ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ: Διοικητικών, Οικονομικών και Κοινωνικών Επιστημών

ΤΜΗΜΑ: Διοίκησης Επιχειρήσεων

ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ "DIGITAL BUSINESS"

Ανάλυση αγορών στο χονδρεμπόριο: Διερεύνηση τεχνικών και προτύπων

ΣΤΕΦΑΝΟΣ ΑΠΟΣΤΟΛΑΚΟΠΟΥΛΟΣ
MBA20004

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΔΗΜΗΤΡΗΣ ΠΑΠΑΚΥΡΙΑΚΟΠΟΥΛΟΣ

Αθήνα, Σεπτέμβριος 2022

UNIVERSITY OF WEST ATTICA
SCHOOL: ADMINISTRATIVE,
FINANCIAL SOCIAL SCIENCES
DEPARTMENT: BUSINESS ADMINISTRATION

“DIGITAL BUSINESS” BUSINESS MANAGEMENT

Wholesale market analysis: Inquiry on techniques and standards

STEFANOS APOSTOLAKOPOULOS
MBA20004

SUPERVISOR: DIMITRIS PAPAKYRIAKOPOYLOS

Athens, September 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ: Διοικητικών, Οικονομικών και Κοινωνικών Επιστημών

ΤΜΗΜΑ: Διοίκησης Επιχειρήσεων

ΤΙΤΛΟΣ ΠΡΟΓΡΑΜΜΑΤΟΣ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Ανάλυση αγορών στο χονδρεμπόριο: Διερεύνηση τεχνικών και προτύπων

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

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

α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	Δημήτρης Παπακυριακόπουλος	Επίκουρος καθηγητής / Επιβλέπων	Dimitrios Papakyriakopoulos Digitally signed by Dimitrios Papakyriakopoulos Date: 2022.09.29 19:37:33 +03'00'
2	Ιωάννης Ψαρομήλιγκος	Καθηγητής / Μέλος	
3	Ιωάννης Ριζομυλιώτης	Επίκουρος καθηγητής / Μέλος	

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η……**Στέφανος Αποστολακόπουλος**…….. του…**Χαράλαμπου**…., με αριθμό μητρώου …**MBA20004**… φοιτητής/τρια του Προγράμματος Μεταπτυχιακών Σπουδών …**DIGIGTAL BUSINESS**… του Τμήματος **..**…**Διοίκησης Επιχειρήσεων**…… της ……**Σχολής Διοικητικών Οικονομικών και Κοινωνικών Επιστημών**…… του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

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

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

*****_Επιθυμώ την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μου μέχρι …… και έπειτα από αίτηση μου στη Βιβλιοθήκη και έγκριση του επιβλέποντα καθηγητή._

Ο Δηλών

Στέφανος Αποστολακόπουλος

*** Ονοματεπώνυμο /Ιδιότητα**

Ψηφιακή Υπογραφή Επιβλέποντα
(Υπογραφή)

**** Εάν κάποιος επιθυμεί απαγόρευση πρόσβασης στην εργασία για χρονικό διάστημα***

6-12 μηνών (embargo), θα πρέπει να υπογράψει ψηφιακά ο/η επιβλέπων/ουσα καθηγητής/τρια, για να γνωστοποιεί ότι είναι ενημερωμένος/η και συναινεί. Οι λόγοι χρονικού αποκλεισμού πρόσβασης περιγράφονται αναλυτικά στις πολιτικές του Ι.Α. (σελ. 6):

https://www.uniwa.gr/wp-content/uploads/2021/01/%CE%A0%CE%BF%CE%BB%CE%B9%CF%84%CE%B9%CE%BA%CE%B5%CC%81%CF%82_%CE%99%CE%B4%CF%81%CF%85%CE%BC%CE%B1%CF%84%CE%B9%CE%BA%CE%BF%CF%85%CC%81_%CE%91%CF%80%CE%BF%CE%B8%CE%B5%CF%84%CE%B7%CF%81%CE%B9%CC%81%CE%BF%CF%85_final.pdf

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά τον επιβλέπον καθηγητή μου κύριο Δημήτρη Παπακυριακόπουλο για την καθοδήγηση, εμπιστοσύνη και υπομονή που μου έδειξε καθ' όλη την διάρκεια εκπόνησης της εργασίας.

Θα ήθελα να ευχαριστήσω τους ανθρώπους της μεγάλης επιχείρησης λιανικής και χονδρικής πώλησης για τα δεδομένα που μου παραχώρησαν και τον χρόνο που αφιέρωσαν να με βοηθήσουν, τα ονόματα των οποίων δεν θα αναφέρω για λόγους εμπιστευτικότητας.

Επίσης θα ήθελα να ευχαριστήσω όλο το ακαδημαϊκό προσωπικό του μεταπτυχιακού προγράμματος για τις γνώσεις και εμπειρίες που μου πρόσφεραν.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για τη στήριξη και συμπαράσταση, καθ' όλη τη διάρκεια των σπουδών μου.

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών MBA με κατεύθυνση Digital Business του πανεπιστημίου δυτικής Αττικής με θέμα «Ανάλυση αγορών στο χονδρεμπόριο: Διερεύνηση τεχνικών και προτύπων».

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι ο εντοπισμός προτύπων από αγοραστικά καλάθια στο χονδρεμπόριο. Εστιάζει κυρίως στις αγορές χονδρικής που κάνουν οι πελάτες μεταποίησης (HO.RE.CA) μεγάλης αλυσίδας λιανικού και χονδρικού εμπορίου. Θα γίνει ανάλυση αγοραστικού καλαθιού (Market Basket Analysis) στα καλάθια τους αλλά και μέσω αυτής της τεχνικής θα εξετασθεί αν είναι εφικτός ο εντοπισμός της επαγγελματική ιδιότητα του πελάτη από τις εμπορικές συναλλαγές του. Η βελτίωση της γνώσης σχετικά με τις προτιμήσεις των εμπορικών πελατών μπορεί να έχει επίδραση στην ποικιλία των προϊόντων που διακινεί μια αλυσίδα χονδρικής-λιανικής και την αναγνώριση ευκαιριών πώλησης. Αυτό θα μπορεί να δώσει πληροφορία στην επιχείρηση για το αν οι πελάτες της αγοράζουν προϊόντα που συμπίπτουν με το επάγγελμα τους και θα μπορέσουν να δημιουργήσουν ένα κατάλληλο πλάνο δράσης για να τους προσεγγίσουν. Τα ερωτήματα που διερευνώνται είναι:

- Μέσα από την ανάλυση αγοραστικού καλαθιού μπορεί η επιχείρηση να εντοπίσει προβλήματα, ευκαιρίες και να βελτιώσει την λήψη αποφάσεων;
- Η ανάλυση του αγοραστικού καλαθιού μπορεί να δώσει πληροφορίες για την επαγγελματική ιδιότητα του πελάτη;

Στο πρώτο κεφάλαιο γίνεται μια αναφορά στην επιχειρηματική ευφυΐα, στην εξόρυξη γνώσης, στα μεγάλα δεδομένα και γενικά στις τάσεις των επιχειρήσεων πάνω σε αυτές τις έννοιες. Παρουσιάζεται αναλυτικά το ερευνητικό περιβάλλον, το κίνητρο και οι στόχοι της εργασίας. Στο δεύτερο κεφάλαιο παρατίθεται η βιβλιογραφική ανασκόπηση με τις τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται στην ακαδημαϊκή και επιχειρηματικοί κοινότητα. Στο τρίτο κεφάλαιο αναλύεται η μεθοδολογία, η τεχνική και τα δεδομένα που χρησιμοποιούνται στην μελέτη. Στο κεφάλαιο τέσσερα δίνεται η εμπειρική μελέτη που εκτελέστηκε χρησιμοποιώντας την μεθοδολογία και την τεχνική που παρουσιάστηκε στο κεφάλαιο τρία της παρούσας. Τέλος στο κεφάλαιο πέντε αναλύονται τα αποτελέσματα της εμπειρικής μελέτης καθώς και τα συμπεράσματα που προέκυψαν μετά την επικοινωνία με ειδικούς στον χώρο των πωλήσεων χονδρικής.

17/09/2022
Στέφανος Αποστολακόπουλος

Περιεχόμενα

Επιτελική Σύνοψη

Με την άνοδο των νέων ψηφιακών τεχνολογιών όπως τα κοινωνικά δίκτυα, τα κινητά, τα μεγάλα δεδομένα κ.λπ., οι εταιρείες σχεδόν σε όλους τους τομείς βιομηχανίας πραγματοποιούν πολλαπλές πρωτοβουλίες για να εξερευνήσουν και να εκμεταλλευτούν τα οφέλη τους. (Reis, Amorim, Melao, & Matos, 2018) Έννοιες όπως επιχειρηματική ευφυΐα (Business Intelligence), εξόρυξη δεδομένων (Data Mining) και μεγάλα δεδομένα (Big Data) βρίσκονται στο επίκεντρο πολλών ερευνών και στη βιβλιογραφία επικρατούν πολλές συζητήσεις και απόψεις σχετικά με τους τρόπους που αυτά προσθέτουν αξία σε μία επιχείρηση. (Davenport, Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, 2014) Παρά τη δημοσιότητα που έχουν πάρει αυτές οι έννοιες, το ποσοστό επιτυχίας των έργων που εμπεριέχονται και η αξία που δημιουργούν είναι ασαφή. Η περισσότερη βιβλιογραφία επικεντρώνεται στο πώς μπορούν αυτές οι έννοιές να χρησιμοποιηθούν για την ενίσχυση των επιχειρήσεων, αλλά πολύ λίγες μελέτες εξετάζουν το αντίκτυπό που έχουν στην επιχειρηματική αξία. (Grover, Chiang, Liang, & Zhang, 2018)

Παρόλο που οι τρόποι με τους οποίους η επιχειρηματική ευφυΐα προσδίδει αξία σε μια επιχείρηση είναι ασαφής, η ανακάλυψη γνώσης από τα δεδομένα έχει αρχίσει να ενσωματώνεται σε διάφορους τομείς όπως Οικονομία (Finance), Λιανικό εμπόριο (Retail), Τηλεπικοινωνίες (Telecommunication), Επιστήμη και μηχανική (Science & Engineering). (Sethi, Malhotra, & Verma, 2016) Η ανακάλυψη γνώσης στις βάσεις δεδομένων είναι μια επαναληπτική διαδικασία η οποία ξεκινώντας από τα διαθέσιμα δεδομένα προσπαθεί να δημιουργήσει νέα γνώση. Το ενδιαφέρον της ακαδημαϊκής κοινότητας για περισσότερο από 20 χρόνια βρίσκεται στην εξόρυξη δεδομένων (Data Mining), που ένα συγκεκριμένο βήμα της διαδικασίας αυτής επιτρέπει την εξαγωγή πολύτιμων πληροφοριών από μεγάλους όγκους δεδομένων. (Sethi, Malhotra, & Verma, 2016)

Το λιανικό εμπόριο είναι ένας κλάδος που έχει βρεθεί στο επίκεντρο πολλών ερευνών αναφορικά με την εξόρυξή γνώσης καθώς πέρα από το ανταγωνιστικό περιβάλλον των επιχειρήσεων αυτού του κλάδου διαθέτουν και πληθώρα δεδομένων. Ένας από τους παλαιότερους στόχους των επιχειρήσεων λιανεμπορίου είναι να ανακαλύψουν τα μοτίβα αγορών των καταναλωτών τους, ώστε να μπορέσουν να προβλέψουν τις αγοραστικές τους συνήθειες και να πάρουν στρατηγικές αποφάσεις.

Η παρούσα εργασία αναλύει τις αγορές στον χονδρεμπόριο μεγάλης εταιρείας εταιρειών λιανεμπορίου – χονδρεμπορίου (fast moving customer goods), με χρήση τεχνικών εξόρυξης δεδομένων (Data Mining) και πιο συγκεκριμένα με ανάλυση αγοραστικού καλαθιού (Market Basket Analysis), στοχεύει στην ανακάλυψη συνδέσεων μεταξύ προϊοντικών κατηγοριών. Μέσα από επικοινωνία με ανθρώπους της επιχείρησης εντοπίστηκε η ανάγκη της να ταξινομεί τους πελάτες της σε επαγγελματικές ιδιότητες. Προέκυψε λοιπόν ένας δεύτερος στόχος που είναι να χρησιμοποιηθεί η ανάλυση αγοραστικού καλαθιού για την πιθανή εξαγωγή πληροφορίας σχετικά με την επαγγελματική ιδιότητα του πελάτη μέσα από τις εμπορικές τους συναλλαγές.

Μέσω της βιβλιογραφικής ανασκόπησης η παρούσα μελέτη διερευνά τις τεχνικές εξόρυξης γνώσης που χρησιμοποιούνται από την ακαδημαϊκή και επιχειρηματική κοινότητα. Στόχος της είναι να εντοπισθεί η καταλληλότερη τεχνική που θα

μπορούσε να αξιοποιηθεί για τον εντοπισμό προτύπων από τα αγοραστικά καλάθια στο χονδρεμπόριο και να διερευνηθεί αν είναι εφικτό να εντοπισθεί η επαγγελματική ιδιότητα ενός πελάτη χονδρικής βάση των αγορών του. Ξεκινά με μια εισαγωγή στην ανακάλυψη γνώσης από τα δεδομένα και την εξόρυξη γνώσης, αναφέρει τους σημαντικότερους τομείς που χρησιμοποιείται η εξόρυξη γνώσης καθώς και τις βασικότερες τεχνικές της. Συνεχίζει με τη διερεύνηση των τεχνικών που χρησιμοποιούνται για την επίλυση διάφορων προβλημάτων εξόρυξης δεδομένων. Η προσέγγιση στις τεχνικές έγινε με βάση το επίκεντρο του εκάστοτε αναλυτή δηλαδή αν εστιάζει στις τεχνικές, στα προϊόντα ή στους πελάτες. Τέλος γίνεται μια αναφορά στις τεχνικές που αξιοποιούνται στον επιχειρηματικό κόσμο αλλά και σε αυτές που παραμένουν μόνο σε ακαδημαϊκό επίπεδο και τους λόγους που μπορεί συμβαίνει αυτό.

Για να μπορέσει να γίνει ανάλυση αγοραστικού καλαθιού χρειάστηκε να μελετηθούν οι τεχνικές και οι μεθοδολογίες που προτείνονται από την ακαδημαϊκή και επιχειρηματική κοινότητα. Όπως αναφέρεται και στο κεφάλαιο τρία της παρούσας, υπάρχουν πολλές μεθοδολογίες για την διαδικασία εξόρυξης γνώσης που μπορεί να ακολουθήσει ένας αναλυτής, μπορεί ακόμα και να μην ακολουθήσει καμιά. Παρόλα αυτά για την μετατροπή του επιχειρηματικού προβλήματος σε εργασία εξόρυξης δεδομένων, στη εμπειρική μελέτη της εργασίας χρησιμοποιήθηκε μια από τις πιο διαδεδομένες μεθοδολογίες η CRISP-DM και μέσω αυτής της μεθοδολογίας αξιοποιήθηκε η τεχνική εξόρυξης κανόνων συσχέτισης (ανάλυση αγοραστικού καλαθιού – Market Basket Analysis) για την παραγωγή ενός μοντέλου που θα μπορούσε να εκτιμηθεί από την εταιρεία με σκοπό την εξόρυξη γνώσης.

Με στόχο την ανακάλυψη συνδέσεων μεταξύ προϊοντικών κατηγοριών και την εξαγωγή πληροφορίας σχετικά με την επαγγελματική ιδιότητα του πελάτη μέσα από τα αγοραστικά του καλάθια, πραγματοποιήθηκε η εμπειρική μελέτη που παρουσιάζεται στην παρούσα εργασία. Έγινε χρήση της μεθοδολογία CRISP-DM όπου:

- Στην επιχειρηματική κατανόηση, καθορίζονται αναλυτικά οι στόχοι και το εύρος τους.
- Στην κατανόηση δεδομένων, δίνεται η αναλυτική περιγραφή των διαθέσιμων δεδομένων ενώ στο κεφάλαιο τρία δίνονται και τα περιγραφικά στατιστικά που παράχθηκαν από αυτά.
- Στην προετοιμασία δεδομένων, παραθέτονται αναλυτικά οι διαδικασίες που ακολουθήθηκαν για τον καθαρισμό, την κωδικοποίηση και την μορφοποίηση των δεδομένων ενώ εμφανίζονται και τα σύνολα των δεδομένων που εξάχθηκαν.
- Στην μοντελοποίηση, παρουσιάζεται αναλυτικά η χρήση της τεχνικής ανάλυσης αγοραστικού καλαθιού και μερικοί από τους κανόνες συσχέτισης που παράχθηκαν από το μοντέλο.
- Στην εκτίμηση, παρουσιάζονται και επεξηγούνται οι παραγόμενοι κανόνες συσχέτισης καθώς και οι διάφοροι προβληματισμοί που προκύπτουν.

Τέλος παρουσιάστηκαν τα βασικά αποτελέσματα και ερωτήματα που διέγειρε η έρευνα σε ανθρώπους της εταιρείας που βρίσκονται εντός του αντικειμένου και ανήκουν στο τμήμα πωλήσεων χονδρικής, με σκοπό την καλύτερη κατανόηση των ευρημάτων, τον εντοπισμό της επιχειρηματικής αξίας και ίσως την μελλοντική αξιοποίηση των τεχνικών ανάλυσης δεδομένων για την καλύτερη λήψη αποφάσεων.

Το κύριο συμπέρασμα που προέκυψε μέσα από αυτή την επικοινωνία ήταν ότι η παραγωγή κανόνων συσχέτισης (ανάλυση αγοραστικού καλαθιού) μπορεί να αξιοποιηθεί και να βοηθήσει στην λήψη αποφάσεων. Αναφορικά με τους κανόνες συσχέτισης μεταξύ προϊόντων θα μπορούσε να βοηθήσει:

- Στις γενικές προσφορές και στις προσφορές συνδυασμών (COMBO).
- Στην παραγωγή προσφορών έκπτωσης με την αγορά επιλεγμένων προϊόντων.
- Στην διαχείριση αποθεμάτων και παραγγελιοδοσίας των καταστημάτων.
- Στην δημιουργία λίστας προϊόντων για πελάτες συμφωνίας που δεν θα βασίζεται μόνο στην εμπειρία.

Αναφορικά με τους κανόνες συσχέτισης μεταξύ προϊόν και επαγγελμάτων θα μπορούσε να βοηθήσει:

- Στον εντοπισμό κωδικών κλειδιά που η εταιρεία μέχρι στιγμής χαρακτηρίζει εμπειρικά σαν κωδικούς που ανήκουν σε ένα επάγγελμα
- Στον έλεγχο αν επιλεγμένα προϊόντα αγοράζονται από την αναμενόμενη επαγγελματική ιδιότητα.

Τα αποτελέσματα της έρευνας αποδεικνύουν ότι η ανάλυση αγοραστικού καλαθιού μπορεί να αξιοποιηθεί και να βοηθήσει στην λήψη αποφάσεων, όμως οι ειδικοί αναφέρουν ότι τα αποτελέσματα θα ήταν πιο αξιοποιήσιμα αν μπορούσαν οι ίδιοι να τα παράγουν, ώστε να εισάγουν τα κριτήρια που επιθυμούν και να εμφανίζουν τις συνδέσεις είτε μεταξύ προϊόντων είτε μεταξύ προϊόντων και επαγγελμάτων που αυτοί έχουν ανάγκη. Αυτό θα μπορούσε να επιτευχθεί μόνο με την δημιουργία εφαρμογών που θα χειριζόταν ο οποιοσδήποτε χρήστης χωρίς εξειδικευμένες γνώσεις εξόρυξης δεδομένων. Με αυτό τον τρόπο η ανάλυση αγοραστικού καλαθιού θα μπορούσε να δώσει μεγαλύτερη αξία στην επιχείρηση και να βοηθήσει στη βελτίωση της ποιότητας των προσφορών, των υπηρεσιών και την εμπειρία πελάτη.

1 Εισαγωγή

Τα τελευταία χρόνια το σύνολο των επιχειρήσεων βρίσκονται σε έναν «αγώνα δρόμου» να ψηφιοποιήσουν τις διαδικασίες τους και να κάνουν χρήση των πιο σύγχρονων τεχνολογιών, για να εξασφαλίσουν καλύτερα επίπεδα παρεχόμενων υπηρεσιών, αποδοτικότερη διαχείριση πόρων, αποτελεσματικότερη επίτευξη στόχων και καλύτερη λήψη αποφάσεων (Davenport, Big Data at Work: Dispelling the Myths, Uncovering the Opportunities, 2014). Την τελευταία δεκαετία, η επιχειρηματική υποδομή έχει γίνει ψηφιακή με αυξημένες διασυνδέσεις μεταξύ προϊόντων, διαδικασιών και υπηρεσιών. Σε πολλές εταιρείες που καλύπτουν διαφορετικούς κλάδους και τομείς, οι ψηφιακές τεχνολογίες (που θεωρούνται συνδυασμοί πληροφοριών, τεχνολογίες υπολογιστών, επικοινωνίας και συνδεσιμότητας) μεταμορφώνουν θεμελιωδώς τις επιχειρηματικές στρατηγικές, τις επιχειρηματικές διαδικασίες, τις δυνατότητες της εταιρείας, τα προϊόντα, τις υπηρεσίες και τις βασικές σχέσεις μεταξύ επιχειρήσεων. (Bharadwaj, El Sawy, Pavlou, & Venkatraman, 2013)

Πολλές επιχειρήσεις έχουν προσαρμοστεί σε αυτά τα δεδομένα και έχουν ψηφιοποιήσει αρκετές από τις διαδικασίες που εκτελούν, για παράδειγμα Logistic, συστήματα διαχείρισης αποθήκης (WMS), εκπαίδευση εργαζομένων, εμπειρία πελατών, customer engagement, e-commerce στρατηγικές κ.α. (Belka, 2022) Αυτό έχει σαν αποτέλεσμα να συλλέγονται δεδομένα σε μεγάλους όγκους, τα οποία με κατάλληλη επεξεργασία και ανάλυση μπορούν να δημιουργήσουν γνώση που μπορεί να παράγει αξία. Ορισμένες επιχειρήσεις σε αυτά τα νέα πλαίσια αποφασίζουν να χρησιμοποιήσουν τις τελευταίες τεχνολογίες για να ανακαλύψουν γνώση μέσα από τα δεδομένα που οι ίδιες ή και άλλες παράγουν. Η ανακάλυψη γνώσης στις βάσεις δεδομένων (Knowledge discovery in databases-KDD) αναφέρεται στην μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, καινοτόμων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων ή σχέσεων μέσα σε ένα σύνολο δεδομένων προκειμένου να ληφθούν σημαντικές αποφάσεις. (Fayyad, Piatesky-Shapiro, & Smyth, 1996)

Στη βιβλιογραφία επικρατούν πολλές συζητήσεις και απόψεις σχετικά με τους τρόπους που η επιχειρηματική ευφυΐα, τα μεγάλα δεδομένα και η εξόρυξη γνώσης δίνουν αξία σε μία επιχείρηση, η οποία δεν θα πρέπει να τα αντιλαμβάνεται σαν διαφορετικές οντότητες αλλά θα πρέπει να ενσωματωθούν και να συνυπάρχουν με όλες τις υπάρχουσες διαδικασίες, προϊόντα και ανθρώπους. (Davenport & Dyché, International Institute for Analyticts, 2013) Υπάρχουν αρκετοί που υποστηρίζουν ότι οι ανθρώπινοι πόροι, οι δυνατότητες διαχείρισης και η οργανωτική κουλτούρα έχουν μεγαλύτερο αντίκτυπο στην επιχειρηματική αξία, ενώ οι τεχνικές πτυχές διαδραματίζουν μικρό ρόλο στη βελτίωση της απόδοσης μιας εταιρείας. Τόσο όμως οι ακαδημαϊκοί όσο και οι επαγγελματίες αναγνωρίζουν ότι μπορούν να ληφθούν καλύτερες αποφάσεις με βάση στοιχεία που βασίζονται σε δεδομένα και όχι στη διαίσθηση. Η λήψη αποφάσεων με γνώμονα τα δεδομένα έχει διαπιστωθεί ότι έχει θετικό αντίκτυπο στην απόδοση μιας εταιρείας, υποδεικνύοντας ότι οι εταιρείες που χρησιμοποιούν δεδομένα και Business Analysis για τη λήψη αποφάσεων επιτυγχάνουν υψηλότερη παραγωγικότητα. (Oesterreich, Anton, & Teuteberg, 2022)

Η ανακάλυψη γνώσης στις βάσεις δεδομένων είναι μια επαναληπτική διαδικασία η οποία ξεκινώντας από τα διαθέσιμα δεδομένα προσπαθεί να δημιουργήσει νέα γνώση. Το ενδιαφέρον της ακαδημαϊκής κοινότητας για περισσότερο από 20 χρόνια βρίσκεται στην εξόρυξη δεδομένων (Data mining), που ένα συγκεκριμένο βήμα της διαδικασίας αυτής επιτρέπει την εξαγωγή πολύτιμων πληροφοριών από μεγάλους όγκους δεδομένων. (Liao, Chu, & Hsiao, 2012) Υπάρχουν πάρα πολλές τεχνικές data mining που μπορούν να χρησιμοποιηθούν, ποια θα επιλεγεί από τον εκάστοτε αναλυτή εξαρτάται από το πρόβλημα που καλείται να μελετήσει, τη φύση των δεδομένων που έχει στη διάθεση του και τη γνώση που προσπαθεί να εξάγει. Οι πιο γνωστές τεχνικές είναι: Κατηγοριοποίηση (Classification), Συσχετίσεις (Associations) και Ομαδοποίηση (Clustering). Όλες οι τεχνικές data mining χρησιμοποιούνται συνήθως σε συνδυασμό και μέσω αυτών των συνδυασμών παράγονται σε πολλές περιπτώσεις νέες προσεγγίσεις και τεχνικές στα διάφορα ζητήματα που διερευνώνται.

Μία ακόμα έννοια που έρχεται συχνά στο προσκήνιο είναι αυτή των μεγάλων δεδομένων (Big Data). Τα μεγάλα δεδομένα ορίζονται ως τεράστιοι όγκοι δεδομένων (Volume) με μεγάλη ποικιλία που παράγονται (Variety), συλλαμβάνονται και επεξεργάζονται με υψηλή ταχύτητα (Velocity). Υιοθετώντας τεχνολογίες μεγάλων δεδομένων, οι οργανισμοί αναμένουν να αποκομίσουν οφέλη σε πολλούς τομείς, όπως το ηλεκτρονικό εμπόριο, η ηλεκτρονική διακυβέρνηση, η επιστήμη, η υγεία και η ασφάλεια .Τα οφέλη που αντιλαμβάνονται οι οργανισμοί ως «αξία» εξαρτάται από τους στρατηγικούς τους στόχους για την υιοθέτηση και χρήση μεγάλων δεδομένων. Ωστόσο, οι υψηλές ελπίδες και η εκτεταμένη δημοσιότητα σχετικά με τα μεγάλα δεδομένα δεν εγγυώνται την απόκτηση πραγματικής αξίας και μπορεί να οδηγήσουν τους οργανισμούς να πιστέψουν ότι μπορούν να κερδίσουν περισσότερη αξία μέσω των μεγάλων δεδομένων από ότι είναι πραγματικά σε θέση να αποκτήσουν. Υπάρχουν πολλά debate ως προς τον τρόπο που μπορεί μια επιχείρηση να παράγει αξία μέσα από τα μεγάλα δεδομένα. (Gunther, Mehrizi, Huysman, & Feldberg, 2017)

1.1 Ερευνητικό Περιβάλλον

Τα τελευταία χρόνια η βιομηχανία των εταιρειών λιανεμπορίου (fast moving customer goods) έχει γίνει πολύ ανταγωνιστική, σαν αποτέλεσμα με διάφορες έρευνες προσπαθούν να εντοπίσουν τρόπους τα καταστήματα τους να γίνουν πιο ανταγωνιστικά. Ένας από τους παλαιότερους στόχους των επιχειρήσεων λιανεμπορίου είναι να ανακαλύψουν τα μοτίβα αγορών των καταναλωτών τους, ώστε να μπορέσουν να προβλέψουν τις αγοραστικές τους συνήθειες και να πάρουν στρατηγικές αποφάσεις. Μέσα από όλον αυτό τον ανταγωνισμό προκύπτουν διάφοροι στόχοι και ερωτήματα που πολλοί ερευνητές καλούνται να αναλύσουν. Κάποιοι από αυτούς προσπαθούν να ανακαλύψουν την κρυμμένη αγοραστική συμπεριφοράς των πελατών τους. (Mostafa, 2015). Άλλοι ερευνητές μέσα από την ανάλυση αγοραστικού καλαθιού προσπαθούν να βρουν έξυπνους και κερδοφόρους τρόπους τιμολόγησης των προϊόντων. (Nanda & Ram, 2006). Ενώ υπάρχει και ενδιαφέρον για τον καθορισμό της εξάρτησης μεταξύ των διαφορετικών προϊοντικών κατηγοριών (cross category dependence). (Russel & Petrsen, 2000). Υπάρχουν και πιο σύνθετα ερωτήματα που ενδιαφέρουν την ακαδημαϊκή κοινότητα όπως, αν οι ποικίλες μορφές απεικόνισης των τιμών επηρεάζουν τη συμπεριφορά του καταναλωτή στην επιλογή καταστήματος. (Bell & Lattin, 1998). Είναι επομένως αναμενόμενο οι εταιρειών λιανεμπορίου να βρίσκονται στο επίκεντρο διάφορων ερευνών καθώς πέρα από το

ανταγωνιστικό περιβάλλον διαθέτουν και πληθώρα δεδομένων. Πολλές από τις έρευνες που γίνονται στα δεδομένα των εταιρειών λιανεμπορίου, όπως και στις προαναφερόμενες χρησιμοποιούν μια τεχνική εξόρυξης γνώσης που ονομάζεται ανάλυση αγοραστικού καλαθιού.

1.2 Ερευνητικό Κίνητρο

Μετά από επικοινωνία με μεγάλη αλυσίδας λιανεμπορίου-χονδρεμπορίου προέκυψε ενδιαφέρον στο να αναλυθούν τα δεδομένα που κατέχει. Η συγκεκριμένη επιχείρηση επειδή ενεργεί και στον τομέα του χονδρεμπορίου διαθέτει ένα μεγάλο αριθμό πελατών χονδρικής. Παράγει τεράστιους όγκους δεδομένων στις καθημερινές της εμπορικές συναλλαγές και μια έρευνα σε αυτές θα μπορούσε να βοηθήσεις στις πελατειακές σχέσεις της εταιρείας. Λόγο του ότι εμπορεύεται κυρίως τρόφιμα το μεγαλύτερο μέρος τον πελατών χονδρικής της είναι πελάτες μεταποίησης η αλλιώς HO.RE.CA. (Hotel, Restaurant, Cafe). Προέκυψε το κίνητρο να αναλυθούν με τεχνικές εξόρυξης δεδομένων (Data Mining) τα αγοραστικά καλάθια αυτών των πελατών με σκοπό την πιθανή ανακάλυψη γνώσης μέσα από αυτά.

Στα πλαίσια της συγκεκριμένης μελέτης έγινε διερεύνηση των τεχνικών εξόρυξης δεδομένων που χρησιμοποιούνται στη βιβλιογραφία είτε από την ακαδημαϊκή κοινότητα είτε από τις επιχειρήσεις ώστε να εντοπισθεί πως θα μπορούσε να προσεγγιστεί το συγκεκριμένο ζήτημα. Η πιο εγγύς και παραδοσιακή τεχνική στην παραπάνω διερεύνηση είναι η ανάλυση αγοραστικού καλαθιού η οποία παραδοσιακά καλύπτεται από τους κανόνες συσχέτισης (association rules) ενώ πρόσφατα έχουν προταθεί και τεχνικές που χρησιμοποιούν γράφους (δίκτυα).

Η ανάλυση αγοραστικού καλαθιού (Market Basket Analysis) ή αλλιώς εξόρυξη κανόνων συσχέτισης (Association Rule Mining) είναι μια τεχνική που προέρχεται από το χώρο του marketing και έχει ως στόχο να αναγνωρίσει σχέσεις μεταξύ προϊόντων, ομάδων προϊόντων και κατηγοριών. Η εξόρυξη κανόνων συσχέτισης έχει φανεί αρκετά χρήσιμη και σε πάρα πολλούς επιστημονικούς τομείς πέρα των επιχειρήσεων όπως είναι η βιοπληροφορική, πυρηνική επιστήμη, φάρμακό-επιδημιολογία, ανοσολογία και γεωφυσική. (Aguinis,, Forcum, & Joo, 2012)

1.3 Ερευνητικός στόχος και ερωτήματα

Με χρήση τεχνικών εξόρυξης δεδομένων (Data Mining), πιο συγκεκριμένα με ανάλυση αγοραστικού καλαθιού (Market Basket Analysis) η συγκεκριμένη μελέτη στοχεύει στην ανακάλυψη συνδέσεων μεταξύ προϊοντικών κατηγοριών. Μέσα από επικοινωνία με ανθρώπους της επιχείρησης εντοπίστηκε η ανάγκη της να ταξινομεί τους πελάτες της σε επαγγελματικές ιδιότητες βάση των εμπορικών τους συναλλαγών για να εξακριβώνεται το επάγγελμα που δηλώνουν. Ο δεύτερος στόχος της μελέτης είναι να χρησιμοποιηθεί η ανάλυση αγοραστικού καλαθιού για την πιθανή εξαγωγή πληροφορίας σχετικά με την επαγγελματική ιδιότητα του πελάτη μέσα από τα αγοραστικά του καλάθια. Προέκυψαν τα παρακάτω ερευνητικά ερωτήματα:

- **Μέσα από την ανάλυση αγοραστικού καλαθιού μπορεί η επιχείρηση να εντοπίσει προβλήματα, ευκαιρίες και να βελτιώσει την λήψη αποφάσεων;** Απαντώντας σε αυτό το ερώτημα θα κατανοηθεί αν η ανάλυση αγοραστικού καλαθιού μπορεί να βοηθήσει την επιχείρηση στη λήψη αποφάσεων, όπως για

παράδειγμα να βοηθήσει στην εξαγωγή νέων προσφορών, να βοηθήσει στον τρόπο που τοποθετούνται τα προϊόντα στο ράφι, να αξιοποιηθεί στα αποθέματα και την παραγγελιοδοσία των καταστημάτων.

- **Η ανάλυση του αγοραστικού καλαθιού μπορεί να δώσει πληροφορίες για την επαγγελματική ιδιότητα του πελάτη;** Απαντώντας σε αυτό το ερώτημα θα κατανοηθεί αν η ανάλυση αγοραστικού καλαθιού μπορεί να βοηθήσει στον εντοπισμό των προϊόντων που αγοράζονται ή όχι από συγκεκριμένου τύπου επαγγέλματα και να βοηθήσει στον εντοπισμό κωδικών κλειδιά για το εκάστοτε επάγγελμα.

Η συγκεκριμένη έρευνα έχει ενδιαφέρον για την επιχείρηση διότι μπορεί να φανερώσει αν η υιοθέτηση τεχνικών ανάλυσης δεδομένων και επιχειρηματικής ευφυΐας μπορεί να αυξήσει την αξίας και το μερίδιο της στην αγορά καθώς και να βελτιώσει τις διαδικασίες της.

1.4 Δομή Εργασίας

Η δομή της εργασίας έχει την επόμενη μορφή:

- Στο κεφάλαιο δύο γίνεται μια βιβλιογραφική ανασκόπηση σχετικά με την εξόρυξη γνώσης από τα δεδομένα, διερευνώνται οι τεχνικές που χρησιμοποιούνται για την επίλυση διάφορων προβλημάτων εξόρυξης δεδομένων και γίνεται μια αναφορά στην οπτική των επιχειρήσεων και της ακαδημίας προς αυτές τις τεχνικές.
- Στο κεφάλαιο τρία δίνεται μια θεωρητική αναφορά της μεθοδολογίας που χρησιμοποιήθηκε για την μετατροπή του επιχειρηματικού μοντέλου σε εργασία εξόρυξης γνώσης καθώς και οι βασικές έννοιες που αφορούν την εξόρυξη κανόνων συσχέτισης. Τέλος δίνεται μια περιγραφή των διαθέσιμων δεδομένων και τα περιγραφικά στατιστικά που εξάγονται από αυτά
- Στο κεφάλαιο τέσσερα παρουσιάζεται όλη η μεθοδολογία που ακολουθήθηκε από την περιγραφή του επιχειρηματικού μοντέλου που αναλύεται μέχρι την εξαγωγή των κανόνων συσχέτισης μεταξύ προϊόντων και μεταξύ προϊόντων και επαγγελμάτων. Στο τέλος κάθε υπό ενότητας παραθέτονται τα εμπόδια και ερωτήματα που εμφανίστηκαν μέσα από τη συγκεκριμένη διαδικασία.
- Στο κεφάλαιο πέντε παρουσιάστηκαν τα βασικά αποτελέσματα και ερωτήματα που διέγειρε η έρευνα σε ειδικούς στον χώρο των πωλήσεων χονδρικής, έγινε εκτίμηση των ευρημάτων και απαντήθηκαν τα ερωτήματα που προκύψαν από την εμπειρική μελέτη. Τέλος παρουσιάζονται τα συμπεράσματα και το πως θα μπορούσε μια τέτοια έρευνα να αξιοποιηθεί ώστε να δώσει αξία στην επιχείρηση.

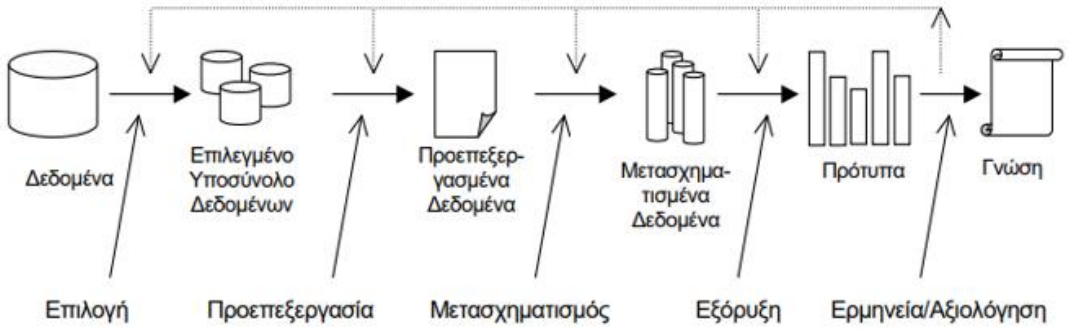
2 Ανασκόπηση Βιβλιογραφίας

Η συγκεκριμένη βιβλιογραφική ανασκόπηση έχει σαν σκοπό να διερευνηθούν οι τεχνικές εξόρυξης γνώσης που χρησιμοποιούνται από την ακαδημαϊκή και επιχειρηματική κοινότητα. Μέσα από αυτή την ανασκόπηση στόχος είναι να εντοπισθεί η καταλληλότερη τεχνική που θα μπορούσε να αξιοποιηθεί για τον εντοπισμό προτύπων από τα αγοραστικά καλάθια στο χονδρεμπόριο και να διερευνηθεί αν είναι εφικτό να εντοπισθεί η επαγγελματική ιδιότητα ενός πελάτη χονδρικής (HO.RE.CA.), βάση των συναλλαγών του με την επιχείρηση.

Ξεκινά με μια εισαγωγή στην ανακάλυψη γνώσης από τα δεδομένα και την εξόρυξη γνώσης, αναφέρει τους σημαντικότερους τομείς που χρησιμοποιείται καθώς και τις βασικότερες τεχνικές της. Συνεχίζει με τη διερεύνηση των τεχνικών που χρησιμοποιούνται για την επίλυση διάφορων προβλημάτων εξόρυξης δεδομένων. Η προσέγγιση στις τεχνικές έγινε με βάση το επίκεντρο του εκάστοτε αναλυτή δηλαδή αν εστιάζει στις τεχνικές, στα προϊόντα ή στους πελάτες. Τέλος γίνεται μια αναφορά ποιες από αυτές τις τεχνικές αξιοποιούνται στον επιχειρηματικό κόσμο αλλά και ποιες παραμένουν μόνο σε ακαδημαϊκό επίπεδο και τους λόγους που συμβαίνει αυτό.

2.1 Εισαγωγή στην ανακάλυψη γνώσης από τα δεδομένα

Η ανακάλυψη γνώσης στις βάσεις δεδομένων αναφέρεται στην μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, καινοτόμων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων ή σχέσεων μέσα σε ένα σύνολο δεδομένων προκειμένου να ληφθούν σημαντικές αποφάσεις (Fayyad, Piatesky-Shapiro, & Smyth, 1996)



Σχήμα 1: Τα βασικά στάδια της διαδικασίας ανακάλυψης γνώσης

Η εξόρυξη δεδομένων (Data Mining) επιτρέπει την αναζήτηση πολύτιμων πληροφοριών σε μεγάλους όγκους δεδομένων. Η εκρηκτική ανάπτυξη των βάσεων δεδομένων έχει δημιουργήσει την ανάγκη ανάπτυξης τεχνολογιών που χρησιμοποιούν τις πληροφορίες και τη γνώση με έξυπνο τρόπο. Ως εκ τούτου, η εξόρυξη δεδομένων

εξελίσσεται σε μια ολοένα και πιο σημαντική ερευνητική περιοχή. (Liao, Chu, & Hsiao, 2012)

Η εξόρυξη δεδομένων στις διάφορες μορφές της χρησιμοποιείται ευρέως σε ποικίλους τομείς. Πολλά ιδρύματα την χρησιμοποιούν για να ανταγωνιστούν το περιβάλλον τους και να παίρνουν γρήγορες και εύκολες αξιολογήσεις των τάσεων της αγοράς.

Οι σημαντικότεροι τομείς που αξιοποιείται είναι:

- Οικονομία (Finance): Για την πρόβλεψη πληρωμών δανείων από πιστωτές, την ανάλυση συγκεκριμένης πιστωτικής πολιτικής πελατών, ταξινόμηση και ομαδοποίηση πελατών για στοχευμένο μάρκετινγκ, ανίχνευση εσόδων από παράνομες δραστηριότητες, οικονομικά εγκλήματα κ.α.
- Λιανικό εμπόριο (Retail): Εντοπισμό συμπεριφοράς πελατών, εντοπισμός προτύπων αγοράς, εντοπισμός προτύπων εφοδιαστικής αλυσίδας κ.λπ.
- Τηλεπικοινωνίες (Telecommunication): εντοπισμός προτύπων, δραστηριοτήτων απάτης, καλύτερη διαχείριση πόρων, βελτίωση ποιότητας των υπηρεσιών κ.α.
- Επιστήμη και μηχανική (Science & Engineering): Παρακολούθηση συστημάτων, βελτίωση συστήματος απόδοσης, απομόνωση σφαλμάτων λογισμικού, ανίχνευση λογοκλοπής λογισμικού, ανάλυσης σφαλμάτων συστήματος υπολογιστή, εντοπισμό εισβολέων και δυσλειτουργιών στα συστήματα δικτύου κ.λπ. (Sethi, Malhotra, & Verma, 2016)

Υπάρχουν πολλές τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται και είναι ευθύνη του αναλυτή να εντοπίσει τις καταλληλότερες για να εξάγει πληροφορία ανάλογα με τι καλείται να αναλύσει.
Οι βασικότερες τεχνικές εξόρυξης δεδομένων είναι:

- Μοτίβα παρακολούθησης (Tracking patterns): Η αναγνώριση μοτίβων στα σύνολα δεδομένων, δηλαδή η αναγνώριση κάποιας εκτροπής στα δεδομένα που συμβαίνει σε τακτά χρονικά διαστήματα ή μια άμπωτη ροή μιας συγκεκριμένης μεταβλητής με την πάροδο του χρόνου.
- Ταξινόμηση (Classification): Συγκέντρωση-συλλογή διαφόρων χαρακτηριστικών σε ευδιάκριτες κατηγορίες, οι οποίες στη συνέχεια χρησιμοποιούνται για να εξαχθούν συμπεράσματα ή να εξυπηρετήσουν κάποια λειτουργία.
- Συσχετίσεις (Associations): Αναζήτηση συγκεκριμένων συμβάντων ή χαρακτηριστικών που σχετίζονται σε μεγάλο βαθμό με ένα άλλο συμβάν ή χαρακτηριστικό.
- Ανίχνευση ακραίων τιμών (Outlier detection): Εντοπισμός ανωμαλιών η ακραίων τιμών στα δεδομένα.
- Ομαδοποίηση (Clustering): Η ομαδοποίηση είναι παρόμοια με την ταξινόμηση, αλλά περιλαμβάνει την ομαδοποίηση τμημάτων δεδομένων μαζί, με βάση τις ομοιότητές τους.
- Παλινδρόμηση (Regression): Χρησιμοποιείται κυρίως ως μορφή σχεδιασμού και μοντελοποίησης, για τον προσδιορισμό της πιθανότητας μιας συγκεκριμένης μεταβλητής, δεδομένης της παρουσίας άλλων μεταβλητών.

- Πρόβλεψη (Prediction): Χρησιμοποιείται για την προβολή των τύπων δεδομένων που θα προκύψουν στο μέλλον. Σε πολλές περιπτώσεις, μόνο η αναγνώριση και η κατανόηση των ιστορικών τάσεων αρκεί για να χαράξουμε μια κάπως ακριβή πρόβλεψη για το τι θα συμβεί στο μέλλον. (Alton, 2017)

Η βιβλιογραφική ανασκόπηση προσεγγίζεται με βάση που εστιάζουν οι ερευνητές:

- Τεχνικές, όπου ο ερευνητής δίνει έμφαση στην έρευνα με βάση συγκεκριμένες τεχνικές και αλγορίθμους, πόσο χρήσιμες είναι για το θέμα που ερευνούν και κατά πόσο μπορούν να βελτιωθούν και να παράγουν τα επιθυμητά αποτελέσματα.
- Πελάτες, όπου ο ερευνητής εστιάζει στους πελάτες και χρησιμοποιεί διάφορες τεχνικές για την ανάλυση και παραγωγή χρήσιμης πληροφορίας.
- Προϊόντα, όπου ο ερευνητής δίνει έμφαση στα προϊόντα, στις πωλήσεις τους και στις συνδέσεις μεταξύ τους για να ανακαλύψει χρήσιμα μοτίβα.

Κάποιες από αυτές τις έρευνες επικεντρώνονται σε περισσότερο από μια έννοιες και καλύπτουν ένα μεγαλύτερο εύρος των παραπάνω προσεγγίσεων, επίσης ο διαχωρισμός αυτών μπορεί να γίνει σε ακόμα μεγαλύτερο βάθος ανάλογα την γενικότερη τεχνική εξόρυξης γνώσης που χρησιμοποιούν οι ερευνητές. Οι πιο συχνές προσεγγίσεις στη συγκεκριμένη βιβλιογραφική ανασκόπηση είναι Association rules, Classification και Clustering που θα αποτελέσουν και τη μονάδα ανάλυσης (Unit of Analysis).

Η εξόρυξη κανόνων συσχέτισης (Association Rule Mining), είναι μια από τις πιο σημαντικές και καλά ερευνημένες τεχνικές εξόρυξης δεδομένων. Στοχεύει στην εξαγωγή ενδιαφέρον συσχετίσεων, συχνών μοτίβων, συσχετισμών ή περιστασιακών δομών μεταξύ συνόλων στοιχείων (itemset) από βάσεις δεδομένων (data bases) συναλλαγών ή άλλα αποθετήρια δεδομένων. Οι κανόνες συσχέτισης χρησιμοποιούνται ευρέως σε διάφορους τομείς όπως τα δίκτυα τηλεπικοινωνιών, τη διαχείριση αγοράς και κινδύνου, τον έλεγχο αποθεμάτων κ.λπ. (Kotsiantis & Kanellopoulos, 2006) Στην πολύ γνωστή δουλεία των Agrawal et al. (1993) αναφέρεται ότι δοθέντος μιας μεγάλης βάσης δεδομένων συναλλαγών πελατών, κάθε συναλλαγή αποτελείται από αντικείμενα που αγόρασε ένας πελάτης σε μια επίσκεψη. Ένα αλγόριθμος association rule δημιουργεί όλους τους σημαντικούς κανόνες συσχέτισης μεταξύ στοιχείων αυτής της βάσης δεδομένων. (Agrawal, Imielinski, & Swami, 1993)

Η ταξινόμηση (Classification) είναι μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται για να προβλέψει μια περίπτωση στοιχείων σε ποια ομάδα ανήκει. Κάποια κύρια είδη μεθόδων ταξινόμησης είναι: decision tree, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy. (Phyi, 2009)

Η ομαδοποίηση (Clustering) είναι μια κοινή τεχνική για την ανάλυση στατιστικών δεδομένων, η οποία χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης. Είναι η διαδικασία ομαδοποίησης παρόμοιων αντικειμένων σε διαφορετικές ομάδες, ή ακριβέστερα, η κατάτμηση ενός συνόλου δεδομένων σε

υποσύνολα, έτσι ώστε τα δεδομένα σε κάθε υποσύνολο να είναι σύμφωνα με κάποιο καθορισμένο μέτρο απόστασης. (Madhulatha, 2012)

2.2 Σχετικές τεχνικές

Μία από τις έρευνες που εστιάζουν στις τεχνικές είναι η δουλεία των Wang et al. (2019) όπου χρησιμοποιούν τον αλγόριθμο Apriori για την ανακάλυψη κανόνων συσχέτισης στα καλάθια πελατών και στη συνέχεια εφαρμόζουν τους κανόνες συσχέτισης και τον αλγόριθμο δέντρου αποφάσεων CART για να αποκαλύψουν τα χαρακτηριστικά της ομάδας πελατών και την ταξινόμηση τους. Η ανακάλυψη κανόνων συσχέτισης στα καλάθια πελατών και η εφαρμογή ενός αλγόριθμού δέντρου απόφασης μπορεί να χρησιμοποιηθεί και στην συγκεκριμένη περίπτωση χονδρεμπορίου. (Wang & Sun, 2019)

Ο Χριστοδουλάκης (2005) χρησιμοποιεί μια γνωστή τεχνική ταξινόμησης πελατών με βάση την RFM ανάλυση. Τα δεδομένα τα οποία χρησιμοποιεί αφορούν πελάτες τράπεζας με έμφαση στο e-banking και μελετάται η βαθμολογία RFM των ενεργών χρηστών e-banking μαζί και η κατάταξη τους σύμφωνα με το μοντέλο της πυραμίδας. Το μοντέλο πυραμίδας ομαδοποιεί τους πελάτες με βάση τα έσοδα που παράγουν. Ουσιαστικά η ανάλυση RFM υποδηλώνει ότι ο πελάτης που παρουσιάζει υψηλή βαθμολογία RFM θα πρέπει να πραγματοποιεί περισσότερες συναλλαγές και να έχει ως αποτέλεσμα υψηλότερα κέρδη για την τράπεζα. (Christodoulakis, 2005).

Η ανάλυση RFM θεωρείται σημαντική και για τις τράπεζες και συγκεκριμένα κομμάτια τους όπως το e-banking. Ένας πελάτης που επισκέφτηκε έναν ιστότοπο ηλεκτρονικής τραπεζικής Πρόσφατα (R) και Συχνά (F) και δημιούργησε πολλή Νομισματική Αξία (M) μέσω πληρωμών και πάγιων εντολών, είναι πολύ πιθανό να επισκεφθεί και να πραγματοποιήσει ξανά πληρωμές. Ουσιαστικά η ανάλυση RFM υποδηλώνει ότι ο πελάτης που παρουσιάζει υψηλή βαθμολογία RFM θα πρέπει κανονικά να πραγματοποιεί περισσότερες συναλλαγές και να έχει ως αποτέλεσμα υψηλότερα κέρδη για την τράπεζα. Η τεχνική RFM παρόλο που μπορεί να ομαδοποιήσει τους πελάτες σε σχέση με τη σημαντικότητα τους δεν βοηθά στην ταξινόμηση τους σε επαγγελματική ιδιότητα ή την ανακάλυψη πληροφορίας σχετικά με τις αγορές τους.

Έχοντας σαν επίκεντρο τις τεχνικές οι Chen et al. (2006) στην έρευνα τους, παρουσιάζουν μια νέα προσέγγιση για τη δόμηση ενός ταξινομητή, με βάση μιας εκτεταμένης τεχνική εξόρυξης κανόνων συσχέτισης. Το πρόβλημα της εξαγωγής κανόνων ταξινόμησης αντιμετωπίζεται σαν ένα πρόβλημα εξαγωγής κανόνων συσχέτισης. Προτείνουν τον αλγόριθμο GARC (Gain based Association Rule Classification) που δημιουργεί έναν ταξινομητή με ικανοποιητική ακρίβεια. Ο συγκεκριμένος αλγόριθμος, πρώτων συνδυάζει τις συμβατικές διαδικασίες

παραγωγής στοιχειό συνόλων και συσχετίσεων, δεύτερον η πληροφορία ενσωματώνεται μόνο για την παραγωγή των στοιχειό συνόλων συμπεριλαμβανομένης της τιμής διαχωρισμού που οδηγεί στη μείωση των υποψήφιων συνόλων. Τρίτων ενσωματώνονται συγκεκριμένες τεχνικές στη διαδικασία εξόρυξης έτσι ώστε να αποφεύγονται περιττοί και αντικρουόμενοι κανόνες. Ως αποτέλεσμα, το σύνολο που προκύπτει είναι πιο συμπυκνωμένο και κατανοητό και από πειράματα δεδομένων στην συγκεκριμένη εργασία, η ακρίβεια της ταξινόμησης αποδεικνύεται ικανοποιητική. Θα ήταν χρήσιμο στην περίπτωση της συγκεκριμένης διπλωματικής το πρόβλημα παραγωγής κανόνων ταξινόμησης σε επαγγελματική ιδιότητα να αντιμετωπιστεί σαν πρόβλημα εξαγωγής συσχετίσεων. (Chen, Liu, Yu, Wei, & Zhang, 2006)

Οι Pandey et al. (2009) στην εργασία τους προτείνουν το Association Rules Network σαν μια δομή για τη σύνθεση, κλάδεμα και ανάλυση μιας συλλογής κανόνων συσχέτισης στην κατασκευή υποψήφιων υποθέσεων. Από την σκοπιά της ανακάλυψής γνώσης τα ARN επιτρέπουν μια ανάλυση με επίκεντρο τον στόχο του ερευνητή. Η βασική ιδέα του ARN είναι ότι οι κανόνες συσχέτισης που ανακαλύπτονται μπορούν να συντεθούν, κλαδευτούν και ενσωματωθούν στο πλαίσιο συγκεκριμένων στόχων ώστε να λύσει το πρόβλημα των γενικών τεχνικών εξόρυξης γνώσης που παράγουν τεράστια πρότυπα υποθέσεων και δυσκολεύουν στις αποφάσεις. Τα ARN έχουν τα παρακάτω χαρακτηριστικά:

1. Κλάδεμα περιεχομένου: Οι κανόνες που παράγονται κλαδεύονται ως προς ένα συγκεκριμένο σκοπό, αν ο σκοπός αλλάξει τότε θα κλαδευτούν και διαφορετικά σύνολα κανόνων.
2. Δομή δικτύου: Παρέχουν ένα μηχανισμό για τον προσδιορισμό της σχέσης δικτύου μεταξύ σχετικών μεταβλητών και στόχου.
3. Παραγωγή υποθέσεων: Το ARN μπορεί να χρησιμεύσει ως γέφυρα μεταξύ των αποτελεσμάτων που παράγονται και της στατιστικής αξιολόγησής τους.

Το Association Rule Network θα μπορούσε να χρησιμοποιηθεί κατά την εξαγωγή κανόνων συσχέτισης ώστε να προκύψει γνώση με επίκεντρο τον στόχο της διπλωματικής αλλά είναι μια χρονοβόρα διαδικασία που στην προκυμμένη περίπτωση δεν θα μπορούσε να αξιοποιηθεί κατάλληλα. (Pandey, Chawla, Poon, Arunasalam, & Davis, 2009)

Οι Liu et al. (1998) στην έρευνά τους ενσωματώνουν τεχνικές classification και association rules. Η ενσωμάτωση πετυχαίνετε εξάγοντας ένα ειδικό υποσύνολο κανόνων συσχέτισης CARs (class association rules) των οποίων η δεξιά πλευρά του κανόνα περιορίζεται στην ιδιότητα της κλάσης ταξινόμησης. Χρησιμοποιούν έναν υπάρχον αλγόριθμο εξόρυξης κανόνων συσχέτιση (Agrawal, Imielinski, & Swami, 1993) ς που παράγει όλα τα CARs και ικανοποιεί το minimum support και minimum confidence. Το προτεινόμενο πλαίσιο associative classification αποτελείται από τρία βήματα: διακριτοποίηση συνεχών χαρακτηριστικών, δημιουργία όλων των class association rules (CARs), και τη δημιουργία ταξινομητή βασιζόμενο στα παραγόμενα CARs. Ο προτεινόμενος αλγόριθμος ονομάζεται CBA (Classification Based on Associations). Αποτελείται από δύο μέρη, μια γεννήτρια κανόνων (που ονομάζεται CBA-RG), η οποία βασίζεται στον αλγόριθμο Apriori για την εύρεση κανόνων

συσχέτισης (Agrawal, Imielinski, & Swami, 1993) , και έναν δημιουργό ταξινομητή (που ονομάζεται CBA-CB). Επισημάνουν ότι η προτεινόμενη μέθοδος όχι μόνο δίνει έναν νέο τρόπο κατασκευής ταξινομητών, αλλά βοηθά επίσης στην επίλυση ορισμένων προβλημάτων που υπάρχουν στα τρέχοντα συστήματα ταξινόμησης. Σε αυτή την εργασία θα χρειαστεί να γίνει συγχώνευση τεχνικών association rule και classification για να μπορέσει να παραχθεί ποιοτική πληροφορία από τα διαθέσιμα δεδομένα. (Liu, Hsu, & Ma, 1998)

Σε συνέχεια του παραπάνω άρθρου αξίζει να αναφερθεί το πακέτο arulesCBA σε γλώσσα R που παρέχει την υποδομή για την ταξινόμηση βάσει κανόνων συσχέτισης, συμπεριλαμβανομένου των αλγορίθμων CBA, CMAR, CPAR, C4.5, FOIL, PART, PRM, RCAR, και RIPPER που δημιουργούν ταξινομητές συσχετιστικής ταξινόμησης. (Hahsler, Johnson, & Giallanza, arulesCBA: Classification Based on Association Rule, 2022)

Οι Hao et al. (2009) στην έρευνα τους αναφέρουν χαρακτηριστικά ότι η ταξινόμηση βασισμένη σε προγνωστικούς κανόνες συσχέτισης (predictive association rules) (CPAR) είναι ένα είδος μεθόδων ταξινόμησης συσχετισμού που συνδυάζει τα πλεονεκτήματα τόσο associative classification όσο και παραδοσιακής rule-based classification. Η CPAR είναι πιο αποτελεσματική από την παραδοσιακή ταξινόμηση βάση κανόνων, επειδή αποφεύγεται πολύς επαναλαμβανόμενος υπολογισμός και μπορούν να επιλεχθούν πολλαπλά χαρακτηριστικά για τη δημιουργία πολλών κανόνων ταυτόχρονα. Η δημιουργία κανόνων του CPAR τον καθιστά έναν από τους πιο αποτελεσματικούς associative classification αλγόριθμους. Παρόλα αυτά παρουσιάζει τα εξής προβλήματα

1. Ο αριθμός των κανόνων κάθε κλάσης μπορεί να κυμαίνεται από αρκετές δεκάδες έως αρκετές χιλιάδες.
2. Στο στάδιο της ταξινόμησης, κάθε κατηγορία αντιμετωπίζεται ομοιόμορφα που αυξάνει την πιθανότητα εσφαλμένης ταξινόμησης.
3. Η CPAR είναι άχρηστη για περιπτώσεις που δεν πληρούν κανέναν κανόνα.

Οι διαδικασίες πρόβλεψης έχουν τις αδυναμίες της ανισόρροπης κατανομής κανόνων (rule distribution imbalance) και της διακοπής εσφαλμένων κανόνων κλάσης (interruption of incorrect class rules). Για την αντιμετώπιση αυτών των προβλημάτων, οι ερευνητές χρησιμοποιούν Class Weighting Adjustment, Center Vector-based Pre classification και Post-processing με Support Vector Machine τεχνικές, που δημιουργούν την ICPAR (Improved Classification Based on Predictive Association Rules). Σύμφωνα με την συγκεκριμένη εργασία το ICPAR πέτυχε μεγαλύτερη ακρίβεια από το CPAR. Παρά τις δυνατότητες της ταξινόμησης βασισμένης σε προγνωστικούς κανόνες συσχέτισης είναι μια χρονοβόρα και πολύπλοκη διαδικασία και δεν θα αξιοποιηθεί στην συγκεκριμένη εργασία. (Hao, Wang, Yao, & Zhang, 2009)

Σε μία άλλη έρευνα οι Thabtah et al. (2005) παρουσιάζουν μια νέα μέθοδος ταξινόμησης που ονομάζεται ταξινόμηση πολλαπλών κλάσεων με βάση τους κανόνες συσχέτισης (MCAR: Multi-class Classification based on Association Rule). Το MCAR χρησιμοποιεί μια αποτελεσματική τεχνική για την ανακάλυψη συχνών στοιχείων και χρησιμοποιεί μια μέθοδο κατάταξης βάση κανόνων που διασφαλίζει ότι

οι λεπτομερείς κανόνες με υψηλά επίπεδα εμπιστοσύνης αποτελούν μέρος του ταξινομητή. Στην εργασία τους αναφέρουν ότι:

- Το MCAR χρησιμοποιεί μια τεχνική για την ανακάλυψη συχνών στοιχείων που απαιτεί μόνο ένα πέρασμα, καταναλώνοντας σημαντικά λιγότερο χώρο αποθήκευσης και χρόνο εκτέλεσης από τις προσεγγίσεις πολλαπλών περασμάτων.
- Το MCAR ανακαλύπτει και δημιουργεί συχνά στοιχεία και κανόνες σε μία φάση. Άλλες μέθοδοι συσχετιστικής ταξινόμησης, όπως το CPAR και το CBA, ανακαλύπτουν συχνά στοιχεία σε μια φάση και στη συνέχεια καθορίζουν ποιο υποσύνολο από αυτά σχηματίζει τον ταξινομητή σε ξεχωριστή φάση.
- Το MCAR εισάγει μια τεχνική κατάταξης κανόνων που ελαχιστοποιεί τη χρήση της τυχαιότητας

Στα πλαίσια της συγκεκριμένης εργασίας και βάση του προβλήματος ανάλυσης των αγορών χονδρικής, η ανακάλυψη κανόνων συσχέτισης είναι απαραίτητη και η χρήση αυτών των κανόνων στη ταξινόμηση των πελατών σε επάγγελμα θα μπορούσε να αξιοποιηθεί ανάλογα. (Thabtah, Cowling, & Peng, 2005)

Στο άρθρο τους οι Niu et al. (2009) προτείνουν μια νέα μέθοδο συσχετιστικής ταξινόμησης (association classification) που βασίζεται σε συμπαγείς κανόνες συσχέτισης, επεκτείνουν τον αλγόριθμο Apriori, και εξετάζουν το ενδιαφέρον, τη σημασία και τις αλληλεπικαλυπτόμενες σχέσεις μεταξύ των κανόνων. Αναφέρουν ότι τα πειραματικά αποτελέσματα τους δείχνουν ότι ο αλγόριθμος έχει καλύτερη ακρίβεια ταξινόμησης σε σύγκριση με τον CBA και το CMAR και είναι εξαιρετικά κατανοητός. Η χρήση του αλγορίθμου Apriori θα παρουσιαστεί και αναλυτικά παρακάτω αφού χρησιμοποιείται στην εργασία για την ανακάλυψη κανόνων συσχέτισης. (Niu, Xia, & Zhang, 2009)

Οι Li et al. (2001) στην εργασία τους αναφέρουν ότι η συσχετιστική ταξινόμηση υποφέρει από το τεράστιο σύνολο εξορυσσόμενων κανόνων και μερικές φορές από μεροληπτική ταξινόμηση ή υπερπροσαρμογή, καθώς βασίζεται σε έναν μόνο κανόνα υψηλής εμπιστοσύνης. Γι' αυτό και προτείνουν μια νέα μέθοδο συσχετιστικής ταξινόμησης, την CMAR, δηλαδή την ταξινόμηση με βάση τους κανόνες πολλαπλών συσχετίσεων. Η μέθοδος αυτή εφαρμόζει μια δομή δέντρου CR για την αποτελεσματική εξόρυξη κανόνων συσχέτισης και κλαδεύει κανόνες με βάση την εμπιστοσύνη και τη συσχέτιση. Η ταξινόμηση γίνεται με βάση μια σταθμισμένη ανάλυση χ^2 χρησιμοποιώντας πολλαπλούς ισχυρούς κανόνες συσχέτισης. Αναφέρουν ότι από τις δοκιμές τους σε 26 βάσεις δεδομένων τα αποτελέσματα δείχνουν ότι η CMAR είναι συνεπής, εξαιρετικά αποτελεσματική στην ταξινόμηση διαφόρων ειδών βάσεων δεδομένων και έχει καλύτερη μέση ακρίβεια ταξινόμησης σε σύγκριση με το CBA και το C4.5. Επιπλέον, η μελέτη απόδοσης τους δείχνει ότι η μέθοδος είναι εξαιρετικά αποδοτική και επεκτάσιμη σε σύγκριση με άλλες αναφερόμενες μεθόδους συσχετιστικής ταξινόμησης. (Li, Han, & Pei, 2001)

Διαπιστώνεται ότι η ανακάλυψη κανόνων συσχέτισης στα καλάθια πελατών και η εφαρμογή ενός αλγόριθμού δέντρου απόφασης θα μπορούσε να χρησιμοποιηθεί και

στην συγκεκριμένη περίπτωση χονδρεμπορίου καθώς και να γίνει μια προσπάθεια ταξινόμησης των πελατών σε κάποιο επάγγελμα όπως συμβαίνει στην δουλεία των Wang et al. (2019). Θα ήταν χρήσιμο το πρόβλημα παραγωγής κανόνων ταξινόμησης σε επαγγελματική ιδιότητα να αντιμετωπιστεί σαν πρόβλημα εξαγωγής συσχετίσεων όπως αναφέρουν οι Chen et al. (2006) επομένως θα μπορούσε να γίνει μια συγχώνευση τεχνικών association rule και classification όπως παρουσιάζεται από τους Liu et al. (1998) για να παραχθεί ποιοτική πληροφορία από τα διαθέσιμα δεδομένα που παρατέθηκαν στο πλαίσιο αυτή της εργασίας. Βάση του προβλήματος ανάλυσης των αγορών χονδρικής η ανακάλυψη κανόνων συσχέτισης είναι απαραίτητη καθώς και η χρήση αυτών των κανόνων στη ταξινόμηση των πελατών σε επάγγελμα Thabtah et al. (2005). Για την ανακάλυψη κανόνων συσχέτισης θα χρησιμοποιηθεί ο αλγόριθμος Apriori (Agrawal, Imielinski, & Swami, 1993), (Niu, Xia, & Zhang, 2009).

Παρατηρείται ότι η τεχνική RFM παρόλο που θα μπορούσε να ομαδοποιήσει τους πελάτες σε σχέση με τη σημαντικότητα τους (Christodoulakis, 2005) δεν θα βοηθούσε στην ταξινόμηση σε επαγγελματική ιδιότητα ή την ανακάλυψη πληροφορίας σχετικά με τις αγορές τους. Το Association Rule Network (Pandey, Chawla, Poon, Arunasalam, & Davis, 2009) θα μπορούσε να αξιοποιηθεί κατά την εξαγωγή κανόνων συσχέτισης ώστε να προκύψει γνώση με επίκεντρο τον στόχο της διπλωματικής αλλά είναι μια χρονοβόρα διαδικασία που στην προκυμμένη περίπτωση δεν είναι χρήσιμη. Η διαδικασία της ταξινόμησης βασισμένης σε προγνωστικούς κανόνες συσχέτισης (Hao, Wang, Yao, & Zhang, 2009) είναι πολύ χρονοβόρα και πολύπλοκη για να αξιοποιηθεί κατάλληλα και όμοια για τη νέα μέθοδος που προτείνουν οι Li et al. (2001) ενώ είναι αρκετά ενδιαφέρον και χρήσιμη δεν μπορεί να χρησιμοποιηθεί.

2.3 Σχετικές τεχνικές που εστιάζουν σε πελάτες και προϊόντα

Έχοντας σαν επίκεντρο τους πελάτες οι Baby et al. (2012) στην εργασία τους αναλύουν έναν τεράστιο όγκο δεδομένων με πελάτες και τους ταξινομούν με βάση τις συμπεριφορές τους ενώ κάνουν και προβλέψεις γι' αυτούς. Ο ταξινομητής προβλέπει σε ποια κατηγορία έχει μεγαλύτερη πιθανότητα να ταξινομηθεί ένας πελάτης. Παράγεται ένα μοντέλο δεδομένων με βάση το ιστορικό των πελατών στην τράπεζα. Στη συνέχεια, τα δεδομένα του δείγματος ταξινομούνται χρησιμοποιώντας τον αλγόριθμο ταξινόμησης Naive Bayesian και τα τοποθετούν στην κατάλληλη κατηγορία βάση την εκ των υστέρων πιθανότητα, βάσει αυτής της πιθανότητας μπορεί να προβλεφθεί το ποσοστό του κινδύνου κύρωσης δανείου για τους πελάτες. (Baby & Priyanka, 2012)

Μια ακόμα εργασία με επίκεντρο τους πελάτες υλοποιήθηκε από τους Park et al. (2009) , σε αυτήν οι συγγραφείς προτείνουν την κατασκευή ενός προφίλ πελατών που βασίζεται σε ατομικές και ομαδικές πληροφορίες συμπεριφοράς, όπως κλικ, εισαγωγές καλαθιού, αγορές και πεδία ενδιαφέροντος. Εφαρμόζουν ένα σύστημα συστάσεων χρησιμοποιώντας το προτεινόμενο μοντέλο και αξιολογούν την απόδοση της σύστασης με βάση γνωστά evaluation metrics. Χωρίζουν τη διαδικασία σε τρείς φάσεις: 1) Υπολογίζουν τα ατομικά ενδιαφέροντα χρησιμοποιώντας προσωπικά δεδομένα συμπεριφοράς. 2) Υπολογίζονται τα ενδιαφέροντα βάση πληροφοριών

συμπεριφοράς της ομάδας. Μια ομάδα είναι ένα σύνολο πελατών που έχουν παρόμοια δημογραφικά δεδομένα. 3) Το προφίλ του πελάτη κατασκευάζεται χρησιμοποιώντας τα χαρακτηριστικά προϊόντων καθώς και ατομικά ή ομαδικά ενδιαφέροντα. (Park & Chang, 2009)

Οι Abirami et al. (2016) στην εργασία τους κάνουν μια νέα προσέγγιση ταξινόμησης πελατών με βάση το μοντέλο RFM, ασχολούνται με τα δεδομένα των πελατών με σκοπό την ανάλυση και πρόβλεψη της συμπεριφοράς τους χρησιμοποιώντας τεχνικές ομαδοποίησης και εξόρυξης κανόνων συσχέτισης. Αναφέρουν ότι κατηγοριοποιούν τους πελάτες σε νέους η επαναλαμβανόμενους, αλλά η αγοραστική συμπεριφορά των νέων δεν προβλέπεται λόγο έλλειψης ιστορικότητας. Για τους επαναλαμβανόμενους πελάτες η πρόβλεψη συμβαίνει με clustering αλγόριθμους, το μοντέλο RFM και association rules. (Abirami & Pattabiraman, 2016)

Παρόλο που η εργασία των Baby et al. (2012) γίνεται σε διαφορετικά πλαίσια η λογική της ταξινόμησης παραμένει ίδια και στο λιανεμπόριο, χονδρεμπόριο και ίσως θα μπορούσε να χρησιμοποιηθεί Naive Bayesian για τον σκοπό της συγκεκριμένης διπλωματικής. Δεν είναι εφικτό το Profiling των πελατών βάση των διαθέσιμων δεδομένων επόμενός η εργασία των Park et al. (2009) δεν είναι χρήσιμη στα συγκεκριμένα πλαίσια. Οι Abirami et al. (2016) χρησιμοποιούν RFM Analysis στο λιανεμπόριο κάτι που είναι παρεμφερές στο χονδρεμπόριο, αλλά ο τρόπος ταξινόμησης σε παλιούς και νέους πελάτες δεν θα μπορούσε να αξιοποιηθεί για τον στόχο αυτής της εργασίας.

Οι Ya-Han et al. (2014) στην εργασία τους έχουν ως στόχο την ανάπτυξη ενός αλγόριθμου για την ανακάλυψη μοτίβων RFM που μπορούν να προσεγγίσουν το σύνολο των μοτίβων RFM-πελατών χωρίς πληροφορίες για αυτούς. Αντί να αξιολογεί τις τιμές των μοτίβων από την άποψη του πελάτη, αξιολογεί τα μοτίβα λαμβάνοντας υπόψη τα χαρακτηριστικά RFM. Προτείνουν μια δομή δέντρου, που ονομάζεται RFM-pattern-tree, για τη συμπίεση και αποθήκευση ολόκληρης της βάσης δεδομένων συναλλαγών και αναπτύσσουν έναν αλγόριθμο βασισμένο στην ανάπτυξη προτύπων, που ονομάζεται RFMP-growth, για την ανακάλυψη όλων των RFM μοτίβων σε ένα RFM -δέντρο. Η ενσωμάτωση RFM ανάλυσης και Association Rule Mining έχει αποδειχθεί πολύ αποτελεσματική για την ανακάλυψη της αγοραστικής συμπεριφοράς ενός καταναλωτή. Ενσωματώνοντας την έννοια της ανάλυσης RFM στην εξόρυξη συχνών προτύπων, η μελέτη αναπτύσσει μια νέα προσέγγιση στην ανακάλυψη προτύπων RFM-πελάτη. (Ya-Han & Tzu-Wei, 2014)

Με επίκεντρο τα προϊόντα οι Mu-Chen et al. (2007) στην μελέτη τους προτείνουν μια προσέγγιση εξόρυξης δεδομένων για τη λήψη αποφάσεων σχετικά με το ποια προϊόντα θα αποθηκευτούν, πόσος χώρος στο ράφι θα διατεθεί στα αποθηκευμένα προϊόντα και πού θα τα εκθέσουν. Εφαρμόζει τη συσχέτιση αντί για την ελαστικότητα χώρου για να διαμορφώσει το μαθηματικό μοντέλο για την ποικιλία προϊόντων. Η προτεινόμενη διαδικασία διαχείρισης χώρου ραφιών, ξεκινά με εξόρυξη κανόνων συσχέτισης πολλαπλών επιπέδων από δεδομένα συναλλαγών, για τη παραγωγή των σχέσεων μεταξύ ειδών, μεταξύ υποκατηγοριών προϊόντων και μεταξύ κατηγοριών προϊόντων. Στη συνέχεια, η διαδικασία προχωρά σε ποικιλία

προϊόντων στην οποία λαμβάνονται υπόψη τα κέρδη των συχνών στοιχειό συνόλων. Αναφέρονται χαρακτηριστικά τα παρακάτω πλεονεκτήματα:

1. Οι κανόνες συσχέτισης λαμβάνονται από απευθείας ανάλυση της βάσης δεδομένων συναλλαγών, είναι αξιόπιστα για διαχείριση χώρου ραφιών.
2. Εξαλείφεται η τεράστια εκτίμηση των παραμέτρων της ελαστικότητας του χώρου και μειώνεται η λάθος εκτίμηση και τα δαπανηρά πειράματα.
3. Οι κανόνες συσχέτισης ανταποκρίνονται γρήγορα στις αλλαγές στην αγορά δεδομένου ότι τα δεδομένα συναλλαγών συλλέγονται έγκαιρα από το σύστημα POS
4. Το μοντέλο συμπεριλαμβάνει τα βασικά προϊόντα για την εικόνα του καταστήματος και τα προβαλλόμενα προϊόντα καθορίζονται χρησιμοποιώντας τις συσχετίσεις μεταξύ ειδών προϊόντων.
5. Με βάση τους κανόνες σύνδεσης πολλαπλών επιπέδων οι κατηγορίες προϊόντων, οι υποκατηγορίες και τα είδη μπορούν να κατανεμηθούν με βάση τις ενώσεις και τα κέρδη τους.

Μια χρήσιμη προσέγγιση στους κανόνες συσχέτισης είναι ότι μπορεί να προκύπτουν και σε επίπεδο προϊοντικής κατηγορίας κάτι που θα μπορούσε να αξιοποιηθεί στην συγκεκριμένη εργασία, ώστε να μειωθεί ο όγκος των συσχετίσεων. (Mu-Chen & Chia-Ping, 2007)

Μια ακόμα μελέτη με επίκεντρο τα προϊόντα είναι το «Προσαρμοσμένο πακέτο πρότασης με βάση κανόνες συσχέτισης προϊοντικών κατηγοριών για online super market των Fang et al. (2018). Ένα customized bundle είναι μια λίστα προϊόντων που προτείνεται στον καταναλωτή όπου μπορούν να επιλέξουν τα αγαπημένα τους προϊόντα σύμφωνα με τις προτιμήσεις τους. Είναι ένας αποτελεσματικός τρόπος να απλοποιηθεί η διαδικασία αγορών του πελάτη και να μειωθεί το κόστος εκπλήρωσης της παραγγελίας για τα διαδικτυακά Super Market. Πραγματοποιείται συνδυάζοντας εξόρυξη κανόνων συσχέτισης, τμηματοποίηση πελατών και recommendation τεχνικών. Με τη χρήση κανόνων συσχέτισης σε επίπεδο προϊοντικής κατηγορίας αποφεύγεται η περιττή μάζα κανόνων σε επίπεδο προϊόντος. Τα προϊόντα που προτείνονται μέσα σε κάθε κατηγορία προκύπτουν από την τμηματοποίηση του πελάτη και την ημερομηνία κατάταξης τους βάση της τμηματοποίηση. Η πρόταση σε πελάτη δεν εμπίπτει στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας αλλά υπάρχει μια χρήσιμη προσέγγιση στην παραπάνω έρευνα που είναι η εξαγωγή κανόνων συσχέτισης στις προϊοντικές κατηγορίες κάτι που θα μπορούσε να αξιοποιηθεί ανάλογα. (Fang, Xia, Wang, & Lan, 2018)

Άλλη μια ενδιαφέρον εργασία έχει τίτλο «Εξόρυξη σπάνιων κανόνων συσχέτισης μέσω συσταδοποίησης συναλλαγών, των Koh et al. (2008) που έκανε την εμφάνιση της στο Αυστραλιανό συνέδριο Data Mining, στην οποία χρησιμοποιείται η προσέγγιση για τη ομαδοποίηση συναλλαγών πριν από την εξόρυξη κανόνων συσχέτισης. Αποδεικνύει ότι η προ επεξεργασία του συνόλου δεδομένων με συσταδοποίηση επιτρέπει σε κάθε συστάδα να εκφράσει τους δικούς της κανόνες συσχέτισης χωρίς να μολύνονται από άλλες υποομάδες που έχουν διαφορετικά μοτίβα σχέσεων. Αναφέρουν ότι οι σπάνιοι κανόνες που παράγονται από κάθε συστάδα είναι

πιο ενημερωτικοί από τους κανόνες συσχέτισης που παράγονται από το σύνολο των δεδομένων. (Koh & Pears, 2008)

Διαπιστώνεται από την μελέτη των Mu-Chen et al.(2007) ότι η πληροφορία που παράγεται από την εξόρυξη γνώσης μπορεί να χρησιμοποιηθεί ακόμα και για την αναδιάταξη των προϊόντων μέσα σε ένα κατάστημα. Μια χρήσιμη προσέγγιση στους κανόνες συσχέτισης είναι ότι μπορεί να προκύπτουν και σε επίπεδο προϊοντικής κατηγορίας κάτι που θα μπορούσε να αξιοποιηθεί στην συγκεκριμένη εργασία, ώστε να μειωθεί ο όγκος των συσχετίσεων. Η ίδια προσέγγιση εμφανίζεται και στην εργασία των Fang et al. (20018) όπου εστιάζουν στους κανόνες συσχέτισης προϊοντικών κατηγοριών ώστε να αποφεύγεται η περιττή μάζα κανόνων που εμφανίζεται σε επίπεδο προϊόντος. Η εργασία των Yan-Han et al. (2014) εστιάζει σε RFM ανάλυση με τη διαφορά ότι δίνει έμφαση στα προϊόντα και στις αγορές των πελατών χωρίς να έχει πληροφορίες γι' αυτούς. Παρόλο που η RFM Ανάλυση είναι μια πολύ καλή τεχνική ο τρόπος που ταξινομεί τους πελάτες δεν είναι χρήσιμος στα πλαίσια της συγκεκριμένης διπλωματικής. Τέλος η συσταδοποίηση των συναλλαγών πριν από την εξόρυξη κανόνων συσχέτισης των Koh et al. (2008) θα μπορούσε να δημιουργήσει νέα πληροφορία και μοναδικούς κανόνες που δεν μολύνονται από άλλους ή από το σύνολο των δεδομένων.

Στον παρακάτω πίνακα εμφανίζονται συγκεντρωτικά οι τεχνικές που χρησιμοποιούνται από τη βιβλιογραφία οι στόχοι και τα ευρήματα τους (Πίνακας 3).

Πίνακας 1: Συγκεντρωτικός πίνακας τεχνικών

Συγγραφείς	Στόχος	Τεχνικές	Ευρήματα
(Wang & Sun, 2019)	Κανόνες συσχέτισης στα καλάθια πελατών. Εφαρμογή κανόνων συσχέτισης. Εφαρμογή δέντρου αποφάσεων CART	Apriori CART	Εμφάνιση χαρακτηριστικών ομάδων πελατών. Εμφάνιση προϊόντων που αγοράζονται μαζί. Προτάσεις ως προς την τοποθέτηση των προϊόντων.
(Christodoulakis, 2005)	RFM βαθμολόγηση των ενεργών χρηστών e-banking. Κατάταξη χρηστών σύμφωνα με το μοντέλο της πυραμίδας.	RFM Pyramid Model	Πελάτης με υψηλή βαθμολογία RFM θα πρέπει να πραγματοποιεί περισσότερες συναλλαγές. Η τράπεζα θα πρέπει να εστιάσει στους πελάτες που ανήκουν στο 20% των πιο σημαντικών.
(Chen, Liu, Yu, Wei, & Zhang, 2006)	Δημιουργία ταξινομητή με βάση μιας εκτεταμένης τεχνικής εξόρυξης κανόνων συσχέτισης	GARC	Συμπυκνωμένο και κατανοητό σύνολο κανόνων. Ικανοποιητική ακρίβεια ταξινόμησης.
(Pandey,	Επίλυση του πρόβλημα των	ARN	Σύνθεση, κλάδεμα και ανάλυση μιας

Chawla, Poon, Arunasalam, & Davis, 2009)	γενικών τεχνικών εξόρυξης γνώσης που παράγουν τεράστια πρότυπα υποθέσεων και δυσκολεύουν στις αποφάσεις		συλλογής κανόνων συσχέτισης στην κατασκευή υποψήφιων υποθέσεων. Επικέντρωση σε συγκεκριμένους στόχους.. Διευκόλυνση αποφάσεων.
(Liu, Hsu, & Ma, 1998)	Δημιουργία των class association rules (CARs). Δημιουργία ταξινομητή βασιζόμενο στα παραγόμενα CARs	CBA, Apriori	Νέος τρόπος κατασκευής ταξινομητών Επίλυση προβλημάτων στα τρέχοντα συστήματα ταξινόμησης.
(Hao, Wang, Yao, & Zhang, 2009)	Εξέλιξη της CPAR στην ICPAR (Improved Classification Based on Predictive Association Rules)	Class Weighting Adjustment, Center Vector-based Pre classification, Support Vector Machine	Αντιμετώπιση προβλημάτων της CPAR. Εξισορρόπηση ικανότητας ταξινόμησης κάθε κατηγορίας.. Φόρτωση κλάσεων με υψηλό επίπεδο ενδιαφέροντος. Απόρριψη παραδειγμάτων που δεν πληρούν κανένα κανόνα.
(Thabtah, Cowling, & Peng, 2005)	Ταξινόμηση σε πολλαπλές κλάσεις με βάση κανόνες συσχέτισης	MCAR	Χρησιμοποιεί μια αποτελεσματική τεχνική για την ανακάλυψη συχνών στοιχείων. Χρησιμοποιεί μια μέθοδο κατάταξης βάση κανόνων. Διασφαλίζει ότι οι κανόνες με υψηλά επίπεδα εμπιστοσύνης αποτελούν μέρος του ταξινομητή.
(Niu, Xia, & Zhang, 2009)	Έρευνα μεθόδου συσχετιστικής ταξινόμησης	Apriori, CBA, CMAR	Καλύτερη ακρίβεια ταξινόμησης συγκριτικά με τους CBA και CMAR. Εξαιρετικά κατανοητός.
(Li, Han, & Pei, 2001)	Έρευνα μεθόδου συσχετιστικής ταξινόμησης CMAR	CMAR, CR-tree, FP-growth,	Συνεπής και εξαιρετικά αποτελεσματική στην ταξινόμηση διαφόρων ειδών βάσεων δεδομένων. Καλύτερη μέση ακρίβεια ταξινόμησης σε σύγκριση με CBA και C4.5.

			Εξαιρετικά αποδοτική και επεκτάσιμη σε σύγκριση με άλλες μεθόδους συσχετιστικής ταξινόμησης
(Baby & Priyanka, 2012)	Ταξινόμηση πελατών με βάση συμπεριφορές και προβλέψεις	Naive Bayesian	Μοντέλο δεδομένων με βάση το ιστορικό των πελατών στην τράπεζα. Τοποθέτηση στην κατάλληλη κατηγορία βάση την εκ των υστέρων πιθανότητα. Πρόβλεψη ποσοστού κινδύνου κύρωσης δανείου για τους πελάτες.
(Park & Chang, 2009)	Κατασκευή προφίλ πελατών βασισμένο σε ατομικές και ομαδικές πληροφορίες συμπεριφοράς (κλικ, εισαγωγές καλαθιού, αγορές ,πεδία ενδιαφέροντος)	Product profile model Customer profile model RS_IP RS_IB RS_IGB	Υπολογισμός ατομικών ενδιαφέροντών χρησιμοποιώντας προσωπικά δεδομένα συμπεριφοράς. Υπολογισμός ενδιαφέροντών βάση πληροφοριών συμπεριφοράς ομάδας. Κατασκευή προφίλ πελάτη χρησιμοποιώντας χαρακτηριστικά προϊόντος καθώς και μεμονωμένα και ομαδικά ενδιαφέροντα.
(Abirami & Pattabiraman, 2016)	Ταξινόμησης πελατών βάση μοντέλου RFM, ανάλυση και πρόβλεψη συμπεριφοράς	RFM, K-means	Κατηγοριοποίηση πελατών σε νέους ή επαναλαμβανόμενους. Αδυναμία πρόβλεψης αγοραστικής συμπεριφοράς των νέων λόγο έλλειψης ιστορικότητας. Πρόβλεψη με clustering αλγόριθμους, μοντέλο RFM και association rules για επαναλαμβανόμενους πελάτες.
(Ya-Han & Tzu-Wei, 2014)	Ανάπτυξη αλγόριθμου για την ανακάλυψη μοτίβων RFM που προσεγγίζουν το σύνολο μοτίβων RFM-πελάτη χωρίς πληροφορίες για τους πελάτες.	RFM RFM-pattern-tree RFMP-growth	Αξιολόγηση βαθμολογίας RFM συχνών προτύπων, Ορισμός μοτίβων RFM ως προσεγγίσεις στα πρότυπα RFM-πελάτη. Πρόταση δομής δέντρου (RFM-pattern-tree)για την αποθήκευση ολόκληρης της βάσης δεδομένων. Ανάπτυξη νέας μεθόδου (RFMP-

			growth) για την αποτελεσματική ανακάλυψη ενός πλήρους συνόλου μοτίβων RFM.
(Mu-Chen & Chia-Ping, 2007)	Εξόρυξης δεδομένων για τη λήψη αποφάσεων σχετικά με το ποια προϊόντα θα αποθηκευτούν, πόσος χώρος στο ράφι θα διατεθεί στα αποθηκευμένα προϊόντα και πού θα τα εκθέσουν	Multi-level association rules	Κανόνες συσχέτισης από απευθείας ανάλυση της βάσης δεδομένων συναλλαγών. Εξάλειψη της τεράστιας εκτίμησης των παραμέτρων της ελαστικότητας του χώρου. Μείωση λάθος εκτίμησης από δαπανηρά πειράματα. Γρήγορη ανταπόκριση κανόνων συσχέτισης. Τα προβαλλόμενα προϊόντα καθορίζονται χρησιμοποιώντας τις συσχετίσεις μεταξύ ειδών προϊόντων. Κατανομή υποκατηγοριών και ειδών με βάση τις ενώσεις και τα κέρδη τους.
(Fang, Xia, Wang, & Lan, 2018)	Δημιουργία customized bundle - λίστα προϊόντων που προτείνεται στον καταναλωτή	Traditional clustering methods, Apriori	Αποτελεσματικός τρόπος προτάσεων προϊόντων προς τους καταναλωτές. Αποφυγή περιττής μάζας κανόνων συσχέτισης. Πρόταση προϊόντων βάση τη τμηματοποίηση του πελάτη και την ημερομηνία κατάταξης.
(Koh & Pears, 2008)	Έρευνα κανόνων συσχέτισης που προκύπτουν από συστάδες σχετικά με το αν είναι πιο αποτελεσματικοί από αυτούς στο σύνολο των δεδομένων.	Apriori-Inverse (Koh &Rountree 2005)	Συσταδοποίηση του συνόλου δεδομένων και εξαγωγή κανόνων συσχέτισης σε κάθε μια συστάδα. Οι κανόνες που παράγονται από κάθε συστάδα δεν μολύνονται από άλλες υποομάδες με διαφορετικά μοτίβα σχέσεων. Οι κανόνες που παράγονται από κάθε συστάδα είναι πιο ενημερωτικοί από

	τους κανόνες που παράγονται στο σύνολο των δεδομένων.

2.4 Αξιοποίηση τεχνικών

Όλες οι παραπάνω τεχνικές χρησιμοποιούνται σε ερευνητικό επίπεδο και παράγουν πολύτιμη πληροφορία. Υπάρχουν όμως κάποιες από αυτές που εφαρμόζονται σε επίπεδο επιχείρησης και ξεφεύγουν από τα ακαδημαϊκά πλαίσια.

2.4.1 Η οπτική των επιχειρήσεων

Οι Wang et al. (2019) κάνουν Market Basket Analysis χρησιμοποιώντας Apriori και δέντρα απόφασης και παρόλο που οι συγκεκριμένοι το κάνουν σε ερευνητικό επίπεδο η τεχνική αυτή είναι πολύ γνωστή και χρησιμοποιείται εδώ και χρόνια σε διάφορες επιχειρήσεις κυρίως στο λιανεμπόριο (π.χ. Amazon). Ο Χριστοδουλάκης (2005) χρησιμοποιεί τη γνωστή τεχνική ταξινόμησης πελατών με βάση την RFM ανάλυση, σε επιχειρηματικό επίπεδο όπου τα δεδομένα τα οποία χρησιμοποιεί αφορούν πραγματικούς πελάτες τράπεζας με έμφαση στο e-banking. Οι Liu et al. (1998) στην έρευνά τους Integrating Classification and Association Rule Mining ενσωματώνουν τεχνικές classification και association rules. Τέτοιου είδους ενσωμάτωση είναι ένα πολύ συχνό φαινόμενο κατά τη διαδικασία εξόρυξης γνώσης από τα δεδομένα και πολλές επαγγελματικές εργασίες την αξιοποιούν. Οι τεχνικές ταξινόμησης και πρόβλεψης που αναλύουν οι Baby et al. (2018) είναι πολύ διαδεδομένες και κλασικές τεχνικές που χρησιμοποιούνται συχνά από τις επιχειρήσεις όταν κάνουν Data Mining – Analysis. Αναφορικά με το Customer Profiling πολλές εταιρείες κυρίως στο χώρο του e-commerce το χρησιμοποιούν για να δίνουν στον πελάτη εξατομικευμένες προτάσεις, οι Park et al. (2009) αναφέρουν μια τέτοια τεχνική η οποία βασίζεται και σε ατομικά αλλά και ομαδικά δεδομένα. Οι Abirami et al. (2016) χρησιμοποιούν την RFM Ανάλυση σε καταστήματα λιανικής για να ταξινομήσουν τους πελάτες (νέους, επαναλαμβανόμενους) και να αξιοποιήσουν την πληροφορία από τα αποτελέσματα, η RFM analysis είναι μια κλασική τεχνική που αξιοποιείται συχνά από τις εταιρείες λιανεμπορίου. Οι Fang et al. (2018) δημιουργούν ένα προσαρμοσμένο πακέτο πρότασης με βάση τους κανόνες συσχέτισης μεταξύ προϊοντικών κατηγοριών, τέτοιου είδους προτάσεις χρησιμοποιούνται από πολλές επιχειρήσεις και οργανισμούς στα online καταστήματά τους. Πολλοί ερευνητές στον επιχειρηματικό κόσμο προ επεξεργάζονται τα δεδομένα σε συστάδες και μετά εξάγουν κανόνες συσχέτισης από αυτές όπως μας παρουσιάζουν οι Koh et al. (2008) στην εργασία τους.

2.4.2 Η οπτική της ακαδημίας

Οι Chen et al. (2006) προσπαθούν να δομήσουν έναν ταξινομητή, με βάση μια τεχνική εξόρυξης κανόνων συσχέτισης (classification based on association rule mining), αποτελεί εργασία αποκλειστικά για ερευνητικούς σκοπούς και δεν χρησιμοποιείται σε επαγγελματικό επίπεδο. Το Association Rules Network των Pandey et al. (2009) που είναι μια καλή τεχνική για τη σύνθεση, κλάδεμα και ανάλυση μιας συλλογής κανόνων συσχέτισης στην κατασκευή υποψήφιων υποθέσεων δεν χρησιμοποιείται από τις επιχειρήσεις κυρίως λόγο της πολυπλοκότητας και της μεροληψίας του ερευνητή. Η τεχνική Improved

Classification Based on Predictive Association Rules όπου συμβαίνει μια ταξινόμηση βασισμένη σε προγνωστικούς κανόνες συσχέτισης (Hao, Wang, Yao, & Zhang, 2009), είναι πολύ διαδεδομένη αλλά έχει μείνει μόνο σε ερευνητικό επίπεδο. Όμοια παρέμεινε σε ερευνητικό επίπεδο και η MCAR: Multi-class Classification based on Association Rule (Thabtah, Cowling, & Peng, 2005). Παρόλο που στην εργασία των Niu et al. (2009) προτείνεται μια νέα μέθοδο συσχετιστικής ταξινόμησης (association classification) που βασίζεται σε συμπαγείς κανόνες συσχέτισης, και μέσα σε αυτή επεκτείνεται ο αλγόριθμο Apriori, δεν έχει γίνει καμία αναφορά για χρήση σε επιχειρήσεις. Αυτό μάλλον γιατί εξετάζουν το ενδιαφέρον, τη σημασία και τις αλληλεπικαλυπτόμενες σχέσεις μεταξύ των κανόνων παρά την παραγωγή πολύτιμης για την αγορά πληροφορίας. Η έρευνα των Li et al. (2001) CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules γίνετε για να βελτιώσουν την συσχετιστική ταξινόμηση που υποφέρει από το τεράστιο σύνολο εξορυσσόμενων κανόνων και μερικές φορές μεροληπτική ταξινόμηση ή υπερπροσαρμογή, δεν χρησιμοποιείται από της επιχειρήσεις και αυτό οφείλεται στο ότι βελτιώνει υπάρχον διαδικασίες παρά δημιουργεί πληροφορία. Οι Ya-Han et al. (2014) χρησιμοποιούν RFM Analysis για να ανακαλύψουν μοτίβα πελατών χωρίς τις πληροφορίες τους, παρόλο που η ανάλυση RFM χρησιμοποιείται σε επιχειρηματικό επίπεδο, η συγκεκριμένη τεχνική (χωρίς πληροφορίες πελατών) δεν φαίνεται να είναι διαδεδομένη στις επιχειρήσεις καθώς συνήθως διαθέτουν πληροφορίες και στοιχεία πελατών και προτιμούν τη αξιοποίηση τους για να εξάγουν συμπεράσματα και να λαμβάνουν αποφάσεις. Οι Mu-Chen et al. (2007) επικεντρωμένοι στα προϊόντα προτείνουν μια προσέγγιση εξόρυξης δεδομένων για τη λήψη αποφάσεων σχετικά με το ποια προϊόντα θα αποθηκευτούν, πόσος χώρος στο ράφι θα διατεθεί στα αποθηκευμένα προϊόντα και πού θα τα εκθέσουν. Η συγκεκριμένη τεχνική έχει υψηλές προσδοκίες αλλά δεν χρησιμοποιείται λόγο της αδυναμίας πολλών επιχειρήσεων να συλλέξουν και αποθηκεύσουν την απαραίτητη πληροφορία καθώς και την έλλειψη οικονομικών πόρων για τέτοιου είδους μαζικές αλλαγές.

Πίνακας 2: Χρήση τεχνικών σε επιχειρηματικό επίπεδο

Συγγραφείς	**Τεχνική**	**Χρήση σε επιχειρηματικό επίπεδο**
(Wang & Sun, 2019)	Market Basket Analysis	NAI
(Christodoulakis, 2005)	RFM Analysis	NAI
(Chen, Liu, Yu, Wei, & Zhang, 2006)	Classification based on association rule mining	OXI
(Pandey, Chawla, Poon, Arunasalam, & Davis, 2009)	Association Rules Network	OXI
(Liu, Hsu, & Ma, 1998)	Integrating Classification and Association Rule Mining	NAI
(Hao, Wang, Yao, &	Improved Classification Based on	OXI

Zhang, 2009)	Predictive Association Rules	
(Thabtah, Cowling, & Peng, 2005)	MCAR: Multi-class Classification based on Association Rule	OXI
(Niu, Xia, & Zhang, 2009)	Association Classification	OXI
(Li, Han, & Pei, 2001)	CMAR: Classification Based on Multiple Class-Association Rules	OXI
(Baby & Priyanka, 2012)	Customer Classification And Prediction	NAI
(Park & Chang, 2009)	Customer Profiling	NAI
(M. Abirami ..., 2016)	RFM Analysis in Retail	NAI
(Abirami & Pattabiraman, 2016)	RFM analysis without customer identification information	OXI
(Mu-Chen & Chia-Ping, 2007)	Product assortment and shelf space allocation mining approach	OXI
(Fang, Xia, Wang, & Lan, 2018)	Customized Bundle Recommendation	NAI
(Koh & Pears, 2008)	Clustering and Association rules	NAI

Παρατηρείται λοιπόν ότι σε επίπεδο επιχείρησής χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων ή συνδυασμός αυτών που είναι κλασικές, όπως Market Basket Analysis, RFM Analysis, Classification κ.λπ. Αυτό μας φανερώνει ότι ο επιχειρηματικός κόσμος δεν είναι έτοιμος να επενδύσει οικονομικά ή δεν εμπιστεύεται ακόμα τις ποιο ιδιαίτερες και περίπλοκες τεχνικές που έχουν ερευνηθεί στο ακαδημαϊκό επίπεδο.

3 Μεθοδολογία

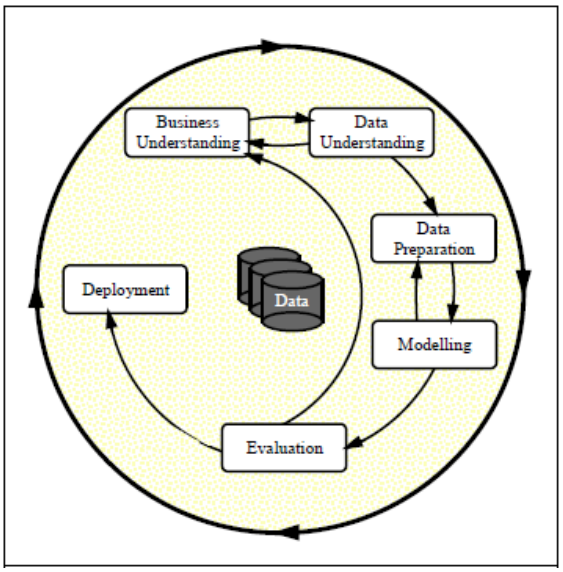
Το συγκεκριμένο κεφάλαιο κάνει μια ομαλή μετάβαση μεταξύ της διερεύνησης των τεχνικών της βιβλιογραφικής ανασκόπησης και της εμπειρικής μελέτης. Υπάρχουν πολλές μεθοδολογίες για την διαδικασία εξόρυξης γνώσης που μπορεί να ακολουθήσει ένας αναλυτής, μπορεί ακόμα και να μην ακολουθήσει καμιά. Παρόλα αυτά για την μετατροπή του επιχειρηματικού προβλήματος σε εργασία εξόρυξης γνώσης στη συγκεκριμένη εμπειρική μελέτη χρησιμοποιήθηκε μια από τις πιο διαδεδομένες μεθοδολογίες η CRISP-DM.

Σε αυτό το κεφάλαιο δίνεται μια αναλυτική επεξήγηση της CRISP-DM και αναφέρεται περιληπτικά η χρήση της στην εμπειρική μελέτη. Παραθέτετε επίσης μια θεωρητική αναφορά στην εξόρυξη κανόνων συσχέτισης, τις βασικές έννοιες και τα μέτρα αξιολόγησης τους. Και τέλος περιγράφονται τα διαθέσιμα δεδομένων και τα περιγραφικά στατιστικά που προκύπτουν από αυτά.

3.1 Μεθοδολογία CRISP-DM

Η εξόρυξη δεδομένων χρειάζεται μια προσέγγιση που να βοηθάει στη μετατροπή των επιχειρηματικών προβλημάτων σε εργασίες εξόρυξης δεδομένων, να προτείνει κατάλληλους μετασχηματισμούς δεδομένων και τεχνικές εξόρυξης δεδομένων, και να παρέχει μέσα για την αξιολόγηση των αποτελεσμάτων και την τεκμηρίωση της γνώσης που προκύπτει. Η CRISP-DM (Cross Industry Standard Process for Data Mining) διαχειρίζεται αυτή τη μετατροπή με ένα μοντέλο διαδικασιών που παρέχει ένα πλαίσιο για την διεξαγωγή εξόρυξης δεδομένων ανεξάρτητο τόσο από τον κλάδο της βιομηχανίας που ερευνάται όσο και από την τεχνολογία που χρησιμοποιείται. Αυτό το μοντέλο μπορεί να χρησιμεύσει ως κοινό σημείο αναφοράς και να αυξήσει την κατανόηση των κρίσιμων ζητημάτων εξόρυξης δεδομένων από όλους τους συμμετέχοντες ειδικά από την πλευρά των πελατών.

Ο κύκλος ζωής ενός έργου εξόρυξης δεδομένων αναλύεται σε έξι φάσεις. Όπως φαίνεται στο **σχήμα 2**. Η σειρά των φάσεων δεν είναι αυστηρή. Τα βέλη υποδεικνύουν μόνο τις πιο σημαντικές και συχνές εξαρτήσεις μεταξύ των φάσεων αλλά εξαρτάται από το αποτέλεσμα κάθε φάσης ποια φάση πρέπει να εκτελεστεί στη συνέχεια



Σχήμα 2: Μεθοδολογία Crisp-DM

Ο εξωτερικός κύκλος συμβολίζει την κυκλική φύση της ίδιας της εξόρυξης γνώσης η οποία δεν ολοκληρώνεται μόλις αναπτυχθεί μια λύση αλλά τα διδάγματα που αντλήθηκαν κατά τη διάρκεια της διαδικασίας μπορούν να προκαλέσουν νέα, συχνά πιο εστιασμένα επιχειρηματικά ερωτήματα.

Περιγραφή φάσεων μεθοδολογίας Crisp-DM:

- **Business understanding (Επιχειρηματική κατανόηση):** Αυτό είναι το αρχικό στάδιο του project και των απαιτήσεων από την πλευρά της επιχείρησης, ο αναλυτής εντοπίζει το πρόβλημα που πρέπει να λύσει και τις ερωτήσεις που πρέπει να απαντήσει και το μετατρέπει σε πρόβλημα εξόρυξης δεδομένων.
- **Data Understanding (Κατανόηση δεδομένων):** Σε αυτή τη φάση ξεκινάει η αρχική συλλογή δεδομένων, ο αναλυτής πρέπει να εξοικειωθεί με τα δεδομένα που έχει στη διάθεση του, να εντοπίσει προβλήματα ποιότητας δεδομένων και να ανιχνεύσει υποσύνολα ή κρυφές πληροφορίες που μπορεί να έχουν ενδιαφέρον. Υπάρχει στενή σχέση μεταξύ της Επιχειρηματικής Κατανόησης και της Κατανόησης Δεδομένων.
- **Data Preparation (Προετοιμασία δεδομένων):** Αυτή η φάση περιλαμβάνει όλες τις διαδικασίες για την κατασκευή του τελικού συνόλου δεδομένων (δηλαδή των δεδομένων που θα τροφοδοτήσουν τα εργαλεία μοντελοποίησης) από τα αρχικά ακατέργαστα δεδομένα. Οι εργασίες προετοιμασίας δεδομένων πιθανόν να εκτελούνται πολλές φορές και όχι με μια συγκεκριμένη σειρά και περιλαμβάνουν την επιλογή πινάκων, εγγραφών και χαρακτηριστικών, τον καθαρισμό δεδομένων, την κατασκευή νέων χαρακτηριστικών και τον μετασχηματισμό των δεδομένων.
- **Modeling (Μοντελοποίηση):** Σε αυτή τη φάση, επιλέγονται και εφαρμόζονται διάφορες τεχνικές μοντελοποίησης, υπάρχουν πολλές τεχνικές για τον ίδιο τύπο προβλήματος εξόρυξης δεδομένων. Υπάρχει στενή σύνδεση μεταξύ της προετοιμασίας δεδομένων και της μοντελοποίησης γιατί πολλές φορές κάποιος αντιλαμβάνεται προβλήματα στα δεδομένα κατά τη μοντελοποίηση ή παίρνει ιδέες για την κατασκευή νέων δεδομένων.
- **Evaluation (Εκτίμηση):** Σε αυτή τη φάση ο αναλυτής έχει δημιουργήσει ένα ή περισσότερα μοντέλα που φαίνεται να έχουν υψηλή ποιότητα από πλευράς ανάλυσης δεδομένων. Πριν προχωρήσει στην τελική ανάπτυξη του μοντέλου

είναι σημαντικό να το αξιολογήσει πιο διεξοδικά και να αναθεωρήσει τα βήματα που εκτελέστηκαν, ώστε να βεβαιωθεί ότι επιτυγχάνει σωστά τους επιχειρηματικούς στόχους. Ένας βασικός στόχος είναι να προσδιοριστεί εάν υπάρχει κάποιο σημαντικό επιχειρηματικό ζήτημα που δεν έχει εξεταστεί επαρκώς. Στο τέλος αυτής της φάσης, θα πρέπει να ληφθεί απόφαση σχετικά με τη χρήση των αποτελεσμάτων της εξόρυξης δεδομένων.

- **Deployment (Ανάπτυξη):** Η δημιουργία του μοντέλου δεν είναι το τέλος του έργου. Η γνώση που αποκτάται θα πρέπει να οργανωθεί και να παρουσιαστεί με τρόπο που να μπορεί να χρησιμοποιηθεί από τον πελάτη. Ανάλογα με τις απαιτήσεις, η φάση της ανάπτυξης μπορεί να είναι τόσο απλή όσο η δημιουργία μιας αναφοράς ή τόσο περίπλοκη όσο η εφαρμογή μιας επαναλαμβανόμενης διαδικασίας εξόρυξης δεδομένων. Σε πολλές περιπτώσεις, ο χρήστης και όχι ο αναλυτής δεδομένων θα είναι αυτός που θα πραγματοποιήσει τα βήματα ανάπτυξης. Σε κάθε περίπτωση, είναι σημαντικό να κατανοηθούν εκ των προτέρων ποιες ενέργειες θα πρέπει να γίνουν προκειμένου να αξιοποιηθούν πραγματικά τα μοντέλα που δημιουργήθηκαν. (Wirth & Hipp, 2000)

3.2 Χρήση της Crisp-DM

Πίνακας 3: Περίληψη χρήσης της Crisp-DM στην συγκεκριμένη ανάλυση

Στάδιο Crisp-DM	Διαδικαστικά	Λεπτομέρειες
Επιχειρηματική κατανόηση	Καθορισμός Στόχου, Βελτίωση εύρους στόχου	Εντοπισμός πρότυπων μοτίβων αγορών από τα αγοραστικά καλάθια των πελατών χονδρικής. Εντοπισμός μοτίβων αγορών στις επαγγελματικές κατηγορίες των πελατών και αξιολόγηση αν αυτά τα μοτίβα μπορούν να υποδηλώσουν επάγγελμα.
Κατανόηση δεδομένων	Ανάλυση και περιγραφή των διαθέσιμων δεδομένων	Ανάλυση των διαθέσιμων δεδομένων που θα εξαχθούν από την βάση, ανάλυση και περιγραφή των πινάκων και των πεδίων. Εξαγωγή περιγραφικών στατιστικών και γενικών στοιχείων.
Προετοιμασία δεδομένων	Καθαρισμός δεδομένων, Κωδικοποίηση δεδομένων, Μορφοποίηση δεδομένων, Εξαγωγή συνόλων, δεδομένων (data set)	Καθαρισμός των κενών στοιχείων και διαγραφή προϊόντων που δεν αφορούν αγορές. Ενσωμάτωση καλαθιών με μοναδικό προϊόν σε καλάθι τού ίδιου πελάτη στην ίδια μέρα και αν δεν υπάρχει

		στην ίδια μέρα σε εύρος δύο ημερών. Διαγραφή των καλαθιών που δεν καλύπτουν τα παραπάνω κριτήρια. Χρήση μόνο των πελατών που έχουν καλάθια αγορών εντός των διαθέσιμων ημερομηνιών. Κωδικοποίηση των ονομάτων για τα προϊόντα. Δημιουργία 3 συνόλων δεδομένων: • Προϊόντων • Προϊόντων και επαγγελμάτων • Προϊόντων ομαδοποιημένων καλαθιών πελάτη Μορφοποίηση των δεδομένων στη μορφή καλαθιού για χρήση από τον αλγόριθμο apriori
Μοντελοποίηση	Παραγωγή κανόνων συσχέτιση: Προϊόντων, Προϊόντων με επάγγελμα, Προϊόντων σε ομαδοποιημένα καλάθια πελατών	Χρήση της R γλώσσας προγραμματισμού, των βιβλιοθηκών της και του αλγόριθμου apriori για την παραγωγή των κανόνων συσχέτισης και στα τρία πειράματα. Δοκιμές ποικίλων μεταβλητών για την εξαγωγή διαφορετικών αποτελεσμάτων με σκοπό την εύρεση του ικανοποιητικότερου. Χρήση των παραγόμενων κανόνων συσχέτισης για την εμφάνιση διαγραμμάτων.
Εκτίμηση	Εκτίμηση των παραγόμενων κανόνων συσχέτισης	Επεξήγηση, παρουσίαση και εκτίμηση των κανόνων συσχέτισης με χρήση των μέτρων εκτίμησης (evaluation measures), Support, Confidence και Lift.

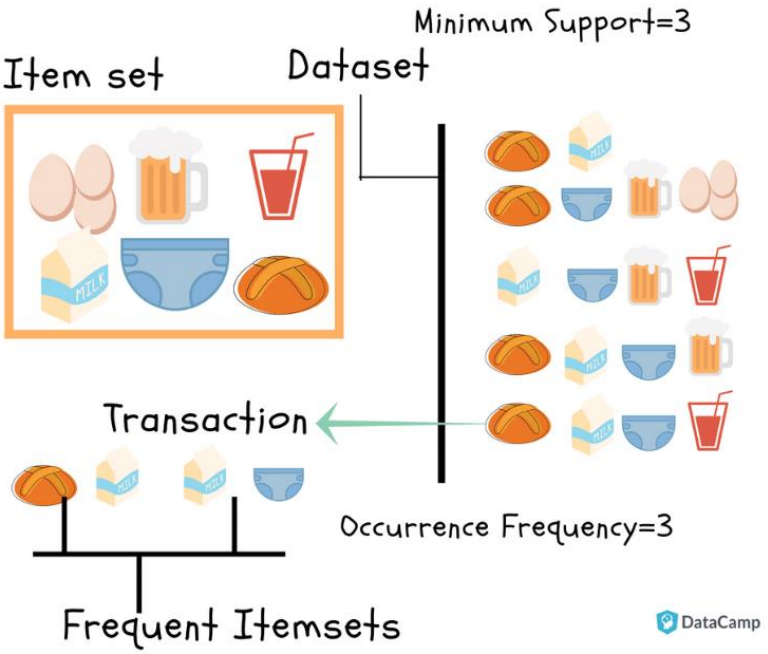
3.3 Εξόρυξη κανόνων συσχέτισης (Association rules mining)

Η εξόρυξη κανόνων συσχέτισης, στοχεύει στην εξαγωγή συνδέσεων με ενδιαφέρον,

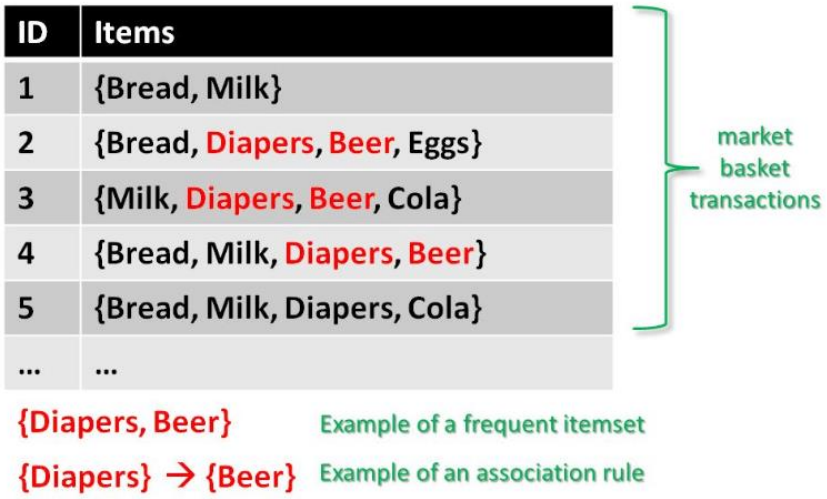
συχνών μοτίβων, συσχετίσεων ή περιστασιακών δομών μεταξύ συνόλων στοιχείων από βάσεις δεδομένων συναλλαγών ή άλλους χώρους αποθήκευσης δεδομένων. Η εξόρυξη κανόνων συσχέτισης είναι η εύρεση εκείνων των κανόνων που ικανοποιούν το προκαθορισμένο ελάχιστο Support και Confidence (αναλύονται παρακάτω) από μια βάση δεδομένων. Η εξόρυξη κανόνων συσχέτισης συνήθως αναλύεται σε δύο υποκατηγορίες ενεργειών. Η μια αφορά την εύρεση εκείνων των στοιχείο-συνόλων (itemset) των οποίων οι εμφανίσεις υπερβαίνουν ένα προκαθορισμένο όριο στη βάση δεδομένων, αυτά τα στοιχεία ονομάζονται συχνά στοιχειοσύνολα (frequent itemset) ή μεγάλα στοιχείο-σύνολα (large itemset). Η δεύτερη υποκατηγορία ενεργειών αφορά τη δημιουργία των κανόνων συσχέτισης από αυτά τα μεγάλα στοιχειοσύνολα με τον προκαθορισμό μιας ελάχιστης εμπιστοσύνης (support).

Σε πολλές περιπτώσεις, οι αλγόριθμοι δημιουργούν έναν εξαιρετικά μεγάλο αριθμό κανόνων συσχέτισης, συχνά σε χιλιάδες ή και εκατομμύρια. Είναι σχεδόν αδύνατο για τους τελικούς χρήστες να κατανοήσουν ή να επικυρώσουν τόσο μεγάλο αριθμό περίπλοκων κανόνων συσχέτισης, περιορίζοντας έτσι τη χρησιμότητα του αποτελέσματος εξόρυξης δεδομένων. Έχουν προταθεί αρκετές στρατηγικές για τη μείωση του αριθμού των κανόνων συσχέτισης, όπως η δημιουργία μόνο «ενδιαφερόντων» κανόνων, η δημιουργία μόνο «μη περιττών» κανόνων ή η δημιουργία μόνο κανόνων που ικανοποιούν ορισμένα άλλα κριτήρια, όπως coverage, leverage, lift, strength.

Βασικές έννοιες κανόνων συσχέτισης: Έστω το $I = I_1, I_2, \ldots, I_m$ ένα σύνολο m διακριτών στοιχείων, το T είναι μια συναλλαγή που περιέχει ένα σύνολο στοιχείων τέτοια ώστε $T \subseteq I$, έστω D είναι μια βάση δεδομένων με διαφορετικές εγγραφές T_s. Ένας κανόνας συσχέτισης είναι μια συνεπαγωγή της μορφής $X \Rightarrow Y$ όπου τα $X,Y \subset I$ είναι σύνολα στοιχείων που ονομάζονται στοιχειοσύνολα, και $X \cap Y = \emptyset$. Το X ονομάζεται προγενέστερο ή αριστερή πλευρά του κανόνα (LHS) και το Y ονομάζεται συνέπεια ή δεξιά πλευρά του κανόνα (RHS), και αυτός ο κανόνας σημαίνει ότι το X υποδηλώνει Y.



Σχήμα 3: Βασικές έννοιες κανόνων συσχέτισης



Σχήμα 4: Παράδειγμα συχνού συνόλου δεδομένων

(Jabbeen, 2018)

Τα δύο σημαντικά μέτρα για τους κανόνες συσχέτισης είναι η υποστήριξη (Support) και η εμπιστοσύνη (Confidence).

Support είναι μια ένδειξη (ποσοστό) που δηλώνει πόσο συχνά εμφανίζεται το στοιχειοσύνολο (συνδυασμός προϊόντων) στο σύνολο των δεδομένων.

$$\text{Support} = P(X \cap Y) = \frac{\textit{αριθμός συναλαγών που περιέχουν το X και το Y}}{\textit{σύνολο των συναλαγών}}$$

Υποθέτοντας ότι το Support ενός στοιχείου είναι 0,1%, σημαίνει ότι μόνο το 0,1 τοις εκατό των συναλλαγών περιέχει την αγορά αυτού του είδους.

Confidence είναι το μέτρο μέτρησης της ισχύος ενός κανόνα συσχέτισης και είναι η ένδειξη (ποσοστό) των συναλλαγών που περιέχουν το X∩Y προς τον αριθμό των συναλλαγών που περιέχουν μόνο το X.

$$\text{Conf}(X=>Y) = P(X|Y) = \frac{Supp(X \cap Y)}{Supp(X)} = \frac{P(X \cap Y)}{P(X)} = \frac{\textit{αριθμός συναλαφών που περιέχουν το X και το Y}}{\textit{αριθμός συναλλαγών που περιέχουν το X}}$$

Υποθέτοντας ότι το confidence του κανόνα X ⇒ Y είναι 80% αυτό σημαίνει ότι το 80% των συναλλαγών που περιέχουν το X περιέχουν επίσης και το Y.

Lift είναι μια πρόταση επιλογής ανάλογα με το πιθανό ενδιαφέρον τους για τον χρήστη, μειώνει επίσης το κόστος και το χρόνο της διαδικασίας εξόρυξης. Υπολογίζει αν η εμφάνιση του X και η εμφάνιση του Y στην ίδια συναλλαγή είναι ανεξάρτητα γεγονότα.

$$Lift = \frac{P(X \cap Y)}{P(X) * P(Y)}$$

Κατά τον υπολογισμό του Lift έχουμε τρείς πιθανότητες:

1. Lift > 1 τότε η συσχέτιση του κανόνα είναι θετική
2. Lift < 1 τότε η συσχέτιση του κανόνα είναι αρνητική
3. Lift = 1 τότε η συσχέτιση είναι ανεξάρτητη

Επομένως όσο πιο μεγάλο είναι το Lift τόσο πιο ισχυρή μπορεί να θεωρηθεί μια συσχέτιση.

(Hussein, Alashqur, & Sowan, 2015)

3.4 Διαθεσιμότητα δεδομένων

Τα δεδομένα που χρησιμοποιούνται στη συγκεκριμένη ανάλυση δόθηκαν μετά από συνάντηση που έγινε με τον γενικό διευθυντή της γενικής διεύθυνσης οργάνωσης και πληροφορικής μεγάλης εταιρείας λιανικής και χονδρικής πώλησης (Super Market). Η εταιρεία που παρείχε τα δεδομένα ανήκει και στον κλάδο του χονδρεμπόριού και εστιάζει στις αγορές που κάνουν οι επιχειρήσεις – πελάτες της από αυτή και όχι στις αγορές του απλού καταναλωτή. Βάση της ανάλυσης και του στόχου της εργασίας εξάχθηκαν μόνο οι πληροφορίες που αφορούν τους πελάτες μεταποίησης (HO.RE.CA.) δηλαδή τις επιχειρήσεις πελάτες που αγοράζουν προϊόντα με σκοπό να τα επεξεργασθούν για να παράγουν ένα νέο.

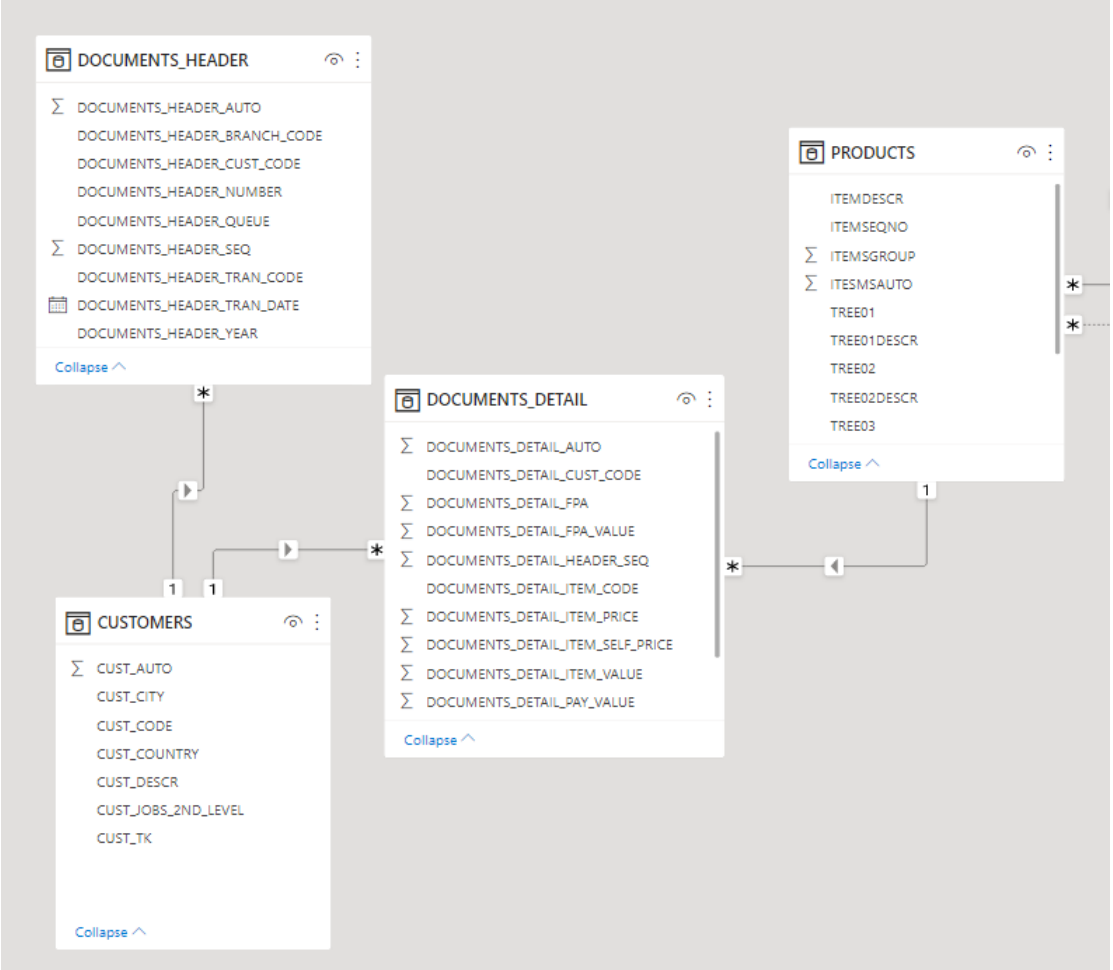
Το χρονικό πλαίσιο των δεδομένων αναφέρεται σε τέσσερα έτη (2018, 2019, 2020, 2021) και αυτό γιατί δεδομένα παλαιότερων ετών μπορεί να μην δώσουν ποιοτική πληροφορία αφού οι συνθήκες τις αγοράς αλλάζουν συνεχώς. Το 2020 και 2021 ήταν δύο έτη που πολλές επιχειρήσεις βρίσκονταν σε αναστολή κατά περιόδους λόγο έξαρσης του ιού Covid-19 και επομένως οι αγορές μέσα σε αυτά μπορεί να είναι μειωμένες για κάποια προϊόντα. Έγινε αναφορά από την ίδια την επιχείρηση ότι μέσα σε αυτά τα χρόνια υπήρξε αύξηση των πωλήσεων κυρίως στα προϊόντα που υπήρχαν στο ηλεκτρονικό της κατάστημα ενώ όσα δεν εμφανίζονταν σε αυτό παρουσίασαν μείωση. Στον παρακάτω πίνακα περιγράφονται οι πίνακες και το σύνολο τον δεδομένων που εξάχθηκαν από την βάση της εταιρείας.

Πίνακας 4: Επεξήγηση δεδομένων που αντλήθηκαν

Πίνακας	Περιγραφή	Πεδία	Σύνολο
Customers (Πελάτες)	Πίνακας με τους ενεργούς πελάτες HO.RE.CA.	Κωδικός πελάτη Περιγραφή πελάτη Περιοχή πελάτη Ταχυδρομικός κώδικας Κωδικός επαγγέλματος	61.184 ενεργοί πελάτες HO.RE.CA.
Jobs_Top_Level (Επίπεδο 1 Επαγγελμάτων)	Πίνακας επαγγελματικών κατηγοριών ομαδοποιημένων στη γενικότερη μορφή τους	Κωδικός κατηγορίας Περιγραφή κατηγορίας	5 γενικές κατηγορίες επαγγελμάτων
Jobs_First_Level (Επίπεδο 2 Επαγγελμάτων)	Πίνακας επαγγελματικών κατηγοριών ομαδοποιημένων σε λεπτομερή μορφή	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον Jobs_Top_Level	24 κατηγορίες επαγγελμάτων

Jobs_Second_Level (Επίπεδο 3 Επαγγελμάτων)	Πίνακας επαγγελματικών κατηγοριών ομαδοποιημένων στην πιο λεπτομερή μορφή τους	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον Jobs_First_Level	442 κατηγορίες επαγγελμάτων
Documents (Αποδείξεις)	Πίνακας αποδείξεων (Header) με γενικές πληροφορίες για το παραστατικό αγοράς, χαρακτηρίζεται και ως καλάθι	Κωδικός συναλλαγής Κωδικός Καταστήματος Κωδικός Πελάτη (σύνδεση πίνακα πελατών) Ημερομηνία συναλλαγής Σειρά Παραστατικού Αριθμός Παραστατικού Έτος έκδοσης	168.52 Αποδείξεις (καλάθια) πελατών HO.RE.CA
Documents Details (Προϊόντα αποδείξεων)	Πίνακας (Detail) που περιέχει τα προϊόντα εντός των αποδείξεων (καλαθιών)	Κωδικός Πελάτη (σύνδεση με πίνακα πελατών) Ημερομηνία συναλλαγής Κωδικός είδους (σύνδεση πίνακα προϊόντων) Συνολική ποσότητα είδους Τιμή Ραφιού Τιμή μονάδας Καθαρή αξία Ποσοστό ΦΠΑ Αξία ΦΠΑ της κίνησης Πληρωτέα αξία Κωδικός συναλλαγής (σύνδεση με πίνακα αποδείξεων)	2.032.484 αγορασμένα προϊόντα εντός των αποδείξεων
Products (Προϊόντα)	Πίνακας που περιέχει τα προϊόντα και την πληροφορία τους	Κωδικός προϊόντος Περιγραφή προϊόντος Προϊοντική κατηγορία επιπέδου 1 Προϊοντική κατηγορία	112.075 προϊόντα

		επιπέδου 2 Προϊοντική κατηγορία επιπέδου 3 Προϊοντική κατηγορία επιπέδου 4 Προϊοντική κατηγορία επιπέδου 5	
Tree_Category_01	Προϊοντικές κατηγορίες επιπέδου 1	Κωδικός κατηγορίας Περιγραφή κατηγορίας	85 προϊοντικές κατηγορίες
Tree_Category_02	Προϊοντικές κατηγορίες επιπέδου 2	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον πίνακα Tree_Category_01	540 προϊοντικές κατηγορίες
Tree_Category_03	Προϊοντικές κατηγορίες επιπέδου 3	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον πίνακα Tree_Category_02	2786 προϊοντικές κατηγορίες
Tree_Category_04	Προϊοντικές κατηγορίες επιπέδου 4	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον πίνακα Tree_Category_03	9711 προϊοντικές κατηγορίες
Tree_Category_05	Προϊοντικές κατηγορίες επιπέδου 5	Κωδικός κατηγορίας Περιγραφή κατηγορίας Κωδικός σύνδεσης με τον πίνακα Tree_Category_04	37281 προϊοντικές κατηγορίες



Σχήμα 5: Βασικοί πίνακες δεδομένων

3.4.1 Περιγραφικά Στατιστικά

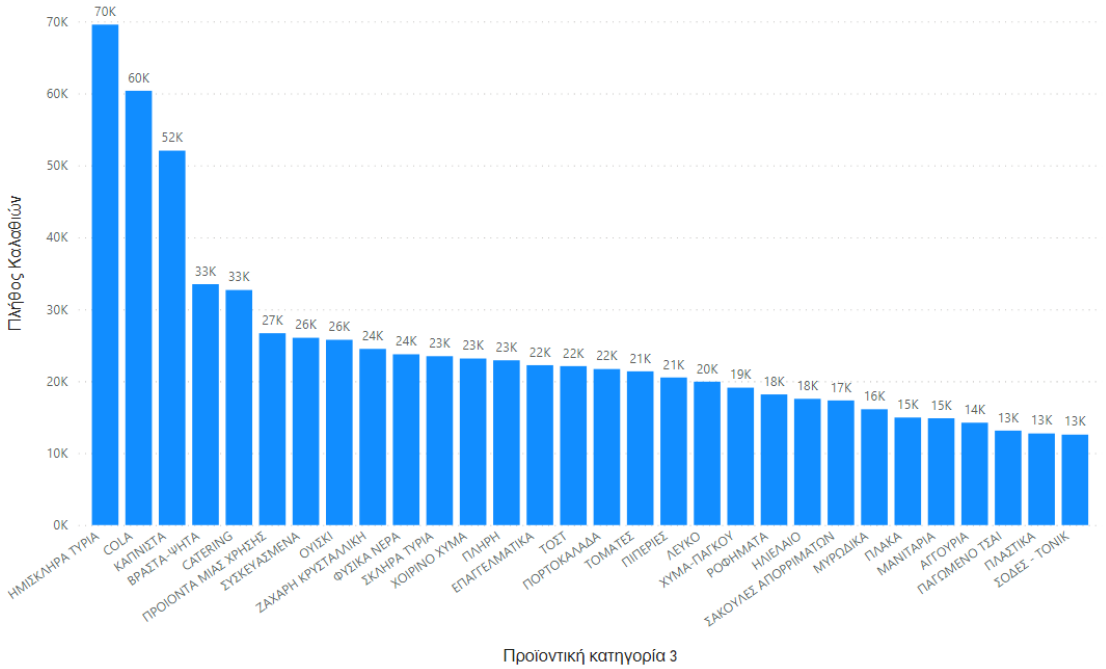
Για την δημιουργία και παρουσίαση των παρακάτω περιγραφικών στατιστικών έγινε χρήση της T-SQL γλώσσας προγραμματισμού καθώς επίσης και του εργαλείου PowerBI (λογισμικό οπτικοποίησης δεδομένων – data visualization software).

Πίνακας 5: Συχνότερα προϊόντα στα καλάθια (HO.RE.CA. 2018-2021)

Προϊοντική κατηγορία	**Προϊόν**	**Συνολικές συναλλαγές (transactions)**
ΑΝΑΨΥΚΤΙΚΑ	*Επώνυμο Προϊόν* (COLA) 330ML 4/(6ΤΕΜ.)	15218
ΣΥΣΚΕΥΑΣΜΕΝΑ-ΤΥΠΟΠΟΙΗΜΕΝΑ	*Επώνυμο Προϊόν* GOUDA ΣΕ ΦΕΤΕΣ 1KG (Τ.Κ)	14666
ΛΑΧΑΝΙΚΑ	ΑΓΓΟΥΡΙΑ ΕΓΧΩΡΙΑ (ΤΙΜΗ ΤΕΜ).	13456
ΖΑΧΑΡΗ	*Επώνυμο Προϊόν* ΖΑΧΑΡΗ ΛΕΥΚΗ ΚΡΥΣΤΑΛ.ΕΙΣ.10/1KG	11782
ΛΑΧΑΝΙΚΑ	ΤΟΜΑΤΕΣ ΕΓΧΩΡΙΕΣ ΠΟΙΟΤΗΤΑ Α (ΤΙΜΗ ΚΙΛΟΥ)	11676

Πίνακας 6: Συχνότερα προϊόντα προσφοράς στα καλάθια (HO.RE.CA. 2018-2021)

Προϊοντική κατηγορία	Προϊόν	Συνολικές συναλλαγές (transactions)
ΜΠΥΡΕΣ STANDARD	*Επώνυμο Προϊόν* ΜΠΙΡΑ ΚΟΥΤΙ 330ML 4/(5+1 ΔΩΡΟ)	3231
ΑΝΑΨΥΚΤΙΚΑ	*Επώνυμο Προϊόν* ΠΟΡΤΟΚΑΛΑΔΑ ΦΙΑΛΗ 330ML (5+1ΔΩΡΟ)	2302
ΑΝΑΨΥΚΤΙΚΑ	*Επώνυμο Προϊόν* ΛΕΜΟΝΑΔΑ ΦΙΑΛΗ 330ML (5+1ΔΩΡΟ)	2259
ΝΕΡΑ	*Επώνυμο Προϊόν* ΦΥΣΙΚΟ ΜΕΤΑΛΛΙΚΟ ΝΕΡΟ 1,5L (5+1)	1911
ΑΝΑΨΥΚΤΙΚΑ	*Επώνυμο Προϊόν* ΠΟΡΤΟΚΑΛΑΔΑ ΜΠΛΕ 330ML (5+1)	1868



Σχήμα 6: Πλήθος καλαθιών ανά προϊοντική κατηγορία (3)

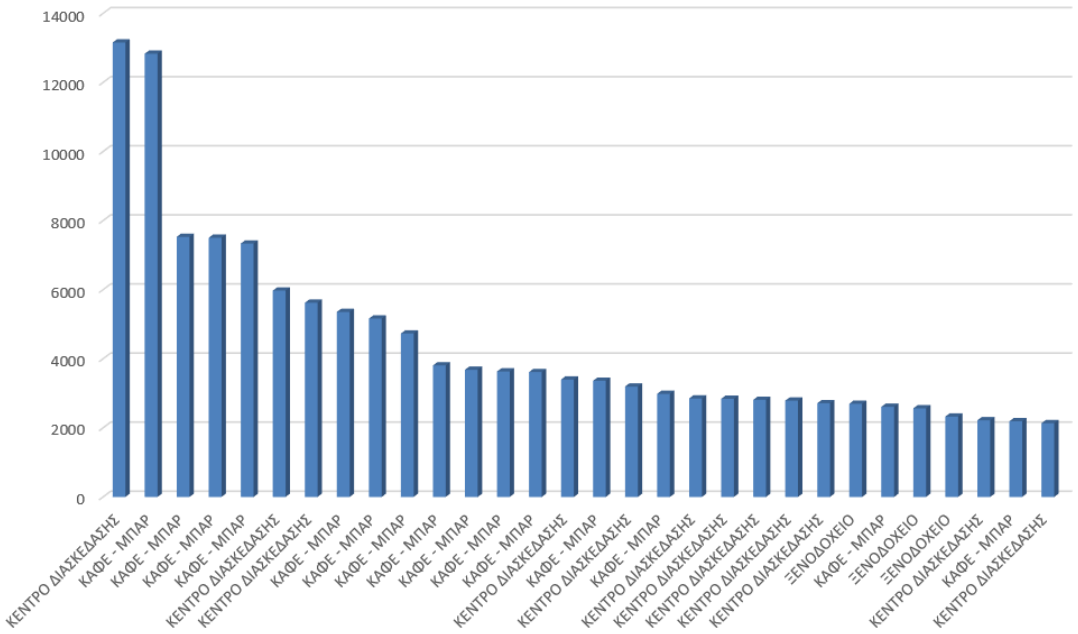
Πίνακας 7: Καλάθια με τα περισσότερα προϊόντα (HO.RE.CA. 2018-2021)

Κωδικός καλαθιού	Κωδικός πελάτη	Ιδιότητα	Συνολικά προϊόντα εντός καλαθιού
6144557	680366	ΞΕΝΟΔΟΧΕΙΟ	142
15427016	271440	ΨΗΤΟΠΩΛΕΙΟ	140
43723155	604633	ΚΑΦΕΤΕΡΙΑ	139
21685210	423679	ΞΕΝΟΔΟΧΕΙΟ	122
4372920	680366	ΞΕΝΟΔΟΧΕΙΟ	121

Πίνακας 8: Καλάθια με το μεγαλύτερο κόστος (HO.RE.CA. 2018-2021)

Κωδικός καλαθιού	Ημερομηνία Αγοράς	Κωδικός πελάτη	Ιδιότητα	Κόστος καλαθιού (με ΦΠΑ)

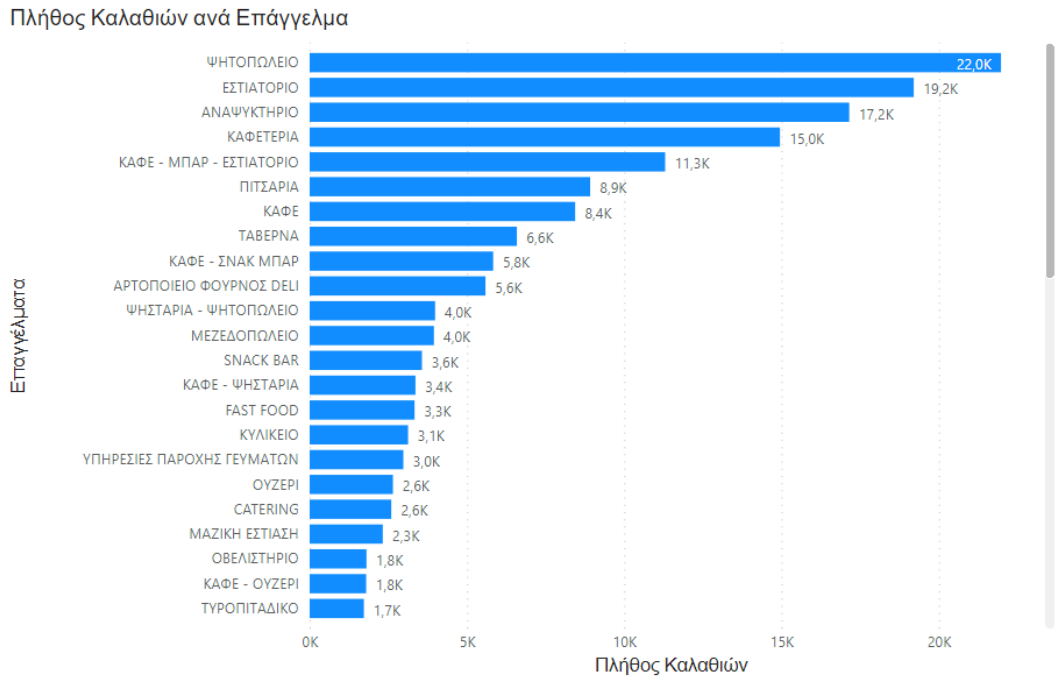
9763525	05/04/2018	641831	ΚΕΝΤΡΟ ΔΙΑΣΚΕΔΑΣΗΣ	13.158
9797431	26/06/2018	685708	ΚΑΦΕ – ΜΠΑΡ	12.833
95229275	24/08/2019	664575	ΚΑΦΕ – ΜΠΑΡ	7.529
129124356	25/01/2020	664575	ΚΑΦΕ – ΜΠΑΡ	7.504
113078195	15/11/2019	664575	ΚΑΦΕ – ΜΠΑΡ	7.335



Σχήμα 7: Καλάθια-Πελάτες με το μεγαλύτερο κόστος

Πίνακας 9: Επαγγέλματα με τα περισσότερα καλάθια

Δηλωμένο Επάγγελμα	**Σύνολο καλαθιών**
ΨΗΤΟΠΩΛΕΙΟ	21977
ΕΣΤΙΑΤΟΡΙΟ	19203
ΑΝΑΨΥΚΤΗΡΙΟ	17152
ΚΑΦΕΤΕΡΙΑ	14950
ΚΑΦΕ - ΜΠΑΡ - ΕΣΤΙΑΤΟΡΙΟ	11301



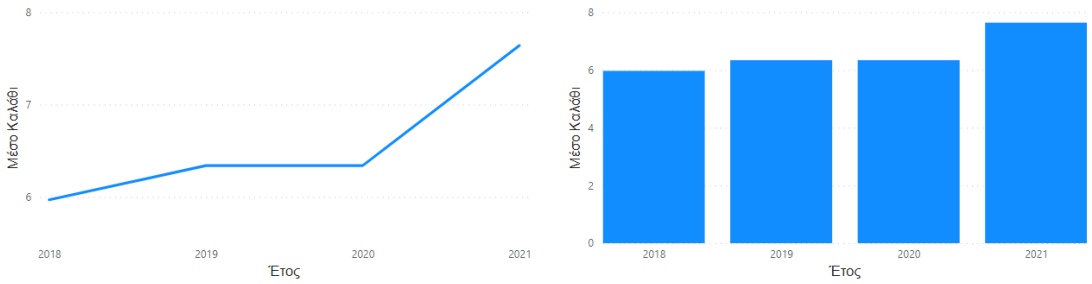
Σχήμα 8: Πλήθος Καλαθιών ανά Επάγγελμα

Πίνακας 10: Πελάτες με το μεγαλύτερο κόστος αγορών (HO.RE.CA. 2018-2021)

Κωδικός πελάτη	Ιδιότητα - Περιοχή	Κόστος καλαθιού (χωρίς ΦΠΑ)	Κόστος καλαθιού (με ΦΠΑ)
338509	ΨΗΤΟΠΩΛΕΙΟ	185.994	217.209
101619	ΠΙΤΣΑΡΙΑ	137.205	15.8023
417087	ΚΑΦΕΤΕΡΙΑ	96.127	11.7313
701120	CATERING	92.827	10.6029
641831	ΚΕΝΤΡΟ ΔΙΑΣΚΕΔΑΣΗΣ	76.086	94.305

Πίνακας 11: Μέσο καλάθι ανά έτος (HO.RE.CA. 2018-2021)

Έτος	Μέσο καλάθι (Χωρίς ΦΠΑ)	Μέσο Καλάθι (Με ΦΠΑ)
2018	5,97	7,12
2019	6,34	7,46
2020	6,34	7,33
2021	7,64	8,77
2018-2021	6,13	7,27



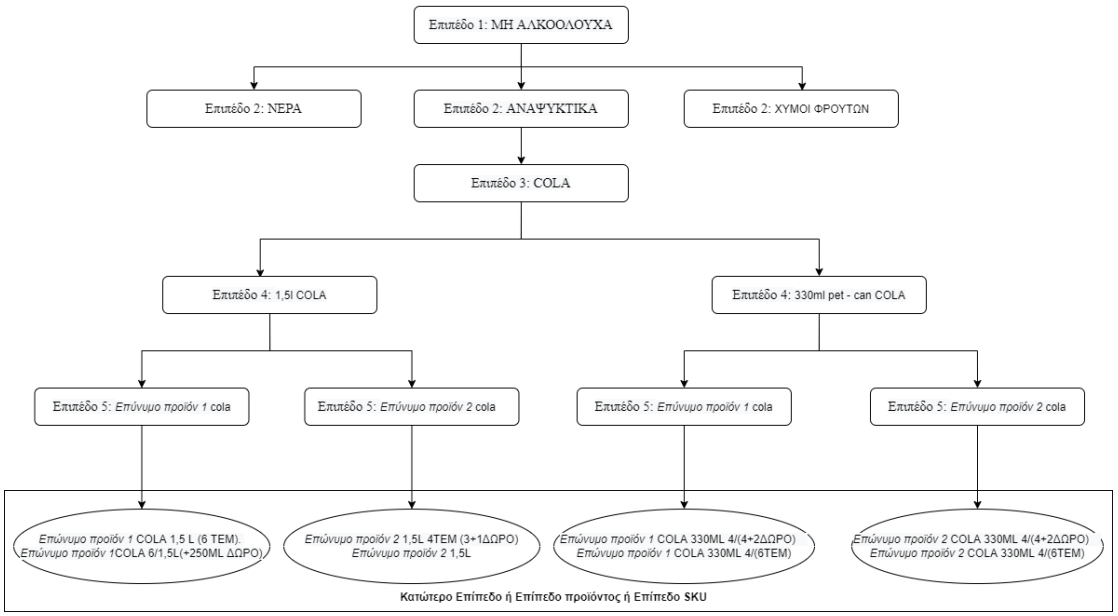
Σχήμα 9: Μέσο καλάθι ανά έτος

4 Εμπειρική Μελέτη

Η παρακάτω εμπειρική μελέτη έχει ως στόχο να χρησιμοποιηθούν η μεθοδολογία Crisp-DM, η τεχνική ανάλυση αγοραστικού καλαθιού (Market Basket Analysis) και τα δεδομένα που αναφέρθηκαν στο κεφάλαιο τρία ώστε να αναλυθούν οι αγορές των πελατών (HO.RE.CA.) της επιχείρησης. Έγινε εξαγωγή μόνο των συγκεκριμένων πελατών μαζί με τις αποδείξεις τους (εμπορικές συναλλαγές) με βάση τα παρακάτω:

- Να είναι HO.RE.CA
- Να είναι ενεργοί
- Να έχουν κάνει τουλάχιστον μια αγορά στα τελευταία 4 χρόνια

Οι εμπορικές συναλλαγές (αποδείξεις συναλλαγών) εμπεριέχουν τα προϊόντα που αγοράστηκαν, όμως η ανάλυση γίνεται σε επίπεδο προϊοντική κατηγορίας. Ως προϊοντική κατηγορία ορίζεται η κατηγορία στην οποία έχει δηλωθεί ένα προϊόν. Η επιχείρηση έχει πέντε συνολικά τέτοιες κατηγορίες και το επίπεδο έξι είναι το ίδιο το προϊόν (SKU):



Σχήμα 10: Παράδειγμα προϊοντικών κατηγοριών

Οι λόγοι που επιλέχθηκε να εκτελεσθεί Market Basket Analysis σε επίπεδο προϊοντικής κατηγορίας (Unit of Analysis for association rules) είναι:

- Ένα προϊόν έχει πολλαπλούς κωδικούς από διαφορετικούς προμηθευτές (π.χ. αναψυκτικά που έχουν τον ίδιο τύπο προϊόντος με διαφορετικό κωδικό-προμηθευτή)
- Οι κανόνες συσχέτισης θα εστιάσουν στον τύπο των προϊόντων (π.χ. Πορτοκαλάδα) παρά στη μάρκα τους.
- Οι κανόνες συσχέτισης θα έχουν υψηλότερη ποιότητα και θα αποφευχθεί ο θόρυβος στην πληροφορία.

Παρακάτω περιγράφεται όλη η διαδικασία της τεχνικής CRISP-DM που ακολουθήθηκε για την παραγωγή κανόνων συσχέτισης μεταξύ προϊόντων αλλά και επαγγελμάτων.

4.1 Επιχειρηματική κατανόηση (Business understanding)

Για τον εντοπισμό και κατανόηση του επιχειρηματικού στόχου της συγκεκριμένης ανάλυσης έγινε συνάντηση με τον γενικό διευθυντή της γενικής διεύθυνσης οργάνωσης και πληροφορικής μεγάλης εταιρείας λιανικής και χονδρικής πώλησης (Super Market). Μετά το πέρας της συνάντησης προέκυψαν τα παρακάτω στοιχεία.

Ο επιχειρηματικός σκοπός της συγκεκριμένη εργασίας είναι να εντοπισθούν πρότυπα και μοτίβα αγορών από τα αγοραστικά καλάθια των πελατών χονδρικής HO.RE.CA. (Hotel Restaurant Cafe). (Wang & Sun, 2019) Ο επιμέρους στόχος είναι να μελετηθεί αν είναι εφικτός ο εντοπισμός μια επαγγελματικής ιδιότητας μέσα από τα καλάθια των πελατών. Με αυτόν τον στόχο θα μπορούσαν να εντοπισθούν οι πελάτες χονδρικής της εταιρείας που δεν αγοράζουν προϊόντα συναφή με το δηλωμένο τους επάγγελμα. (Chen, Liu, Yu, Wei, & Zhang, 2006)

Για παράδειγμα ένας πελάτης έχει δηλωμένο επάγγελμα καφετέρια αλλά αγοράζει μόνο τα απορρυπαντικά του από την εταιρεία και καθόλου καφέ, γάλα ή ζάχαρη, επομένως οι αγορές του δεν αντιστοιχούν στη δηλωμένη του ιδιότητα. Με αυτή την

πληροφορία η εταιρεία μπορεί να εντοπίσει αν ένας πελάτης αγοράζει ή όχι τις πρώτες ύλες του από αυτήν, το να μην αγοράζει τα βασικά προϊόντα που επαγγέλλεται μπορεί να σημαίνει ότι τα αγοράζει από τον ανταγωνισμό ή ότι δεν τον καλύπτει η ποιότητας. Όποιος και να είναι ο λόγος μπορεί να εντοπισθεί από νέες έρευνες και να γίνει μια εκ' νέου προσέγγιση στον συγκεκριμένο πελάτη. Με την εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα η αλυσίδα χονδρεμπορίου μπορεί να επιδράσει στην ποικιλία των προϊόντων που διακινεί και να αναγνωρίσει νέες ευκαιρίες.

Για να μπορέσει να συμβεί αυτό θα χρειαστεί να δημιουργηθεί μια σύνδεση μεταξύ των προϊόντων και των επαγγελμάτων, επομένως να ανακαλυφθούν ποια προϊόντα και συνδυασμοί προϊόντων αγοράζονται περισσότερο από την εκάστοτε επαγγελματική ιδιότητα και να αξιολογηθεί αν είναι εφικτό το πλήθος των αγορών ενός πελάτη να τον ταξινομήσει σε ένα επάγγελμα.

4.2 Κατανόηση Δεδομένων (Data understanding)

Η προσέγγιση που ακολουθείται στην εξόρυξη γνώσης από τα δεδομένα μπορεί να αναλυθεί σε τρεις περιπτώσεις, την εξεύρεση κανόνων συσχέτισης μεταξύ των προϊοντικών κατηγοριών, την εξεύρεση κανόνων συσχέτισης μεταξύ προϊοντικών κατηγοριών και επαγγελμάτων και την εξεύρεση κανόνων συσχέτισης προϊοντικών κατηγοριών και επαγγελμάτων από ομαδοποιημένα καλάθια πελατών.

Από τα διαθέσιμα δεδομένα για να μπορέσουν να παραχθούν οι κανόνες συσχέτισης είναι απαραίτητα τα στοιχεία από τις εμπορικές συναλλαγές των πελατών (πίνακας Documents) μαζί με τα προϊόντα που περιέχουν (πίνακας DocumentsDetail), τα προϊόντα (πίνακας Products) και οι προϊοντικές κατηγορίες οι οποίες εμπεριέχονται μέσα στον πίνακα προϊόντων. Επίσης για να μπορέσει να υπάρξει σύνδεση μεταξύ επαγγελμάτων και προϊόντων είναι απαραίτητοι οι πελάτες (πίνακας Customers) και τα επαγγέλματα που έχει σαν δηλωμένα γι' αυτούς η επιχείρηση (πίνακας Jobs_First_Level και Jobs_Second_Level). Στο κεφάλαιο 3.4, πίνακας 3 αναφέρονται αναλυτικά οι πίνακες, τα πεδία και οι συνολικές εγγραφές που εξάχθηκαν.

4.3 Προετοιμασία Δεδομένων (Data preparation)

Για την προετοιμασία, τη διαχείριση και την επεξεργασία τους τα δεδομένα έγιναν εισαγωγή σε μια νέα βάση δεδομένων, χρησιμοποιήθηκε ο SQL Server και η γλώσσα προγραμματισμού T-SQL.

Από τους διαθέσιμους πίνακες δημιουργήθηκε ένας νέος που περιέχει όλη την απαραίτητη πληροφορία συγκεντρωτικά, ονομάστηκε BASKET_ITEMS στον οποίο συνδέθηκαν τα δεδομένα από τον πίνακας αποδείξεων και προϊόντων.
Τα πεδία του BASKET_ITEMS είναι:

- Κωδικός καλαθιού
- Κωδικός πελάτη
- Ημερομηνία συναλλαγής
- Κωδικός προϊόντος
- Περιγραφή προϊόντος
- Προϊοντική κατηγορία επιπέδου 1
- Προϊοντική κατηγορία επιπέδου 2

- Προϊοντική κατηγορία επιπέδου 3
- Προϊοντική κατηγορία επιπέδου 4
- Προϊοντική κατηγορία επιπέδου 5

Ο συγκεκριμένος πίνακας με διαφοροποιήσεις στα δεδομένα που περιέχει, χρησιμοποιήθηκε και στα τρία πειράματα που ενεργήθηκαν και οι διαφορές του σε κάθε περίπτωση αναφέρονται αναλυτικά στα επόμενα κεφάλαια.

4.3.1 Καθαρισμός δεδομένων (Data Cleansing):

Σε αυτή τη φάση της διαδικασίας γίνεται καθαρισμός των δεδομένων ώστε να μπορέσουν να αξιοποιηθούν σωστά από τον αλγόριθμο που θα χρησιμοποιηθεί για την παραγωγή των κανόνων συσχέτισης.

Από τα προϊόντα του πίνακα BASKET_ITEMS αφαιρέθηκαν όλα όσα ανήκαν στην προϊοντική κατηγορία ΚΕΝΑ, αφορούν τις κενές φιάλες (συνήθως μπύρας) και ανήκουν στα παρελκόμενα, χρεώνονται σαν ξεχωριστό προϊόν αλλά προστίθενται αυτόματα από το σύστημα με την αγορά συγκεκριμένων προϊόντων, επομένως αποτελούν χρέωση αλλά όχι αγορά ξεχωριστού προϊόντος. Αφαιρέθηκαν όλα τα προϊόντα που ανήκουν στην προϊοντική κατηγορία ΕΙΔΗ ΠΡΟΜΗΘΕΙΩΝ και EXTRA ΠΩΛΗΣΕΙΣ διότι περιλαμβάνουν σακούλες και δεν προσδίδουν πραγματική πληροφορία ως προς τις αγορές του πελάτη και τις συνδέσεις μεταξύ προϊόντων αφού η σακούλα αποτελεί προϊόν αναγκαιότητας και όχι προτίμησής.

Βρέθηκαν 13.547 καλάθια τα οποία είχαν ένα και μοναδικό προϊόν, από αυτά τα 5320 παρατηρείται ότι ο πελάτης έκανε αγορά την ίδια μέρα και γι’ αυτό το λόγο γίνεται η παραδοχή ότι η απόδειξη με το μοναδικό προϊόν έγινε σε δεύτερο χρόνο σαν συμπληρωματική της βασικής. Γι’ αυτό το λόγο προστέθηκε το μοναδικό προϊόν στο καλάθι που είχε αγοράσει ο πελάτης την ίδια μέρα, και αν υπάρχουν πάνω από ένα καλάθια προστέθηκε στο τελευταίο της ημέρας.

Μετά από αυτόν τον καθαρισμό απομένουν 8227 καλάθια με ένα και μοναδικό προϊόν στο οποίο δεν έγινε άλλη αγορά την ίδια μέρα. Ελέγχοντας αν υπάρχει απόδειξη για τον ίδιο πελάτη εντός δύο ημερών (μία μέρα πριν και μία μετά) παρατηρήθηκε ότι έχει προκύψει τουλάχιστον μια αγορά στο επιλεγμένο εύρος ημερομηνιών για 3826 καλάθια. Γι’ αυτό το λόγο αποφασίστηκε να προστεθεί το συγκεκριμένο προϊόν στο καλάθι της τελευταίας αγοράς θεωρώντας το συμπληρωματικό.

Τα εναπομείναντα 4401 καλάθια με ένα και μοναδικό προϊόν θεωρήθηκαν μεμονωμένες αγορές και αφαιρέθηκαν από το σύνολο των καλαθιών διότη από άποψη κανόνων συσχέτισης δεν είναι αξιοποιήσιμα αφού δεν παράγουν κανένα κανόνα.

Οι δηλωμένοι πελάτες HO.RE.CA ανέρχονται στους 61.184 πελάτες αλλά από αυτούς δεν έχουν όλοι αγοράσει προϊόντα τα τελευταία 4 χρόνια. Για αυτό το λόγο βάση των δοθέντων αποδείξεων θα κρατηθούν μόνο οι πελάτες που έχουν αγοράσει τουλάχιστον ένα προϊόν αυτά τα χρόνια και ανέρχονται στους 3.557.

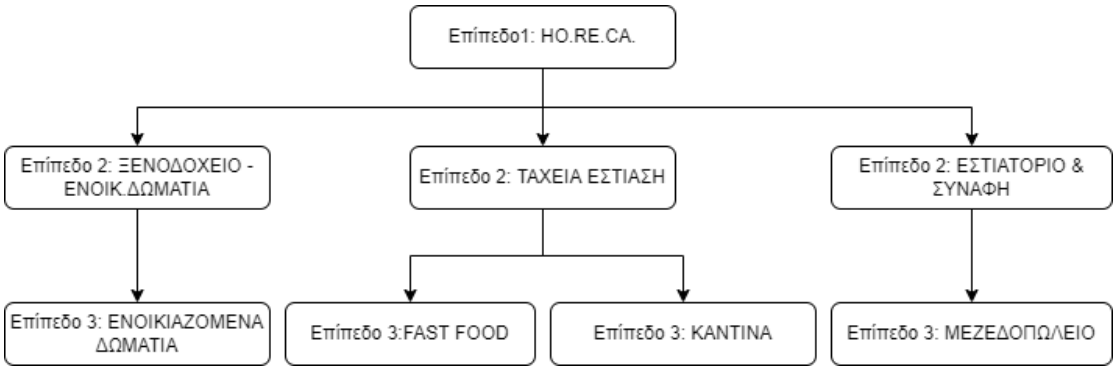
4.3.2 Μορφοποίηση δεδομένων - κανόνες συσχέτισης προϊόντων

Όπως προαναφέρθηκε οι κανόνες συσχέτισης που θα δημιουργηθούν θα είναι σε επίπεδο προϊοντικής κατηγορίας και όχι σε επίπεδο προϊόντος ώστε η πληροφορία να είναι συγκεντρωτική, ποιοτική και να αποφευχθεί ο θόρυβος που δημιουργείται σε χαμηλό επίπεδο με δυσνόητα αποτελέσματα (Σχήμα 10). Δεν θα χρησιμοποιηθεί επίσης η κατηγορία επιπέδου 5 διότι είναι πιο κοντά σε επίπεδο προμηθευτή ενώ η πληροφορία που ενδιαφέρει στην προκείμενη φάση είναι συγκεντρωτικά οι τύποι των προϊόντων και όχι η μάρκα ή ο προμηθευτής (Σχήμα 10).

Επομένως σαν περιγραφή προϊόντος θα αναγράφεται η περιγραφή της προϊοντικής κατηγορίας επιπέδου τέσσερα η οποία παρόλο που είναι αρκετά περιγραφική σε μερικά προϊόντα δεν δίνει πληροφορία κατανοητή στον απλό αναγνώστη. Για αυτόν το λόγο αποφασίστηκε να γίνει μια κωδικοποίηση που θα συνδυάζει τις περιγραφές των προϊοντικών κατηγοριών επιπέδου δύο, τρία και τέσσερα.
BASK_TREEDESCR2 [BASK_TREEDESCR3 (BASK_TREEDESCR4)]
π.χ. ΑΝΑΨΥΚΤΙΚΑ [COLA (1,5l COLA)]
Με αυτό τον τρόπο μοναδικοποιούνται οι περιγραφές των προϊόντων στο επίπεδο τέσσερα και είναι κατανοητές ως προς τον τύπο του προϊόντος που αναφέρονται.

Εντός των περιγραφών παρατηρήθηκε ότι υπάρχει ο χαρακτήρας κόμμα «,» και λόγο ότι ο αλγόριθμος Apriori που θα χρησιμοποιηθεί ξεχωρίζει τα προϊόντα που περιέχονται στα καλάθια διαχωρίζοντας τα με αυτόν, μπορεί να δημιουργήσει πρόβλημα αν υπάρχει μέσα στην περιγραφή. Για αυτό το λόγο έγινε αντικατάσταση του κόμμα «,» με τελεία «.» (π.χ. ΑΝΑΨΥΚΤΙΚΑ [COLA (1.5l COLA)]).

Από τους 3.557 πελάτες παίρνοντας συγκεντρωτικά τις κατηγορίες επαγγελμάτων τους, προκύπτουν συνολικά 66 επαγγέλματα. Η εταιρεία κρατάει μια δενδροειδή κατηγοριοποίηση για τα επαγγέλματα, τα οποία χωρίζει σε τρία επίπεδα όπως φαίνεται και στο σχήμα 11.



Σχήμα 11: Παράδειγμα κατηγοριοποίησης επαγγελμάτων

Επειδή οι κατηγορίες επαγγελμάτων επιπέδου τρία είναι πολύ ειδικές και βγαίνουν συνολικά 66 ελέγχεται η κατηγοριοποίηση επιπέδου 2 η οποία ανέρχονται σε συνολικά 7 κατηγορίες, αυτές με τη σειρά τους είναι πολύ γενικές. Για το λόγο αυτό έγινε μια νέα κατηγοριοποίηση των επαγγελμάτων σε μια ενδιάμεση κατάσταση ώστε να προκύψει μια πιο κατανοητή και εύκολη ταξινόμηση. Συνολικά επιλέχθηκαν 12 κατηγορίες: *Bar – Club, Catering, Αρτοποιείο – ζαχαροπλαστείο, Εργαστήρια γευμάτων, Εστιατόριο, Εσωτερικά μαγειρειά, Καφετέρια – Bar, Καφετέρια - Bar –*

εστιατόριο, Ξενοδοχείο, Ταβέρνα, Ταχεία εστίαση, Ψητοπωλείο

Στην συγκεκριμένη ανάλυση χρησιμοποιείται ο αλγόριθμος Apriori για την δημιουργία των συσχετίσεων, ο οποίο για να λειτουργήσει σωστά πρέπει τα προϊόντα μίας απόδειξης να είναι σε μορφή καλαθιού και όχι λίστας. Χρειάστηκε να γίνει ο παρακάτω μετασχηματισμός για να μπορέσει να λειτουργήσει ορθά ο αλγόριθμος.

Πίνακας 12: Προϊόντα αποδείξεων σε μορφή λίστας

Κωδικός συναλλαγής	Περιγραφή προϊόντος
1	Ψωμί
1	Αυγά
1	Σφουγγάρι
2	Ψωμί
2	Σαπούνι

Πίνακας 13: Προϊόντα αποδείξεων σε μορφή καλαθιού

Κωδικός συναλλαγής	Περιγραφή προϊόντος
1	Ψωμί, Αυγά, Σφουγγάρι
2	Ψωμί, Σαπούνι

Συνεχίζοντας με τον μετασχηματισμό των δεδομένων και για να μπορέσει να εκτελεστεί ο αλγόριθμος έγινε εξαγωγή των δεδομένων σε csv και ξανά εισαγωγή αλλά αυτή τη φορά σαν transaction (object της R) που περιέχει έναν πολύ αραιό πίνακα όπου οι στήλες είναι τα προϊόντα και οι γραμμές τα καλάθια. Πιο συγκεκριμένα στην περίπτωση αυτή ο πίνακας έχει 158763 γραμμές (transactions, baskets, καλάθια) και 9241 στήλες (items, προϊόντα).

Transaction ID	Products				
	Ψωμί	Αυγά	Σφουγγάρι		Σαπούνι
000001	1	1	1	…	0
000002	1	0	0		1
. . .			. . .		
158763	0	0	0	…	0

Όταν ένα προϊόν υπάρχει μέσα σε ένα καλάθι τότε στη στήλη αυτού του προϊόντος και στη γραμμή του καλαθιού μπαίνει η τιμή 1. Από τη στιγμή που τα δεδομένα βρίσκονται σε transaction μορφή μπορούν πλέον να αξιοποιηθούν από τον apriori.

4.3.3 Μορφοποίηση δεδομένων - κανόνες συσχέτισης προϊόντα - επαγγέλματα

Όπως προαναφέρθηκε και στον σκοπό της διπλωματικής εργασίας, το επίκεντρο ενδιαφέροντος πέρα από τα προϊόντα είναι οι πελάτες χονδρικής HO.RE.CA. και αν

είναι εφικτή η ανακάλυψη της επαγγελματικής τους ιδιότητας από τις αγορές που κάνουν. Για να μπορέσει να γίνει ανάλυση των αγορών ενός πελάτη είναι απαραίτητο να ερευνηθεί η σύνδεση της επαγγελματικής ιδιότητας με τα προϊόντα, και ένας τρόπος για να προκύψει μια τέτοια σύνδεση είναι η παραγωγή κανόνων συσχέτισης.

Στα καλάθια των προϊόντων υπάρχει δηλωμένος ο κωδικός πελάτη, μέσα από τον κωδικό πελάτη και την κατηγοριοποίησης επαγγελμάτων που έχει ορισθεί, προκύπτει η επαγγελματική ιδιότητα που αναφέρεται το συγκεκριμένο καλάθι. Για να μπορέσουν να δημιουργηθούν κανόνες συσχέτισης χρειάστηκε να εισαχθεί στο κάθε καλάθι ένα «πλασματικό προϊόν» το οποίο στην ουσία περιέχει την επαγγελματική ιδιότητα που αγόρασε το συγκεκριμένο καλάθι. Με αυτό τον τρόπο και χρησιμοποιώντας τον αλγόριθμο apriori παράγεται ένα νέο σύνολο από κανόνες συσχέτισης με τα επαγγέλματα (πλασματικά προϊόντα) μέσα σε αυτές. Επομένως τα καλάθια μεταμορφώθηκαν στη μορφή:

Πίνακας 14: Προϊόντα μαζί με το επάγγελμα του καλαθιού σε μορφή λίστας

Κωδικός συναλλαγής	Περιγραφή προϊόντος
1	Κάρβουνα
1	Πατάτες
1	Μπριζόλες
1	Ψητοπωλείο
2	Βότκα
2	Μπύρες
2	Καφετέρια – Bar

Ακολουθώντας την ίδια διαδικασία ώστε να έρθουν οι αποδείξεις σε μορφή καλαθιού, που είναι απαραίτητη για να χρησιμοποιηθεί ο apriori μετατρέπονται όπως φαίνεται παρακάτω:

Πίνακας 15: Προϊόντα μαζί με το επάγγελμα του καλαθιού τους σε μορφή καλαθιού

Κωδικός συναλλαγής	Περιγραφή προϊόντος
1	Κάρβουνα, Πατάτες, Μπριζόλες, Ψητοπωλείο
2	Βότκα, Μπύρες, Καφετέρια – Bar

Με τη νέα αυτή αλλαγή τα δεδομένα των καλαθιών αλλάζουν αφού δεν υπάρχει κάποιο που να μην περιέχει το νέο «πλασματικό προϊόν» με την επαγγελματική ιδιότητα. Δίνονται ενδεικτικά κάποια στοιχεία των νέων καλαθιών:

Πίνακας 16: Συχνότερα προϊόντα στα καλάθια

Προϊόν	Προϊοντική κατηγορία	Συνολικές αγορές
ΚΑΦΕΤΕΡΙΑ-BAR		37006
ΨΗΤΟΠΩΛΕΙΟ		25383
ΚΑΦΕΤΕΡΙΑ-BAR-ΕΣΤΙΑΤΟΡΙΟ		23485

ΕΣΤΙΑΤΟΡΙΟ		18210
COCA COLA 330ML 4/(6ΤΕΜ.)	ΑΝΑΨΥΚΤΙΚΑ	15172

Η διαδικασία είναι ίδια με αυτή των προϊόντων, δημιουργήθηκε ένα transaction (object της R) που περιέχει έναν πολύ αραιό πίνακα όπου οι στήλες είναι τα προϊόντα και οι γραμμές τα καλάθια, με την διαφορά ότι αυτή τη φορά στις στήλες έχουν προστεθεί και τα επαγγέλματα που υπάρχει σίγουρα ένα σε κάθε καλάθι.

Μορφοποίηση δεδομένων - κανόνες συσχέτισης προϊόντα, επαγγέλματα σε υψηλότερο επίπεδο προϊοντικής κατηγορίας.

Για να μπορέσει να γίνει μια ολοκληρωμένη ανάλυση των κανόνων συσχέτισης που παράγονται δημιουργήθηκε ένα δεύτερο σύνολο δεδομένων (dataset) που θα έχει την πληροφορία των προϊόντων σε ακόμα υψηλότερο επίπεδο προϊοντικής κατηγορίας και πιο συγκεκριμένα στο επίπεδο 2 (σχήμα 10).
Αυτό έχει σαν αποτέλεσμα πολλά προϊόντα που βρίσκονται στο ίδιο καλάθι να είναι όμοια και να ομαδοποιηθούν σε ένα δηλαδή να αφαιρεθούν τα διπλά, τριπλά κοκ. Για παράδειγμα σε ένα καλάθι υπάρχουν «ΚΡΕΜΙΔΙΑ» και «ΠΑΤΑΤΕΣ», η ανάλυση γίνεται σε επίπεδο προϊοντικής κατηγορίας 2 επόμενος αυτά γίνονται «ΛΑΧΑΝΙΚΑ», επειδή «ΛΑΧΑΝΙΚΑ» θεωρούνται το ίδιο προϊόν μοναδικοποιούνται.

Επομένως το δεύτερο σύνολο δεδομένων θα είναι ακριβώς όπως το από πάνω με τη διαφορά οι ονομασίες των προϊόντων θα έχουν μόνο το όνομα της κατηγορίας που είναι δηλωμένα. Επίσης αφού τα προϊόντα στα καλάθια θα ομαδοποιηθούν (στην κατηγορία) θα είναι πολύ λιγότερα. Αυτό φαίνεται και όταν δημιουργείται το transaction που χρησιμοποιείται από το apriori όπου σε αυτή την περίπτωση ο πίνακα έχει 158763 γραμμές (transactions, baskets, καλάθια) και 339 στήλες (items, προϊόντα)

Πίνακας 17: Συχνότερα προϊόντα στα καλάθια (Προϊοντική κατηγορία 2)

Προϊόν	Συνολικές εμφανίσεις
ΑΝΑΨΥΚΤΙΚΑ ΣΥΣΚΕΥΑΣΜΕΝΑ-ΤΥΠΟΠΟΙΗΜΕΝΑ	54565
ΧΥΜΑ-ΠΑΓΚΟΥ	50895
ΛΑΧΑΝΙΚΑ	49813
ΚΑΦΕΤΕΡΙΑ-BAR	36976

4.3.4 Μορφοποίηση δεδομένων – κανόνες συσχέτισης ομαδοποιημένων καλαθιών

Εστιάζοντας το ενδιαφέρον στους πελάτες HO.RE.CA. και στην προσπάθεια να εντοπισθεί αν είναι εφικτή η ανακάλυψη της επαγγελματικής ιδιότητας από τις αγορές τους γίνεται ένα ακόμα πείραμα παραγωγής κανόνων συσχέτισης αλλά αυτή τη φορά έχουν ομαδοποιηθεί όλα τα καλάθια του εκάστοτε πελάτη σε ένα. Αυτό έχει σαν αποτέλεσμα τα προϊόντα που έχουν αγορασθεί από τον πελάτη και είναι ίδια μεταξύ τους σε διαφορετικές ημερομηνίες να ομαδοποιούνται σε ένα μοναδικό

προϊόν. Ορίζεται σαν κωδικός καλαθιού ο κωδικός πελάτη και όλα τα καλάθια που είναι δηλωμένα στον συγκεκριμένο κωδικό ενοποιούνται και γίνονται μια μεγάλη συνολική αγορά.

Πίνακας 18: Καλάθια ενοποιημένα σε κωδικό πελάτη

Κωδικός πελάτη	Περιγραφή προϊόντος
6559	Κάρβουνα
6559	Πατάτες
6559	Μπριζόλες
6559	Ψητοπωλείο
1056	Βότκα
1056	Μπύρες
1056	Καφετέρια – Bar
1056	Πάνες
1056	Καφετέρια – Bar

Με αυτό τον τρόπο όλες οι συναλλαγές – καλάθια ομαδοποιούνται σε ένα και στο νέο σύνολο δεδομένων υπάρχουν 3.557 καλάθια, όσοι είναι και οι πελάτες HO.RE.CA. Αυτό αποδεικνύεται και από τον πίνακα που περιλαμβάνει το transaction που αποτελείται από έχει 3.557 γραμμές (transactions, baskets, καλάθια) και 7195 στήλες (items, προϊόντα).

4.4 Μοντελοποίηση (Modeling)

Το συγκεκριμένο μοντέλο ανάλυσης περιλαμβάνει τον εντοπισμό σχέσεων μεταξύ των προϊόντων, δηλαδή κανόνων συσχέτισης. Η προσέγγιση όμως στο πρόβλημα που αναλύεται, απαιτεί και τη δημιουργία κανόνων συσχέτισης επαγγελμάτων με προϊόντα, δηλαδή μπορεί η αγορά κάποιον προϊόντων ή συνδυασμό προϊόντων να υπονοήσει ένα συγκεκριμένο επάγγελμα.

Για τη διαδικασία της μοντελοποίησης χρησιμοποιήθηκε το εργαλείο η γλώσσα προγραμματισμού R καθώς επίσης βιβλιοθήκες και αλγόριθμοι της R.
Βιβλιοθήκες: RODBC, PLYR, DPLYR, STRINGR, ARULES, ARULESVIZ
Αλγόριθμοι - Functions: apriori, ddply, read.transactions, itemFrequencyPlot κ.α.
Αξίζει να αναφερθεί το πακέτο Arules της γλώσσας R παρέχει μια βασική υποδομή για τη δημιουργία και τον χειρισμό συνόλων δεδομένων (data sets), για την ανάλυση των σύνολο-στοιχείων (itemset) και κανόνων που προκύπτουν. (Hahsler & Grun, arules – A Computational Environment for Mining Association Rules and Frequent Item Sets, 2005) Καθώς επίσης και το πακέτο aruleViz που εφαρμόζει πολλές διάφορες τεχνικές οπτικοποίησης για εξερεύνηση των κανόνων συσχέτισης.

4.4.1 Μοντελοποίηση κανόνων συσχέτισης προϊόντων

Όπως αναφέρεται στη βιβλιογραφία ο αλγόριθμος apriori χρησιμοποιεί τα μέτρα αξιολόγησης Support και Confidence που πρέπει να δοθούν σαν παράμετροι για να μπορέσει να δημιουργήσει association rules. (Agrawal, Imielinski, & Swami, 1993) (Kotsiantis & Kanellopoulos, 2006)

Για να επιλεχθεί ο καταλληλότερος συνδυασμός Support και Confidence, που θα παράγει ικανοποιητικούς κανόνες από πλευρά πλήθους και ισχύος έγινε χρήση των παρακάτω συνδυασμών:

- Support = 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05
- Confidence = 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5

Δίνεται το σχήμα με το πλήθος κανόνων που παράγονται ανά συνδυασμό support και confidence, το πλήθος έχει χωρισθεί σε κλάσεις των χιλίων, για παράδειγμα ο συνδυασμός support = 0.01 και confidence = 0.02 δίνει από 0 έως 1000 κανόνες συσχέτισης.

Σχήμα 12: Πλήθος κανόνων ανά συνδυασμό Support και Confidence

Συγκεντρωτικά δημιουργήθηκαν 196 σύνολα κανόνων εκ' των οποίον:

- Support = 0.05 και οποιοδήποτε από τα παραπάνω Confidence έδιναν μηδενικούς κανόνες
- Support < 0.05 και Support >= 0.007 και οποιοδήποτε από τα παραπάνω Confidence έδιναν λίγους ισχυρούς κανόνες και όσο το confidence πλησίαζε το ένα αυτοί μειώνονταν αλλά γίνονταν πιο ισχυροί.
- Support <= 0.04 παράγει μεγάλο πλήθος κανόνων που περιλαμβάνει και τους πολύ αδύναμους, ρυθμίζοντας το Confidence να πλησιάζει πιο κοντά στο ένα αυτοί οι κανόνες μειώνονται.

Για να παραχθεί ένα ικανοποιητικό πλήθος κανόνων που θα μπορέσει να δώσει πληροφορία με ενδιαφέρον και οι κανόνες συσχέτισης θα είναι συμπαγείς, η ανάλυση κατέληξε σε ένα **Support = 0.004 και Confidence = 0.5** ώστε να περιλαμβάνονται μέσα σε αυτό κανόνες σχετικά ισχυροί (σύνολο κανόνων 143). Αφαιρώντας τους περιττούς απομένουν 140. Ορισμένοι από τους κανόνες είναι αντίστροφοι έχοντας το ίδιο Support και Lift δηλαδή:

Πίνακας 19: Αντίστροφοι κανόνες

LHS	RHS	Measures

		Supp	Conf	Lift
ΓΛΥΚΙΖΟΝΤΑ ΠΟΤΑ [ΑΠΕΡΙΤΙΦ (ΑΠΕΡΙΤΙΦ)]	ΟΥΙΣΚΙ [ΟΥΙΣΚΙ (STANDARD)]	0.0052	0.53	7.28
ΟΥΙΣΚΙ [ΟΥΙΣΚΙ (STANDARD)]	ΓΛΥΚΙΖΟΝΤΑ ΠΟΤΑ [ΑΠΕΡΙΤΙΦ (ΑΠΕΡΙΤΙΦ)]	0.0052	0.07	7.28

Επιλέχθηκε να κρατηθεί μόνο ο ένας από τους δύο αντίστροφους κανόνες ώστε να μειωθεί το πλήθος και να είναι πιο κατανοητοί, και καθαρίζοντας τους απέμειναν 123 κανόνες συσχέτισης.

4.4.2 Μοντελοποίηση κανόνων συσχέτισης προϊόντων - επαγγελμάτων

Μετά τις απαραίτητες προσαρμογές στα καλάθια ώστε να περιλαμβάνεται και η επαγγελματική ιδιότητα σαν «εικονικό προϊόν», ακολουθήθηκε η ίδια διαδικασία που χρησιμοποιήθηκε και στην περίπτωση των προϊόντων για να επιλεχθεί ο καταλληλότερος συνδυασμός Support και Confidence, που παράγει έναν ικανοποιητικό αριθμό κανόνων οι οποίοι θα δίνουν πληροφορία:

- Support = 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05
- Confidence = 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5

Συγκεντρωτικά δημιουργήθηκαν 196 σύνολα κανόνων εκ’ των οποίον:

- Support = 0.05 και οποιοδήποτε από τα παραπάνω Confidence έδιναν μηδενικούς κανόνες
- Support < 0.05 και Support >= 0.008 και οποιοδήποτε από τα παραπάνω Confidence έδιναν λίγους ισχυρούς κανόνες και όσο το confidence πλησίαζε το ένα αυτοί μειώνονταν αλλά γίνονταν πιο ισχυροί.
- Support <= 0.03 παράγει μεγάλο πλήθος κανόνων που περιλαμβάνει και τους πολύ αδύναμους κανόνες, ρυθμίζοντας το Confidence να πλησιάζει πιο κοντά στο ένα αυτοί οι κανόνες μειώνονται.

Για να παραχθεί ένα ικανοποιητικό πλήθος κανόνων η ανάλυση κατέληξε σε ένα **Support = 0.004 και Confidence = 0.5** ώστε να περιλαμβάνονται μέσα σε αυτό κανόνες σχετικά ισχυροί αλλά και μη προβλέψιμοι (σύνολο κανόνων 297) αφαιρώντας τους περιττούς απομένουν 292. Όπως και με τα προϊόντα επιλέχθηκε να καθαριστούν οι αντίστροφοι κανόνες επομένως μετά την αφαίρεση 655 αντίστροφων κανόνων απέμεινα 242.

Πίνακας 20: Κανόνες συσχέτισης που περιλαμβάνουν επαγγέλματα

LHS	**RHS**	**Measures**		
		Supp	Conf	Lift
ΟΥΙΣΚΙ [ΟΥΙΣΚΙ (DELUXE)]	ΚΑΦΕΤΕΡΙΑ-BAR	0.013	0.52	2.29
ΧΥΜΟΙ - ΣΑΛΤΣΕΣ [ΣΑΛΤΣΕΣ (1.01-5KG ΣΑΛΤΣΕΣ)]	ΤΑΧΕΙΑ ΕΣΤΙΑΣΗ	0.0051	0.540	6.75
ΚΑΦΕΔΕΣ ΣΤΙΓΜΙΑΙΟΙ [ΧΩΡΙΣ ΚΑΦΕΙΝΗ	ΚΑΦΕΤΕΡΙΑ-BAR	0.0051	0.49	2.14

(NESCAFE)]				

Μοντελοποίηση κανόνων συσχέτισης προϊόντων – επαγγελμάτων σε υψηλότερο επίπεδο προϊοντικής κατηγορίας.

Ακολουθώντας ακριβώς την ίδια διαδικασία για να επιλεχθούν τα Support και Confidence που θα χρησιμοποιηθούν παρατηρείται ότι το πλήθος των κανόνων αυξήθηκε ραγδαία. Αυτό οφείλεται στο ότι τα προϊόντα ομαδοποιήθηκαν σε μεγαλύτερο επίπεδο και μέσα στο εκάστοτε καλάθι είναι πολύ πιθανότερο να βρεθεί ένα τέτοιο προϊόν.

Έγινε επιλογή των Support = 0.01 και Confidence = 0.4 και χρησιμοποιήθηκε το transaction που δημιουργήθηκε από το σύνολο στοιχείων (dataset) προϊοντικής κατηγορίας επιπέδου δύο. Συνολικά δημιουργήθηκαν 2300 κανόνες όπου αφαιρώντας τους περιττούς και αντίστροφους απομένουν 1484.

4.4.3 Μοντελοποίηση κανόνων συσχέτισης ομαδοποιημένων καλαθιών

Μετά την ομαδοποίηση των καλαθιών στους πελάτες και την δημιουργία του transaction μπορεί να εκτελεσθεί ο αλγόριθμος Apriori, δοκιμάζοντας διάφορα support και ένα σταθερό υψηλό confidence (0.5) παρατηρείται ότι αυτό το σύνολο δεδομένων επιστρέφει έναν πολύ μεγάλο όγκο κανόνων. Αυτό οφείλεται στο ότι τα ομαδοποιημένα καλάθια εμπεριέχουν τεράστιο πλήθος προϊόντων και επομένως περισσότερες συνδέσεις μεταξύ τους.

Για να μπορέσουν να μειωθούν οι κανόνες αλλά να είναι αρκετά ισχυροί χρησιμοποιήθηκε ο συνδυασμός support = 0.2 και confidence = 0.5 ο οποίος δημιουργεί 169 κανόνες από τους οποίους αφαιρώντας τους τους περιττούς και αντίστροφους απομένουν 95 κανόνες.

4.5 Εκτίμηση (Evaluation)

Μετά την παραγωγή του μοντέλου έρχεται η εκτίμηση των αποτελεσμάτων που παράγει. Το σημαντικό σε αυτό το σημείο είναι να εντοπισθεί η αξία που μπορεί να δώσει στην επιχείρηση (business value) και να προσδιοριστεί εάν υπάρχει κάποιο σημαντικό επιχειρηματικό ζήτημα (business issue) που δεν έχει εξεταστεί επαρκώς. (Wirth & Hipp, 2000)

4.5.1 Εκτίμηση κανόνων συσχέτισης προϊόντων

Από τους 123 κανόνες που παράχθηκαν κάνοντας χρήση του Lift, τρίτου μέτρου αξιολόγησης κανόνων συσχέτισης, δημιουργήθηκε ένα υποσύνολο με 20 αρκετά ισχυρούς κανόνες που έχουν Lift > 10. Ταξινομώντας τους βάση του Lift αναφέρονται ενδεικτικά οι τρεις ποιο ισχυροί από αυτούς.

Πίνακας 21: Ισχυρότεροι κανόνες βάση Lift

LHS	RHS	Measures

		Supp	Conf	Lift
ΑΝΑΨΥΚΤΙΚΑ [ΛΕΜΟΝΑΔΑ (250ml γυάλινο ΛΕΜΟΝΑΔΑ)]	ΑΝΑΨΥΚΤΙΚΑ [ΠΟΡΤΟΚΑΛΑΔΑ (250ml γυάλινο ΠΟΡΤΟΚΑΛΑΔΑ)]	0.0059	0.62	41.4
ΑΝΑΨΥΚΤΙΚΑ [ΛΕΜΟΝΑΔΑ (330ml γυάλινο ΛΕΜΟΝΑΔΑ)]	ΑΝΑΨΥΚΤΙΚΑ [ΠΟΡΤΟΚΑΛΑΔΑ (330ml γυάλινο ΠΟΡΤΟΚΑΛΑΔΑ)]	0.0090	0.84	23.04
Επώνυμο προϊόν[ΑΛΚΟΟΛΟΥΧΑ ΠΟΤΑ (ΑΚΟΟΛΟΥΧΑ ΠΟΤΑ)]	ΛΕΥΚΑ ΠΟΤΑ [ΒΟΤΚΑ (ΒΟΤΚΑ ΦΥΣΙΚΗ)]	0.0053	0.53	19.66

Οι συγκεκριμένοι κανόνες είναι πολύ ισχυροί, που σημαίνει ότι οι συνδυασμοί αυτών των προϊόντων συμβαίνουν συχνά μέσα στα καλάθια των πελατών. **Παρατηρείτε ότι τα προϊόντα που περιέχονται σε αυτούς τους κανόνες ανήκουν συνήθως στην ίδια προϊοντική κατηγορία. Αυτό διεγείρει τα εξής ερωτήματα, οι αγορές αυτές οφείλονται στο ότι τα συγκεκριμένα προϊόντα τοποθετούνται κοντά το ένα με το άλλο;** Ή ο πελάτης έχει έρθει με στόχο να καλύψει τις ανάγκες του καταστήματος του από τη επιχείρηση; Για παράδειγμα ένα εστιατόριο που κάνει τον παραπάνω συνδυασμό έχει έρθει με στόχο να καλύψει τις ανάγκες του σε αναψυκτικά ή ο συνδυασμός αυτός οφείλεται στο γεγονός ότι βρίσκονται στον ίδιο χώρο και έχει εύκολη πρόσβαση και στα δύο.

Εξάγοντας κανόνες με χαμηλότερο Lift (< 1) κάποιες συνδέσεις μεταξύ κατηγοριών μπορεί να έχουν ενδιαφέρον αλλά τα μέτρα αξιολόγησης τους υποδηλώνουν αδύναμους κανόνες που ο συνδυασμός τους συμβαίνει σπάνια επομένως είναι δύσκολο να φανερώσουν μοτίβα αγορών, ενδεικτικά αναφέρονται:

Πίνακας 22: Αδύναμοι Κανόνες με Lift < 1

LHS	RHS	Measures		
		Supp	Conf	Lift
ΞΥΝΑ [ΛΕΜΟΝΙΑ (ΛΕΜΟΝΙΑ ΕΓΧ.)]	ΑΝΑΨΥΚΤΙΚΑ [COLA (330ml pet - can COLA)]	0.008	0.13	0.98
ΓΑΛΑ ΥΨΗΛΗΣ ΠΑΣΤΕΡΙΩΣΗΣ [ΛΕΥΚΟ (ΠΛΗΡΕΣ)]	ΣΠΟΡΕΛΑΙΑ [ΗΛΙΕΛΑΙΟ (10l ΗΛΙΕΛΑΙΟ)]	0.0062	0.086	0.98
ΨΩΜΙ ΣΥΣΚ/ΝΟ [ΤΟΣΤ (ΦΟΡΜΑ ΕΠΑΓΓΕΛΜΑΤΙΚΗ)]	ΝΕΡΑ [ΦΥΣΙΚΑ ΝΕΡΑ (500ml ΝΕΡΑ)]	0.006	0.087	0.98

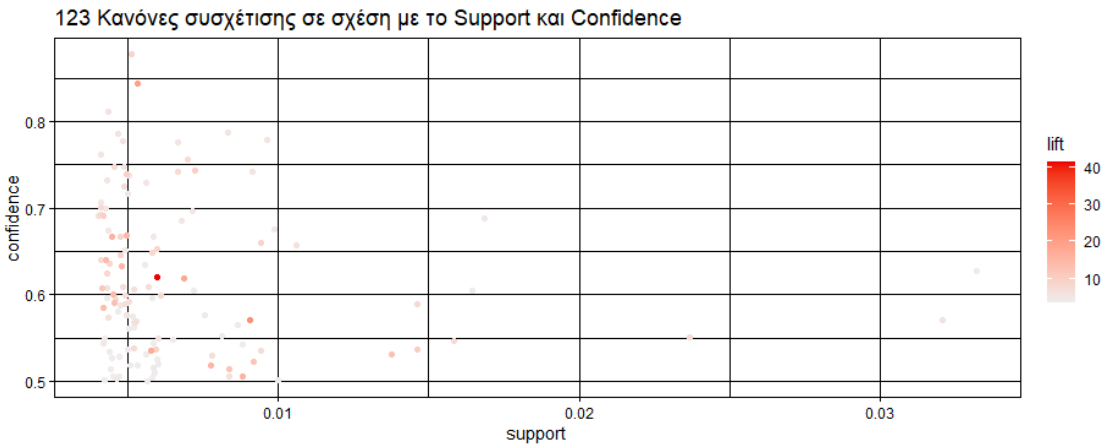
Κρατώντας ένα Lift μεταξύ του 3 και του 5 παίρνουμε κάποιους κανόνες που είναι σχετικά ισχυροί και θα μπορούσαν να θεωρηθούν χρήσιμοι αφού αρκετοί από αυτούς δεν αφορούν προϊόντα που ανήκουν στις ίδιες προϊοντικές κατηγορίες. Δημιουργήθηκαν συνολικά 43 κανόνες συσχέτισης και δίνονται ενδεικτικά μερικοί:

Πίνακας 23: Κανόνες με Lift μεταξύ του 3 και του 5

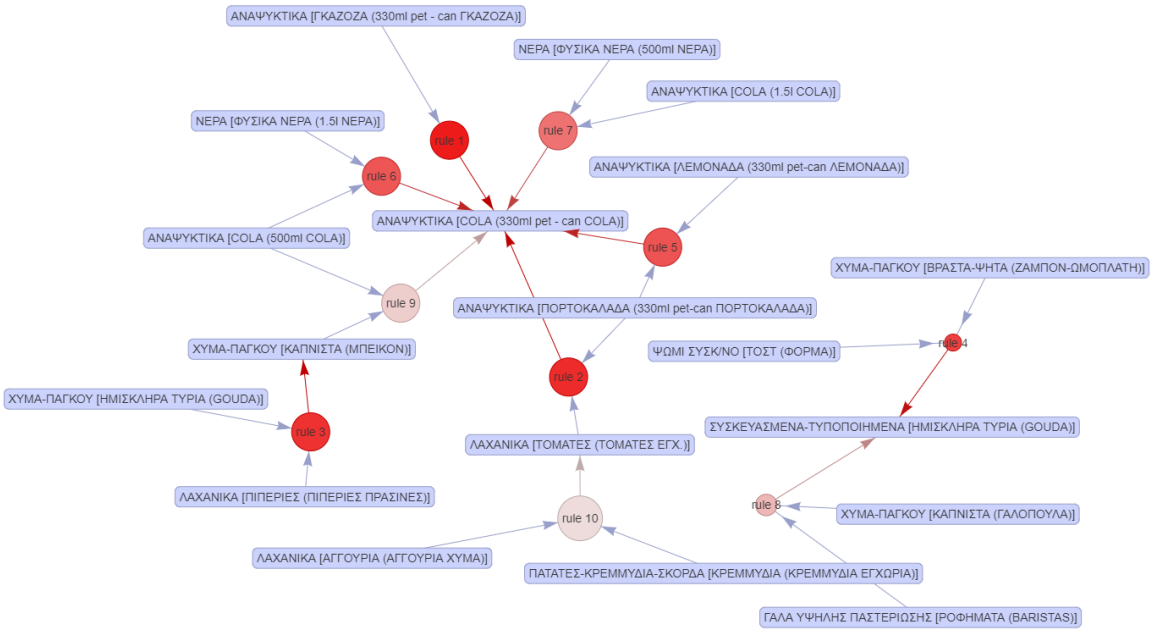
LHS	RHS	Measures		
		Supp	Conf	Lift
ΝΕΡΑ [ΦΥΣΙΚΑ ΝΕΡΑ (500ml ΝΕΡΑ)], ΧΥΜΑ-ΠΑΓΚΟΥ [ΚΑΠΝΙΣΤΑ (ΜΠΕΙΚΟΝ)]	ΑΝΑΨΥΚΤΙΚΑ [COLA (*Επώνυμο προϊόν*)]	0.0058	0.504	3.66
ΓΑΛΑ ΥΨΗΛΗΣ ΠΑΣΤΕΡΙΩΣΗΣ [ΡΟΦΗΜΑΤΑ (*Επώνυμο προϊόν*)], ΧΥΜΑ-ΠΑΓΚΟΥ [ΚΑΠΝΙΣΤΑ (ΓΑΛΟΠΟΥΛΑ)]	ΣΥΣΚΕΥΑΣΜΕΝΑ-ΤΥΠΟΠΟΙΗΜΕΝΑ [ΗΜΙΣΚΛΗΡΑ ΤΥΡΙΑ (GOUDA)]	0.0058	0.515	4.67
ΧΥΜΑ-ΠΑΓΚΟΥ [ΒΡΑΣΤΑ-ΨΗΤΑ (ΖΑΜΠΟΝ-ΩΜΟΠΛΑΤΗ), ΨΩΜΙ ΣΥΣΚ/ΝΟ [ΤΟΣΤ (ΦΟΡΜΑ)]	ΣΥΣΚΕΥΑΣΜΕΝΑ-ΤΥΠΟΠΟΙΗΜΕΝΑ [ΗΜΙΣΚΛΗΡΑ ΤΥΡΙΑ (GOUDA)]	0.0042	0.543	4.93

Οι συγκεκριμένοι κανόνες είναι σχετικά ισχυροί, επομένως συμβαίνουν συχνά στα καλάθια των πελατών και διεγείρουν το ερώτημα αν δημιουργούνται από συμπληρωματικά προϊόντα. Για παράδειγμα, όπως δίνεται και στον πίνακα 23, υπάρχει σύνδεση μεταξύ, «Βραστά πάγκου, ζαμπόν ωμοπλάτη», «Ψωμί του τοστ» και «Συσκευασμένα τυριά Gouda», αυτός ο συνδυασμός θα μπορούσε κανείς να πει ότι υποδηλώνει κατασκευή τοστ. **Θεωρούνται αυτά τα προϊόντα συμπληρωματικά;** Θα μπορούσαν να αναγνωρισθούν τα συμπληρωματικά προϊόντα; Ποια θα ήταν τα κριτήρια για να θεωρηθούν τα προϊόντα ενός κανόνα συμπληρωματικά; Θα μπορούσε να αναγνωρισθεί το προϊόν στο οποίο θα μεταποιηθούν; Θα μπορούσαν αυτά τα προϊόντα να μπουν στον ίδιο χώρο για να αυξηθούν οι πωλήσεις;

Παρακάτω παρουσιάζονται κάποια διαγράμματα από τους κανόνες συσχέτισης που παράχθηκαν, ώστε να γίνει περισσότερο κατανοητή ή έννοια της σύνδεσης μεταξύ προϊόντων και να δοθεί μια πιο πλήρη εικόνα των συνδέσεων και των δυνατοτήτων.



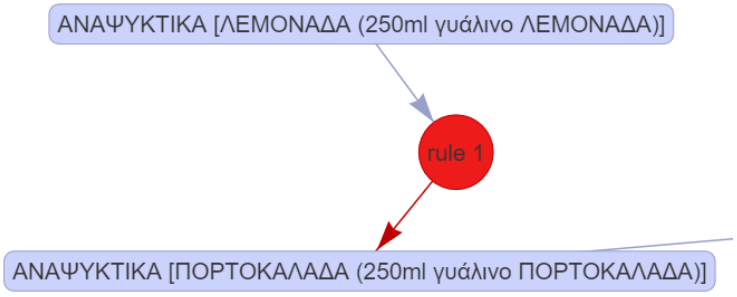
Σχήμα 13: Διάγραμμα κανόνων συσχέτισης



Σχήμα 14: Οι 10 κανόνες συσχέτισης με το μεγαλύτερο Lift

Επιλέχθηκε ο πιο ισχυρός κανόνας βάση των κριτηρίων αξιολόγησης κανόνων (support και confidence) από τους γενικούς κανόνες συσχέτισης για παρουσίαση και περαιτέρω ανάλυση:

- **Rule1 (Supp: 0.005, Conf: 0.62, Lift: 41.4)**{ ΑΝΑΨΥΚΤΙΚΑ [ΛΕΜΟΝΑΔΑ (250ml γυάλινο ΛΕΜΟΝΑΔΑ)]=> ΑΝΑΨΥΚΤΙΚΑ [ΠΟΡΤΟΚΑΛΑΔΑ (250ml γυάλινο ΠΟΡΤΟΚΑΛΑΔΑ)]}



Σχήμα 15: Ισχυρότερος κανόνας συσχέτισης

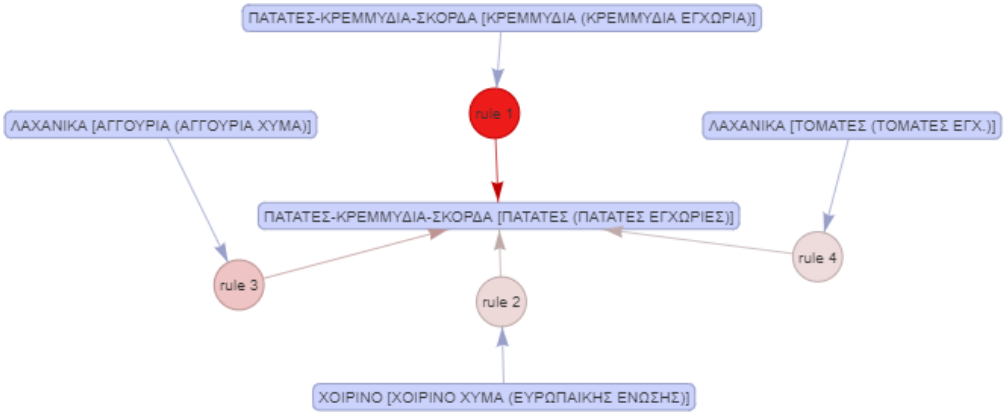
Ο αλγόριθμος apriori δίνει την δυνατότητα να ορίσει ο χρήστης το προϊόν που θέλει να διερευνήσει τοποθετώντας το σε οποιοδήποτε πλευρά του κανόνα δίνοντας την δυνατότητα να απομονωθεί η πληροφορία μόνο στο επιλεγμένο προϊόν και να εντοπισθούν οι συνδέσεις του με τα υπόλοιπα. Όπως αναφέρεται και στο Κεφάλαιο 3.3, ένας κανόνας συσχέτισης είναι μια συνεπαγωγή της μορφής $X \Rightarrow Y$, το X ονομάζεται προγενέστερο ή αριστερή πλευρά του κανόνα (LHS) και το Y ονομάζεται συνέπεια ή δεξιά πλευρά του κανόνα (RHS), σημαίνει ότι το X υποδηλώνει Y. (Jabbeen, 2018). Στις παρακάτω σελίδες δίνονται μερικά τέτοια παραδείγματα. (Hahsler & Grun, 2005)

Παράδειγμα εξόρυξης κανόνων συσχέτισης με δηλωμένο προϊόν στην δεξιά πλευρά

του κανόνα το «ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]» (σύνολο κανόνων 4):

Πίνακας 24:Κανόνες συσχέτισης με συγκεκριμένο RHS

LHS	RHS	Measures		
		Supp	Conf	Lift
ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΚΡΕΜΜΥΔΙΑ (ΚΡΕΜΜΥΔΙΑ ΕΓΧΩΡΙΑ)]	ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	0.0050	0.11	5.25
ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]	ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	0.0058	0.065	2.99
ΛΑΧΑΝΙΚΑ [ΑΓΓΟΥΡΙΑ (ΑΓΓΟΥΡΙΑ ΧΥΜΑ)]	ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	0.0063	0.073	3.35
ΛΑΧΑΝΙΚΑ [ΤΟΜΑΤΕΣ (ΤΟΜΑΤΕΣ ΕΓΧ.)]	ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	0.008	0.064	2.95



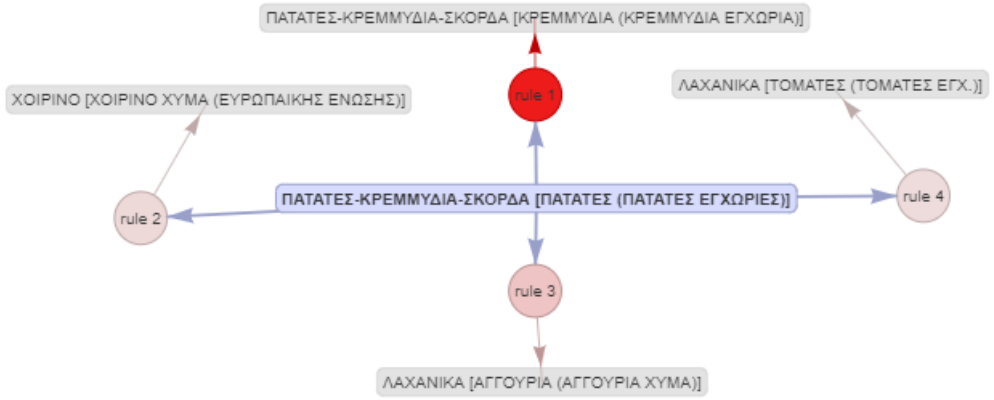
Σχήμα 16: Κανόνες συσχέτισης με δηλωμένο LHS

Παράδειγμα εξόρυξης κανόνων συσχέτισης με δηλωμένο προϊόν στην αριστερή πλευρά του κανόνα το «ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]»:

Πίνακας 25: Κανόνες συσχέτισης με δηλωμένο LHS

LHS	RHS	Measures		
		Supp	Conf	Lift
ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ	ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΚΡΕΜΜΥΔΙΑ	0.0050	0.22	5.25

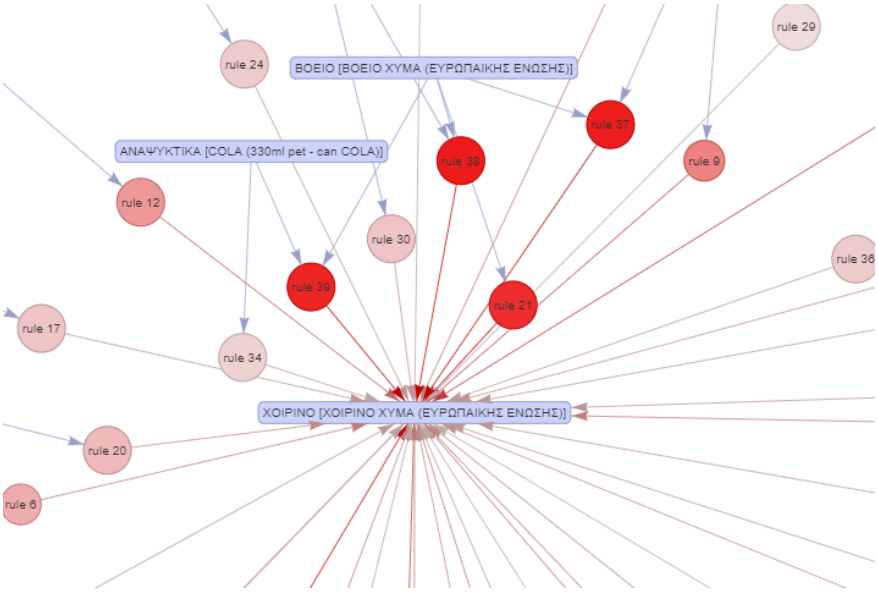
ΕΓΧΩΡΙΕΣ)]	(ΚΡΕΜΜΥΔΙΑ ΕΓΧΩΡΙΑ)]			
ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]	0.0058	0.26	2.99
ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	ΛΑΧΑΝΙΚΑ [ΑΓΓΟΥΡΙΑ (ΑΓΓΟΥΡΙΑ ΧΥΜΑ)]	0.0063	0.29	3.35
ΠΑΤΑΤΕΣ-ΚΡΕΜΜΥΔΙΑ-ΣΚΟΡΔΑ [ΠΑΤΑΤΕΣ (ΠΑΤΑΤΕΣ ΕΓΧΩΡΙΕΣ)]	ΛΑΧΑΝΙΚΑ [ΤΟΜΑΤΕΣ (ΤΟΜΑΤΕΣ ΕΓΧ.)]	0.008	0.37	2.95



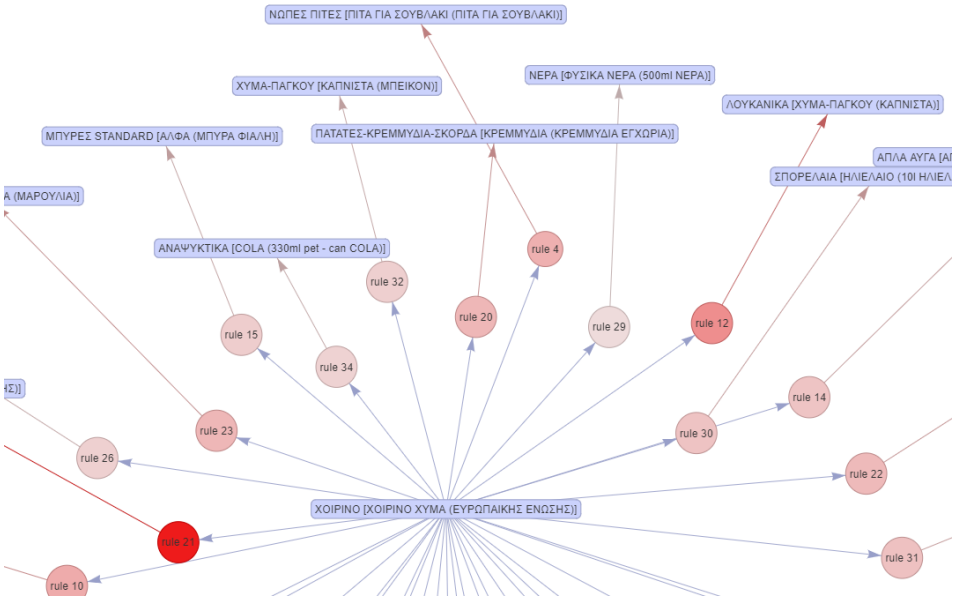
Σχήμα 17: Κανόνες συσχέτισης με δηλωμένο LHS

Παρατηρείται και στις δύο περιπτώσεις, δηλαδή το προϊόν είτε να είναι δηλωμένο στην δεξιά πλευρά του κανόνα είτε στην αριστερή, ότι πολλοί από τους κανόνες είναι αντίστροφοι με διαφορές μόνο στο ένα από τα τρία κριτήρια αξιολόγησης (Confidence).

Στο παραπάνω παράδειγμα το πλήθος των κανόνων συσχέτισης και στις δύο περιπτώσεις είναι το ίδιο, αυτό δεν σημαίνει ότι θα είναι πάντα το ίδιο με αντίστροφους κανόνες. Όπως φαίνεται και στην παρακάτω δοκιμή όπου το προϊόν «ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)» δηλώνεται σε μια από τις δύο πλευρές του κανόνα, παρατηρείται ότι το πλήθος των εξορυσσόμενων κανόνων δεν είναι το ίδιο καθώς μπορεί αυτό το προϊόν να συνδυάζεται και με άλλα προϊόντα.



Σχήμα 18: RHS – 40 κανόνες συσχέτισης χοιρινού



Σχήμα 19: LHS – 36 κανόνες συσχέτισης χοιρινού

Στην περίπτωση που το προϊόν βρίσκεται στην δεξιά πλευρά του κανόνα (RHS) οι τέσσερεις επιπλέον κανόνες προκύπτουν από συνδυασμό προϊόντων δεν παράγουν αντίστροφους όταν το προϊόν δηλωθεί στην αριστερή πλευρά (LHS).

Πίνακας 26: Κανόνας συσχέτισης από συνδυασμό προϊόντων

LHS	RHS	Measures		
		Supp	Conf	Lift
ΒΟΕΙΟ [ΒΟΕΙΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)], ΚΟΤΟΠΟΥΛΟ [ΜΕΡΗ ΚΟΤΟΠΟΥΛΟΥ ΧΥΜΑ	ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]	0.0052	0.60	6.81

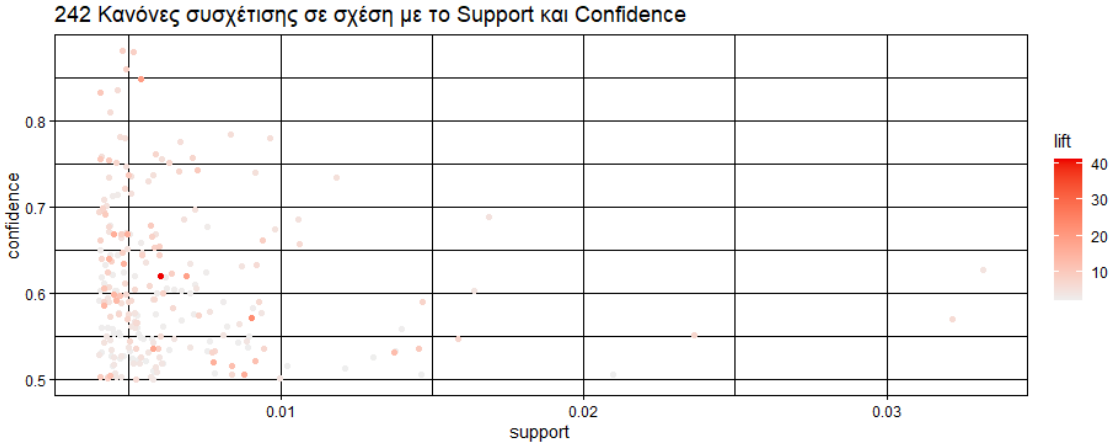
(ΕΛΛΗΝΙΚΟ)				
ΒΟΕΙΟ [ΒΟΕΙΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)], ΣΠΟΡΕΛΑΙΑ [ΗΛΙΕΛΑΙΟ (10l ΗΛΙΕΛΑΙΟ)]	ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]	0.0056	0.60	6.85
…				

Παρατηρήσεις:
Αναλύοντας τη διαδικασία εύρεσης κανόνων συσχέτισης μεταξύ προϊόντων παρατηρείται ότι οι κανόνες μπορούν να χωρισθούν σε δύο κατηγορίες, τους κανόνες ομογενοποίηση, όπου αφορούν συνδέσεις μεταξύ προϊόντων ίδιας προϊοντικής κατηγορίας και τους κανόνες συμπληρωματικότητας όπου αφορούν συνδέσεις μεταξύ προϊόντων διαφορετικών προϊοντικών κατηγοριών αλλά υψηλής συσχέτισης. Η συγκεκριμένη ανάλυση δημιουργεί τα παρακάτω ερωτήματα:

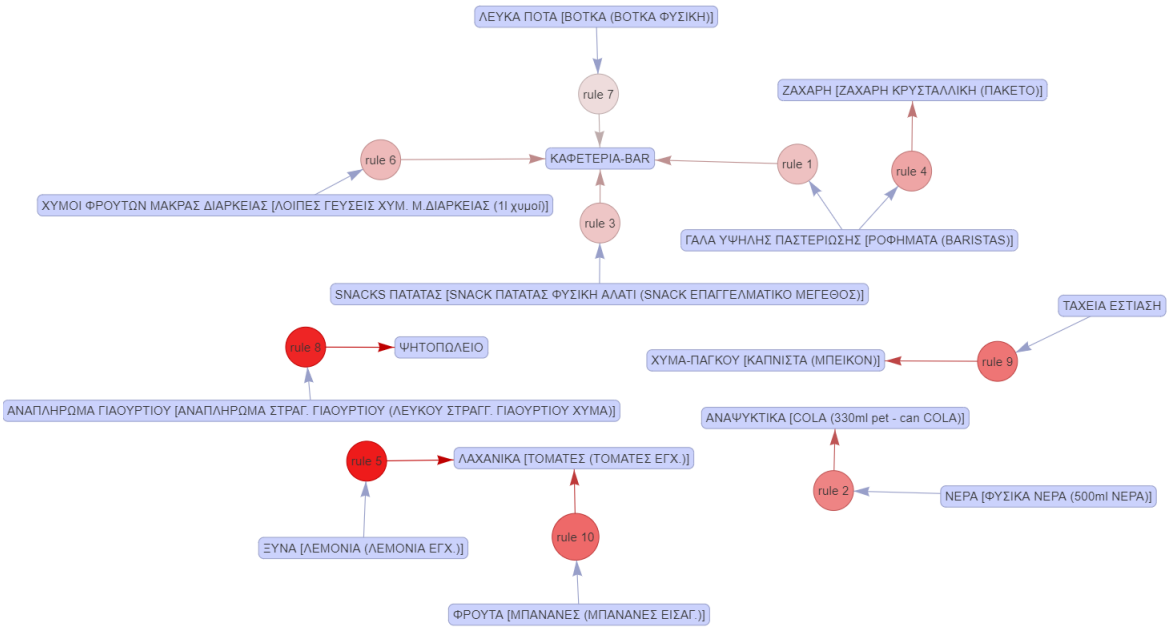
- Σε τι οφείλεται η υψηλή συσχέτιση μεταξύ προϊόντων ίδιας προϊοντικής κατηγορίας;
- Θα μπορούσε μέσω των κανόνων συμπληρωματικότητας να θεωρηθεί ότι τα προϊόντα που συμπληρώνουν το ένα το άλλο μεταποιούνται σε νέο;
- Θα μπορούσε η επιχείρηση να αξιοποιήσει τους κανόνες συμπληρωματικότητας για να δημιουργήσει συνδυασμούς προσφορών;
- Θα μπορούσε η επιχείρηση να εστιάσει σε περιόδους εποχικότητας προϊόντων (π.χ. Πάσχα, Χριστούγεννα) για να εντοπίσει κανόνες συσχέτισης συμπληρωματικών προϊόντων και να διαχειριστεί τα αποθέματα καλύτερα;

4.5.2 Εκτίμηση κανόνων συσχέτισης προϊόντων - επαγγελμάτων

Συνεχίζοντας με την εκτίμηση των κανόνων συσχέτισης προϊόντων και επαγγελμάτων, όπως απεικονίζεται και στα παρακάτω διαγράμματα πλέον οι κανόνες συσχέτισης περιέχουν πέρα από προϊόντα και επαγγελματικές ιδιότητες. Παρακάτω απεικονίζονται οι 242 κανόνες που παράχθηκαν και εμπεριέχουν τα επαγγέλματα.



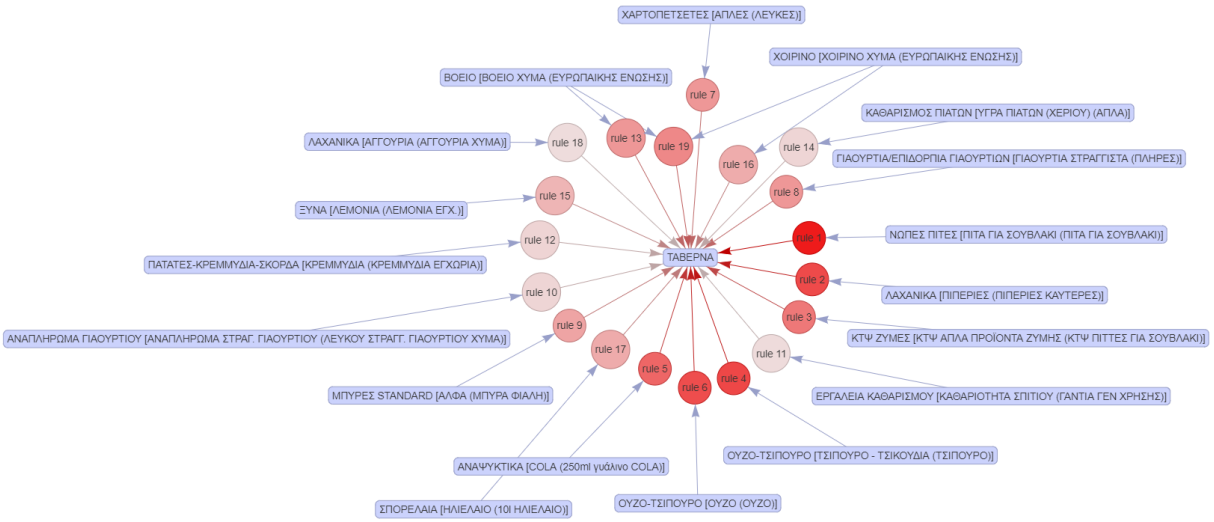
Σχήμα 20: Κανόνες συσχέτισης



Σχήμα 21: Κανόνες συσχέτισης και επαγγελμάτων

Στόχος αυτής της προσέγγισης είναι να προκύψει σύνδεση μεταξύ επαγγέλματος και προϊόντος, για να συμβεί αυτό η παραγωγή των κανόνων συσχέτισης θα πρέπει να εμπίπτει σε συγκεκριμένα πλαίσια. Δηλαδή πλέον και για το νέο σύνολο δεδομένων το ενδιαφέρον εστιάζει στους κανόνες συσχέτισης μεταξύ προϊόντων με επαγγέλματα. Γι' αυτό το λόγο παράγονται οι κανόνες συσχέτισης όπου στην δεξιά πλευρά τους ορίζεται το επάγγελμα και θα δώσουν τα προϊόντα ή τον συνδυασμό προϊόντων που το υποδηλώνουν.

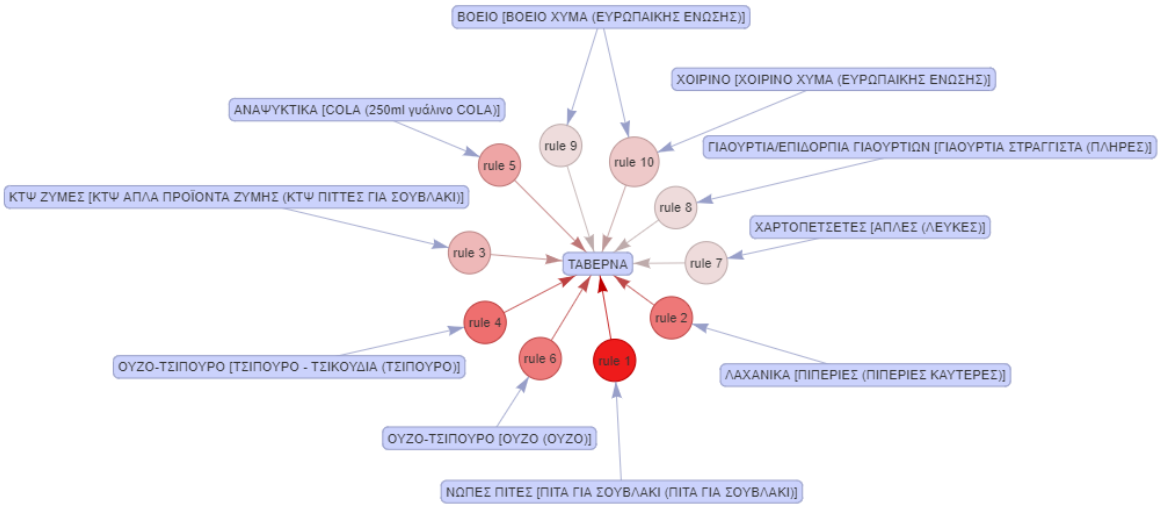
Χρησιμοποιώντας και πάλι τον αλγόριθμο apriori, και κρατώντας τις παραμέτρους που χρησιμοποιήθηκαν στην παραγωγή των κανόνων συσχέτισης προϊόντων, Support = 0. 005 και Confidence = 0.1, ορίστηκε ως δεξί μέρος του κανόνα το επάγγελμα που διερευνάται. Παρακάτω παρουσιάζεται η διερεύνηση της επαγγελματικής ιδιότητας «Ταβέρνα» και των προϊόντων ή ο συνδυασμός προϊόντων που θα μπορούσαν να την υποδηλώσουν. Παράχθηκαν συνολικά δεκαεννέα κανόνες συσχέτισης.



Σχήμα 22: Κανόνες συσχέτισης με δηλωμένο επάγγελμα στο RHS

Παίρνοντας τους πιο ισχυρούς κανόνες από τους παραπάνω (Lift > 2) δίνονται τα προϊόντα ή ο συνδυασμός προϊόντων που αγοράζονται κυρίως από πελάτες με την ιδιότητα «Ταβέρνα».

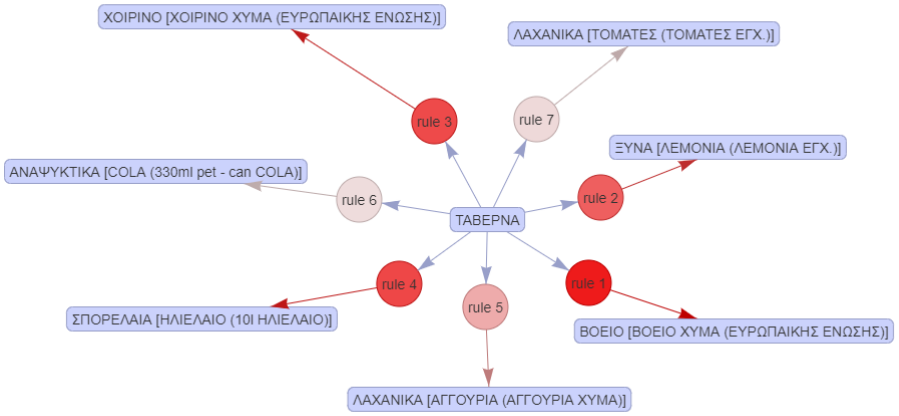
- ΝΩΠΕΣ ΠΙΤΕΣ [ΠΙΤΑ ΓΙΑ ΣΟΥΒΛΑΚΙ (ΠΙΤΑ ΓΙΑ ΣΟΥΒΛΑΚΙ)
- ΛΑΧΑΝΙΚΑ [ΠΙΠΕΡΙΕΣ (ΠΙΠΕΡΙΕΣ ΚΑΥΤΕΡΕΣ)
- ΚΤΨ ΖΥΜΕΣ [ΚΤΨ ΑΠΛΑ ΠΡΟΪΟΝΤΑ ΖΥΜΗΣ (ΚΤΨ ΠΙΤΤΕΣ ΓΙΑ ΣΟΥΒΛΑΚΙ)
- ΟΥΖΟ-ΤΣΙΠΟΥΡΟ [ΤΣΙΠΟΥΡΟ - ΤΣΙΚΟΥΔΙΑ (ΤΣΙΠΟΥΡΟ)
- ΑΝΑΨΥΚΤΙΚΑ [COLA (250ml γυάλινο COLA)
- ΟΥΖΟ-ΤΣΙΠΟΥΡΟ [ΟΥΖΟ (ΟΥΖΟ)
- ΧΑΡΤΟΠΕΤΣΕΤΕΣ [ΑΠΛΕΣ (ΛΕΥΚΕΣ)
- ΓΙΑΟΥΡΤΙΑ/ΕΠΙΔΟΡΠΙΑ ΓΙΑΟΥΡΤΙΩΝ [ΓΙΑΟΥΡΤΙΑ ΣΤΡΑΓΓΙΣΤΑ (ΠΛΗΡΕΣ)]}
- ΒΟΕΙΟ [ΒΟΕΙΟ ΧΥΜΑ (ΕΥΡΩΠΑΪΚΗΣ ΕΝΩΣΗΣ)
- ΒΟΕΙΟ [ΒΟΕΙΟ ΧΥΜΑ (ΕΥΡΩΠΑΪΚΗΣ ΕΝΩΣΗΣ)],ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΪΚΗΣ ΕΝΩΣΗΣ)



Σχήμα 23: Κανόνες συσχέτισης με μεγαλύτερο Lift και δηλωμένο επάγγελμα στο RHS

Τοποθετώντας την επαγγελματική ιδιότητα στην αριστερή πλευρά του κανόνα (LHS)

παράγονται μόνο επτά κανόνες συσχέτισης:

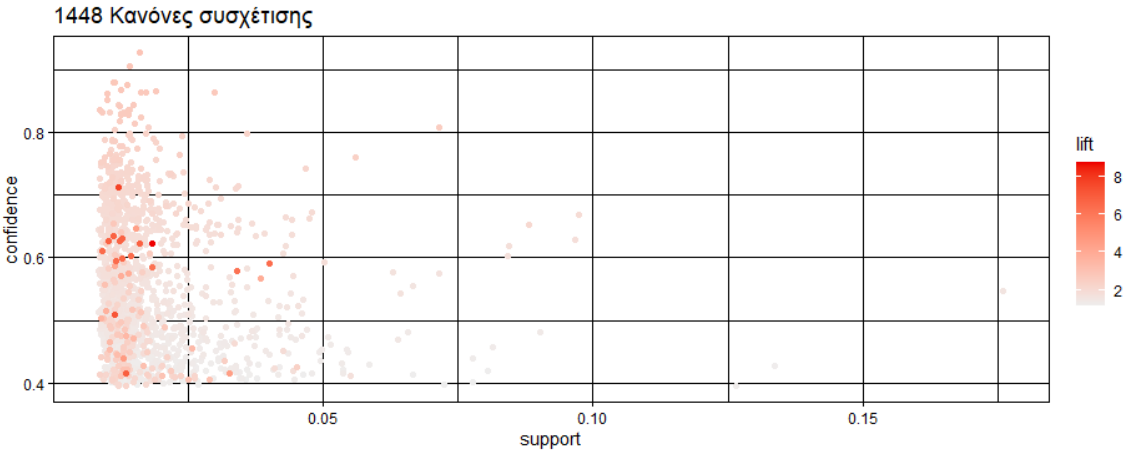


Σχήμα 24: Κανόνες συσχέτισης με δηλωμένο επάγγελμα στο LHS

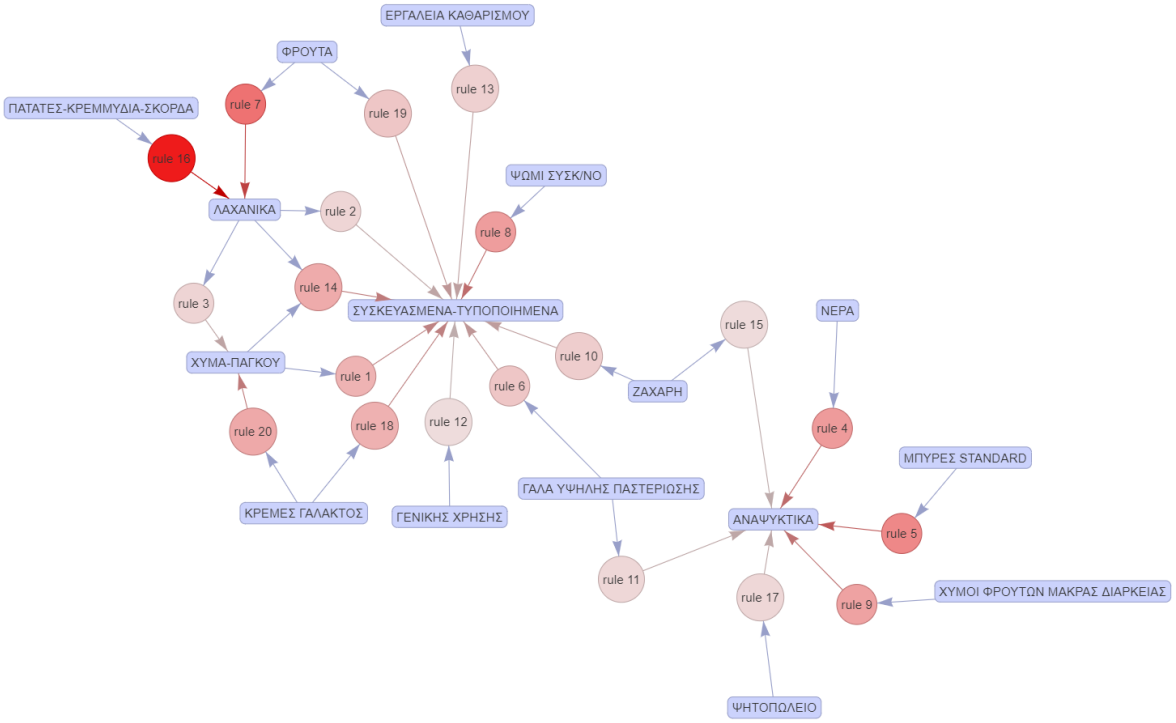
Χρησιμοποιώντας τις ίδιες παραμέτρους (Support = 0.005 και Confidence = 0.1) σε όλες τις κατηγορίες επαγγελμάτων παρατηρήθηκε ότι σε πολλές από αυτές δεν δημιουργήθηκε κανένας κανόνας. Μετά από δοκιμές χρειάστηκε να μειωθούν αρκετά τα κριτήρια επιλογής και αξιολόγησης κανόνων (Support και Confidence) ώστε να ξεκινήσουν να παράγονται κανόνες συσχέτισης για όλα τα επαγγέλματα. Αυτό είχε σαν αποτέλεσμα να υπάρχουν μεγάλες διαφορές στο πλήθος και την ισχύ των κανόνων για διαφορετικά επαγγέλματα.

Για παράδειγμα συγκρίνοντας τους κανόνες συσχέτισης που παράγονται με δεξί μέρος του κανόνα τις επαγγελματικές ιδιότητες «ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ» και «ΤΑΒΕΡΝΑ» με ίδια κριτήρια (Support = 0.002 και Confidence = 0.05), δημιουργήθηκαν για την πρώτη 5 κανόνες και για την δεύτερη 124. Βάση των κριτήριων αξιολόγησης κάποιοι κανόνες της πρώτη ιδιότητας μπορεί να είναι ισχυρότεροι σε σχέση με πολλούς της δεύτερης αλλά λόγο μεγάλης διαφοράς στο πλήθος των κανόνων αν ένας πελάτης ήταν να ταξινομηθεί σε μια ιδιότητα βάση αυτών θα ήταν πολύ πιο πιθανό να ενταχθεί στη δεύτερη.

Δημιουργία κανόνων συσχέτισης σε υψηλότερο επίπεδο προϊοντικής κατηγορίας. Αναλύοντας τους 1448 κανόνες που παράχθηκαν από το σύνολο δεδομένων προϊοντικής κατηγορίας επιπέδου δύο με Support = 0.01 και Confidence = 0.4, παρατηρείται μεγάλη συνδεσιμότητα μεταξύ των ίδιων των κατηγοριών και λιγότερο μεταξύ των κατηγοριών και των επαγγελμάτων. Επομένως οι πιο ισχυροί κανόνες συσχέτισης δεν αφορούν επαγγέλματα.

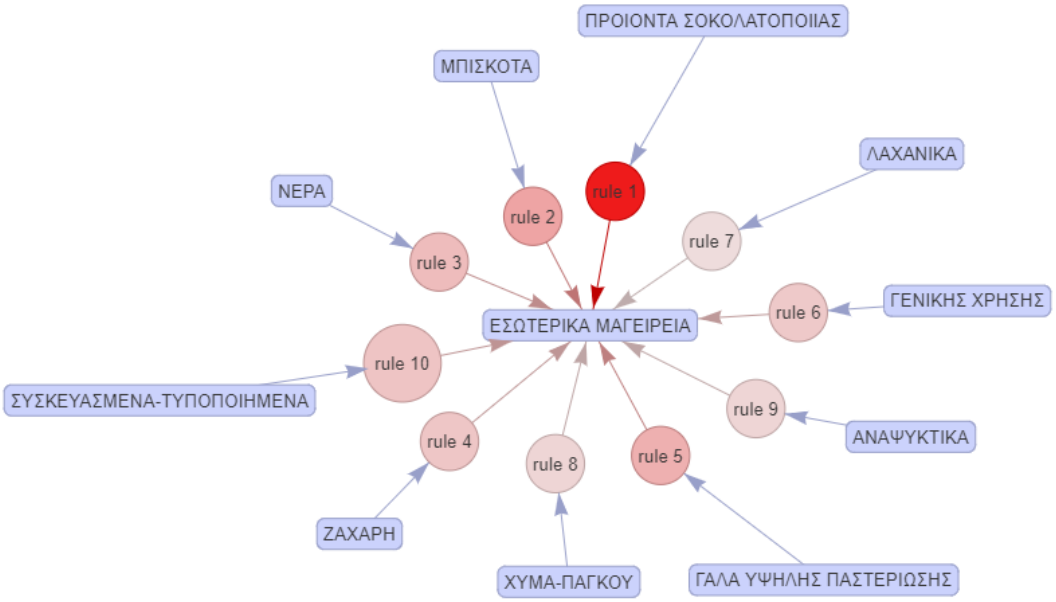


Σχήμα 25: Κανόνες συσχέτισης σε επίπεδο προϊοντικής κατηγορίας

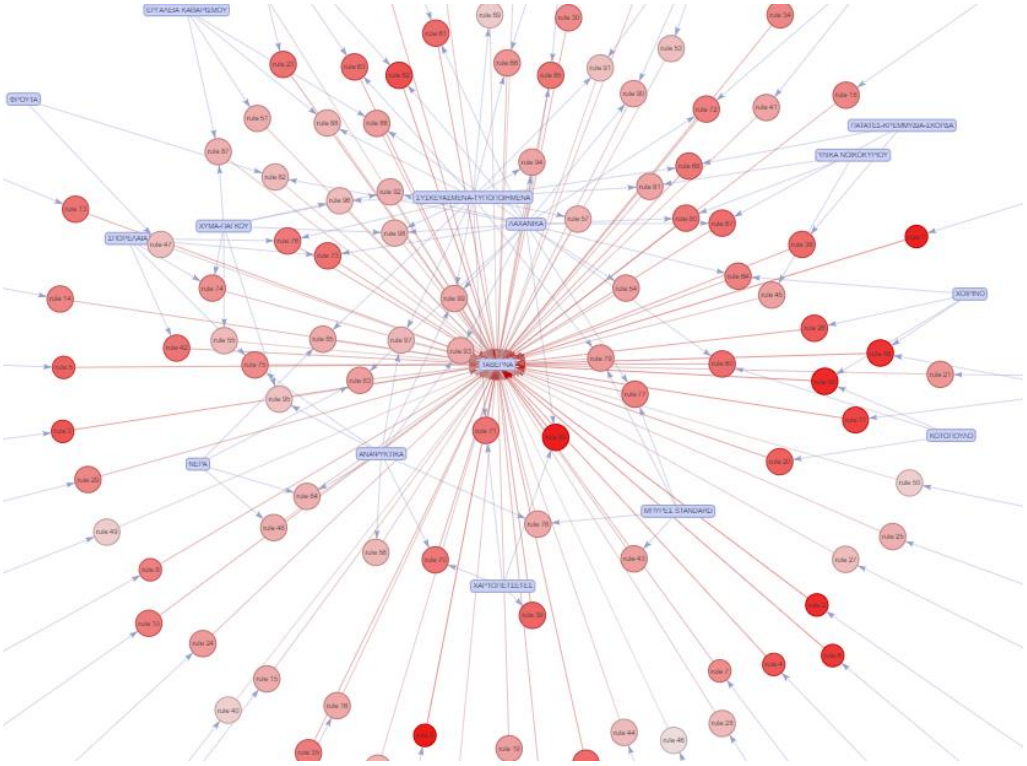


Σχήμα 26: Κανόνες συσχέτισης σε επίπεδο προϊοντικής κατηγορίας

Ορίζοντας σαν αριστερή πλευρά του κανόνα τα επαγγέλματα με Support = 0.01 και Confidence = 0.4 παράγονται πολύ λίγοι κανόνες για να μπορέσουν να αναλυθούν. Δοκιμάζοντας διάφορα support και confidence για να παραχθούν κανόνες συσχέτισης για όλα τα επαγγέλματα παρουσιάζεται μεγάλη απόκλιση στο πλήθος κανόνων ανά επάγγελμα. Αυτό συμβαίνει γιατί τα καλάθια κάποιων επαγγελμάτων είναι πολύ περισσότερα από κάποια άλλων. Για παράδειγμα για το επάγγελμα «ΤΑΒΕΡΝΑ» δημιουργήθηκαν 99 κανόνες συσχέτισης ενώ για το «ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ» 10.



Σχήμα 27: Κανόνες συσχέτισης με RHS "ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ"



Σχήμα 28: Κανόνες συσχέτισης με RHS "ΤΑΒΕΡΝΑ"

Χρησιμοποιόντας διαφορετικό Support και Confidence για τις διαφορετικές επαγγελματικές ιδιότητες παρατηρήθηκε ότι ενώ η διαφορά στο πλήθος μειώνεται υπάρχει μεγάλη απόκλιση στην ισχύ των κανόνων.

Στο παραπάνω παράδειγμα με την ιδιότητα «ΤΑΒΕΡΝΑ» και «ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ» χρησιμοποιόντας το ίδιο Support = 0.005 αλλά διαφορετικό Confidence, για την ακρίβεια χρησιμοποιήθηκε Confidence = 0.1 για το πρώτο και

Confidence = 0.01 για το δεύτερο παράχθηκαν 60 και 10 κανόνες αντίστοιχα. Παρόλο που υπάρχει ακόμα διαφορά στο πληθος το πρόβλημα στη προκειμένη περίπτωση είναι το γεγονός ότι οι κανόνες που αφορούν την ιδιότητα «ΤΑΒΕΡΝΑ» είναι πολύ ισχυρότεροι από αυτούς της «ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ». Επίσεις πολλοί από τους κανόνες συσχέτισης αφορούν τα ίδια προϊόντα όπως φαίνεται παρακάτω.

Πίνακας 27: Κανόνες συσχέτισης διαφορετικών επαγγελμάτων με όμοιο προϊόν

LHS	RHS	Measures		
		Supp	Conf	Lift
ΛΑΧΑΝΙΚΑ	ΤΑΒΕΡΝΑ	0.034	0.11	1.26
ΛΑΧΑΝΙΚΑ	ΕΣΩΤΕΡΙΚΑ ΜΑΓΕΙΡΕΙΑ	0.005	0.016	0.69

Σχήμα 29: Κανόνες συσχέτισης διαφορετικών επαγγελμάτων με όμοιο προϊόν

Παρατηρήσεις

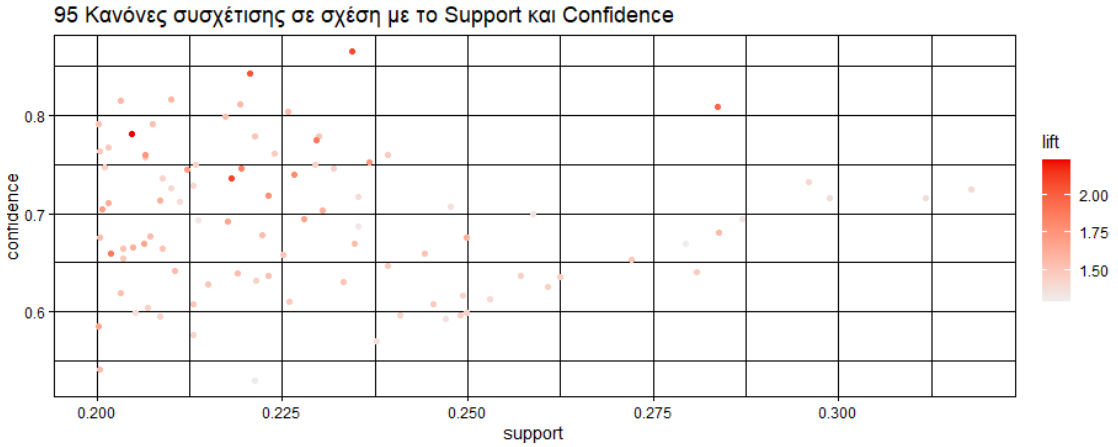
Αναλύοντας τη διαδικασία εύρεση κανόνων συσχέτισης προϊόντων με επαγγέλματα παρατηρείται ότι δημιουργούνται αρκετοί κανόνες είτε αυτοί είναι σε επίπεδο προϊόντος είτε σε επίπεδο προϊοντικής κατηγορίας. Απομονώνοντας την ανάλυση στους κανόνες που αφορούν τις επαγγελματικές ιδιότητες ώστε να εξετασθεί αν είναι εφικτή η ταξινόμηση ενός πελάτη σε επάγγελμα μέσω των αγορών του, προκύπτουν οι κανόνες συσχέτισης προϊόντων και επαγγελμάτων. Μέσω των συγκεκριμένων κανόνων εμφανίζονται προϊόντα η συνδυασμοί προϊόντων που υπονοούν ένα επάγγελμα, όμως σε αυτούς τους κανόνες παρουσιάζονται οι παρακάτω δυσκολίες.

1) Τα καλάθια απευθύνονται σε διαφορετικές επαγγελματικές ιδιότητες, επομένως είναι φυσικό να υπάρχουν διαφορές στο πλήθος των αγορών ανά ιδιότητα, αφού ορισμένες εμφανίζονται πιο συχνά από άλλες.
2) Παράγοντας κανόνες για όλες τις επαγγελματικές ιδιότητες υπάρχει μεγάλη απόκλιση στο πλήθος και την ισχύ των κανόνων ανά κατηγορία.
3) Η αγορά κάποιον προϊόντων μπορεί να υποδηλώνει πολλαπλές επαγγελματικές ιδιότητες, όμως οι συνδέσεις με μερικά επαγγέλματα είναι πιο ισχυρές από κάποιες άλλες. Επομένως υπάρχει δυσκολία στην επιλογή του επαγγέλματος που θα μπορούσε να ενταχθεί ένας πελάτη βάση των αγορών του.

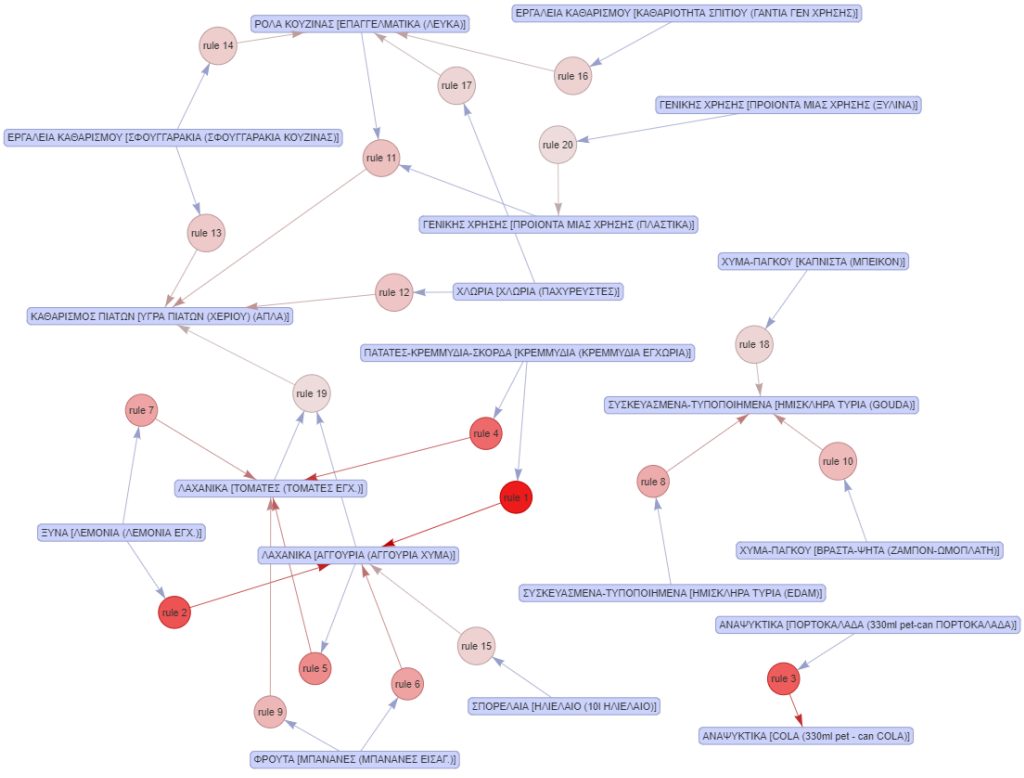
4.5.3 Εκτίμηση κανόνων συσχέτισης ομαδοποιημένων καλαθιών

Συνεχίζοντας με το τρίτο πείραμα που εκτελέσθηκε στη συγκεκριμένη ανάλυση, ομαδοποιώντας τα καλάθια ανά πελάτη και παίρνοντας τους ισχυρότερους κανόνες

συσχέτισης που παράχθηκαν από support = 0.2 και confidence = 0.5 παρατηρείται ότι υπάρχουν πολύ λίγοι κανόνες που αφορούν σύνδεση προϊόντος με επάγγελμα. Αντιθέτως υπάρχει μεγαλύτερη σύνδεση μεταξύ των προϊόντων, αυτό συμβαίνει διότι τα καλάθια είναι πολύ μεγαλύτερα από πριν και περιέχουν πολλά περισσότερα προϊόντα. Οι νέοι κανόνες συσχέτισης προκύπτουν από το σύνολο των αγορών των πελατών και όχι από το κάθε ένα καλάθι ξεχωριστά. Παρακάτω παρουσιάζονται οι 95 πιο ισχυροί κανόνες συσχέτισης:



Σχήμα 30: Κανόνες συσχέτισης στο σύνολο των αγορών



Σχήμα 31: Ισχυρότεροι κανόνες συσχέτισης στο σύνολο τω αγορών

Τα 3.557 καλάθια που δημιουργήθηκαν και περιείχαν όλα τα προϊόντα που αγοράστηκαν από τους πελάτες HO.RE.CA. δημιούργησαν όμοιους κανόνες συσχέτισης με αυτούς που εμφανίζονται στην περίπτωση των ξεχωριστών καλαθιών.

Αυτό φανερώνει ότι ορισμένοι πολύ ισχυροί κανόνες συσχέτισης παραμένουν αναλλοίωτοι και στις 3 περιπτώσεις πειραμάτων

5 Συζήτηση - Συμπεράσματα

Η μεθοδολογία CRISP-DM που ακολουθήθηκε στη παραπάνω εμπειρική μελέτη ολοκληρώνεται με το στάδιο της ανάπτυξης (Deployment), όπου η γνώση που αποκτάται πρέπει να οργανωθεί και παρουσιαστεί με τρόπο που να μπορεί να χρησιμοποιηθεί από τον πελάτη και αυτός θα αποφασίσει πως θα ενεργήσει με βάση αυτά τα δεδομένα.

5.1 Παρουσίαση αποτελεσμάτων

Παρουσιάστηκαν τα βασικά αποτελέσματα και ερωτήματα που διέγειρε η έρευνα σε ανθρώπους της εταιρείας που βρίσκονται είναι γνώστες του αντικειμένου και ανήκουν στο τμήμα πωλήσεων χονδρικής, με σκοπό την καλύτερη κατανόηση των ευρημάτων, τον εντοπισμό της επιχειρηματικής αξίας και ίσως την μελλοντική αξιοποίηση των τεχνικών ανάλυσης δεδομένων για την καλύτερη λήψη αποφάσεων.

Αναφορικά με τις συνδέσεις μεταξύ προϊόντων, τα αποτελέσματα δείχνουν ότι οι κανόνες χωρίζονται σε δύο κατηγορίες, τους κανόνες ομογενοποίησης όπου αφορούν συνδέσεις μεταξύ προϊόντων ίδιας κατηγορίας και τους κανόνες συμπληρωματικότητας όπου αφορούν συνδέσεις μεταξύ προϊόντων διαφορετικών κατηγοριών αλλά υψηλής συσχέτισης. Οι γνώστες στις πωλήσεις χονδρικής της εταιρείας υποδεικνύουν ότι:

- Οι συνδέσεις ομογενοποίησης οφείλονται στον γεγονός ότι τα δεδομένα απευθύνονται σε πελάτες χονδρικής δηλαδή επιχειρηματίες. Αυτοί οι πελάτες όταν αγοράζουν από τα καταστήματα χονδρικής προσπαθούν να δημιουργήσουν απόθεμα των προϊόντων που πωλούν. Επίσης προσπαθούν να κάνουν μεγάλες αγορές ώστε να αποφεύγουν τις συχνές μετακινήσεις από και προς τα καταστήματα. Αυτό δημιουργεί καλάθια που περιέχουν πολλά προϊόντα με κοινές προϊοντικές κατηγορίες και έτσι προκύπτουν οι κανόνες ομογενοποίησης.
- Οι συνδέσεις συμπληρωματικότητας, είναι πιθανότερο να οφείλονται σε επιχειρηματική δραστηριότητα εστιατορίων και υπογραμμίζουν ότι τα προϊόντα που αφορούν αυτές τις συνδέσεις είναι πιθανότερο να μεταποιούνται σε ένα νέο προϊόν, κυρίως τα αλλοιώσιμα (αλλαντικά, τυριά, κρέατα) ενώ σε αντίθετη περίπτωση συνδέσεις με προϊόντα όπως αναψυκτικά, αλκοόλ κ.α. είναι λιγότερο πιθανό να υποδεικνύουν μεταποίηση σε νέο προϊόν.

Αναφορικά με τις συνδέσεις μεταξύ προϊόντων και επαγγελμάτων, παράγονται αρκετοί κανόνες συσχέτισης μεταξύ προϊόντων ή συνδυασμών προϊόντων που μπορούν να υπονοήσουν ένα επάγγελμα. Έγινε μια παρουσίαση των αποτελεσμάτων και προβληματισμών στους ειδικούς, ανέφεραν χαρακτηριστικά ότι η εταιρεία αυτή τη στιγμή δεν έχει μια ομαδοποιημένη κατηγοριοποίηση, τα επαγγέλματα που έχει καταχωρημένα είναι αυτά που δηλώνει ο πελάτης στο εμπορικό επιμελητήριο και θα ήταν αρκετά χρήσιμο να υπάρχει μια πιο γενικευμένη κατηγοριοποίηση.

5.2 Περιορισμοί Έρευνας

Επισημάνθηκε στους ειδικούς ότι τα διαθέσιμα δεδομένα αφορούσαν μόνο τους πελάτες μεταποίησης και το ημερολογιακό εύρος των τεσσάρων τελευταίων ετών. Οι ίδιοι ανέφεραν ότι μπορεί οι πωλήσεις να έχουν διαφοροποιήσεις σε σχέση με την κανονική τους ροή λόγο των δύο ετών πανδημίας που εμπεριέχουν, ενώ επισήμαναν ότι υπήρξε μεγάλη αύξηση των πωλήσεων στα προϊόντα που εμφανίζοντας στο e-shop και μείωση σε αυτά που υπήρχαν μόνο στο κατάστημα. Θα πρέπει να ληφθεί υπόψη ότι μια τέτοια ανάλυση μπορεί να δώσει πολύ διαφορετικά αποτελέσματα ανάλογα την φύση , το μέγεθος και το πεδίο εφαρμογής της εκάστοτε επιχείρησης.

Η κανόνες συσχέτισης που παράγονται από την διαδικασία εξόρυξης γνώσης παρόλο που θα είναι χρήσιμοι για την λήψη αποφάσεων, δεν παρέχουν άμεση και στοχευμένη στις ανάγκες των ειδικών πληροφορία. Χρειάζεται να εκτελεσθούν διαδικασίες που είναι αρκετά χρονοβόρες και επομένως έχουν κόστος για την επιχείρηση.

Η συγκεκριμένη έρευνα φανέρωσε ότι υπάρχει μεγάλη διαφορά στο πλήθος των καλαθιών μεταξύ των επαγγελμάτων, κάτι που επιβεβαιώθηκε και από την επικοινωνία με το τμήμα των πωλήσεων χονδρικής. Παρόλο που είναι δύσκολο βάση της συγκεκριμένης προσέγγισης (μέσα από τα καλάθια αγορών τους) να ταξινομηθεί ένας πελάτης σε επάγγελμα, μια τέτοιου είδους ανάλυση μπορεί να δώσει τα προϊόντα που συνδέονται περισσότερο με ένα επάγγελμα.

5.3 Συμπεράσματα

Οι ειδικοί στον τομέα των πωλήσεων αναφέρουν την αξία που μπορεί να προσφέρει μια τέτοια έρευνα στη συγκεκριμένη επιχείρηση και το πως μπορούν οι κανόνες συσχέτισης μεταξύ προϊόντων να βοηθήσουν στην λήψη αποφάσεων.

- Αυτή τη στιγμή οι γενικές προσφορές και αυτές συνδυασμών προϊόντων (COMBO) εξάγονται εμπειρικά, μια αλγοριθμικής προσέγγισης θα μπορούσε να βοηθήσει σε αυτή τη διαδικασία καθώς και να δώσει περιπτώσεις που ο άνθρωπος δεν μπορεί να εντοπίσει. Βεβαίως την τελική εκτίμηση θα την κάνει ο ανθρώπινος παράγοντας για το αν μια προσφορά αξίζει να εκδοθεί.
- Προσπαθούν να δημιουργήσουν προσφορές όπου με την αγορά ενός προϊόντος ο πελάτης θα αποκτά έκπτωση σε μια πληθώρα άλλων προϊόντων, αυτά τα εκπτωτικά προϊόντα θα μπορούσε να επιλεχθούν με την βοήθεια των κανόνων συσχέτισης.
- Σε περιόδους εποχικότητα ο εντοπισμός συνδυασμού προϊόντων θα μπορούσε πέρα από τη δημιουργία προσφορών να αξιοποιηθεί στα αποθέματα και την παραγγελιοδοσία των καταστημάτων.
- Η εταιρεία διαθέτει πελάτες συμφωνίας, οι οποίοι αγοράζουν συμφωνηθέντα προϊόντα σε φθηνότερη τιμή. Κατά τη διαδικασία της συμφωνίας ο πωλητής της επιχείρησης προτείνει μέσα από μια εμπειρική λίστα προϊόντων, βάση επαγγέλματος του πελάτη που προσεγγίζει. Θα μπορούσαν οι κανόνες συσχέτισης να εντοπίσουν συνδυασμούς προϊόντων για τα συγκεκριμένα επαγγέλματα και να δημιουργήσουν μια λίστα που δεν θα βασίζεται αποκλειστικά στην εμπειρία αλλά και στα δεδομένα.

Όπως ανέφεραν οι ειδικοί, οι κανόνες συσχέτισης θα αξιοποιηθούν καλύτερα αν

εκτελεσθούν στοχευμένα σε συγκεκριμένες επαγγελματικές ιδιότητες. Με αυτό τον τρόπο θα εξαχθούν συνδυασμοί που αφορούν αποκλειστικούς πελάτες και επαγγελματικές ιδιότητες. Το ιδανικό για την επιχείρηση θα ήταν η ανάπτυξη μιας εφαρμογής που να μπορεί ο ίδιος ο χρήστης να ορίζει την περίοδο, επάγγελμα ή προϊόν και το λογισμικό να παράγει τους ισχυρότερους κανόνες συσχέτισής βάσει αυτών των κριτηρίων.

Μια έρευνα στις συνδέσεις μεταξύ προϊόντων και επαγγελμάτων θα μπορούσε:

- Να βοηθήσει στον εντοπισμό κωδικών κλειδιά που η εταιρεία μέχρι στιγμής χαρακτηρίζει εμπειρικά σαν κωδικούς που ανήκουν σε ένα επάγγελμα.
- Θα μπορούσε να βοηθήσει σε έρευνες όπου το τμήμα πωλήσεων ελέγχει αν επιλεγμένα προϊόντα αγοράζονται από τα αναμενόμενα επαγγέλματα.
- Σε περίπτωση που τα επιλεγμένα προϊόντα δεν δίνουν συνδέσεις με ένα αναμενόμενο επάγγελμα ή οι συνδέσεις είναι πολύ αδύναμες τότε μπορεί το τμήμα πωλήσεων να προβεί σε περεταίρω έρευνα ως προς τους λόγους που μπορεί να συμβαίνει αυτό.

Όπως και στις συνδέσεις μεταξύ προϊόντων οι expert δίνουν μεγαλύτερη αξία στην στοχευμένη ανάλυση των επαγγελματικών ιδιοτήτων. Οι ειδικοί ανέφεραν ότι ήταν πιο χρήσιμο να βλέπουν τις αγορές των επαγγελμάτων σαν ποσοστά, π.χ. η ιδιότητα «ΚΑΦΕ - BAR» αγοράζει από την προϊοντική κατηγορία «ΚΑΦΕ» κατά 51% στο σύνολο των αγορών της. Επομένως θεωρούν ότι μια τέτοια έρευνα θα μπορούσε να είναι η αρχή για την δημιουργία νέων εφαρμογών που θα διαχειρίζονται οι ίδιοι οι χρήστες με τα κριτήρια που επιθυμούν.

Τα αποτελέσματα της έρευνας αποδεικνύουν ότι η ανάλυση αγοραστικού καλαθιού μπορεί να αξιοποιηθεί και να βοηθήσει στην λήψη αποφάσεων, όμως οι ειδικοί αναφέρουν ότι τα αποτελέσματα θα ήταν πιο αξιοποιήσιμα αν μπορούσαν οι ίδιοι να τα παράγουν, ώστε να εισάγουν τα κριτήρια που επιθυμούν και να εμφανίζουν τις συνδέσεις είτε μεταξύ προϊόντων είτε μεταξύ προϊόντων και επαγγελμάτων που αυτοί έχουν ανάγκη. Αυτό θα μπορούσε να επιτευχθεί μόνο με την δημιουργία εφαρμογών που θα χειριζόταν ο οποιοσδήποτε χρήστης χωρίς εξειδικευμένες γνώσεις εξόρυξης δεδομένων. Με αυτό τον τρόπο η ανάλυση αγοραστικού καλαθιού θα μπορούσε να δώσει μεγαλύτερη αξία στην επιχείρηση και να βοηθήσει στη βελτίωση της ποιότητας των προσφορών, των υπηρεσιών και την εμπειρία πελάτη.

Τέλος, βάση της συζήτησης πάνω στα αποτελέσματα αλλά και της βιβλιογραφίας φαίνεται ότι κοινωνικές πτυχές όπως οι ανθρώπινοι πόροι, οι δυνατότητες διαχείρισης και η οργανωτική κουλτούρα παίζουν σημαντικό ρόλο στη διαδικασία δημιουργίας επιχειρηματικής αξίας, ενώ τεχνολογικοί παράγοντες, όπως τα τεχνικά περιουσιακά στοιχεία και η ποιότητα των δεδομένων, είναι λιγότερο σημαντικοί. Πέρα από τα πλαίσια της έρευνας αυτό που θα φέρει τα τελικά αποτελέσματα και αποφάσεις είναι ο ανθρώπινος παράγοντας αξιολογώντας την φύση και στρατηγική της εταιρείας. (Trieu V.-H. , 2016)

Βιβλιογραφία

Abirami, M., & Pattabiraman, V. (2016). Data Mining Approach for Intelligent Customer Behavior Analysis for a Retail Store. *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges*, (σσ. 283-291).

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, (σσ. 207-2016).

Aguinis,, H., Forcum, L., & Joo, H. (2012, Νοέμβριος 30). Using Market Basket Analysis in Management Research. *Journal of Management 2013*, σσ. 1799-1821.

Alton, L. (2017, Δεκέμβριος 22). *Data Science Central*. Ανάκτηση από Data Science Central: https://www.datasciencecentral.com/the-7-most-important-data-mining-techniques/

Baby, N., & Priyanka, L. (2012, Δεκέμβριος 12). Customer Classification and Prediction Based On Data Mining Technique. *International Journal of Emerging Technology and Advanced Engineering*, σσ. 314-318.

Belka, A. (2022, Απρίλιος 29). *Boldare*. Ανάκτηση από Digital product design company Boldare: https://www.boldare.com/blog/5-examples-of-digital-transformation/#what-is-digital-transformation?-examples-of-digital-transformation-are:

Bell, D., & Lattin, J. (1998). Shopping Behavior and Consumer Preference for Store Price Format: Why "Large Basket" Shoppers Prefer EDLP. *Marketing Science, 1*(17), σσ. 66-88. Ανάκτηση από http://dx.doi.org/10.1287/mksc.17.1.66

Bharadwaj, A., El Sawy, O., Pavlou, P., & Venkatraman, N. (2013, Ιούνιος). Digital Business Strategy: Toward a Next Generation of Insights. *MIS Quarterly*, σσ. 471-482.

Chen, G., Liu, H., Yu, L., Wei, Q., & Zhang, X. (2006). A new approach to classification based on association rule mining. *Decision Support Systems*, σσ. 674-689.

Christodoulakis, D. (2005). *Customer Clustering using RFM analysis.* Πάτρα: Computer Engineering and Informatics Department University of Patras.

Davenport, T. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities.* Harvard Business Press.

Davenport, T., & Dyché, J. (2013, Μάιος). International Institute for Analyticts. *Big Data in Big Companies*, σσ. 1-31.

Fang, Y., Xia, X., Wang, X., & Lan, H. (2018). Customized Bundle Recommendation by Association Rules of Product Categories for Online Supermarkets. *IEEE Third International Conference on Data Science in Cyberspace*, (σσ. 472-475).

Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceedings of the second international conference on knowledge discovery and data mining.*

Grover, V., Chiang, R., Liang, T.-P., & Zhang, D. (2018, Μαϊου 15). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Management Information Systems*, σσ. 388-423.

Gunther, W. A., Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, 191-209.

Hahsler, M., & Grun, B. (2005, Οκτώμβριος). arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*(15).

Hahsler, M., Johnson, I., & Giallanza, T. (2022, Μαίος 30). *arulesCBA: Classification Based on Association Rule.* Ανάκτηση από https://github.com/mhahsler/arulesCBA

Hao, Z., Wang, X., Yao, L., & Zhang, Y. (2009). Improved Classification Based on Predictive Association Rules. *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*, (σσ. 1165-1170). San Antonio, TX, USA.

Hussein, N., Alashqur, A., & Sowan, B. (2015). Using the interestingness measure lift to generate association rules. *Journal of Advanced Computer Science & Technology*, σσ. 156-162.

Jabbeen, H. (2018, Αύγουστος 21). *DataCamp*. Ανάκτηση από Τοποθεσία Web της DataCamp: https://www.datacamp.com/tutorial/market-basket-analysis-r

Koh, Y., & Pears, R. (2008). Association Rule Mining via Transaction Clustering. *Proceeding of the 2008 Australasian Data Mining Conference*, (σσ. 87-94).

Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering,*, σσ. 71-82.

Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. *EEE International Conference on Data Mining (ICDM)*, (σσ. 369-376).

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert systems with applications*, σσ. 11303-11311.

Liu, B., Hsu, W., & Ma, Y. (1998). *Integrating Classification and Association Rule Mining.* Singapore: Department of Information Systems and Computer Science National University of Singapore.

Madhulatha, S. (2012, Απρίλιος). An Overview on Clustering Methods. *IOSR Journal of Engineering*, σσ. 719-725.

Mostafa, M. (2015). Knowledge discovery of hidden consumer purchase behaviour: a market basket analysis. *Knowledge discovery of hidden consumer purchase, 7*(4), σσ. 384-405.

Mu-Chen, C., & Chia-Ping, L. (2007). A data mining approach to product assortment and shelf space allocation. *Expert Systems with Applications*, σσ. 976-986.

Nanda, K., & Ram, R. (2006). Research Note—Using Basket Composition Data for Intelligent Supermarket Pricing. *Marketing*(25), σσ. 188-199.

Niu, Q., Xia, S.-X., & Zhang, L. (2009). Association Classification based on Compactness of Rules. *Second International Workshop on Knowledge Discovery and Data Mining (WKDD)*, (σσ. 245-247).

Oesterreich, T. D., Anton, E., & Teuteberg, F. (2022). What translates big data into bussiness value? A meta-analysis of the impact of business analytics on firm performance. *Information & Management*, 1-47.

Pandey, G., Chawla, S., Poon, S., Arunasalam, B., & Davis, J. (2009, Ιανουάριος 22). Association Rules Network: Definition and Applications. Στο *Statistical Analysis and Data Mining: The ASA Data Science Journal* (σσ. 2060-279). Wiley. Ανάκτηση από Published in Wiley InterScience: www.interscience.wiley.com

Park, Y.-J., & Chang, K. (2009). Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications*, σσ. 1932-1939.

Phyi, T. (2009). Survey of Classification Techniques in Data. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I.* Hong Kong.

Reis, J., Amorim, M., Melao, N., & Matos, P. (2018, Μάριος). Digital Transformation: A Literature Review and Guidelines for Future Research. *Trends and Advances in Information Systems and Technologies*, σσ. 411-421.

Russel, G., & Petrsen, A. (2000). Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing,, 76*, σσ. 367-392.

Sethi, S., Malhotra, D., & Verma, N. (2016, Απρίλιος). Data Mining: Current Applications & Trends. *International Journal of Innovations in Engineering and Technology (IJIET)*, σσ. 667-673.

Thabtah, F., Cowling, P., & Peng, Y. (2005). MCAR: Multi-class Classification based on Association Rule. *The 3rd ACS/IEEE International Conference onComputer Systems and Applications.*

Trieu, V.-H. (2016, 10 11). Getting value from Business Intelligence systems: A review and. *Decision Support Systems*, σσ. 111-124.

Trieu, V.-H. (2017, January). Getting value from Business Intelligence systems: A review and research agenda. *Decision Support Systems*, σσ. 111-124.

Wang, L., & Sun, J. (2019). Market Basket Analysis based on Apriori and CART. *5th International Conference on Education Technology, Management and Humanities Science (ETMHS 2019)*, (σσ. 1456 - 1461).

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*

Ya-Han, H., & Tzu-Wei, Y. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based Systems*, σσ. 76-88.

Παράρτημα

Παράρτημα 1: Κώδικας στην γλώσσα R

```
install.packages("arules")
install.packages("arulesViz")
install.packages("RODBC")
install.packages("plyr")
install.packages("dplyr")
install.packages("stringr")

library(arules)
library(arulesViz)
library(RODBC)
library(plyr)
library(dplyr)
library(stringr)

#Create a connection to the database
dbhandle <-odbcDriverConnect ('driver = {SQLServer}; server = ********; database=*****;
trusted_connection = true')

#Get the data from the database into a data frame
bask_list <- sqlQuery(dbhandle, 'SELECT
                      BASK_ID,
                      Replace(TRIM(BASK_TREEDESCR2) + \' [\' +
TRIM(BASK_TREEDESCR3) + \' (\'+ TRIM(BASK_TREEDESCR4) + \')] \', \',\', \'.\')
                    FROM BASK_ITEMS')
names(bask_list) <- c("TransId","ItemDescr")
close(dbhandle)

#Add names to th data frame columns
names(bask_list) <- c("TransId","ItemDescr")

#Turn data frame to the basket format grouping products via TransId into baskets
baskets <- ddply(bask_list,c("TransId"),
                 function(df1)paste(df1$ItemDescr,
                                    collapse = ","))

#Add names to basket data frame and remove the column with the basket id's we will create
new basket id's
names(baskets) <- c("TransId","Basket")
baskets$TransId <- NULL

#Extract the basket into a csv (giving it a row number as basket id)
write.csv(baskets,"baskets.csv", quote = FALSE, row.names = TRUE)

#Create the transaction object via the csv file
basketTrans = read.transactions ( file="baskets.csv", rm.duplicates= TRUE, format =
"basket", sep = "," , cols = 1)
basketTrans@itemInfo$labels <- gsub("\"","",basketTrans@itemInfo$labels)

#Create the association rules, passing parameters Support and Confidence
basket_rules <- apriori(basketTrans,parameter = list(sup = 0.008, conf = 0.5, target="rules"))
```

```
inspect(basket_rules[1:50])
summary(basket_rules)
length(basket_rules)

#Remove redundant rules
#Remove inverted (reverse/duplicate) rules
basket_rules <- basket_rules[!is.redundant(basket_rules)]
gi <- generatingItemsets(basket_rules)
d <- which(duplicated(gi))
basket_rules <- basket_rules[-d]

#Use lift as another form of measure, get the rules the have Lift over 2 and bellow 3
subRules <- subset(basket_rules, subset = lift < 3)
subRules <- subset(subRules, subset = lift > 2)

#Sort the subrules by Lift
subRules <- sort(subRules, decreasing=TRUE, by=c("lift"))

# Filter rules with confidence greater than 0.4 or 40%
subRules <- basket_rules[quality(basket_rules)$confidence>0.4]

#Plot SubRules
dev.off()
plot(subRules)
plot(subRules, method = "graph",  engine = "htmlwidget")

#Get the rules with the given support and confidence with the specified rhs
rhsProduct.rules <- apriori( basketTrans, parameter = list( sup = 0.005, conf = 0.03, minlen =
2, target="rules"), appearance = list( default="lhs",rhs="ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ
(ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]"))
inspect(rhsProduct.rules)

#Get the rules with the given support and confidence with the specified lhs
lhsProduct.rules <-
  apriori(basketTrans, parameter = list(sup = 0.005, conf = 0.03, minlen = 2, target="rules"),
        appearance = list(lhs="ΧΟΙΡΙΝΟ [ΧΟΙΡΙΝΟ ΧΥΜΑ (ΕΥΡΩΠΑΙΚΗΣ ΕΝΩΣΗΣ)]",
default="rhs"))
inspect(lhsProduct.rules)
#Create 169 different combinations of support and confidence
#Put each one into a new variable to check the amount of rules in each compination
sup = list(0.001,0.002,0.003,0.004,0.005,0.006,0.007,0.008,0.009,0.01,0.02,0.03,0.04,0.05)
conf = list(0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.2,0.3,0.4,0.5)
for(i in sup) {
  for (j in conf) {
    bask = str_replace_all( toString (cbind ("Bask_Sup_", toString(i), "_Conf_", toString(j))),
",", "")
    assign ( bask, apriori( basketTrans,parameter = list(sup = i, conf = j, target="rules", minlen
= 2)))
  }
}

# extract the global environment variables for inspection
glblst = as.list(.GlobalEnv)
capture.output(summary(glblst), file = "glblst.txt")
```