, είναι ο υπολογισμός των λεγόμενων scores [6]. Με αυτόν τον τρόπο εκφράζουμε

απλουστευμένα το ταίριασμα ή τον υπολογισμό κάποιας συμπεριφοράς μέσω των δεδομένων, χρησιμοποιώντας απλώς έναν αριθμό. Δηλαδή, έχοντας υιοθετήσει μία αριθμητική συνάρτηση που επεξεργάζεται τα δεδομένα εισόδου, θα μπορούσαμε να πάρουμε για κάθε μία διαφορετική καταγραφή που περιλαμβάνεται στα δεδομένα μας, μία τιμή με βάση την οποία δύναται πλέον να ταξινομηθούν οι διάφορες οντότητες με έναν απλό και εύκολα αντιληπτό τρόπο από τον οποιονδήποτε. Ανάλογα λοιπόν με την εφαρμογή που βρισκόμαστε κάθε φορά και με βάση τη συνάρτηση και το κριτήριο που αυτή εξετάζει, θα μπορούσαμε να κατατάξουμε τους πελάτες ενός καταστήματος από τον πιο ευκολόπιστο στο λιγότερο δυνατό ή ακόμη και να υπολογίσουμε την πιθανότητα ανταπόκρισης των πελατών σε νέες διαφημιστικές καμπάνιες κ.ο.κ

1.3 Βασικές εργασίες εξόρυξης γνώσης

Σε αυτήν την παράγραφο θα αναφερθούν οι σημαντικότερες εργασίες που δύναται να διεξαχθούν από εφαρμογές της εξόρυξης γνώσης. Για λόγους σχετικότητας με τον αντικειμενικό σκοπό της παρούσας εργασίας, θα γίνει μία ονομαστική καταγραφή των δημοφιλέστερων αυτών λειτουργιών και θα αναλυθούν συντόμως μόνο όσες εξ αυτών χρησιμοποιήθηκαν και αποτέλεσαν αντικείμενο της εργασίας.

Έχουμε λοιπόν τις παρακάτω εργασίες, οι οποίες μπορούν χωρίς βλάβη της γενικότητας να αντιμετωπίσουν το σύνολο των προβλημάτων που θέτουν οι σύγχρονες επιχειρήσεις, βιομηχανίες, οικονομικοί και εκπαιδευτικοί παράγοντες:

- Κατηγοριοποίηση
- Πρόβλεψη
- Συσταδοποίηση
- Περιγραφή και σκιαγράφηση συμπεριφορών
- Παλινδρόμηση
- Ανάλυση χρονοσειρών
- Κανόνες συσχέτισης
- Παρουσίαση συνόψεων

- Ανακάλυψη ακολουθιών

Παρακάτω θα πραγματοποιηθεί μία μικρή περιγραφή της κατηγοριοποίησης και των κανόνων συσχέτισης. Για περισσότερες πληροφορίες σχετικά με τις υπόλοιπες εργασίες μπορεί κανείς να ανατρέξει στις πηγές [1], [3].

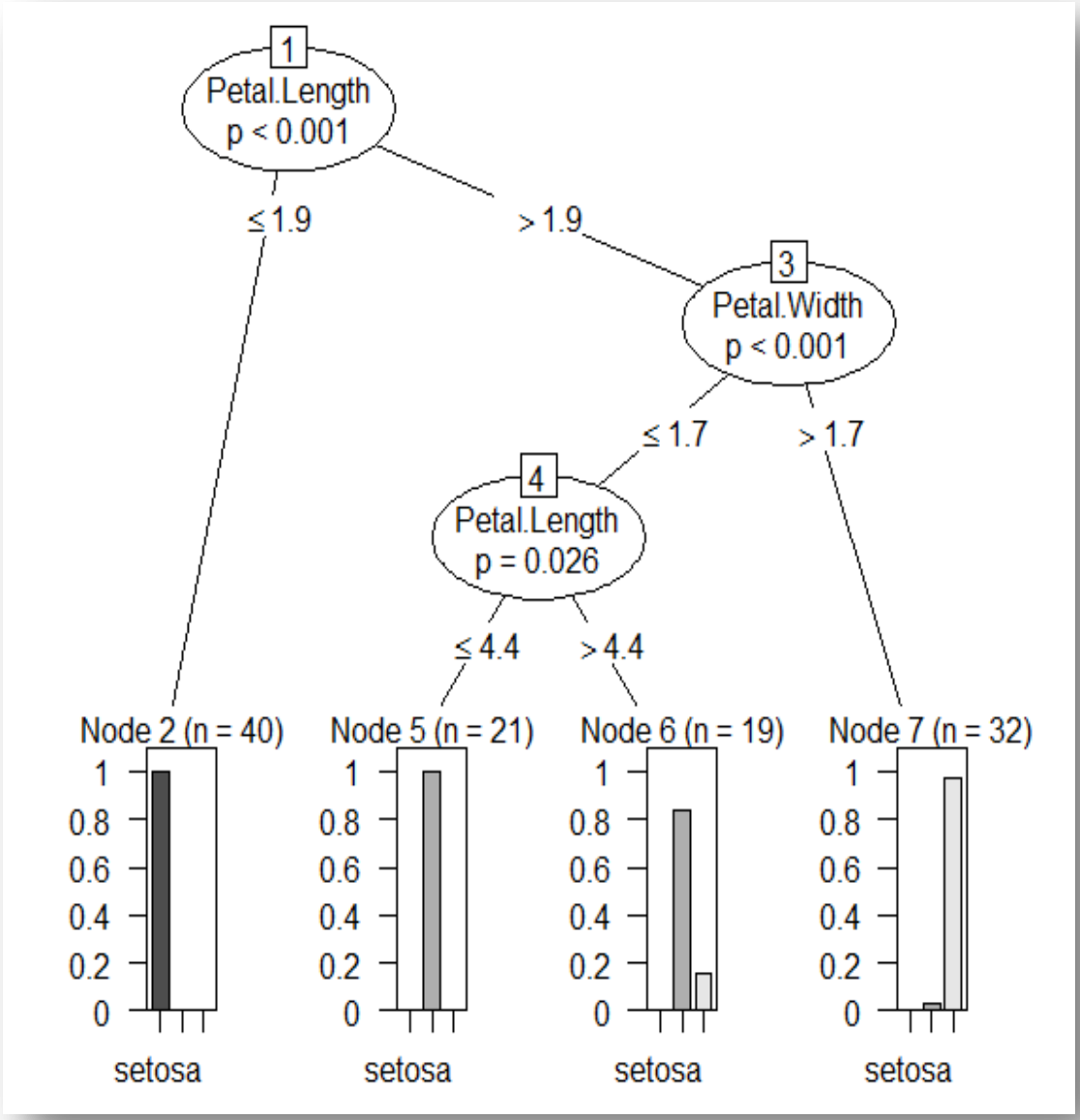
Όσον αφορά το ζήτημα της κατηγοριοποίησης (classification), αυτό αποτελεί μία από τις πιο κοινές λειτουργίες της εξόρυξη γνώσης και μάλιστα φαίνεται να ικανοποιεί μία επιτακτική ανάγκη της ανθρώπινης φύσης. Σκοπός δηλαδή είναι ο διαχωρισμός των εξεταζόμενων αντικειμένων σε κατηγορίες, όπως ακριβώς και ο άνθρωπος κατηγοριοποιεί τους υπόλοιπους ανθρώπουςγύρωτου με βάση το φύλο, το γένος, τη χώρα καταγωγής κ.ο.κ

Με περισσότερο επιστημονικό ύφος όμως, θα μπορούσαμε να αναφέρουμε πως η κατηγοριοποίηση αποτελείται από την εξέταση των καθορισμένων κάθε φορά χαρακτηριστικών ενός νέου αντικειμένου και την ανάθεση του σε μία από τις ήδη υπάρχουσες και προδιαγεγραμμένες κατηγορίες ή - όπως συνηθίζεται να αναφέρεται στο πεδίο της εξόρυξη γνώσης – κλάσεις. Τα αντικείμενα προς κατηγοριοποίηση περιγράφονται αποκλειστικά από τις τιμές που έχουν τα γνωρίσματα τους και πρακτικά αυτό που αναζητάτε είναι η εκτίμηση του αύξοντα αριθμού ή του ονόματος της κλάσης στην οποία αυτό ανήκει. Φυσικά, πολλές φορές αυτή η διεργασία μπορεί αν αστοχήσει, κατηγοριοποιώντας ένα ή περισσότερα αντικείμενα σε λανθασμένη κλάση. Για αυτό συνήθως εκτιμάται και η ακρίβεια (classificationaccuracy) που πετυχαίνει ο εκάστοτε αλγόριθμος που χρησιμοποιείται σε υπάρχοντα δεδομένα, προκειμένου να ελεγχθεί η αξιοπιστία αυτού και το κατά πόσον πληροί τις προϋποθέσεις του αγοραστή.Πρακτικά λοιπόν, υπάρχει όπως αναφέρθηκε και προηγουμένως η ανάγκη δημιουργίας ενός μοντέλου, στο οποίο θα στηριχθεί η λειτουργία της ταξινόμησης. Για τη δημιουργία αυτού, απαιτούνται κάποια αρχικά δεδομένα, συνήθως αποκαλούνται δεδομένα εκπαίδευσης (trainingdataset), τα οποία περιγράφονται από μία σειρά από κοινά χαρακτηριστικά (features) και προσδιορίζουν και τις διαθέσιμες κλάσεις κατηγοριοποίησης. Σκοπός λοιπόν είναι η πρόβλεψη με βάση το δημιουργηθέν μοντέλου, το οποίο στηρίζεται αποκλειστικά στα αρχικά δεδομένα. Συνεπάγεται πως τυχόν πρόβλημα στη διαδικασία της κατηγοριοποίησης μπορεί να προέλθει από την ποιότητα των δεδομένων – π.χ.μπορεί να είναι αναξιόπιστα ή να έχουν συλλεχθεί χωρίς κατάλληλα δειγματοληπτικά κριτήρια – ή

τελικώς το προς αντιμετώπιση πρόβλημα να είναι αρκετά δύσκολο από τη φύση του, αποτρέποντας τη δημιουργία ενός αξιόπιστου μοντέλου.

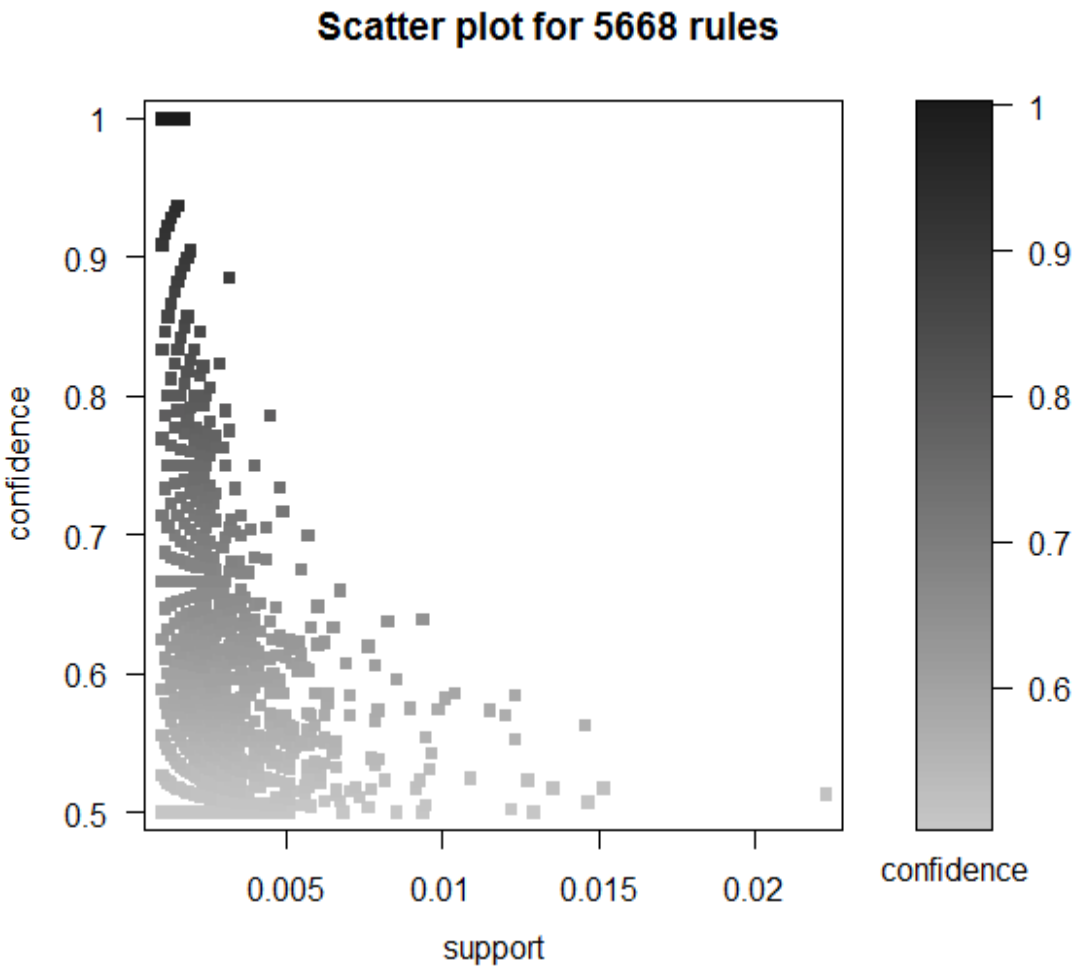
Ενώ η κατηγοριοποίηση ανήκει στην κατευθυνόμενη εξόρυξη γνώσης, το άλλο εξίσου σημαντικό θέμα που διακυβεύεται και η παρούσα εργασία, αυτό της εξαγωγής κανόνων συσχέτισης, ανήκει στη μη-κατευθυνόμενη εξόρυξη γνώσης. Το κύριο λοιπόν ζήτημα σε αυτού του είδους τη λειτουργία αποτελεί η απόφαση για το ποια χαρακτηριστικά μπορούν να συνδυαστούν ή, για να αναφερθούμε και στο τομέα των δεδομένων καλαθιού, να διευκρινιστεί ποια καταναλωτική συνήθεια συνεπάγεται την αγορά ή μη ενός συγκεκριμένου προϊόντος. Το συγκεκριμένο ζήτημα άλλωστε αποτελεί και την καρδιά της εξόρυξη γνώσης από δεδομένα καλαθιού. Έχοντας για παράδειγμα τα υπερκαταστήματα στη διάθεση τους κάποιο μοντέλο που περιγράφει κάποιες συνήθεις προτιμήσεις των καταναλωτών, μπορούν να αναδιατάξουν τα προϊόντα τους μέσα στα καταστήματα τους με τέτοιον τρόπο που κάποιος που θέλει να αγοράσει το Α προϊόν, να έρθει σε οπτική επαφή με το Β προϊόν, το οποίο ενώ δεν σχετίζεται άμεσα με το περιεχομένου του διαδρόμου ή του συγκεκριμένου ραφιού, υπάρχει εκεί για να ωθήσει τον καταναλωτή στην αγορά αυτού. Με αυτή την απλή λογική και με νέες τεχνικές αλγορίθμων και στατιστικών δεικτών, είναι δυνατή η εξαγωγή κανόνων συσχέτισης που προσδιορίζουν συνήθως με ένα ποσοστό εμπιστοσύνης τη πιθανότητα να αγοραστεί ένα προϊόν μαζί με ένα άλλο ή με μία ομάδα από άλλαπροϊόντα, χωρίς η λέξη προϊόν να μη μπορεί αν γενικευθεί φυσικά και για διάφορα άλλα σενάρια.

Έχοντας κατά νου τις παραπάνω λειτουργίες και αναλογιζόμενοι τις επιρροές του πεδίου της εξόρυξη γνώσης από την επιστήμη των υπολογιστών και τη στατιστική, είναι εύκολο να φανταστούμε πως κάθε λειτουργία μπορεί να εξελιχθεί σχετικά εύκολα, είτε αυτόνομα με τη χρήση ενός νέου αλγορίθμου δημιουργίας μοντέλων ή με νέους στατιστικούς δείκτες και κριτήρια που αν επηρεάζουν τη βέλτιστη επιλογή, είτε και σε ένα περισσότερο συνεργατικό πεδίο συνδυάζοντας αυτές τις τεχνικές και πετυχαίνοντας την κάλυψη νέων αναγκών. Παρακάτω ακολουθούν δύο σχήματα από το περιβάλλον της R που μέσω των βιβλιοθηκών arulesVizκαι party έχουν δημιουργηθεί ενδεικτικά οι απεικονίσεις από ένα παράδειγμα κατηγοριοποίησης και ένα με εξαγωγή κανόνων συσχέτισης:



Εικόνα 2: Παράδειγμα κατηγοριοποίηση με δένδρο απόφασης

Τα δεδομένα του συγκεκριμένου παραδείγματος προέρχονται από το datasetiris, το οποίο περιέχει 150 καταγραφές με φυτά και τα κατηγοριοποιεί σε 3 ξεχωριστές κατηγορίες.



Εικόνα 3: παράδειγμα απεικόνισης εξαγωγής κανόνων βασισμένο στα κριτήρια support*και* confidence

Αντίστοιχα με πριν, τα δεδομένα που απεικονίζονται στην Εικόνα 3 προέρχονται από το datasetGroceries, το οποίο περιλαμβάνει 9835 εγγραφές και 169 διαφορετικές κατηγορίες [12]. Περισσότερες λεπτομέρειες σχετικά με τους όρους που αφορούν στατιστικούς δείκτες θα δοθούν σε επόμενο κεφάλαιο.

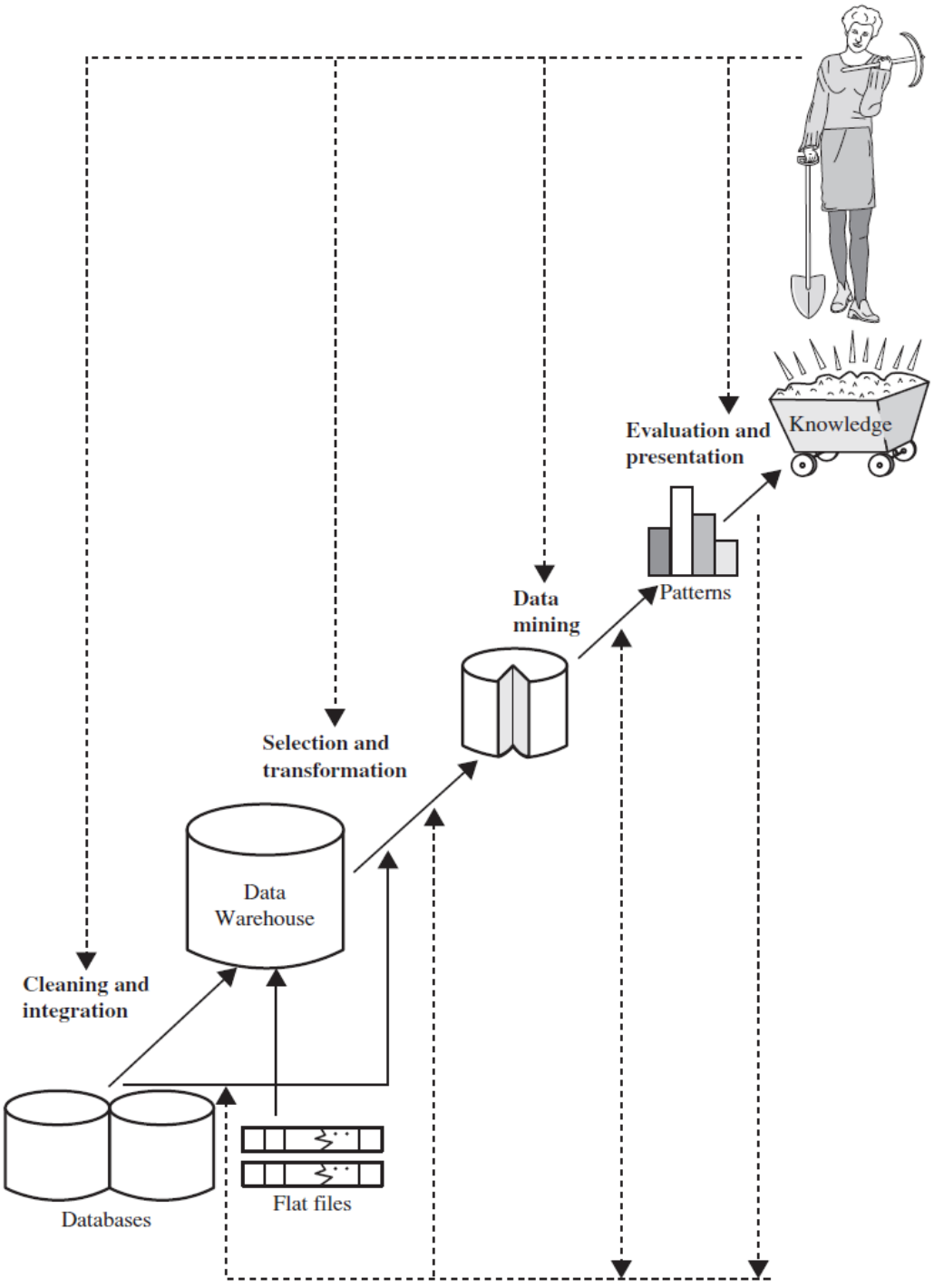
1.4 Εφαρμογής της εξόρυξης γνώσης σε βάσεις δεδομένων

Συχνά όσοι ασχολούνται με τον τομέα της εξόρυξη γνώσης χρησιμοποιούν τους όρους DataMiningκαι KDDεναλλακτικά για την ίδια ακριβώς έννοια. Πέρα από τις παρεμφερείς έννοιες που έχουν συσχετιστεί με την εξόρυξη γνώσης ο όρος KDDχρησιμοποιείται κατά κόρον για να εκφράσει μία διαδικασία που αποτελείται από επτά διαφορετικές φάσεις, μία από τις οποίες είναι ουσιαστικά και η εξόρυξη γνώσης [4]. Οι φάσεις μάλιστα αυτές μοιάζουν με τη διαδικασία που ακολουθείται σε γλώσσες όπως η SQLή ακόμη και σε άλλα μοντέλα καταχώρισης δεδομένων συνδεδεμένα με κάποιο σύστημα ερωταπαντήσεων [1]. Επιπλέον, για να καταστεί χρήσιμη και να ερμηνευθεί σωστά αυτή η διαδραστική και ορισμένες φορές περίπλοκη διαδικασία, απαιτείται η συνεργασία πολλών κλάδων και ειδικών ανά χώρο. Στο παρακάτω χωρίο περιγράφονται εν συντομία οι διαφορετικές φάσεις, η ροή των οποίων απεικονίζεται και στην Εικόνα 4.

Επομένως, η KDDδιαδικασία αναλύεται στις επόμενες φάσεις στις οποίες θεωρούμε σαν είσοδο πως έχουμε μία συλλογή δεδομένων, συνήθως οριοθετημένη όπως στα κοινά μοντέλα των βάσεων δεδομένων [4] :

1. Καθαρισμός δεδομένων (datacleaning): αποτελεί την εισαγωγική φάση της πρόσβασης στα δεδομένα και την αφαίρεση στιγμιότυπων που θεωρούνται είτε ως θόρυβος είτε ως ακραία είτε ως ασυνεπή με το κύριο όγκο δεδομένων ή ακόμη και ανίκανα να προσφέρουν χρήσιμη πληροφορία.
2. Ενοποίηση δεδομένων (dataintegration): σε αυτή τη φάση περιλαμβάνεται η ενοποίηση και ο συνδυασμός δεδομένων από διαφορετικές πηγές. Ουσιαστικά αυτό το βήμα μαζί με το προηγούμενο αποτελούν μία διαδικασία προ-επεξεργασίας των δεδομένων.
3. Επιλογή δεδομένων (dataselection): ανασύρονται τα σχετικά με το κάθε ερώτημα δεδομένα και παρέχονται στο χρήση με κάποιον συνοπτικό και περιγραφικό τρόπο.
4. Μετασχηματισμός δεδομένων (datatransformation): πολλές φορές τα δεδομένα που πρόκειται να εξετασθούν στα μετέπειτα στάδια απαιτούν κάποια κωδικοποίηση – όπως να μετατραπούν σε μία άλλη μορφή – ή να υποστούν

κάποια μαθηματική διεργασία – όπως η διακριτοποίηση με σκοπό τη χρήση τους από συγκριμένους αλγορίθμους που επιθυμούν αυτή τη μορφή – ή ακόμη και να περιοριστούν ορισμένα από αυτά.



Εικόνα 4: Ρόλος της εξόρυξης γνώσης στο KDD

5. Εξόρυξη γνώσης (datamining): σε αυτό το στάδιο εισέρχεται η ευφυία κάποιου αλγορίθμου ή κάποιας άλλης τακτικής με σκοπό την εξαγωγή χρήσιμων και περιγραφικών προτύπων ή κανόνων.
6. Εκτίμηση προτύπων (patternevaluation): έχοντας ανακαλύψει τα πρότυπα που ικανοποιούν κάποιο σχετικά ανεκτικό threshold, σειρά έχει η εκτίμηση των σημαντικότερων και πιο κατάλληλων προτύπων εξετάζοντας τα με τη βοήθεια στατιστικών δεικτών ή άλλων μετρικών και κριτηρίων.
7. Αναπαράσταση γνώσης (knowledgepresentation): στο τελικό στάδιο, το στάδιο της απεικόνισης, προτιμάτε και επιλέγεται κάθε φορά ο πιο ουσιώδης και περιγραφικός τρόπος αναπαράστασης των αποτελεσμάτων που προέκυψαν από τα προηγούμενα βήματα. Παρά το γεγονός πως αυτό το βήμα φαίνεται εύκολο και συνήθως δεν παρέχεται η κατάλληλη σημασία, αποτελεί αναπόσπαστο κομμάτι αυτής της αλυσίδας. Όσο καλή και παραγωγική δουλειά να έχει επιτευχθεί προηγουμένως, αν τα αποτελέσματα που εξάχθηκαν δεν παρουσιαστούν με έναν αρμονικό τρόπο υπάρχει η περίπτωση να μην εκτιμηθεί ή και να κατανοηθεί η χρησιμότητα των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 2

Εξόρυξη δεδομένων σε δεδομένα καλαθιού

2.1 Λόγοι ραγδαίας εξέλιξης της εξόρυξης γνώσης στα δεδομένα καλαθιού

«Η ανάγκη είναι η μητέρα των ανακαλύψεων» είπε oPlato. Πράγματι, ζούμε σε έναν κόσμο όπου τεράστιες ποσότητες δεδομένων συλλέγονται σε καθημερινή βάση. Η ανάλυση αυτών των δεδομένων έχει πλέον αποδειχθεί μία απαραίτητη ανάγκη. Μία άλλη γνωστή ρήση είναι πως «ζούμε στην εποχή της πληροφορίας». Έχουμε άλλωστε φτάσει σε μονάδες δεδομένων των terabyteskαι petabytes στα δίκτυα των υπολογιστών, στον παγκόσμιο ιστό αλλά σε σκληρούς δίσκους επιχειρήσεων τεχνολογικών, μηχανικών ή και ιατρικώνσυμφερόντων. Αυτή η εκρηκτική ανάπτυξη των διαθέσιμων δεδομένων είναι ένα αποτέλεσμα της βιομηχανοποίησης και της υιοθέτησης του Η/Υ από την κοινωνία έχουν συντελέσει στη γέννηση της εξόρυξη γνώσης. πλέον οι εταιρίες τηλεπικοινωνιών παγκοσμίως μεταφέρουν δεκάδες petabyteδεδομένων μέσα από τα δίκτυα τους. Οι βιομηχανίες υγείας παράγουν τρομακτικές ποσότητες δεδομένων μέσω των ιατρικών θεραπειών, του ελέγχου και της παρακολούθησης των ασθενών και των ιατρικών εικόνων. Εξίσου μεγάλη κυκλοφορία δεδομένων παρατηρείται καθημερινά στις μηχανές αναζήτησης καθώς και σε προσπάθειες καταγραφής εξομοιώσεων σε χημικούς αντιδραστήρες. Αυτές και άλλες πολλές εφαρμογέςαποτελούν αντικείμενο μελέτης του νέου, δυναμικού και πολλά υποσχόμενου πεδίο της εξόρυξη γνώσης [3].

Ένα πολύ καίριο ερώτημα είναι το εξής: Γιατί ενώ οι περισσότεροι αλγόριθμοι εξόρυξη γνώσης υπάρχουν εδώ και χρόνια ή και δεκαετίες σε επιστημονικά βιβλία ή διδάσκονται ακαδημαϊκά, μόλις πρόσφατα υιοθετήθηκαν από τις σύγχρονες αγορές με τόσο έντονο τρόπο; Οι παράγοντες που συνέβαλαν σε αυτό το φαινόμενο είναι αρκετοί και οι σημαντικότεροι αναφέρονται ευθύς αμέσως [5]:

- *Μεγαλύτερη παραγωγή δεδομένων από ποτέ*. Προκειμένου να δημιουργηθούν μοντέλα για εκτίμηση, πρόβλεψη, κατηγοριοποίηση ή άλλες εφαρμογές της εξόρυξη γνώσης, απαιτείται μεγάλος όγκος δεδομένων. Η διαδικασία της εκπαίδευσης άλλωστε αποτελεί το σημαντικότερο κομμάτι και χωρίς την ύπαρξη αντιπροσωπευτικών δεδομένων από όλες τις πιθανές κατηγορίες σε ένα πρόβλημα αυτής της φύσης, δε θα μπορούσαν να επιτευχθούν και παραγωγικές στρατηγικές. Η συλλογή επομένως δεδομένων από πιστωτικές κάρτες, onlineshopping, ηλεκτρονικών μεταφορών κεφαλαίων, αυτόματων πωλητών και άλλων αυτοματοποιημένων διαδικασιών έχει πλέον φθάσει σε ρυθμούς ασύγκριτα μεγαλύτερους σε σχέση με προηγούμενα χρόνια.
- *Αποθήκευση δεδομένων*. Όχι απλά παράγονται τεράστιες ποσότητες δεδομένων, αλλά η εξέλιξη των συστημάτων αποθήκευσης έχουν καταστήσει εφικτή την αποθήκευση και τηνενσωμάτωση τους σε υπάρχονταλειτουργικάσυστήματα. Επίσης, με τις νέες μεθόδους συλλογής και αποθήκευσης, δεδομένα από διαφορετικές πηγές συνδυάζονται και συνυπάρχουν για την αναζήτηση και την εξόρυξη χρήσιμων προτύπων
- *Η απαιτούμενη υπολογιστική ισχύς είναι πλέον εφικτή.*Οι αλγόριθμοι εξόρυξη γνώσης τυπικά απαιτούν πολλαπλά περάσματα σε τεράστιου όγκου δεδομένα. Η συνεχιζόμενη άλλωστε μείωση του κόστους των σκληρών δίσκων, των μηνών και της ισχύς επεξεργασίας έχουν καταστήσει δυνατή την εφαρμογή και την επεξεργασία αλγοριθμικών διαδικασιών που δε ήταν προσιτές στα επιστημονικά και ερευνητικά κέντρα των περισσότερων χωρών. Επιπλέον, η παραλληλοποίηση των αλγορίθμων και οι παράλληλες σχεσιακές βάσεις δεδομένων μεγάλων εταιριών (Oracle, Teradata, IBM) έχουν εξελίξει τους χρόνους απόκρισης στην πλειονότητα των επιστημόνων, παρέχοντας ένα ιδανικό περιβάλλον για μεγάλης κλίμακας εξόρυξη γνώσης.
- *Το ενδιαφέρον προς τις καταναλωτικές συνήθειες έχει ενταθεί*. Στο μεγαλύτερο φάσμα των επιχειρήσεων έχει πλέον διαπιστωθεί πως οι πελάτες είναι το κέντρο ενδιαφέροντος και πως η πρόσβαση στις πληροφορίες που προκύπτουν από αυτούς αποτελεί σημαντικό πλεονέκτημα έναντι των υπόλοιπων ανταγωνιστών. Αλυσίδες ξενοδοχείων

παρατηρούν και καταγράφουν τις προτιμήσεις για δωμάτια καπνιζόντων ή μη με σκοπό τη στοχευμένη διαφήμιση νέων προσφορών. Εταιρίες πώλησης πετρελαίου θέρμανσης παρατηρούν τα διαθέσιμα των πολυκατοικιών, προσφέροντας καλύτερες υπηρεσίες σε έγκαιρο για τους πελάτες χρονικό διάστημα παρά περιμένοντας μία κλήση από τους ενοίκους όταν στερέψουν τα αποθέματα τους. Πλέον όλες οι επιχειρήσεις προσπαθούν να συνοδέψουν τα προϊόντα τους με υπηρεσίες.

- *Η πληροφορία αποτελεί πλέον προϊόν*. Οι πληροφορίες δηλαδή που συλλέγονται από μία επιχείρηση για τους διάφορους πελάτες είναι εξίσου πολύτιμες και για άλλες εταιρίες. Πληροφορίες πλέον ανταλλάσσονται μεταξύ αεροπορικών εταιριών, υπερκαταστημάτων και από τράπεζες που υποστηρίζουν πιστωτικές ή καταναλωτικές κάρτες. Η Googleγια παράδειγμα γνωρίζει ανά πάσα στιγμή που προηγούμαστε στο κυβερνοχώρο, εκτοξεύοντας μας συνέχεια κατάλληλες διαφημίσεις που εγείρουν το ενδιαφέρον μας. Και αντίστοιχα, οι εταιρίες που θέλουν να διαφημιστούν αυξάνουν τις χορηγίες τους προς αυτές τις εταιρίες είτε με μπόνους επι των αγορών είτε με ετήσιες χορηγίες.
- *Εμπορικά προϊόντα εξόρυξη γνώσης είναι πλέον διαθέσιμα*. Δυστυχώς υπάρχει πάντα μία καθυστέρηση μεταξύ της εμφάνισης ενός νέου αλγορίθμου σε ένα επιστημονικό περιοδικό μέχρις ότου να εισέλθει στο εμπόριο με κάποιο στοχευμένο καταναλωτικό κοινό, όμως πλέον όλο και περισσότερες τεχνικές εξετάζονται, τροποποιούνται και βελτιώνονται λόγω της διεύρυνσης του επιστημονικού κοινού. Παρόλα αυτά, έχει φθάσει πλέον η εποχή όπου δεκάδες ιστοσελίδες πωλούν υπηρεσίες άμεσα σχετιζόμενες με το κομμάτι της εξόρυξη γνώσης. παρέχοντας συμβουλευτικές υπηρεσίες που αφορούν προβλέψεις ή τρόπους προσέγγισης των καταναλωτών.

Ψάχνοντας τώρα στη σύγχρονη αγορά για εφαρμογές και υπηρεσίες που εκτελούν και επωφελούνται από τις υπηρεσίες της εξόρυξη γνώσης, μπορεί εύκολα κανείς να διαπιστώσει πως το μέλλον και η αναγκαιότητα των εποχών επιβάλλουν την άμεση υιοθέτηση τέτοιων τεχνικών από οποιονδήποτε επιχειρηματία ή ιδιώτη, ώστε να μη παραγκωνιστεί η επιχείρηση του. Το πιο χαρακτηριστικό παράδειγμα αποτελούν τα

υπερκαταστήματα που υπάρχουν πλέον παγκοσμίως, τα οποία μέσω των ηλεκτρονικών σαρωτών, καταγράφουν κάθε αντικείμενο το οποίο πωλείται καθώς επίσης και με ποια άλλα αγαθά αυτά συνδυάζονται από πελάτες, που συνήθως λόγω των προσωπικών καρτών που κατέχουν και σαρώνονται πριν την πληρωμή, στοιχειοθετούν ένα καταναλωτικό προφίλ στο κάθε πελάτη. Φυσικά, μέσω μεθόδων, οι οποίες δε μπορούν εύκολα να χαρακτηρισθούν ως ανθρωποκεντρικές, υπάρχει η τάση να προσφέρουν και να παρέχουν προσωπικές κάρτες σε όλους τους πελάτες τους. Η πρώτη γνωστή αλυσίδα που διενήργησε με αυτόν τον τρόπο ήταν η εταιρία Safewayστις Η.Π.Α. προσφέροντας ειδικές εκπτώσεις τους πελάτες της, αρκεί να χρησιμοποιούσαν τις ειδικές αυτές κάρτες. Επομένως, σε κάθε επόμενη αγορά, το ιστορικό προτίμησης του καταναλωτή ανανεωνόταν, εμπλουτίζοντας τη βάση δεδομένων της εταιρίας. Οι πληροφορίες λοιπόν που συλλέχθηκαν, δεν επηρέασαν μόνο την ίδια την επιχείρηση, η οποία μπορούσε να προωθήσει στοχευμένα πλέον τα προϊόντα της με τον αποδοτικότερο για αυτήν τρόπο, αλλά παράλληλα παρείχε και πληροφορίες σε εταιρίεςπαραγωγήςπροϊόντων, επηρεάζοντας και τις τιμές αυτών, αλλά και την ποιότητα τους και τις πολιτικέ των προσφορών τους, ακόμη και τις κατασκευαστικές συνήθειες των παρεχόμενων αγαθών [5]. Παρόμοιες συμπεριφορές έχουν παρατηρηθεί και από εταιρίες με χρεωστικές κάρτες, όπως η AmericanExpress, οι οποίες χρέωναν την τοποθέτηση διαφημιστικών μηνυμάτων πάνω στις κάρτες τους, δίνοντας την ευκαιρία στις εταιρίες να προωθήσουν τις υπηρεσίες τους σε συγκεκριμένες ομάδες κατόχων αυτών των καρτών.

Επιπλέον, η εισαγωγή μεθόδων εξόρυξη γνώσης μπορούν να βοηθήσουν τις διάφορες εταιρίες να διατηρήσουν τους πελάτες τους. Ψάχνοντας τις συνήθειες τους, μπορούν να αναγνωρίσουν αν η εκτίμηση προς τα δικά τους προϊόντα έχει μειωθεί ή έχει υποτιμηθεί, και αν συμβαίνει αυτό, πιθανόν να μπορούν να προσδιορίσουν το γιατί. Επειδή γενικά η διατήρησή ενός υπάρχοντος πελάτη είναι φθηνότερη από την απόκτηση ενός νέου πελάτη λόγω διαφημιστικών εξόδων, τέτοιες πληροφορίες είναι ανεκτίμητες για τις εταιρίες. Όπως επίσης από την άλλη πλευρά, η διαδικασία απομάκρυνσης ενός πελάτη ή η άρνηση συνεργασίας με πελάτες που έχουν αρνητικό ιστορικό, χαρακτηριστικό παράδειγμα αποτελούν οι τράπεζες οι οποίες ελέγχουν τη πιστωτική συμπεριφορά ενός νέου πελάτη όταν αυτός έρχεται να ζητήσει ένα δάνειο.

Τέτοιες λοιπόν εφαρμογές, δίνουν την εντύπωση πως οφείλονται για οποιαδήποτε επιχειρηματική δραστηριότητα. Η αλήθεια δεν απέχει πολύ αν ψάξει κανείς σε

σχετικά άρθρα για το πώς η εξόρυξη γνώσης επηρεάζει την παραγωγή και την παροχή υπηρεσιών σήμερα. Το σίγουρο πάντως είναι πως στα απόμενα χρόνια, οι τακτικές που ακολουθούνται και υιοθετούν την εξόρυξη γνώσης θα είναι ο κύριος λόγος άνθησης και κερδοφορίας των νέων επιχειρήσεων [5].

2.2 Δημοφιλείς συλλογές δεδομένων καλαθιού

Σε αυτό το κομμάτι της εργασίας θα αναφερθούμε σε ορισμένες γνωστές συλλογές δεδομένων που σχετίζονται με δεδομένα καλαθιού. Οι 3 πρώτες αναφέρονται με περισσότερες λεπτομέρειες σχετικά με τη ανάλυση και την επεξεργασία που έχουν υποστεί ενώ για την τέταρτη κατά σειρά έχουμε ορισμένα αποτελέσματα που αφορούν ένα χώρο της σύγχρονης τεχνολογίας.

Η πρώτη λοιπόν αναφέρεται σε μία αλυσίδα αρτοποιείων, της οποίας τα καταστήματα βρίσκονται στις δυτικές ακτές των Η.Π.Α. και πιο συγκεκριμένα στη Καλιφόρνια, το Όρεγκον, την Αριζόνα και τη Νεβάδα. Το μενού που υποστηρίζουν αυτά τα καταστήματα περιλαμβάνει 40 είδη αγαθών με ζύμη και 10 διαφορετικά ροφήματα. Τα δεδομένα που έχουν δημοσιευτεί ως ExtendedBAKERYdataset και περιλαμβάνουν πληροφορίες σχετικά με:

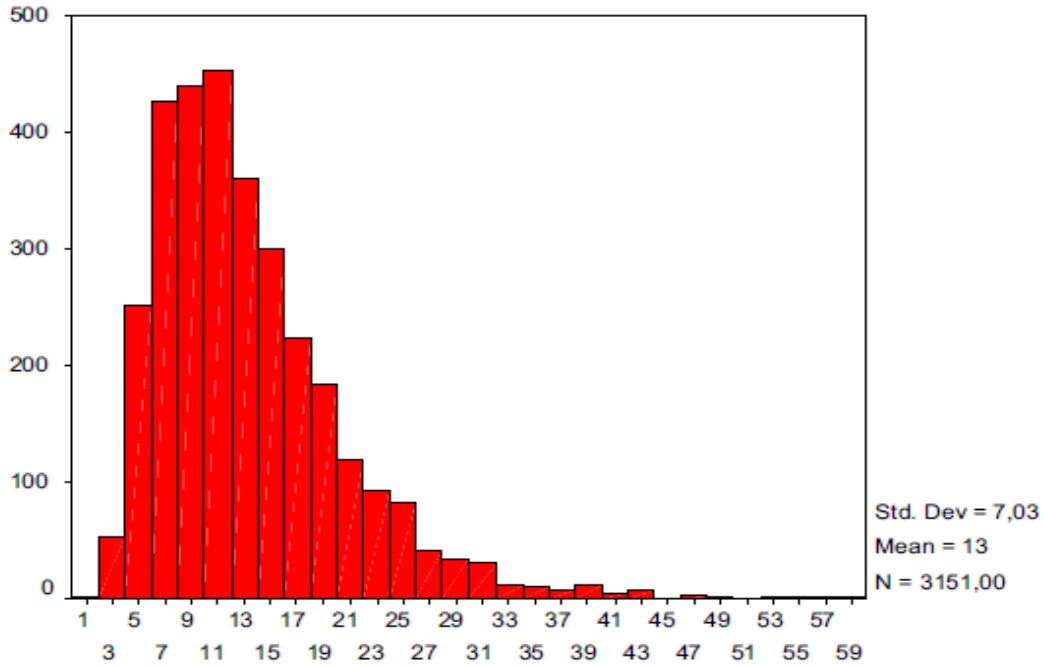
- τα παρεχόμενα αγαθά,
- την τοποθεσία του εκάστοτε καταστήματος,
- τους εργαζόμενους σε κάθε κατάστημα,
- τις αποδείξεις,
- τους συνδυασμούς με τα αγαθά που αγοράστηκαν μαζί.

Συνολικά διατίθενται 4 διαφορετικά dataset με 1.000, 5.000, 20.000 και 75.000 αποδείξεις [7]. Η χρονική περίοδος που έγινε η συλλογή δεδομένων αφορά ένα ολόκληρο έτος και τα δικαιώματα αυτής της συλλογής ανήκουν στο Πολυτεχνικό Πολιτειακό Πανεπιστήμιο της Καλιφόρνια. Αξίζει να αναφερθεί πως για την εξαγωγή κανόνων συσχέτισης χρησιμοποιούν και οι ίδιοι τον αλγόριθμο Apriori.

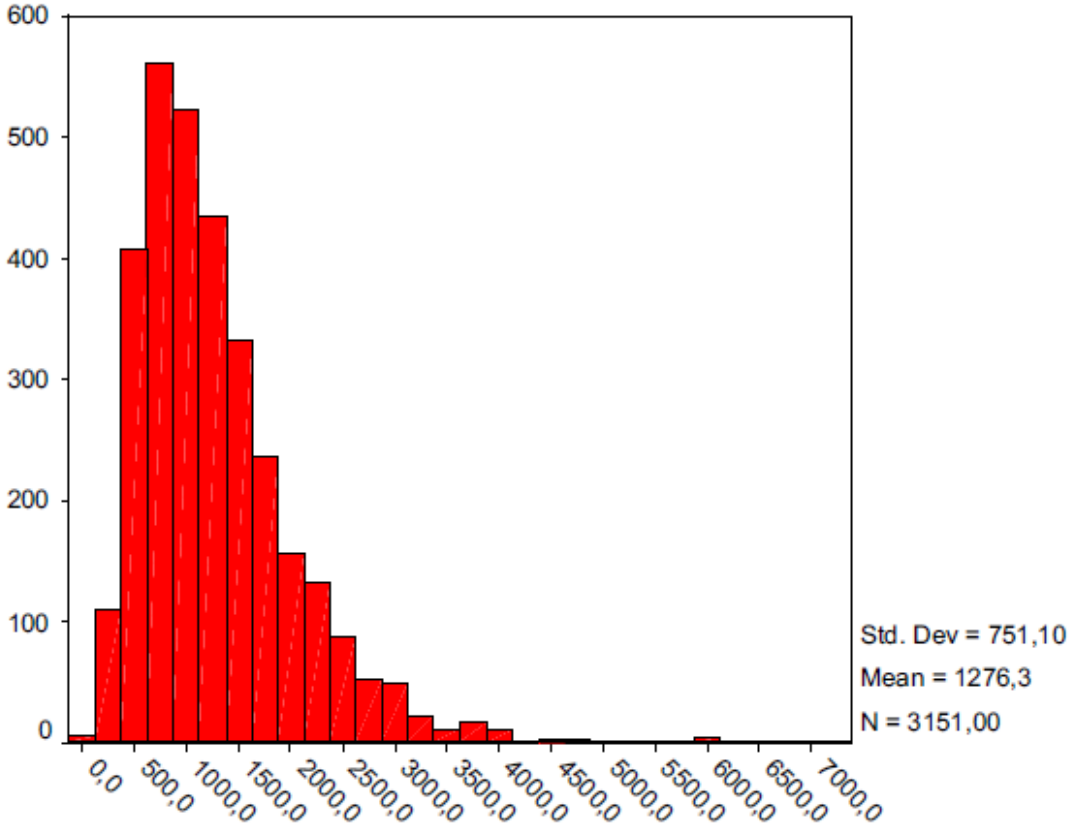
Μία παρόμοια συλλογή δεδομένων καταγράφθηκε το 1999 από τον TomBrijsκαι περιλαμβάνει δεδομένα καλαθιού από ανώνυμο υπερκατάστημα λιανικής στο Βέλγιο [8]. Τα δεδομένα σε αυτήν την περίπτωση συλλέχθηκαν σε 3 μη συνεχόμενες περιόδους, από τα μέσα Δεκέμβρη του 1999 έως τα μέσα Γενάρη του 2000,από τις αρχές Μάη έως τις αρχές Ιούνη του 2000 και από τον Αύγουστο έως το τέλος του Νοέμβρη του 2000 ξανά. Σε πλήθος, οι εγγραφές αγγίζουν τις 88.163. τα χαρακτηριστικά που καταγράφθηκαν είναι τα εξής:

- ημερομηνία
- αριθμός απόδειξης
- αριθμός αντικειμένου
- το πλήθος των αγορασθέντων αγαθών
- ποσό ανά αγορά σε βέλγικα φράγκα
- αριθμός πελάτη

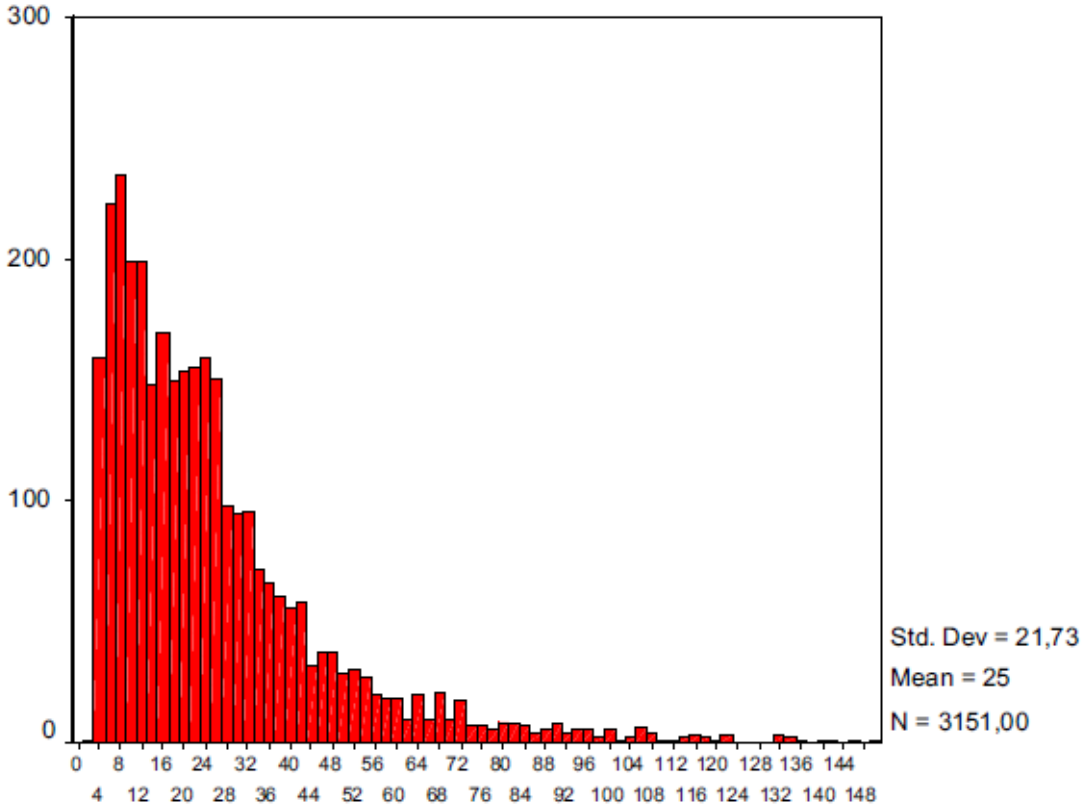
Συνολικά το κατάστημα αυτό προσφέρει 16.470 διαφορετικά αντικείμενα, ορισμένα εκ των οποίων είναι εποχιακά και δεν διατίθενται όλο το χρόνο. Κατά την περίοδο συλλογής δεδομένων 5.133 πελάτες αγόρασαν τουλάχιστον ένα αντικείμενο. Παρακάτω ακολουθούν ορισμένες γραφικές που περιγράφουν κάποια στατιστικά στοιχεία των δεδομένων για τους 5 μήνες που διήρκησε η έρευνα :



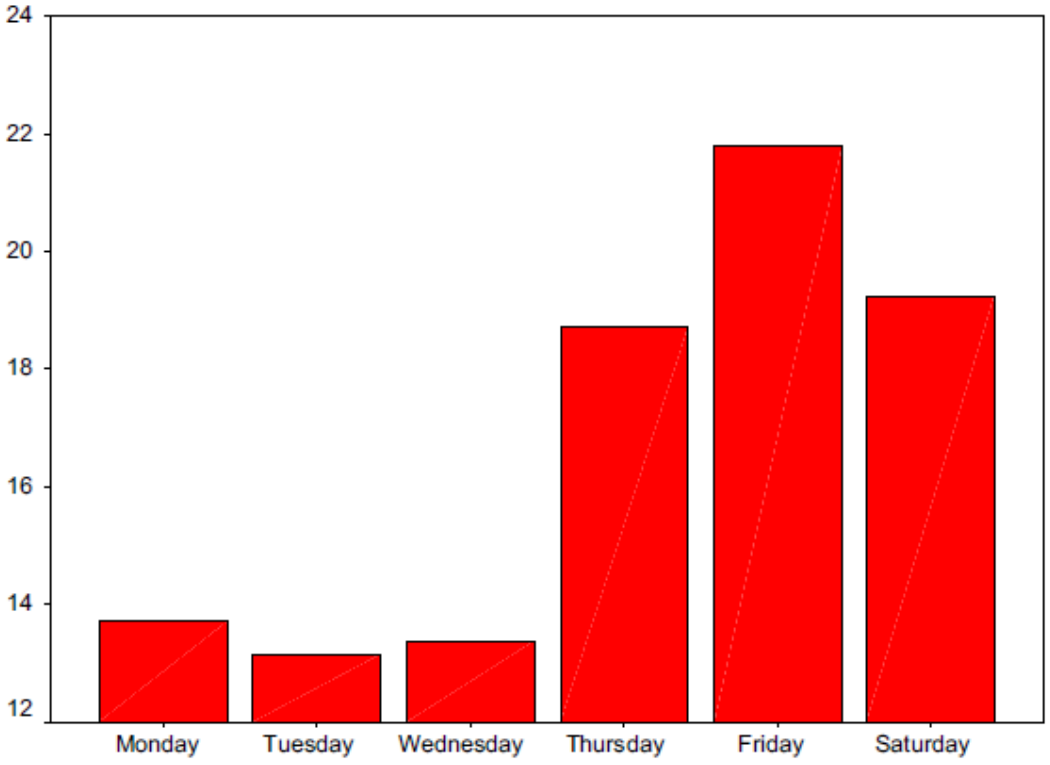
Εικόνα 5:Μέσο πλήθος διακεκριμένων αντικειμένων ανά επίσκεψη



Εικόνα 6: Μέσο εισαχθέν ποσό ανάεπίσκεψη



Εικόνα 7: Συνολικός αριθμός επισκέψεων εντός 24 εβδομάδων



Εικόνα 8: Κατανομή των επισκέψεων ανά ημέρα

Η τρίτη περίπτωση αποτελεί τη συλλογή δεδομένων για διάστημα 30 ημερών σε ένα παντοπωλείο, με σκοπό την διερεύνηση της συμπεριφοράς των στατιστικών δεικτώνconfidenceκαι liftόσον αφορά την εξαγωγή κανόνων συσχέτισης. Τα δεδομένα μάλιστα έχουν ενσωματωθεί στο περιβάλλον της R. συνολικά περιλαμβάνει 9835 εγγραφές από 169 διαφορετικά κατηγορίες προϊόντων, συντελώντας σε ένα μέσορυθμό συναλλαγών ανά ημέρα ίσο με 327.5. Προκειμένου να διερευνηθεί η ακρίβεια των δύο αυτών μέτρων, στήθηκε ένα στατιστικό σενάριο το οποίο παράγει ένα σύνολο δεδομένων μέσω εξομοίωσης και με το οποίο γίνεται η σύγκριση [17]. Έχοντας προσχωρήσει στις κατάλληλες παρατηρήσεις και στατιστικούς ελέγχους, επαλήθευσαν πως η παραγωγή δεδομένων σύμφωνα με το μοντέλο τους παρουσιάζει παρόμοια στατιστική συμπεριφορά με ρεαλιστικά δεδομένα. Επίσης, κατέληξαν στο συμπέρασμα πως οι στατιστικό δείκτες που μελετώνται επηρεάζονται έντονα από τη συχνότητα των αντικειμένων που αγοράστηκαν από τα αντίστοιχα συνολοστοιχεία καθώς και οι κανόνες που προκύπτουν. Πιο συγκεκριμένα, το liftαποδίδει πολύ άσχημα σε θορυβώδη δεδομένα και ακραίες τιμές και πως επιστρέφει υψηλότερες τιμές για κανόνες που περιλαμβάνουν τα λιγότερο συχνά στοιχεία. Επίσης, το μέτρο confidenceεπηρεάζεται συστηματικά από χαμηλής συχνότητας εμφάνισης δεδομένα. Το συμπέρασμά τους λοιπόν είναι πως πρέπει να πραγματοποιούνται περισσότεροι στατιστικοί έλεγχοι προτού διενεργηθούν παρόμοια πειράματα που εξετάζουν κανόνες συσχέτισης.

Η τελευταία περίπτωση αναφέρεται στην inmobi[9], μία εταιρία που ειδικεύεται σε διαδικτυακές πωλήσεις. Μία πρόσφατη έρευνα της δημοσιεύθηκε το 2014 και αναφερόταν στην παγκόσμια αγορά κινητών παιχνιδιών. Μάλιστα, μέσω συλλογής δεομένων που πραγματοποίησαν, ανέλυσαν συνήθειες του gamingκαι το περιεχόμενο των παιχνιδιών που προτιμώνται στις Η.Π.Α. ΣΤΗΝ Κίνα και στην Κορέα. Η έρευνα διεξήχθη ερευνώντας τη συμπεριφορά 1250 gamersστις 3 παραπάνω χώρες το δεύτερο μισού του 2013. Η έρευνα λοιπόν παρουσίασε σε ποσοστά τις προτιμήσεις αυτών ανά κονσόλες διαφορετικές εταιρίας και γενιάς, διαχώρισε τη συμπεριφορά ανδρών και γυναικών καθώς και τις προτιμήσεις ανάλογα με το gameplayτων προτιμώμενων παιχνιδιών. Με βάση αυτή την έρευνα, προσπάθησαν να διαλευκάνουν στις αντίστοιχες εταιρίες ποιοι και που αγοράζονται, παίζονται και ξοδεύονται τα περισσότερα ποσά ανά κατηγορία παιχνιδιών.

ΚΕΦΑΛΑΙΟ 3

Συλλογή δεδομένων και χρήση μεθόδων Εξόρυξης Γνώσης

3.1 Περιγραφή της συλλογής δεδομένων

Για τις ανάγκες υλοποίησης και ολοκλήρωσης της εργασίας, δημιουργήθηκε ένα πρωτότυπο datasetμε δεδομένα που συλλέχθηκαν σε ένα κοινό παντοπωλείο γειτονιάς στην περιοχή της Αγίας Σοφίας. Η διαδικασία της συλλογής πραγματοποιήθηκε στο διάστημα 23/05/2015 έως και 27/06/2015, στο οποίο και περιέχονται 30 εργάσιμες ημέρες. Σε αυτό λοιπόν το διάστημα συλλέχθηκαν 324 διαφορετικές εγγραφές – πελάτες, οδηγώντας σε μία μέση ημερήσια καταγραφή 10.8 πελατών. Προφανώς αυτό το μέγεθος υστερεί αρκετά σχετικά με τις αντίστοιχες εργασίες που αναφέρθηκαν στο $2^{ο}$ κεφάλαιο. Ο λόγος φυσικά είναι πως το συγκεκριμένο datasetαποσκοπεί στην υλοποίηση μίας πτυχιακής εργασία καθώς και στην εισαγωγή του συγγραφέα σε μία άμεση επαφή με το αντικείμενο της εξόρυξη γνώσης και των δεδομένων καλαθιού, ενώ οι αντίστοιχες προσπάθειες που προαναφέρθηκαν αποτελούσαν αντικείμενο συνήθως χρηματοδοτούμενης έρευνας με σκοπό τη δημοσίευση σε διεθνή συνέδρια και περιοδικά καθώς και την επίσημη χρήση τους από λογισμικά σχετικά με γνωστά προγραμματιστικά περιβάλλοντα με αρκετά εκτεταμένο community.

Όσον αφορά το πρακτικό μέρος, η καταγραφή των δεδομένων ξεκινούσε κάθε μέρα με το άνοιγμα του μαγαζιού στις 08.00 π.μ. και ολοκληρωνόταν κάθε μέρα στο κλείσιμο αυτού στις 09.00 μ.μ. Τα στοιχεία που συλλέγονταν για κάθε πελάτη ήταν τα εξής:

- Ημερομηνία
- Ώρα
- Φύλλο (άνδρας ή γυναίκα)
- Ηλικία (Α: 0 έως 18, Β: 19-45, Γ: 46 και πάνω)

- Π1: Σνακ - Κρύα Φαγητά
- Π2: Γλυκά -Σοκολάτες - Κρουασάν
- Π3: Ζυμαρικά -Όσπρια
- Π4: Γαλακτικά Προϊόντα
- Π5: Ανθρακούχα Ποτά - Νερό
- Π6: Αλκοολούχα Ποτά
- Π7: Καθαριστικά Σπιτιού
- Π8: Είδη Υγιεινής - Ατομικής Χρήσης

Όπως φαίνεται και παραπάνω, τα διάφορα προϊόντα που παρέχονται από το συγκεκριμένο κατάστημα έχουν ομαδοποιηθεί σε 8 διαφορετικές κατηγορίες. Οι λόγοι που έγινε αυτή η διαδικασία, είναι αφενός η συγκέντρωση όμοιων προϊόντων σε κοινές ομάδες ώστε να υπάρχει μεγαλύτερη συγκέντρωση δειγμάτων σε λιγότερες κατηγορίες, αφετέρου η έλλειψη ανάγκης να πραγματοποιήσουμε στα πλαίσια της παρούσας εργασίας κάποια ανάλυση προτίμησης μεταξύ προϊόντων. Επίσης, η μικρή συγκέντρωση εγγραφών στο περιορισμένο αυτό χρονικό διάστημα ενός μικρού καταστήματος θα οδηγούσε σε αποτυχία πρόβλεψης οποιασδήποτε σημαντικής και ουσιώδης συσχέτισης μεταξύ περισσότερων και πιο συγκεκριμένων κατηγοριών.

Τα δεδομένα αυτά λοιπόν συλλέχθηκαν και καταγράφθηκαν στο αρχείο proionta.xlsxτου MicrosoftExcel. Επίσης στο συγκεκριμένο αρχείο έχουν υπολογιστεί ορισμένα μερικά αθροίσματα. Στο τέλος κάθε στήλης με τις κατηγορίες των προϊόντων Π1 έως Π8 έχουν αθροιστεί οι αγορές του κάθε προϊόντος ξεχωριστά σε σύνολο 1209 πωληθέντων αντικειμένων. Επιπρόσθετα, στο τέλος κάθε γραμμής έχει επίσης υπολογιστεί το συνολικό άθροισμα προϊόντων που αγόρασε ο κάθε πελάτης. Ένα διάγραμμα τύπου πίτας αναπαριστά το πρώτο από τα δύο μερικά αθροίσματα:

Εικόνα 9: Απεικόνιση των επιμέρους ποσοστών πώλησης των διάφορων προϊόντων ανά κατηγορία.

Εν συνεχεία, τα δεδομένα από τη μορφή του .xlsxθα πρέπει να περάσουν σε μορφή .arffέτσι ώστε να μπορούν να υποστούν επεξεργασία από το ανοιχτού κώδικα λογισμικό WEKA [10]. Το πρόγραμμα αυτό αποτελεί ένα από τα κυριότερα στο χώρο της εξόρυξη γνώσης έχοντας υλοποιημένους σε JAVAκώδικα τους περισσότερους και πιο εύχρηστους ταξινομητές αλλά και αλγορίθμους συσταδοποίησης και αλγορίθμους εξαγωγής κανόνων συσχέτισης. Ο μόνος περιορισμός είναι πως τα αρχεία τα οποία προορίζονται να επεξεργαστούν από το WEKA θα πρέπει εκφραστούν αρχικά στην μορφή ARFF(Attribute-RelationFileFormat). Τα αρχεία αυτού του τύπου αποτελούν αρχεία κειμένου τύπου ASCII και περιγράφουν ένα σύνολο από εγγραφές (γραμμές) οι οποίες μοιράζονται από κοινού μία ομάδα χαρακτηριστικών (στήλες). Προκειμένου να υλοποιηθεί αυτό το βήμα, απαιτείται ο διαχωρισμός των αρχείων σε δύο μέρη, το HEADERκαι το DATA [11].

Το πρώτο λοιπόν μέρος της κεφαλίδας περιλαμβάνει προαιρετικά κάποιες ενημερωτικές γραμμές ως σχόλια (προηγείται ο χαρακτήρας %) που πληροφορούν τον χρήστη για το σύνολο των δεδομένων που έρχεται σε επαφή και πιθανόν προσωπικές πληροφορίες του συγγραφέα ή άλλες βοηθητικές οδηγίες. Ακριβώς από κάτω ακολουθούν οι συσχετίσεις, οι οποίες ξεκινούν με το χαρακτήρα @ και ονομάζουν ουσιαστικά τα διάφορα χαρακτηριστικά καθώς και αναφέρουν

ταυτόχρονα τον τύπο τους. Ένα ενδεικτικό παράδειγμα από το datasetIRIS που περιέχεται στο περιβάλλον της Rδίνεται παρακάτω:

```
% 1. Title: Iris Plants Database
   %
   % 2. Sources:
   %      (a) Creator: R.A. Fisher
   %      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
   %      (c) Date: July, 1988
   %
   @RELATION iris

   @ATTRIBUTE sepallength  NUMERIC
   @ATTRIBUTE sepalwidth   NUMERIC
   @ATTRIBUTE petallength  NUMERIC
   @ATTRIBUTE petalwidth   NUMERIC
   @ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-
virginica}
```

Μεταξύ των σχολίων και των συσχετίσεων, υπάρχει πάντα μία γραμμή με το πρότυπο @RELATIONname και ακολουθείται επίσης από μία κενή γραμμή. Το τελευταίο επίσης attributeφέρει το όνομα classκαι περιλαμβάνει μέσα σε αγκύλες τις κατηγορίες των διαφορετικών κλάσεων στις οποίες αναμένουμε να κατηγοριοποιηθούν οι διαφορετικές εγγραφές των εκάστοτε δεδομένων. Τα υπόλοιπα attributesμετά το όνομα της αντίστοιχης στήλης που ορίζουν αναφέρουν το κωδικό όνομα NUMERIC, υποδηλώνοντας πως όλα τα στοιχεία που ανήκουν σε αυτή τη στήλη αποτελούν αριθμητικά δεδομένα. Στην περίπτωση της εισαγωγής της ημερομηνίας στα δεδομένα μας, απαιτείται η εισαγωγή μίας γραμμής συσχέτισης με το παρακάτω format:

```
@ATTRIBUTE timestamp DATE "dd-M-yyyy"
@ATTRIBUTE timestamp TIME "HH.mm.ss"
```

Όπου οι αγγλικοί όροι εννοούν τα εξής:

- dd: ημέρα
- M: μήνας
- yyyy: χρονολογία
- HH: ώρα
- mm: λεπτά
- ss: δευτερόλεπτα

Στη δικιά μας περίπτωση, το στοιχείο ssπαραλείφθηκε αφού η καταγραφή των δευτερολέπτων δεν προσφέρει κάποια πληροφορία και δεν συμπεριλήφθηκε στη διαδικασία.Αξίζει να αναφερθεί πως υπάρχει δυνατότητα να εισαχθούν δεδομένα άλλων δύο τύπων. Αυτοί είναι ο τύπος stringόπου τα δεδομένα εσωκλείονται εντός αποστρόφων, και των nominalδεδομένων.

Το δεύτερο τώρα μέρος ενός .arffαρχείου αποτελεί το κύριο όγκο των δεδομένων. Προηγείται η γραμμή @Dataκαι από κάτω ακολουθούν τα δεδομένα έχοντας τις διαφορετικές στήλες χωρισμένες με κόμμα. Ένα μικρό δείγμα των δεδομένων του datasettου οποίου το headerαναφέρθηκε προηγουμένως ακολουθεί παρακάτω:

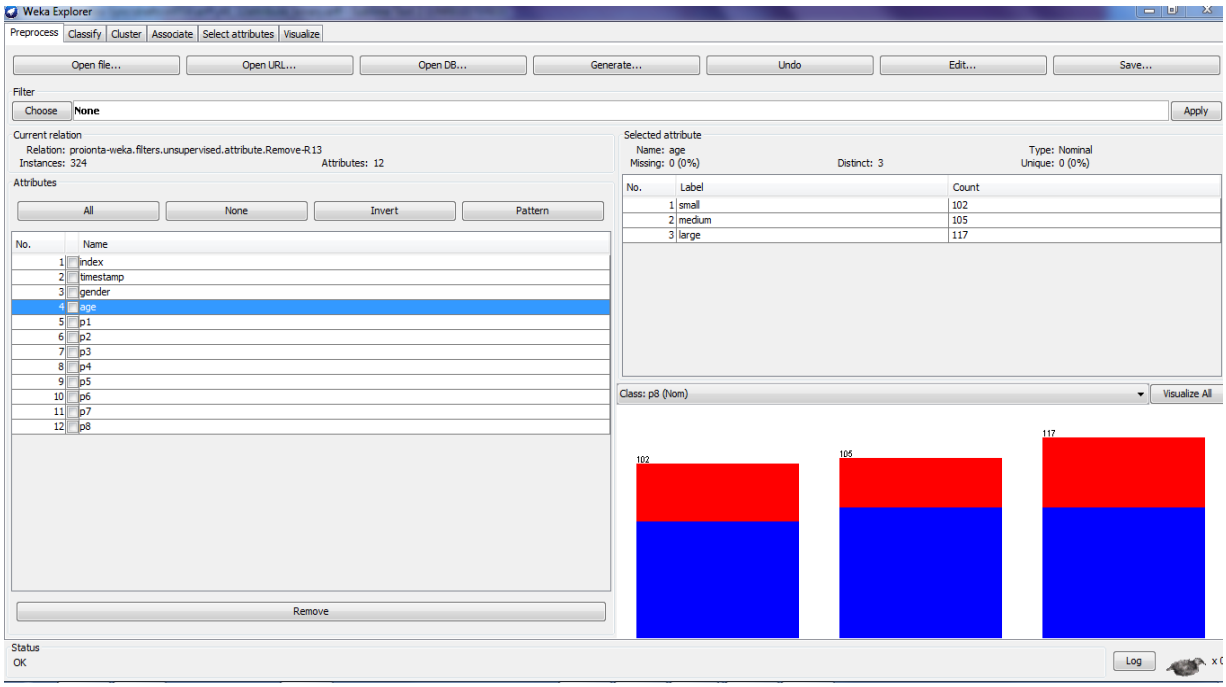
```
@DATA
   5.1,3.5,1.4,0.2,Iris-setosa
   4.9,3.0,1.4,0.2,Iris-setosa
   4.7,3.2,1.3,0.2,Iris-setosa
   4.6,3.1,1.5,0.2,Iris-setosa
   5.0,3.6,1.4,0.2,Iris-setosa
   5.4,3.9,1.7,0.4,Iris-setosa
   4.6,3.4,1.4,0.3,Iris-setosa
   5.0,3.4,1.5,0.2,Iris-setosa
   4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Παρατηρώντας αυτά τα δύο μέρη διαπιστώνεται εύκολα πως κάθε στήλη πρέπει να υπακούει στην περιγραφή που δόθηκε για αυτή στο κομμάτι του header. Σε διαφορετική περίπτωση το WEKAδεν μπορεί να ανοίξει το αρχείο και ενημερώνει το χρήση για την πρώτη ασυμφωνία που βρέθηκε στο αρχείο σε σχέση με το formatπου ορίζει κάθε φορά το κομμάτι του header.

Εισάγοντας το kefalaio3.arffστο WEKAμπορούμε να δούμε ορισμένα συγκεντρωτικά στοιχεία που αναφέρει και το συγκεκριμένο περιβάλλον. Μία σύνοψη του αρχείου και μία απεικόνιση του ακολουθούν ευθύς αμέσως:

```
@relation proionta

@attribute 1 numeric
@attribute timestamp date 'dd-MM-yyyy HH:mm'
@attribute gender {male,female}
@attribute age {small,medium,large}
@attribute 5 numeric
@attribute 6 numeric
```

```
@attribute 7 numeric
@attribute 8 numeric
@attribute 9 numeric
@attribute 10 numeric
@attribute 11 numeric
@attribute 12 numeric

@data
1,'23-05-0015 09:38',male,large,2,0,1,1,0,0,0,0
2,'23-05-0015 10:12',male,small,0,0,0,1,0,1,0,0
3,'23-05-0015 10:25',female,large,1,1,0,0,0,0,0,0
4,'23-05-0015 10:29',male,medium,0,0,1,0,0,0,1,0
5,'23-05-0015 11:01',female,large,2,0,1,0,0,1,0,1
6,'23-05-0015 11:16',female,small,0,1,0,2,0,1,0,0
7,'23-05-0015 12:01',female,small,0,2,0,0,0,1,0,0
```



Εικόνα 10: Ενδεικτική απεικόνιση του περιβάλλοντος WEKAγια ένα τυπικό datasetδεδομένων καλαθιού

Στο δεξί μέρος της Εικόνας 6, παρατηρεί κανείς πως από τους 324 πελάτες οι 102 ανήκουν στη πρώτη ηλικιακή κατηγορία, οι 105 στη δεύτερη και οι υπόλοιποι 117 στην τρίτη και τελευταία κατηγορία. Επίσης, διάσπαρτες μέσα στο ίδιο παράθυρο του Preprocess, εμφανίζονται και άλλες χρήσιμες πληροφορίες ή διαφορετικοί τρόποι απεικόνισης

Έχοντας μορφοποιήσει τα δεδομένα μας στην απαραίτητη μορφή, πραγματοποιήθηκαν ορισμένες αλλαγές για να πραγματοποιηθούν σωστά τα διάφορα πειράματα μας. Η πιο σημαντική αλλαγή ήταν η μετατροπή των datatου arffσε binary μορφή. Δηλαδή, στις στήλες που περιγράφουν τα δεδομένα των 8 διαφορετικών κατηγοριών, όλες οι τιμές διάφορες του μηδενός (προφανώς μιλάμε για μη αρνητικούς αριθμούς) άλλαξαν σε τιμή 1 ενώ οι μηδενικές συνέχισαν ως έχουν. Επιπλέον, οι στήλες που περιείχαν την ημερομηνία, το φύλλο, την ηλικία και τον αύξοντα αριθμό διαγράφηκαν. Αυτές οι αλλαγές μπορούν να γίνουν είτε μέσω του Exceleίτε μέσω ενός προχωρημένου editor (π.χ. SublimeText) είτε και μέσω κάποιων scriptsσε γλώσσες προγραμματισμού που υποστηρίζουν την ανάγνωση, μετατροπή και εγγραφή αρχείων κειμένου. Στη συγκεκριμένη περίπτωση, συνδυάστηκαν και οι 3 αυτοί τρόποι και η γλώσσα που γράφθηκαν τα scriptsήταν ηPython 2.7. Το αρχείο λοιπόν που δημιουργήθηκε και τελικά θα υποστεί την εφαρμογή των αλγορίθμων εξόρυξη γνώσης είναι το j48_attributes_binary.arff.

3.2 Πίνακας Συσχέτισης

Το πρώτο στατιστικό στοιχείο που παρουσιάζουμε για την επίβλεψη του συνόλου δεδομένων που έχουμε στη διάθεσή μας είναι ο πίνακας συσχέτισης (correlationmatrix) συνοδευόμενος από τις p-values [13], [14]. Αρχικά, όταν αναφερόμαστε σε μία τιμή συσχέτισης, εννοούμε μία τιμή στο διάστημα [-1,1] η οποία συμβολίζει την εξάρτηση που παρουσιάζουν μεταξύ τους δύο διαφορετικά σετ δεδομένων. Ένα κλασσικό παράδειγμα είναι το αν το ύψος των ανθρώπων συσχετίζεται με την εξυπνάδα τους ή η σχέση ζήτησης και προσφοράς που περιγράφει ορισμένα αγαθά στις παγκόσμιες ή και τοπικές αγορές. Η μεγάλη χρησιμότητα αυτών των τιμών οφείλεται στο γεγονός πως πληροφορούν με περιγραφικό τρόπο για το αν και κατά πόσο έντονα συνδέονται μεταξύ τους δύο ή και περισσότερα πράγματα, συμπεριφορές ή και χαρακτηριστικά, διευκολύνοντας με αυτόν τον τρόπο την αξιοποίηση της πληροφορίας στην πράξη. Ο λόγος που αναφερόμαστε σε σύνολα δεδομένων και όχι σε μεμονωμένες τιμές προκύπτουν και από τον ορισμό των διαφορετικών τιμών συσχέτισης και από τη γενικότερη φύση της

Στατιστικής, η οποία απαιτεί συνεχώς όλο και μεγαλύτερα δείγματα για πιο αξιόπιστη πρόβλεψη. Και εμπειρικά όμως, δεν έχει νόημα να αναφερόμαστε σε συσχέτιση μεταξύ δύο μεμονωμένων τιμών, μιας και η απάντηση δε θα προσφέρει κάποια ουσιαστική πληροφορία. Για αυτό άλλωστε και όλα τα στατιστικά πακέτα που παρέχουν τέτοιους υπολογισμούς, απαιτούν τουλάχιστον 5 διαφορετικές τιμές για κάθεσύνολο δεδομένων.

Στην περίπτωση των δεδομένων καλαθιού, αντί για μία τιμή, προκύπτει ένας πίνακας συσχέτισης ο οποίος είναι συμμετρικός και σε όλη η κύρια διαγώνιος ισούται με τη μονάδα. Ένας τέτοιος πίνακας ουσιαστικά υπολογίζει για κάθε δυάδα διαφορετικών στηλών τη συσχέτιση μεταξύ του συνόλου των τιμών της κάθε μίας. Ο λόγος που η κύρια διαγώνιο έχει παντού τη μονάδα είναι το γεγονός πως αν συγκριθεί ένα σύνολο δεδομένων με τον εαυτό του τότε έχουμε πλήρη συσχέτιση και γενικότερα όταν παίρνουμε τιμή 1 για τη συσχέτιση, ονομάζεται τέλεια θετική συσχέτιση. Όσο πιο κοντά είναι η τιμή στη μονάδα, τόσο πιο σίγουροι είμαστε για να πούμε πως οι τιμές που εξετάστηκαν τείνουν σε θετική γραμμική συσχέτιση. Ακριβώς τα αντίστοιχα ισχύουν για τις αρνητικές τιμές με το -1 να αποκαλείται τέλεια αρνητική συσχέτιση. Η αρνητική συσχέτιση εμφανίζεται όταν οι τιμές των εξεταζόμενων συνόλων παρουσιάζουν διαφορετική συμπεριφορά, δηλαδή όταν οι μεν πρώτες αυξάνονται οι δε δεύτερες μειώνονται. Χαρακτηριστικό παράδειγμα αποτελεί η συσχέτιση της ηλικίας ενός ανθρώπου με το πλήθος των παιχνιδιών που κατέχει στη διάθεση του. Τέλος, όταν η αντίστοιχη τιμή είναι 0, δεν μπορεί να εξαχθεί κάποιο συμπέρασμα για τη συσχέτιση των προς εξέταση τιμών.

$$\begin{pmatrix} 1 & \cdots & \alpha \\ \vdots & \ddots & \vdots \\ \alpha & \cdots & 1 \end{pmatrix}$$

Πίνακας 1: ***Τυπική μορφή ενός πίνακα συσχέτισης με α ϵ [-1,1]***

Το πιο σύνηθες μέτρο της ανεξαρτησίας μεταξύ δύο ή περισσοτέρων ποσοτήτων είναι το Pearsonproduct-momentcorrelationcoefficientή συντελεστής συσχέτισης του Pearson [13]. Ο τρόπος που υπολογίζεται για κάθε ζευγάρι δεδομένων περιγράφεται από την παρακάτω σχέση:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E([(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

Όπου έχουμε τις εξής παραμέτρους:

- E: αναμενόμενη τιμή
- Cov: συνδιακύμανση
- μ_X, μ_Y: αναμενόμενη τιμή των μεταβλητών X και Y
- σ_X, σ_Y: τυπική απόκλιση των μεταβλητών X και Y
- $\rho_{X,Y}$: συντελεστής συσχέτισης μεταξύ των μεταβλητών X και Y

Όσον αφορά την αξιολόγηση σημαντικότητας αυτών των τιμών, η δημοφιλέστερο μέθοδος είναι ο υπολογισμός των p-values [14]. Αυτές οι τιμές κυμαίνονται μεταξύ του μηδέν και της μονάδας και υποδεικνύουν το κατά πόσον μπορεί να θεωρηθεί αξιόπιστη η τιμή που εξετάζεται, εν προκειμένω οι τιμές συσχέτισης. Προκειμένου να πραγματοποιηθεί αυτή η διαδικασία, ορίζεται μία nullhypothesisη οποία υποθέτει πως τα εξεταζόμενα σύνολα δεδομένων είναι πλήρως ασυσχέτιστα. Επομένως, οι p-valuesεξετάζουν με βάση ορισμένους ευρέως γνωστούς στατιστικούς πίνακες τα δείγματα και το πλήθος αυτόν και εκφράζουν αν και κατά πόσον αξιόπιστες είναι οι τιμές αυτοσυσχέτισης. Για παράδειγμα, αν η τιμή συσχέτισης έχει υπολογιστεί ίση με 0.254 για δύο σύνολα δεδομένων πλήθους 100, τότε μία p-valueίση με 0.02 σημαίνει πως υπάρχουν 2% πιθανότητες να εμφανιστεί αυτή η τιμή, αν τα δεδομένα ήταν ασυσχέτιστα.

3.3 Κατηγοριοποίηση και δένδρα απόφασης: Ταξινομητής C4.5

Το αντικέιμενο της κατηγοριόποιησης αποτελεί πιθανόν το πιο γνωστό και ευρύτερα διαδεδομένο κομμάτι της εξόρυξη γνώσης. Συχνές εφαρμογές που συναντάται και απαιτείται η χρήση της είναι η αυτόματη αναγνώριση σημάτων ομιλίας, ήχου και εικόνας για ιατρικές και όχι μόνο σκοπούς, η ανίχνευση αστοχιών σε βιομηχανικές διεργασίες, η εξέταση και πρόβλεψη οικονομικών τάσεων, η αναγνώριση καταναλωτικών συνηθειών μέσω της χρήσης των προσωπικών λογαριασμών και των κινήσεων αυτών, και προφανώς και η ανεύρεση προτύπων σε δεδομένα καλαθιού, ηλεκτρονικών και μη καταστημάτων. Για λόγους αποσαφήνισης, αρκεί να αναφερθεί πως οι διαδικασίες πρόβλεψης και εκτίμησης μπορούν να θεωρηθούν ως ειδικοί τύποι κατηγοριοποίησης [1]. Ουσιαστικά δηλαδή, ακόμα και σε περιπτώσεις πρόβλεψης, π.χ. στοιχηματικών αγώνων, η ε.γ μπορεί να βοηθήσει το χρήση να αυξήσει τις πιθανότητες επιτυχίας. Αξιοποιώντας στοιχεία όπως η προϊστορία, η φόρμα μίας ομάδας, το πλήθος τραυματιών και πιθανώς και με άλλες πληροφορίες, η εξόρυξη γνώσης προσπαθεί να αυξήσει την τελείως τυχαία πιθανότητα πρόβλεψης του νικητή μεταξύ των τριών διαφορετικών αποτελεσμάτων 1,X, ή 2 με πιθανότητα 33,3% εκάστη, σε κάτι καλύτερο. Φυσικά, αξιοποιώντας παρόμοια εργαλεία και οι εταιρίες παροχής αυτών των υπηρεσιών προσπαθούν να μειώσουν τις πιθανότητες νίκες του παίκτη.

Προσπαθώντας να δώσουμε ένα πιο επιστημονικό ορισμό ο οποίος να καλύπτει τις παραπάνω περιπτώσεις και όχι μόνο, μπορούμε να περιγράψουμε τη διαδικασία της κατηγοριοποίησης ως εξής:

Ορισμός: θεωρώντας ένα δεδομένο σύνολο δεδομένων $T = \{(x_i, c_i), i = 1 \dots N\}$με το x_i να συμβολίζει ένα διάνυσμα χαρακτηριστικών, διάστασης pγενικά, για το $i - οστό$ στοιχείο των δεδομένων και αντίστοιχα το $c_i \, \epsilon \, C = \{1, \dots K\}$όπου το K συμβολίζει τις διαφορετικές κλάσεις στις οποίες μπορεί να ανήκει το κάθε στοιχείο, η κατηγοριοποίηση ορίζεται ως η διαδικασία απεικόνισηςκάθε x_i σεμία από τις K κλάσεις [4].

Όσον αφορά ορισμένες έννοιες που χρησιμοποιήθηκαν για τον ορισμό, αξίζει να αναφερθεί πως σαν Tαναφέρουμε συνήθως ένα σύνολο δεδομένων εκπαίδευσης

(trainingdataset)το οποίο είτε καταγράφεται από τους ενδιαφερόμενους με σχετικά απλό τρόπο είτε απαιτείται ορισμένες φορές και κάποιος ειδικός ενός επιστημονικού χώρου να αναγνωρίσει την κατηγορία στην οποία ανήκουν τα στιγμιότυπα. Έχοντας στη διάθεση μας αυτό το σετ, μπορούμε είτε να εκτιμήσουμε την αξιοπιστία των εκάστοτε ταξινομητών είτε να το χρησιμοποιήσουμε σαν οδηγό για την πρόβλεψη της κλάσης σε νέα άγνωστα δεδομένα εισόδου (testset), ανάλογα και με το είδος της εφαρμογής. Επιπλέον, οι κλάσεις οι όποιές αναφέρθηκαν προηγουμένως είναι διακεκριμένες μεταξύ τους και δεν υπάρχει καμία επικάλυψη.

Για τη διαδικασία της κατηγοριοποίησης, ανάλογα με τη μέθοδο που επιλέγεται κάθε φορά, ένας ταξινομητής δέχεται το trainingset και υπολογίζει για κάθε στοιχείο της βάσης μία τιμή εκτίμησης (score) για κάθε $k \in C$. Στο τέλος αυτής της φάσης, η κλάση για την οποία έχει εκτιμηθεί η πιο ισχυρή τιμή είναι αυτή στην οποία κατηγοριοποιείται το κάθε στιγμιότυπο. Οι μέθοδοι που χρησιμοποιούνται για τη σωστή κατηγοριοποίηση, μπορούν να ομαδοποιηθούν στις επόμενες κατηγορίες:

- Καθορισμός των ορίων
- Χρήση κατανομών πιθανότητας
- Χρήση εκ των υστέρων πιθανοτήτων

Για περισσότερες τεχνικές λεπτομέρειες με αυτές τις μεθόδους, συνίσταται η ανάγνωση των [1], [3]. Όσον αφορά τα χαρακτηριστικά της κατηγοριοποίησης, αυτό που θα συζητηθεί εκτενέστερα εδώ είναι η μέτρηση της επίδοσης των εκατοστέ ταξινομητών, ώστε να μπορεί να κατανοηθούν τα ποσοστά που θα εμφανιστούν στο κεφάλαιο των αποτελεσμάτων. Ο πιο συνήθης τρόπος μέτρησης της επίδοσης, αποτελεί η ακρίβεια της κατηγοριοποίησης, όπως αυτή προσμετράτε ανάλογα με το ποια στιγμιότυπα ταξινομήθηκαν στις σωστές κλάσεις. Αξίζει να αναφερθεί πως πολλές φορές χρησιμοποιούνται και κάποια πολλαπλασιαστικά βάρη για τις διάφορες κατηγορίες που θα δούμε ευθύς αμέσως, αυξομειώνοντας τη σημαντικότητα ορισμένων επιτυχών ή ανεπιτυχών κατατάξεων. Στο παράδειγμα που ακολουθεί με δύο μόνο κατηγορίες, θα αρκεστούμε στην καταμέτρηση τα των δεδομένων χωρίς να αναθέσουμε βάρος σε καμία από τις 4 κατηγορίες.

Συνεπώς, για ένα τυπικό παράδειγμα που υπάρχουν μόνο 2 κατηγορίες, Α και Β, παρουσιάζεται παρακάτω ο λεγόμενος confusionmatrix:

2-Class Problem		Κλάση πρόβλεψης	
		A	B
Πραγματική Κλάση	A	f_{11}	f_{10}
	B	f_{01}	f_{00}

Πίνακας 2: ***Confusionmatrix για ένα πρόβλημα δύο κατηγοριών***

Διακρίνονται οι παρακάτω παράμετροι:

1. f_{11}: αληθώς θετικό
2. f_{01}: ψευδώς θετικό
3. f_{00}: αληθώς αρνητικό
4. f_{10}: ψευδώς αρνητικό

Οι δύο λοιπόν δείκτες που χρήζουν ιδιαίτερου ενδιαφέροντος είναι οι εξής:

$$Accuracy = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}} \quad (2)$$

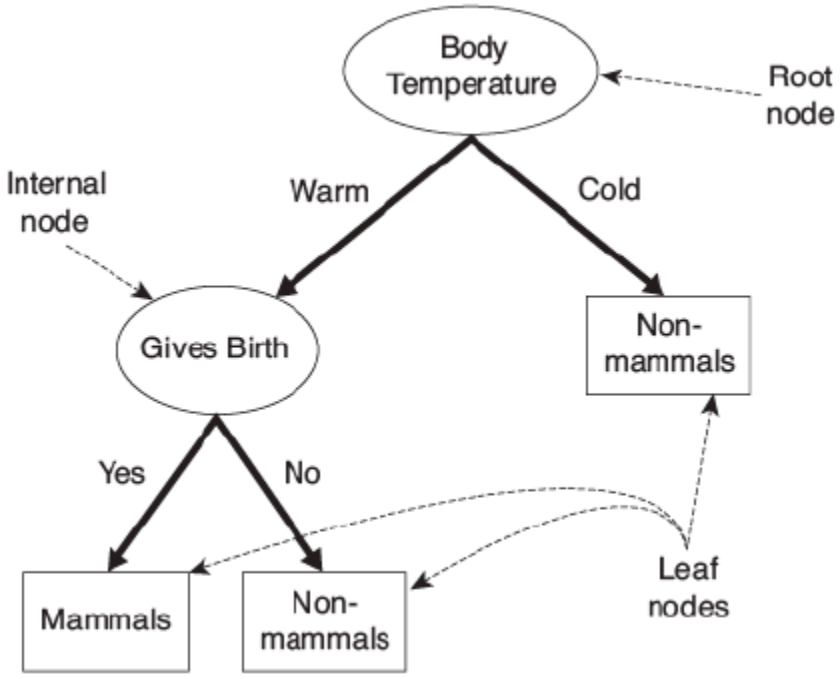
$$Error\ Rate = \frac{f_{01} + f_{10}}{f_{11} + f_{00} + f_{01} + f_{10}} \quad (3)$$

Για το δείκτη ακριβείς λοιπόν προσμετρώνται οι σωστά κατηγοριοποιημένες καταχωρήσεις προς το σύνολο αυτών, ενώ για το δείκτη σφάλματος οι λανθασμένες προς το σύνολο αντίστοιχα.

Στα πλαίσια τηςσυγκεκριμένης εργασίας, αποφασίστηκε πως θα δοθεί ιδιαίτερο βάρος στα δένδρα απόφασης, οπότε δε θα αναλυθούν οι άλλες οικογένειες αλγορίθμων που χρησιμοποιούνται επίσης κατά κόρον για την επίλυση προβλημάτων εξόρυξη γνώσης. Τα ΄δένδρα απόφασης επομένως αποτελούν μία πολύ χρήσιμη προσέγγιση αυτού του είδους των προβλημάτων. Η λογική που ακολουθείται είναι αυτή των συνεχών ερωτήσεων. Ξεκινώντας δηλαδή με κάποιο κριτήριο από τη σημαντικότερη κατηγορία που διέπει τη βάση δεδομένων μας, ρωτάμε για το είδος

αυτής της κατηγορίας, περιμένοντας ένα ναι ή ένα όχι αν αντιμετωπίζουμε ένα δυαδικό πρόβλημα, ή περισσότερες απαντήσεις για ένα nominal πρόβλημα. Έχοντα πάρει τις διαφορετικές απαντήσεις κάθε φορά συνεχίζουμε να ρωτάμε για επόμενα σημαντικότερα χαρακτηριστικά, μέχρις ότου να καταλήξουμε σε ένα συμπεράσματα για σχετικά με την κλάση στην οποία ανήκει το εκάστοτε στοιχείο. Αυτή η περιγραφή οδηγεί στην ανάπτυξη δένδρων αποφάσεων, τα οποία αποτελούν ορισμένες ιεραρχικά δομημένες δομές με που περιλαμβάνουν κόμβους και κατευθυνόμενες ακμές. Τα 3 διαφορετικά είδη κόμβων που συναντώνται είναι τα εξής [3]:

1. Rootnode: κόμβος με καμία εισερχόμενη ακμή και καμία ή περισσότερες εξερχόμενες
2. Internalnodes: κόμβοι με μία εισερχόμενη ακμή και δύο ή περισσότερες εξερχόμενες ακμές
3. Leafή Terminalnode: κόμβοι με καμία εξερχόμενη ακμή και ακριβώς μία εισερχόμενη.



Εικόνα 11: τυπικό δένδρο απόφασης για κατηγοριοποίηση σε θηλαστικά ή μη

Προκειμένου να ολοκληρωθεί η επίλυση ενός προβλήματος κατηγοριοποίησης με τη βοήθεια δένδρα απόφασης εφαρμόζονται τα ακόλουθα βήματα:

- Επαγωγή δένδρου απόφασης: χτίσιμο ενός δένδρα απόφασης με τη χρήση των trainingdata.
- Εφαρμογή του δένδρα απόφασης για κάθε στιγμιότυπο με σκοπό τη πρόβλεψη της κλάσης στην οποία αυτό ανήκει.

Τα πλεονεκτήματα που προκύπτουν από τη χρήσηδένδρα απόφασης για το ζήτημα της κατηγοριοποίησης είναι η εύκολη χρήση και η ικανοποιητική συνήθως αποτελεσματικότητα που παρέχουν. Επιπλέον, λόγω της αύξησης της υπολογιστικής ισχύς τα τελευταία χρόνια, είναι αρκετά πιθανό ακόμα και για χιλιάδες εγγραφές η δημιουργία ενός μοντέλου σε πορισμένα δευτερόλεπτα, εξασφαλίζοντας πολύτιμο χρόνο και παρέχοντας γρήγορα μία εικόνα για την ποιότητα της ακρίβειας του εκάστοτεδένδρα απόφασης. Επίσης, έχοντας χτίσει ένα δένδρο, ο χρόνος που απαιτείται για να εκτιμηθεί η κλάση μίας νέα εγγραφής είναι ανάλογος με το ύψος του δένδρου, το οποίο είναι δεδομένο για κάθε αλγόριθμο και επομένως εκτελείται σε σταθερό χρόνο. Τέλος, λόγω της εύκολης απεικόνισης που συνοδεύει τα δένδρα απόφασης, είναι σχετικά απλή η παρατήρηση και η κατανόηση των κανόνων με βάση τους οποίους χτίστηκε το κάθε δένδρο, ακόμα και από άτομα με μικρή εμπειρία στο χώρο. Τα μειονεκτήματα από την άλλη είναι η αδυναμία χειρισμού συνεχών δεδομένων και η ελλαττωματική αντιμετώπιση των ελλιπών δεδομένων [1].

Τα κύρια χαρακτηριστικά των δένδρα απόφασης μπορούν να αναζητηθούν στις πηγές [1], [3] και [4]. Σε αυτό το σημείο θα αναφερθούν τα κύρια χαρακτηριστικά του αλγορίθμου C4.5 [15], ο οποίος μάλιστα θα χρησιμοποιηθεί στην πειραματική διαδικασία που θα ακολουθήσει αυτού του κεφαλαίου. Τα καινοτόμα στοιχεία που εισήγαγε ο συγκεκριμένος αλγόριθμος σε σχέση με τα υπάρχοντα δένδρα απόφασης είναι τα επόμενα:

- Ελλιπή δεδομένα: αγνόηση αυτών των δεδομένων κατά το χτίσιμο του δένδρα απόφασης. Για την πρόβλεψη μίας καταχώρησης με ελλιπή δεδομένα, γίνεται με βάση τις γνωστές για αυτό πληροφορίες σε σχέση με τις υπόλοιπες εγγραφές.
- Συνεχή δεδομένα: χωρισμός των δεδομένων σε κατάλληλα διαστήματα με βάση τις ιδιαιτερότητες των γνωρισμάτων.

- Κλάδεμα: είτε με την αντικατάσταση υποδένδρου (subtreereplacement) είτε με την ανύψωση δένδρου (subtreeraising).
- Κανόνες: απλούστευση των κανόνων που εξάγονται για μονοπάτια με κοινές συμπεριφορές.
- Διάσπαση: προκειμένου να αποφευχθεί η υπερπροσαρμογή, εφαρμόζεται το GainRatioτο οποίο ορίζεται ως εξής:

$$GainRatio\ (D,C) = \frac{Gain(D,C)}{H(\frac{|D_1|}{|D|}, \dots, \frac{|D_c|}{|D|})} \quad (4)$$

$$Gain(D,C) = H(D) - \sum_{I=1}^{C} P(D_i) * H(D_i) \quad (5)$$

$$H(p_1, p_{2,\ \dots} p_c) = \sum_{i=1}^{c} p_i * \log(1/p_i) \quad (6)$$

Όπου το Dαναπαριστά μία κατάσταση της βάσης δεδομένων, ενώ το H(D) ορίζει ένα μέτρο της τάξης. Με αυτήν την τακτική, λαμβάνεται υπόψιν η πληθικότητα της κάθε διάσπασης, αποτελώντας ένα πιο δίκαιο τρόπο κατηγοριοποίησης.

3.4 Κανόνες συσχέτισης: Αλγόριθμος Apriori

Η εξόρυξη γνώσης μέσω κανόνων συσχέτισης αποτελεί μία σημαντική τεχνική για την ανακάλυψη ουσιαστικών προτύπων σε βάσεις δεδομένων, όπως αυτές δημιουργούνται σε περιπτώσεις καταγραφής συναλλαγών. Ένας κανόνας συσχέτισης δεν είναι τίποτε άλλο παρά ένας κανόνας της μορφής $X \Rightarrow Y$, όπου τα X,Yδύο ξένα μεταξύ τους σύνολα αντικειμένων (itemsets), δηλαδή $X \cap Y = \emptyset$ και φυσικά τα X,Y αποτελούν υποσύνολα του συνόλου των δεδομένων. Με την έννοια κανόνας εννοούμε πως αν βρεθούν όλα τα στοιχεία του X σε μία συναλλαγή, είναι αρκετά πιθανό σε αυτή τη συναλλαγή να περιλαμβάνονται και τα αντικείμενα του Y.

Μία τυπική εφαρμογή της εξόρυξης κανόνων συσχέτισης είναι το παράδειγμα της ανάλυσης δεδομένων καλαθιού με σκοπό την ανεύρεση συσχετίσεων για την τελική απόφαση όσον αφορά τη διάταξη των αντικειμένων σε καταστήματα λιανικής, για προσωποκεντρικές συστάσεις προϊόντων και για την υιοθέτηση νέων προωθητικών τάσεων ή και απόφαση περί συνέχισης η διακοπής μίας καμπάνιας. Φυσικά υπάρχουν και άλλα πεδία που βρίσκουν εφαρμογή αυτές οι τακτικές, όπως τα web-basedσυστήματα, όπου εκεί αναζητούνται συσχετίσεις μεταξύ ιστοσελίδων, αρχείων, άρθρων και διαφημίσεων [17].

Για την επιλογή τώρα των κανόνων συσχέτισης από το σύνολο όλων των πιθανών κανόνων χρησιμοποιούνται κριτήρια και δείκτες που καθορίζονται με βάση το επίπεδο ενδιαφέροντος που υπάρχει, πάντα σε ένα πλαίσιο στατιστικής εμπιστοσύνης. Το πρωταρχικό στοιχείο μέτρησης σημαντικότητας αποτελεί το support, το οποίο ορίζεται ως το κλάσμα των όλων των συναλλαγών που περιέχουν έστω και ένα αντικείμενο ενδιαφέροντος προς το συνολικό πλήθος των συναλλαγών στο προς εξέταση πείραμα:

$$supp(X \Rightarrow Y) = supp(X \cup Y) = \frac{count(X \cup Y)}{m} \quad (7)$$

Ένα εξίσου σημαντικό μέτρο για τους κανόνες συσχέτισης αποτελεί το confidenceή αλλιώς strength, το οποίο ορίζεται ως το κλάσμα του αριθμού των συναλλαγών που περιέχουν το $X \cup Y$ προς τον αριθμό των συναλλαγών που περιέχουν το X μόνο:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} = P(Y|X) \quad (8)$$

Μιλώντας γενικά για τους αλγορίθμους κανόνων συσχέτισης, απαιτείται κάθε φορά ορισμός ενός ελάχιστου ορίου υποστήριξης (minimumsupportthreshold) το οποίο χρησιμοποιείται για την επιλογή των πιο συχνών και πιθανότατα σημαντικών συνδυασμών αντικειμένων, συχνά αποκαλούμενο ως frequentitemsets. Ένα σημαντικό ζήτημα που καλούνται όλοι αυτοί οι αλγόριθμοι να επιλύσουν είναι το πώς η διαδικασία εύρεσης frequentitemsets σε μεγάλες βάσεις δεδομένων είναι η αυξημένη, στη χειρότερη μάλιστα περίπτωση εκθετικά. πολυπλοκότητα που

προκύπτει σε αυτά. Μία από τις πιο διαδεδομένες και επαρκής αντιμετώπιση αυτού του ζητήματος, δημιούργησε τον αλγόριθμο Apriori [16], ο οποίος περιγράφεται παρακάτω.

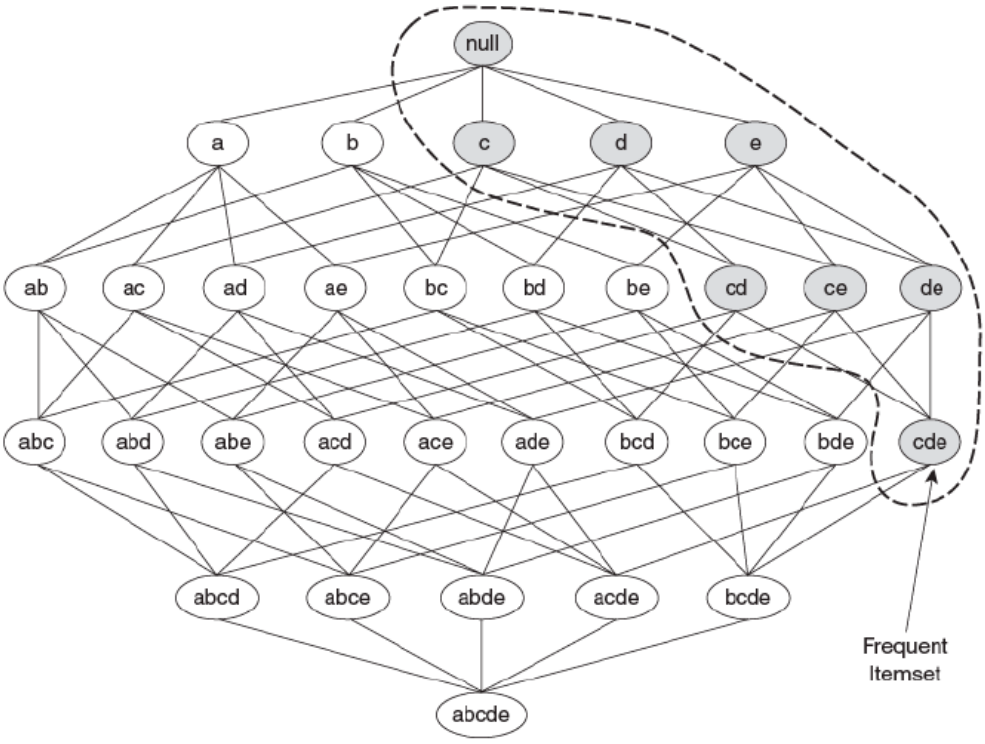
Ο αλγόριθμος Aprioriαποτελεί ένα ισχυρό αλγόριθμό για ανεύρεση και εξόρυξη frequentitemsets για Booleanκανόνες συσχέτισης. Τα βασικά χαρακτηριστικά αυτού είναι τα ακόλουθα:

- Εκτελεί στην χειρότερη περίπτωση τόσα περάσματα όσο το πλήθος των διαφορετικών αντικειμένων.
- Κάθε υποσύνολο ενός συχνού συνολοστοιχείου πρέπει υποχρεωτικά να είναι και αυτό συχνό (όπως ορίστηκε προηγούμενα).
- Σε κάθε διαδοχική προσπέλαση χρησιμοποιούνται τα συχνά συνολοστοιχεία του προηγούμενου περάσματος με στόχο να δημιουργηθούν καινούργια συνολοστοιχεία.
- Λειτουργία συνένωσης: για να δημιουργηθεί ένα νέο υποσύνολο στο κ-οστό βήμα L_k, πρέπει να δημιουργηθούν κ υποψήφια συνολοστοιχεία τα οποία συνενώνονται με το L_{k-1} για να εξετασθούν αν ικανοποιούν το κριτήριο του συχνού συνολοστειχίου.

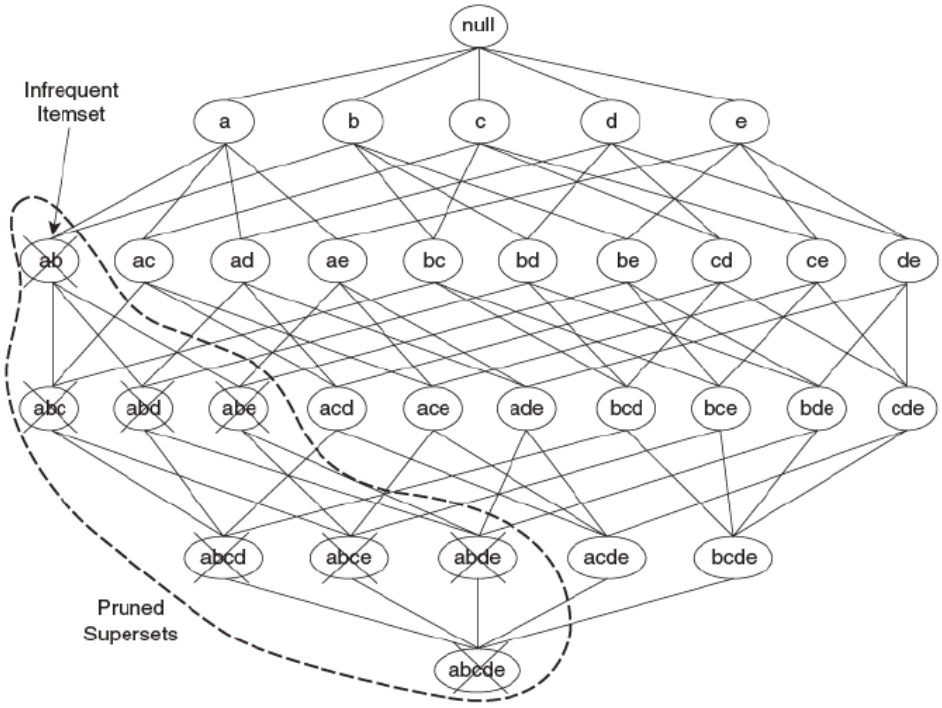
Το όνομα του αλγορίθμου καθιερώθηκε γιατί, όπως άλλωστε φαίνεται και από την περιγραφή του, χρησιμοποιεί εκ των προτέρων γνώση για τα εξεταζόμενα συνολοστοιχεία. Αποτελεί δηλαδή έναν επαναληπτικό αλγόριθμο, ο οποίος με βάση κ δεδομένα συνολοστοιχεία, εξετάζει κ+1 συνολοστοιχεία. Ξεκινάει δηλαδή από τα σύνολα μεγέθους ίσο με τη μονάδα (L_1) και στηριζόμενος σε αυτά υπολογίζει τα L_2. Έπειτα με βάση τα L_2 προχωράει στα L_3κ.ο.κ επίσης, προκειμένου να μειωθεί το μέγεθος των συνόλων και προφανώς και των αναζητήσεων σε κάθε βήμα, επιλέγεται μία διαδικασία κλαδέματος (pruning). Η βασική ιδέα που διακατέχει αυτή την ενέργεια είναι πως κάθε υπερσύνολο ενός μη συχνού στοιχεισυνόλου, δε μπορεί να χαρακτηρισθεί ως συχνό σύνολοστοιχειο. Έτσι σχηματίζονται τα C_iτα οποία είναι υπερσύνολα όλων των frequentitemsetsσε κάθε βήμα, χωρίς όμως να μην περιέχονται και μη συχνά στοιχεισυνολά σε αυτά. Ο ψευδοκώδικας που περιγράφει τον αλγόριθμο Aprioriείναι ο ακόλουθος:

```
procedure Apriori (T, minSupport)

{
 //T is the database and minSupport is the minimum support

L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++)
{
      Ck= candidates generated from Lk-1

      //that is cartesian product Lk-1 x Lk-1 and
      //eliminating any k-1 size itemset that is not
      //frequent

      for each transaction t in database do
      {
            increment the count of all candidates in Ck
            that are contained in t

            Lk = candidates in Ck with minSupport

      }//end for each
}//end for

return ;

}
```

Επομένως, η παράμετρος minSupportεφαρμόζεται για την εύρεση των frequentitemsetsκαι εν συνεχεία αυτά τα σύνολα μαζί με τον περιορισμό περί ελάχιστης εμπιστοσύνης χρησιμοποιούνται για το σχηματισμό κανόνων. Ο αλγόριθμοςσταματά όταν δεν υπάρχουν περαιτέρω επιτυχείς εκτάσεις των συνολοστοιχείων. Για την αναζήτηση χρησιμοποιείται τακτική αναζήτησης κατά πλάτος και δομή δένδρου για τα υποψήφια συνολοστοιχεία. Ακολουθούν δύο χαρακτηριστικές εικόνες για τις περιπτώσεις που προαναφέρθηκαν:



Εικόνα 12: Απεικόνιση ενός συχνού συνολοστοιχείου με όλα τα υποσύνολα του να είναι επίσης συχνά



Εικόνα 13: Απεικόνιση ενός μη συχνού συνολοστοιχείου με όλα τα υπερσύνολα του να είναι επίσης μη συχνά

ΚΕΦΑΛΑΙΟ 4

Αποτελέσματα

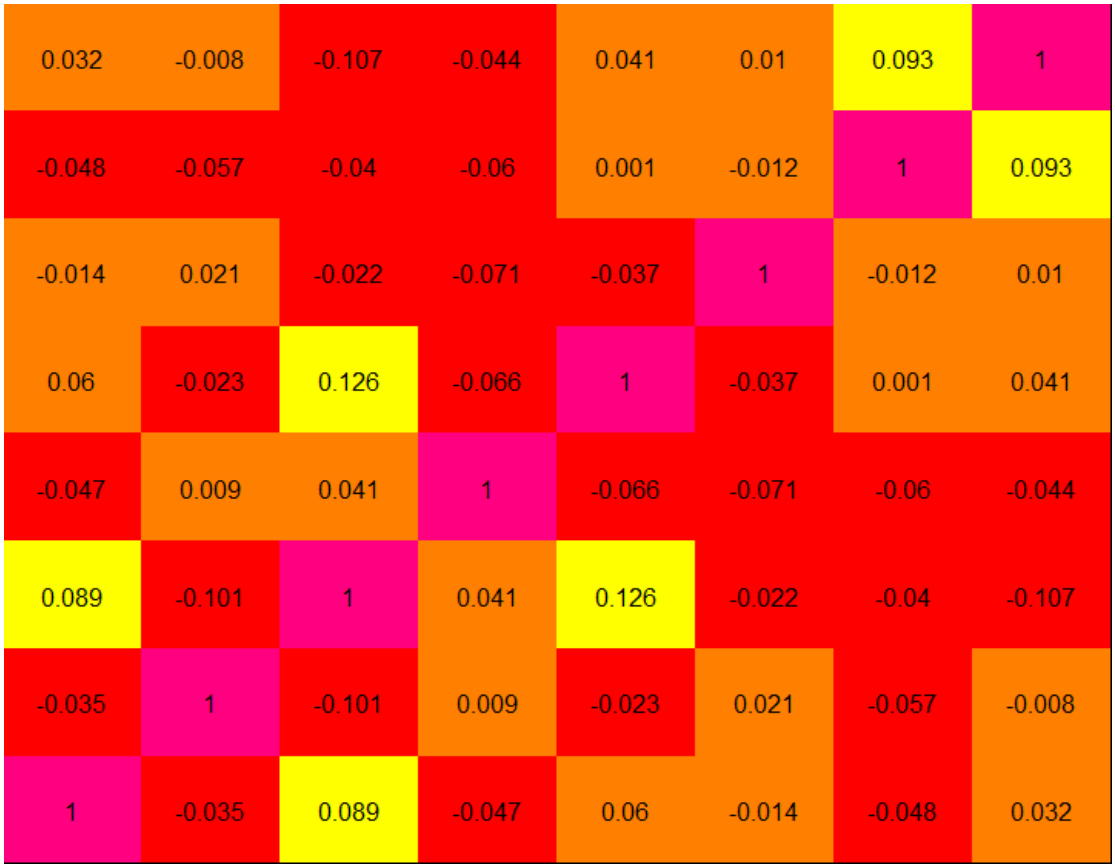
4.1 Υπολογισμός Πίνακα Συσχέτισης

Ο υπολογισμός του correlationmatrix μαζί με τις p-values έγινε στο περιβάλλον της Rκαι πιο συγκεκριμένα με τη βοήθεια του opensourceplatform τηςR, τοRStudio. Οι βιβλιοθήκες που χρειάστηκαν για να πραγματοποιηθούν οι υπολογισμοί αλλά και το γραφικό κομμάτι είναι οι εξής:

- corrplot
- Hmisc
- arules
- arulesViz
- rrcov
- PerformanceAnalytics
- plyr
- Rcpp
- RcolorBrewer
- igraph
- irlba
- colorspace

Για την εισαγωγή των δεδομένων στο RStudioχρησιμοποιήθηκε η .csvμορφή των δεδομένων μας και η εντολή read.csvτης R. Ενώνοντας τις στήλες των δεδομένων μας σε μορφή dataframeυπολογίσαμε το correlationmatrixόπως φαίνεται και στην εικόνα παρακάτω, προσθέτοντας αποχρώσεις του κόκκινου για τις διάφορες τιμές.Κοιτώντας

αυτόν τον πίνακα, παρατηρούμε πως η κύρια διαγώνιος είναι γεμάτη με άσσους καθώς και πως ο πίνακας είναι συμμετρικός. Στο κάτω αριστερό μέρος του πίνακα βρίσκεται η κατηγορία Π1 ενώ στο πάνω δεξιά η κατηγορία Π8.



Πίνακας 3: ***Πίνακας συσχέτισης με χρήση colormap***

Όπως αναφέρθηκε και στο θεωρητικό κομμάτι, η παρουσίαση μόνο των τιμών συσχέτισης δεν αποτελεί αξιόπιστο μέτρο. Για αυτό το λόγο θα χρησιμοποιηθούν και οι p-valuesγια την αποσαφήνιση της εμπιστοσύνης προς τη συμπεριφορά των δεδομένων. Παρακάτω παρουσιάζονται οι τιμές αυτές καθώς και κάποια διαγράμματα σχετικά με την εμπιστοσύνη αυτών:

p - values

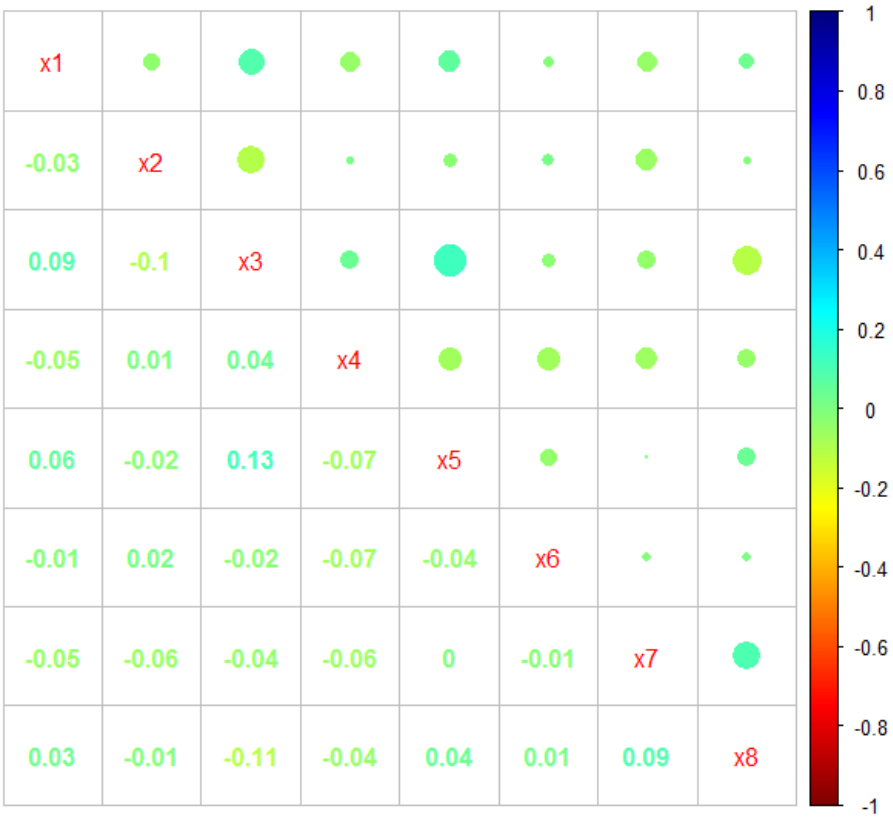
	Π1	Π2	Π3	Π4	Π5	Π6	Π7	Π8
Π1		0.5323	0.1099	0.4040	0.2826	0.8050	0.3876	0.5655
Π2	0.5323		0.0690	0.8750	0.6865	0.7044	0.3111	0.8898
Π3	0.1099	0.0690		0.4672	0.0235	0.6989	0.4775	0.0536
Π4	0.4040	0.8750	0.4672		0.2377	0.2013	0.2788	0.4289

```
      П5 0.2826 0.6865 0.0235 0.2377        0.5083 0.9880 0.4638
      П6 0.8050 0.7044 0.6989 0.2013 0.5083        0.8362 0.8536
      П7 0.3876 0.3111 0.4775 0.2788 0.9880 0.8362        0.0943
П8 0.5655 0.8898 0.0536 0.4289 0.4638 0.8536 0.0943
```

*Πίνακας 4:**p-values για το σύνολο των προϊόντων***

A/A	**Γραμμή**	**Στήλη**	**ΣυντελεστήςΣυσχέτισης**	**p-value**
1	x1	x2	-0.03487419	0.53228217
2	x1	x3	0.089130439	0.10985538
3	x2	x3	-0.101307675	0.06901158
4	x1	x4	-0.046590492	0.40397635
5	x2	x4	0.008788568	0.87497643
6	x3	x4	0.04059859	0.46715423
7	x1	x5	0.059960172	0.28264102
8	x2	x5	-0.022543639	0.68647794
9	x3	x5	0.125982612	0.0235473
10	x4	x5	-0.065880694	0.23771642
11	x1	x6	-0.013786981	0.80503708
12	x2	x6	0.021186473	0.7044399
13	x3	x6	-0.021602102	0.69892109
14	x4	x6	-0.071290016	0.20128548
15	x5	x6	-0.03693508	0.50831934
16	x1	x7	-0.04823095	0.38761059
17	x2	x7	-0.056539904	0.31105176
18	x3	x7	-0.039659299	0.47752985
19	x4	x7	-0.060436588	0.27882719
20	x5	x7	0.000839797	0.98800467
21	x6	x7	-0.01155056	0.83617302
22	x1	x8	0.032096122	0.56546058
23	x2	x8	-0.00773857	0.88981207
24	x3	x8	-0.107485928	0.05362371
25	x4	x8	-0.044170227	0.42885951
26	x5	x8	0.040905081	0.46379578
27	x6	x8	0.010306354	0.85360973
28	x7	x8	0.093242124	0.09434388

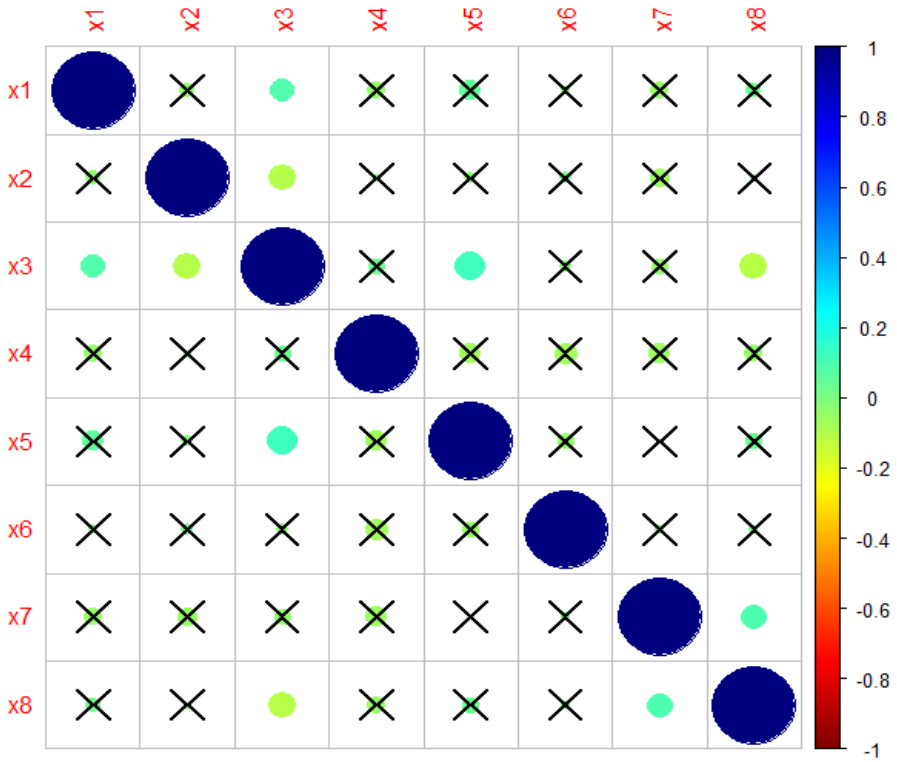
*Πίνακας 5: **Καταγραφή των συντελεστών συσχέτισης και των p-values***



Πίνακας 6: ***Μεικτός πίνακας συσχέτισης με τιμές και κύκλους διαφορετικής ακτίνας για αναπαράσταση***

Ακολουθούν δύο τελευταίαδιαγράμματα, στα οποίο έχοντας υπολογίσει τις p-values, παραβλέπουμε όσες ξεπερνούν το επιθυμητό επίπεδο εμπιστοσύνης. Για λόγους απεικόνισης, το οποίο είναι το τελευταίο στάδιο στην αλυσίδα του KDD, μπορεί κανείς χρησιμοποιώντας τις βιβλιοθήκες της Rνα παραμετροποιήσει τα διαγράμματα του καθώς και να επιλέξει μεταξύ μίας μεγάλης γκάμας γραφικών προτύπων. Ψάχνοντας άλλωστε κανείς στη βιβλιογραφία θα παρατηρήσει πω υπάρχουν αρκετές διαφορετικές τεχνοτροπίες παρουσίασης σχετικών αποτελεσμάτων, ανάλογα με το πλήθος των δειγμάτων αλλά και το ύφος και το χαρακτήρα της παρουσίασης και του επιστημονικού επιπέδου του ενδιαφερόμενου κοινού.

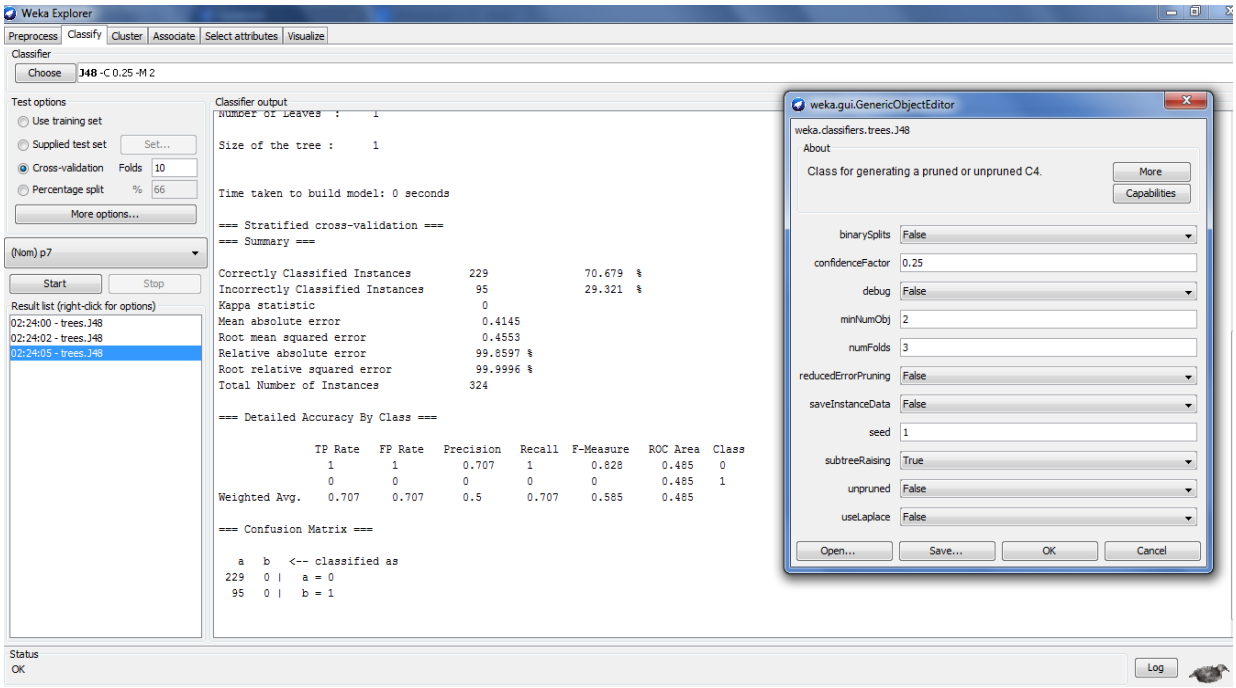
Πίνακας 7: ***Απογραφή των ζευγαριών που έχουν p-value πάνω από 0.5***



Πίνακας 8: ***Απόρριψη ζευγαριών που παρουσιάζουν p-value πάνω από 0.20.***

4.2 Κατηγοριοποίηση των διαφορετικών κλάσεων

Για του συγκεκριμένο κομμάτι χρησιμοποιήθηκαν τα arffαρχεία μέσω του λογισμικού WEKA.Σε αυτό επιλέγοντας στην καρτέλα Classificationτην κατηγορίας treeskαι έπειτα τον J48 [15]έχουμε τη δυνατότητα να ταξινομήσουμε τα δεδομένα μας στις διαφορετικές κλάσεις μέσω του αλγορίθμου C4.5 να δημιουργήσουμε ένα prunedunpruneddecisiontree. Στην παρακάτω εικόνα φαίνονται οι παράμετροι του αλγορίθμου. Σαν κριτήριο αξιολόγησης χρησιμοποιείται το 10-crossvalidation.

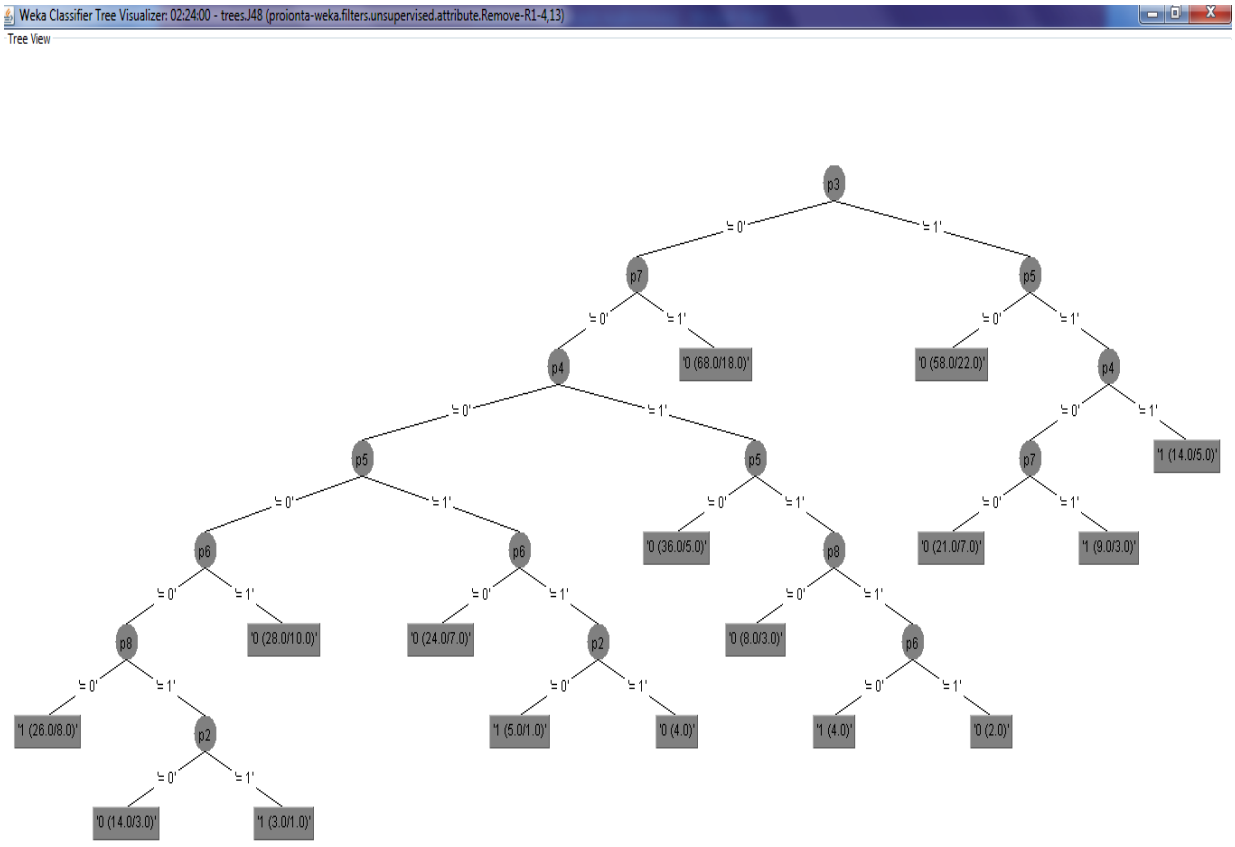


Εικόνα 14: Απεικόνιση του περιβάλλοντος WEKAγια εφαρμογή του C4.5

Τρέχοντας κάθε φορά το ίδιο πείραμα με τις defaultτιμές του ταξινομητή J48 και αλλάζοντας την κλάση για την οποία ψάχνουμε την πρόβλεψη, καταγράψαμε τις τιμές που φαίνονται στον επόμενο πίνακα για την ακρίβεια πρόβλεψης. Ακολουθεί επίσης η απεικόνιση του δένδρου απόφασης που σχηματίστηκε για την κλάση Π3, όπως αυτό δημιουργείται από το WEKA.

Κατηγορία προς πρόβλεψη	Ποσοστό ακρίβειας (%)
Π1	63.5802
Π2	60.4938
Π3	67.5926
Π4	65.7407
Π5	62.345
Π6	71.9135
Π7	70.6079
Π8	66.0494

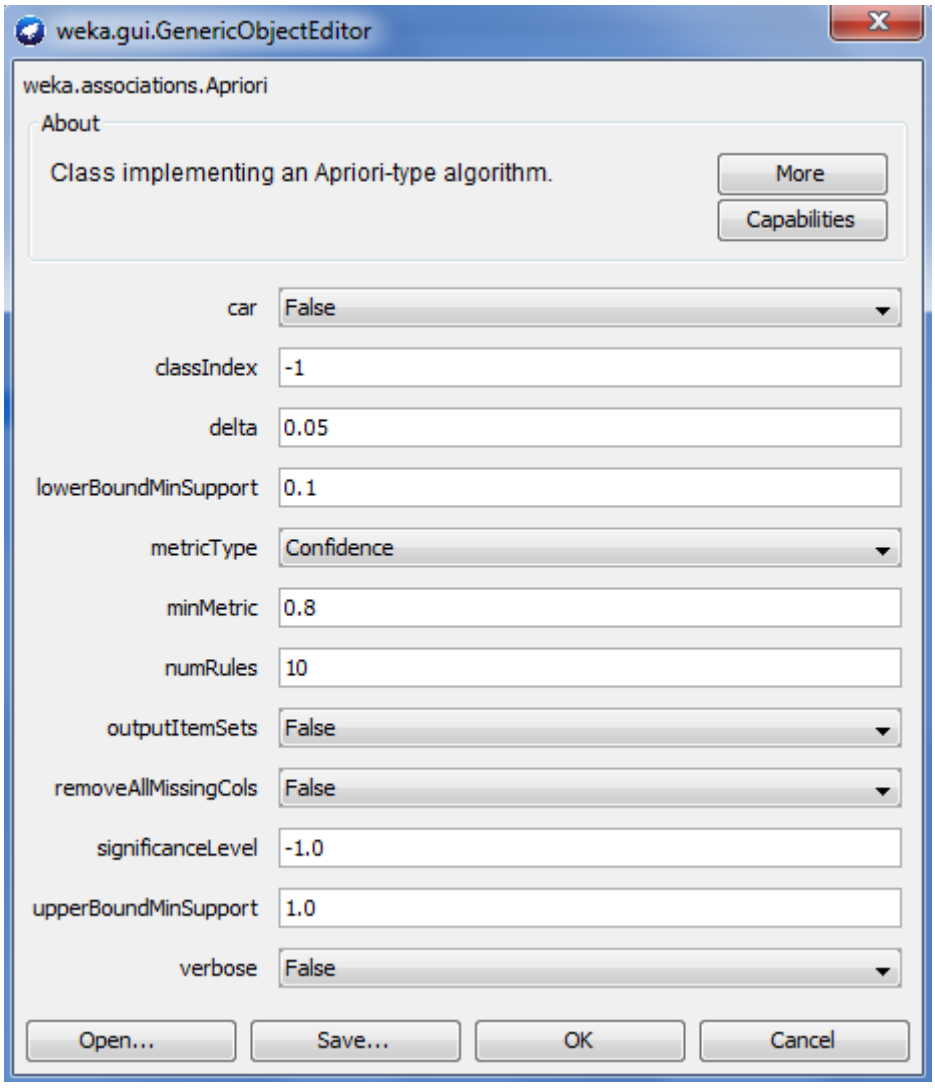
Πίνακας 9: ***Ποσοστά επιτυχίας για κάθε κλάση του C4.5***



Εικόνα 15: Δένδρο απόφασης για την κατηγορία Π3

4.3 Εξαγωγή κανόνων συσχέτισης

Για την εξαγωγή κανόνων συσχέτισης, χρησιμοποιώντας το ίδιο arffαρχείο και επιλέγοντας αυτή τη φορά την καρτέλα Associate, επιλέγουμε τον αλγόριθμο Apriori [16]. Το παράθυρο με τις διάφορες επιλογές που προσφέρονται φαίνεται παρακάτω. Η μόνη τροποποίηση που κάναμε είναι η αλλαγή του minMetricστην τιμή 0.80, έτσι ώστε να αποκτήσουμε περισσότερους κανόνες, μιας και υψηλότερες τιμές δεν επέστρεφαν αρκετό πλήθος ή και καθόλου κανόνες, έτσι ώστε να περιγραφθεί η συσχέτιση των δεομένων μας. Ακολουθεί επίσης το σχετικό μήνυμα του WEKA, το οποίο μας πληροφορεί για τους κανόνες και το επίπεδο εμπιστοσύνης που αυτή πετυχαίνουν:



Εικόνα 16: Παράμετροι επιλογής του Aprioriαλγορίθμου

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: proionta-weka.filters.unsupervised.attribute.Remove-R1-4,13
Instances: 324
Attributes: 8
p1
p2
p3
p4
p5
p6
p7
p8
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.15 (49 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 84

Size of set of large itemsets L(3): 140

Size of set of large itemsets L(4): 76

Best rules found:

1. p3=1 p6=0 p7=0 57 ==> p8=0 49 conf:(0.86)
2. p1=1 p6=0 p8=0 59 ==> p7=0 49 conf:(0.83)
3. p3=1 p7=0 p8=0 60 ==> p6=0 49 conf:(0.82)
4. p3=1 p6=0 75 ==> p8=0 61 conf:(0.81)
5. p1=0 p8=1 64 ==> p3=0 52 conf:(0.81)
6. p6=0 p8=1 74 ==> p3=0 60 conf:(0.81)
7. p2=1 p6=0 p8=0 63 ==> p7=0 51 conf:(0.81)
8. p1=1 p7=0 p8=0 61 ==> p6=0 49 conf:(0.8)
9. p3=1 p6=0 p8=0 61 ==> p7=0 49 conf:(0.8)
10. p4=1 p7=0 80 ==>p6=0 64 conf:(0.8)

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα

5.1 Σχολιασμός των αποτελεσμάτων

Στο συγκεκριμένο κομμάτι θα συνοψιστούν τα σημαντικότερα συμπεράσματα που προέκυψαν κατά την έρευνα μας πάνω στο κομμάτι των δεδομένων καλαθιού και την ανεύρεση γνώσης σε αυτά. Επομένως έχουμε:

- Η μικρή συλλογή δεδομένων αποτέλεσε ανασταλτικό παράγοντα για την εξαγωγή σημαντικών συσχετίσεων, μιας και κανένας συνδυασμός δεν επέφερε δείκτη εμπιστοσύνης μεγαλύτερο του 90%.
- Η συνεχόμενη περίοδος μάλιστα συλλογής δεδομένων αποτελεί επίσης σημαντικό ανασταλτικό παράγοντα, μιας και οι αντίστοιχες έρευνες που έχουν γίνει προσπαθούν να αποφύγουν την εξάρτηση από περιόδους εορταστικές ή περιόδους που κοινωνικά γεγνότα μπορούν να επηρεάσουν το πείραμα.
- Η μικρή επίσης ποικιλία των προϊόντων δεν μπορεί να συμβάλει στον ανταγωνισμό του γειτονικού καταστήματος στο οποίο συλλέχθηκαν τα δεδομένα με τα μεγαλύτερα και πιο διαφημισμένα καταστήματα που προσφέρουν παρόμοια και ίδια αγαθά. Άλλωστε σε παρόμοιες και πιο πετυχημένες συλλογές παρατηρούμε τεράστια διαφορά ως προς την οικονομική ισχύ των εταιριών τους σε σχέση με το κατάστημα το οποίοαποτέλεσε πηγή συλλογής πληροφοριών για εμάς.
- Ο στατιστικόςέλεγχος που έγινε απέδειξε άλλωστε και την ασθενή συσχέτιση που παρουσιάζουν τα δεδομένα.
- Οι 2 κατηγορείς προϊόντων που παρουσίασαν μεγαλύτερα ποσοστά πρόβλεψης της αγοράς ή μη των προϊόντων τους ήταν οι Π6 & Π7 (Αλκοολούχα Ποτά και Καθαριστικά Σπιτιού αντίστοιχα).

- Τα χαμηλά ποσοστά επιτυχίας (≈71%) αποδεικνύουν τα αποτελέσματα του στατιστικού ελέγχου περί χαμηλής συσχέτισης.
- Οι κατηγορίες που προσέφεραν την υψηλότερη πληροφορία για τους διαχωρισμούς των δένδρων που δημιουργήθηκαν είναι η Π3 (Ζυμαρικά – όσπρια) και οι Π6, Π7 που αναφέρθηκαν προηγουμένως.
- Ο πιο ισχυρός κανόνας που εμφανίσθηκε ήταν ο εξής: αν κάποιος/α έχει αγοράσει προϊόν/ντα της κατηγορίας Π3 και δεν αγοράσει από τις κατηγορίες Π6, Π7 τότε δε θα προβείς την αγορά από την κατηγορία Π8 με επίπεδο εμπιστοσύνης 86%.
- Ο κατά σειρά δεύτεροισχυρότερος κανόνας είναι πως αν κάποιος/α αγοράσει από την κατηγορία Π1 και δεν αγοράσει από τις κατηγορίες Π6, Π8 τότε με εμπιστοσύνη 83% δε θα προβεί στην αγορά προϊόντων από την κατηγορία Π7.

5.2 Ανοικτά θέματα προς διερεύνηση

Στα πλαίσια της έρευνας και της αναζήτησης και ανάγνωσης των βιβλιογραφικών πηγών που αναφέρθηκαν καθόλη τη διάρκεια του κειμένου και βρίσκονται και συγκεντρωμένες στο πεδίο της Βιβλιογραφίας, προέκυψαν διάφορα ερωτήματα και προτάσεις από ειδικούς του χώρου σχετικά με θέματα που αφορούν το πεδίο της εξόρυξης γνώσης και τα δεδομένα καλαθιού. Παρακάτω συνοψίζονται ορισμένεςμελλοντικέςπροτάσεις που μπορούν αν αποτελέσουν πρόταση προς περαιτέρω έρευνα και αναζήτηση νέας χρήσιμης γνώσης:

- Η αξιοποίηση της χρονικής συμπεριφοράς των στιγμιοτύπων θα μπορούσε να δημιουργήσει χρησιμότερες συσχετίσεις.
- Η εισήγηση χρονικών πιθανοτικών μοντέλων θα μπορούσε να προσομοιάσει με μεγαλύτερη ακρίβεια τη ρεαλιστική συμπεριφορά των καταναλωτών.
- Η καταγραφή και η αξιοποίηση προσφορών και διαφορές τιμών μεταξύ συνεχόμενων ημερών η εβδομάδων πιθανόν θα μπορούσε να βοηθήσει

στην εξόρυξη πληροφορίας σχετικά με την επιρροή του καταναλωτικού κοινού από αυτές.

- Η διαδικασία μαζικής συλλογής δεδομένων από ηλεκτρονικά καταστήματα με περισσότερες κατηγορίες και οικονομικά.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]. Dunham M. H.: Data Mining Introductory and Advanced Topics, Pearson Education Inc., 2004

[2]. https://en.wikipedia.org/wiki/Data_mining

[3]. Tan P.-N., Steinbach M. and Kumar V.: Introduction to Data Mining, Pearson Education Inc., 2005

[4]. Han J., Kamber M. and Pei J.: Data Mining: Concepts and Techniques, Elsevier Inc., 2000

[5]. Berry M. J. A. and Linoff G. S.: Data Mining Techniques For Marketing, Sales and Customer Relationship Management, Wiley Inc., 2004

[6]. March N. and Reutterer T.: Building an Association Rules Framework forTarget Marketing. Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007, pp 439-446, 2004

[7]. https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery

[8]. Brijis T., Swinnen G., Vanhoof K. and Wets G.: Using Association Rules for Product Assortment Decisions: A Case Study. Knowledge Discovery and Data Mining, pp. 254-260, 1999

[9]. http://www.inmobi.com/

[10]. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, 2009

[11]. http://www.cs.waikato.ac.nz/ml/weka/arff.html

[12]. Agrawal R. and Srikant R.: Fast Algorithms for Mining Association Rules inLarge Databases. Proceedings of the 20th International Conference on Very LargeData Bases (VLDB), Santiago, Chile, September. 1994.

[13]. https://en.wikipedia.org/wiki/Correlation_and_dependence

[14]. Fenton N. and Neil M.: Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, 2012

[15]. QuinlanR.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993

[16]. Liu B., Hsu W. and Ma Y.: Integrating Classification and Association Rule Mining. In: Fourth International Conference on Knowledge Discovery and Data Mining, 80-86, 1998.

[17]. Hahsler M., Hornik K., and Reutterer T.: Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pages 598–605. Springer-Verlag, 2006

Παραρτήματα

Παρακάτω παρατίθενται τα αποτελέσματα από τα πειράματα που εκτελέστηκαν στο Weka:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4
Instances: 324
Attributes: 8
p1
p2
p3
p4
p5
p6
p7
p8
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
---

p3 = 0
| p7 = 0
| | p4 = 0
| | | p5 = 0
| | | | p6 = 0

| | | | | p8 = 0: 1 (26.0/8.0)
| | | | | p8 = 1
| | | | | | p2 = 0: 0 (14.0/3.0)
| | | | | | p2 = 1: 1 (3.0/1.0)
| | | | p6 = 1: 0 (28.0/10.0)
| | | p5 = 1
| | | | p6 = 0: 0 (24.0/7.0)
| | | | p6 = 1
| | | | | p2 = 0: 1 (5.0/1.0)
| | | | | p2 = 1: 0 (4.0)
| | p4 = 1
| | | p5 = 0: 0 (36.0/5.0)
| | | p5 = 1
| | | | p8 = 0: 0 (8.0/3.0)
| | | | p8 = 1
| | | | | p6 = 0: 1 (4.0)
| | | | | p6 = 1: 0 (2.0)
| p7 = 1: 0 (68.0/18.0)
p3 = 1
| p5 = 0: 0 (58.0/22.0)
| p5 = 1
| | p4 = 0
| | | p7 = 0: 0 (21.0/7.0)
| | | p7 = 1: 1 (9.0/3.0)
| | p4 = 1: 1 (14.0/5.0)

Number of Leaves : 16

Size of the tree : 31

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	206	63.5802 %
Incorrectly Classified Instances	118	36.4198 %
Kappa statistic	0.0565	
Mean absolute error	0.4482	
Root mean squared error	0.4858	
Relative absolute error	96.7225 %	
Root relative squared error	100.9385 %	
Total Number of Instances	324	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.937	0.89	0.648	0.937	0.766	0.56	0
	0.11	0.063	0.5	0.11	0.181	0.56	1
Weighted Avg.	0.636	0.589	0.594	0.636	0.553	0.56	

=== Confusion Matrix ===

a	b	<-- classified as
193	13	a = 0
105	13	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

---

p3 = 0

| p8 = 0

| | p7 = 0

| | | p6 = 0

| | | | p4 = 0: 1 (43.0/15.0)

| | | | p4 = 1: 0 (26.0/10.0)

| | | p6 = 1: 0 (33.0/12.0)

| | p7 = 1

| | | p4 = 0: 0 (29.0/8.0)

| | | p4 = 1: 1 (12.0/5.0)

| p8 = 1: 0 (79.0/29.0)

p3 = 1: 0 (102.0/32.0)

Number of Leaves : 7

Size of the tree : 13

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	196	60.4938 %
Incorrectly Classified Instances	128	39.5062 %
Kappa statistic	0.0319	
Mean absolute error	0.4735	
Root mean squared error	0.4965	
Relative absolute error	99.5764 %	
Root relative squared error	101.8384 %	
Total Number of Instances	324	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.924	0.897	0.618	0.924	0.741	0.474	0
	0.103	0.076	0.464	0.103	0.169	0.474	1
Weighted Avg.	0.605	0.578	0.558	0.605	0.518	0.474	

=== Confusion Matrix ===

a	b	<-- classified as
183	15	a = 0
113	13	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

: 0 (324.0/102.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	219	67.5926 %
Incorrectly Classified Instances	105	32.4074 %
Kappa statistic	-0.0111	

Mean absolute error 0.4335

Root mean squared error 0.4693

Relative absolute error 100.3717 %

Root relative squared error 101.0323 %

Total Number of Instances 324

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.982	0.99	0.683	0.982	0.806	0.472	0
	0.01	0.018	0.2	0.01	0.019	0.472	1
Weighted Avg.	0.676	0.684	0.531	0.676	0.558	0.472	

=== Confusion Matrix ===

a	b	<-- classified as
218	4	a = 0
101	1	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

---

p6 = 0

| p7 = 0

| | p5 = 0

| | | p1 = 0

| | | | p8 = 0: 1 (40.0/16.0)

| | | | p8 = 1

| | | | | p2 = 0: 0 (17.0/5.0)

| | | | | p2 = 1: 1 (6.0/1.0)

| | | p1 = 1: 0 (40.0/10.0)

| | p5 = 1

| | | p1 = 0: 0 (35.0/7.0)

| | | p1 = 1

| | | | p2 = 0: 1 (15.0/6.0)

| | | | p2 = 1: 0 (11.0/4.0)

| p7 = 1: 0 (69.0/18.0)

p6 = 1: 0 (91.0/25.0)

Number of Leaves : 9

Size of the tree : 17

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	213	65.7407 %
Incorrectly Classified Instances	111	34.2593 %
Kappa statistic	0.0473	
Mean absolute error	0.4379	
Root mean squared error	0.4785	
Relative absolute error	98.8858 %	
Root relative squared error	101.7259 %	
Total Number of Instances	324	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.926	0.888	0.679	0.926	0.784	0.505	0
	0.112	0.074	0.429	0.112	0.178	0.505	1
Weighted Avg.	0.657	0.619	0.596	0.657	0.584	0.505	

=== Confusion Matrix ===

a	b	<-- classified as
201	16	a = 0
95	12	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

: 0 (324.0/112.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	202	62.3457 %
Incorrectly Classified Instances	122	37.6543 %
Kappa statistic	-0.0544	
Mean absolute error	0.4634	
Root mean squared error	0.4935	
Relative absolute error	102.3505 %	
Root relative squared error	103.7567 %	

Total Number of Instances 324

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.948	0.991	0.644	0.948	0.767	0.455	0
0.009	0.052	0.083	0.009	0.016	0.455	1
Weighted Avg. 0.623	0.666	0.45	0.623	0.508	0.455	

=== Confusion Matrix ===

a	b	<-- classified as
201	11	a = 0
111	1	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

: 0 (324.0/91.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	233	71.9136 %
Incorrectly Classified Instances	91	28.0864 %
Kappa statistic	0	
Mean absolute error	0.404	
Root mean squared error	0.4494	
Relative absolute error	99.8383 %	
Root relative squared error	99.9995 %	
Total Number of Instances	324	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.719	1	0.837	0.491	0
	0	0	0	0	0	0.491	1
Weighted Avg.	0.719	0.719	0.517	0.719	0.602	0.491	

=== Confusion Matrix ===

```
  a   b   <-- classified as
233   0 |   a = 0
 91   0 |   b = 1
```

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     proionta-weka.filters.unsupervised.attribute.Remove-R13-
weka.filters.unsupervised.attribute.Remove-R1-4
Instances:    324
Attributes:   8
              p1
              p2
              p3
              p4
              p5
              p6
              p7
              p8
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: 0 (324.0/95.0)

Number of Leaves  :         1

Size of the tree :    1
```

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	229	70.679 %
Incorrectly Classified Instances	95	29.321 %
Kappa statistic	0	
Mean absolute error	0.4145	
Root mean squared error	0.4553	
Relative absolute error	99.8597 %	
Root relative squared error	99.9996 %	
Total Number of Instances	324	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.707	1	0.828	0.485	0
	0	0	0	0	0	0.485	1
Weighted Avg.	0.707	0.707	0.5	0.707	0.585	0.485	

=== Confusion Matrix ===

a	b	<-- classified as
229	0	a = 0
95	0	b = 1

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: proionta-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Remove-R1-4

Instances: 324

Attributes: 8

p1

p2

p3

p4

p5

p6

p7

p8

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

: 0 (324.0/104.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	214	66.0494 %
Incorrectly Classified Instances	110	33.9506 %
Kappa statistic	-0.0293	

Mean absolute error	0.4401
Root mean squared error	0.4769
Relative absolute error	100.8615 %
Root relative squared error	102.1485 %
Total Number of Instances	324

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.968	0.99	0.674	0.968	0.795	0.457	0
	0.01	0.032	0.125	0.01	0.018	0.457	1
Weighted Avg.	0.66	0.683	0.498	0.66	0.545	0.457	

=== Confusion Matrix ===

a	b	<-- classified as
213	7	a = 0
103	1	b = 1