ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Ταξινόμηση Συστάσεων ARM κατά OMST

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΟΛΓΑ ΛΟΪΑ

Επιβλέπων : Δημήτριος Δέρβος
Καθηγητής, ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Φεβρουάριος 2022

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Ταξινόμηση Συστάσεων ARM κατά OMST

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΟΛΓΑ ΛΟΪΑ

Επιβλέπων : Δημήτριος Δέρβος
Καθηγητής ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις Choose a date.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο	Όνομα Επώνυμο	'Ονομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Choose a date

(Υπογραφή)

.....

Click here to enter text.

Click here to enter text.

Περίληψη

Το διαδίκτυο άλλαξε ριζικά τα μέσα και την ταχύτητα μετάδοσης πληροφοριών και έχει δημιουργήσει νέες προοπτικές σε διάφορους τομείς. Μια από αυτές είναι και η ανάγκη παραγωγής προβλέψεων-συστάσεων τόσο από την πλευρά του χρήστη όσο και από την πλευρά των επιχειρήσεων. Από την πλευρά του ο χρήστης στην πλατφόρμα που χρησιμοποιεί κάνει χρήση της επιλογής «προτεινόμενα» που θα βρει διαθέσιμη για να βοηθηθεί στον τεράστιο όγκο δεδομένων που μπορεί να του παρέχονται. Από την άλλη η επιχείρηση από την πλευρά της είτε για στατιστικούς λόγους είτε για λόγους προώθησης είναι σημαντική η υπηρεσία αυτή.

Για την παραγωγή συστάσεων – προβλέψεων χρησιμοποιείται πληθώρα τεχνικών. Στην παρούσα διπλωματική εργασία έχοντας ως θεωρητικό υπόβαθρο την Εξόρυξη Δεδομένων και συγκεκριμένα την παραγωγή συνδυαστικών κανόνων, η παραγωγή συνδυαστικών κανόνων γίνεται με τη βοήθεια του αλγορίθμου Apriori. Εφόσον ολοκληρωθεί η εξαγωγή των κανόνων τότε προκύπτουν από αυτούς οι συστάσεις – προβλέψεις.

Στόχος της διπλωματικής εργασίας συνιστά η διαμόρφωση, η αξιολόγηση στην πράξη και η τεκμηριωμένη διατύπωση μίας νέας πρότασης για ένα νέα πλαίσιο ιεραρχικής ταξινόμησης των συστάσεων που προκύπτουν στην έξοδο ενός Συστήματος Παραγωγής Συστάσεων (Recommender System, RS) που βασίζεται στην Εξόρυξη Συνδυαστικών Κανόνων (Association Rule Mining, ARM). Κι ενώ ο κλασικός τρόπος παραγωγής συστάσεων θέλει να ταξινομούνται οι κανόνες με βάση το μέτρο lift πρώτα και το μέτρου του confidence μετά, το μοντέλο της νέας πρότασης βασίζεται στην ταξινόμηση των συστάσεων στην εφαρμογή σχήματος με το οποίο καταρτίζεται ο πίνακας κατάταξης των χωρών βάσει των μεταλίων που κατακτούν οι αθλητές τους στους Ολυμπιακούς Αγώνες (Olympic Medal Standings Table, OMST).

Το νέο πλαίσιο ιεράρχησης συστάσεων ARM πρόκειται να υλοποιηθεί σε περιβάλλον R/RStudio και να αξιολογηθεί σε σύγκριση με παραλλαγές της κλασικής προσέγγισης όπου οι συστάσεις ταξινομούνται με τη χρήση των αριθμητικών τιμών των μέτρων αξίας των συνδυαστικών κανόνων: lift, conviction, confidence, και support.

Η εργασία περιλαμβάνει και το στάδιο της αξιολόγησης και σύγκρισης των αποτελεσμάτων που προκύπτουν σε σύγκριση με την κλασική μέθοδο αλλά και από δυο άλλες, αυτές που η ταξινόμηση των κανόνων βασίζεται πρώτα στο lift και μετά στο support αλλά και σε αυτή που ταξινομούνται πρώτα ως προς το conviction και μετά ως προς το support. Για την αξιολόγηση γίνεται χρήση νέων μέτρων αξίας των κανόνων. Σημαντικό είναι το γεγονός ότι ο κώδικας μπορεί να τροφοδοτηθεί στην είσοδο με οποιουδήποτε δεδομένα- συναλλαγές, γεγονός που τον καθιστά ευέλικτο και εξελίξιμο.

Λέξεις Κλειδιά: Συνδυαστικοί Κανόνες, Μέθοδος Μεταλλίων, Παραγωγή Συστάσεων, Αλγόριθμος Apriori,, Ανάλυση Δεδομένων

Abstract

The internet has radically changed the means and speed of information transmission and has created new perspectives in various fields. One of them is the need to produce forecasts-recommendations at both the user and the business side. For his part, the user makes use of the "suggested" option that he will find available to help with the huge amount of data that can be provided to him. The service is important at the business side for statistical and sales promotion reasons.

A variety of techniques are used to produce recommendations - forecasts. In the present dissertation having as theoretical background the Data Mining and specifically the production of combinatorial rules, the production of combinational rules is done with the help of the Apriori algorithm. Once the extraction of the rules is completed then the recommendations - predictions emerge from them.

The aim of the dissertation is to formulate, evaluate in practice and document a proposal for a new framework for the hierarchical classification of recommendations resulting from the output of a Recommender System (RS) based on Association Rules Mining, ARM). And while the classic way of making recommendations is to classify the rules based on the lift measure first and the confidence measure afterwards, the model of the new proposal is based on classifying the recommendations in the form application that compiles the country ranking table based on variables won by their athletes at the Olympic Games (Olympic Medal Standings Table, OMST)

The new ARM recommendation hierarchy is to be implemented in an R / RStudio environment and evaluated in comparison with variants of the classical approach where recommendations are classified using the numerical values of the value measures of the combined rules: lift, conviction, confidence, and support.

The proposed model includes the stage of evaluating and comparing the results obtained in comparison with the classical method but also of two others, those in which the classification of the rules is based first on the lift and then on the support but also on that which are classified first in terms of conviction and then in terms of support. New measures are used for evaluation. Important is the fact that the code can be fed to the input with any data-transactions, which makes it flexible and extendable.

Keywords: association rules, medal standings, recommendations, data mining, Apriori, data analytics

Ευχαριστίες

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω τους ανθρώπους που διαδραμάτισαν καθοριστικό ρόλο στην εκπόνηση της παρούσα διπλωματικής εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, τον κύριο Δέρβο Δημήτριο για την εμπιστοσύνη που μου έδειξε με την ανάθεση της εργασίας, τις στοχευμένες παρατηρήσεις του και την άψογη συνεργασία σε όλη την διάρκεια της διπλωματικής μου εργασίας. Η καθοδήγησή του, η άμεση ανταπόκρισή του σε οποιαδήποτε ανάγκη όταν αυτή προέκυπτε ήταν σημαντική.

Δε θα μπορούσα να παραλείψω να ευχαριστήσω τους γονείς μου οι οποίοι υπήρξαν πάντα ένα ανεκτίμητο στήριγμα για μένα και στους οποίους οφείλω όλη τη διαδρομή των σπουδών μου, τα αδέρφια μου και τον σύζυγό μου για την άμετρη συμπαράσταση, ενθάρρυνση, τη στήριξη και την υπομονή τους καθόλη τη διάρκεια των μεταπτυχιακών μου σπουδών.

Πίνακας περιεχομένων

1 *Εισαγωγή*

1.1 Γενικά

Το Internet έχει εξαπλωθεί κατά πολύ συνάμα με την εξέλιξη της τεχνολογίας και η πρόσβαση σε αυτό έγινε προσιτή για πολλούς ανθρώπους και κάθε μέρα σε ολοένα και περισσότερους. Μαζί με το Internet αυξήθηκε η ανάπτυξη ιστοτόπων και αυτό με τη σειρά του αύξησε την ανάγκη για βάσεις δεδομένων τις οποίες χρησιμοποιούν για την αποθήκευση των δεδομένων. Το πλήθος των δεδομένο είναι πολύ μεγάλο και καθημερινά αυξάνεται με εκθετικό ρυθμό. Λόγω του μεγάλου μεγέθους των δεδομένων που συγκεντρώνεται σωρευτικά στις βάσεις δεδομένων δεν μπορούν να αξιοποιηθούν όπως είναι. Είναι απαραίτητο πριν της αξιοποίησή του να γίνουν ενέργειες προκειμένου να τους δοθεί η κατάλληλη δομή.

Η τεχνολογία Εξόρυξης Δεδομένων είναι ένας όρος που χρησιμοποιείται για να περιγραφεί μια διαδικασία σχετικά αυτοματοποιημένη ώστε να μπορέσει να αναλυθεί ο όγκος δεδομένων ενός προβλήματος οποιασδήποτε φύσεως είτε εμπορικού είτε επιστημονικού ενδιαφέροντος. Η εξόρυξη δεδομένων, αν και αρχικά ενδιέφερε συγκεκριμένους κλάδους ερευνητικής φύσης, τα τελευταία χρόνια έχει επεκταθεί σε πολυάριθμα πεδία και οι τακτικές που προσφέρει δίνουν λύσεις σε προβλήματα διαφόρων μορφών. Ιδιαίτερα χρήσιμη κρίνεται για τους κλάδους που απαιτούν διαχείριση μεγάλου όγκου δεδομένων αλλά και γρήγορη λήψη αποφάσεων. Αντικειμενικός σκοπός της οποιαδήποτε διαδικασίας εξόρυξης δεδομένων είναι η υποβοήθηση της λήψης αποφάσεων. Οι μέθοδοι εξόρυξης δεδομένων είναι συνυφασμένοι με την στατιστική χωρίς ωστόσο να μπορούν να συγκριθούν με τις διαδικασίες περιγραφικής στατιστικής. Η εξόρυξη δεδομένων οδηγεί σε ένα είδος διερευνητικής ανάλυσης, ξεπερνώντας τα στεγανά της στατικής, επιβεβαιωτικής ανάλυσης, της απλής στατιστικής. Η βάση της εξόρυξης δεδομένων είναι τα εργαλεία τεχνητής νοημοσύνης, όπως τα νευρωνικά δίκτυα, τα δέντρα απόφασης και οι τακτικές μηχανικής μάθησης, αλλά κυρίως η βασική στατιστική και οι μεθοδολογίες

ανάλυσης παραγόντων , ομαδοποίησης, κλπ. Επομένως και η εξέλιξη της είναι αποτέλεσμα όλων αυτών των παραγόντων [1].

Η εξόρυξη δεδομένων έκανε την εμφάνισή της στα τέλη της δεκαετίας του 1980 όπου κυρίαρχο ρόλο είχαν τα συστήματα βάσεων δεδομένων που χρησιμοποιούνταν για λόγους αποθηκευτικούς σε επιχειρήσεις και οργανισμούς έτσι ώστε να μπορούν να οργανωθούν και να διαχειριστούν καλύτερα τα δεδομένα τους. Γρήγορα όμως ήρθε η ανάγκη να μπορούν να εκμεταλλεύονται τα δεδομένα τους καλύτερα και αυτό οδήγησε στην ανάπτυξη νέων εργαλείων τα οποία επέτρεπαν στους διευθυντές πωλήσεων, προϊσταμένους τμημάτων προώθησης προϊόντων κλπ να βρίσκουν και να ανακαλύπτουν νέα πρότυπα εξερεύνησης δεδομένων. Για παράδειγμα ένας χρήστης μπορούσε να υποβάλει ερωτήματα σχετικά με το ύψος των πωλήσεων των καταστημάτων μιας αλυσίδας καταστημάτων σε μια πόλη με σκοπό να βρει τα καταστήματα με τις χαμηλότερες πωλήσεις, ποιο προϊόν ίσως ευθύνεται για τις χαμηλές πωλήσει και χρειάζονται ενέργειες προκειμένου να βελτιωθούν οι πωλήσεις του, ποια προϊόντα προτιμούν οι πελάτες και πολλά ακόμα ερωτήματα.

Η αυτοματοποίηση της διαδικασίας για την ανακάλυψη της γνώσης έγινε με τη χρήση τεχνικών μεθοδολογιών και εργαλείων τα οποία αναπτύχθηκαν στο πλαίσιο της Εξόρυξης Δεδομένων. Για παράδειγμα, αντί ο ίδιος ο χρήστης να ζητάει μία αναφορά για να δει ποιο είναι το κατάστημα με τις λιγότερες πωλήσεις τον προηγούμενο μήνα, θα μπορούσε να εκτελεί στο σύστημα ερωτήματα που αφορούν γενικότερα στις πωλήσεις των καταστημάτων.

Η μεγάλη άνθιση στον τομέα της Εξόρυξης Δεδομένων έλαβε χώρα σταδιακά και ήταν άμεσα εξαρτημένη από τη δυνατότητα που δόθηκε για συλλογή και καταγραφή τεράστιων ποσοτήτων δεδομένων, διαφορετικών μορφών και τύπων, μέσω της ανάπτυξης γρήγορων δικτυακών υποδομών, πάνω στις οποίες μπορούσαν να υποστηριχτούν αξιόπιστες εμπορικές εφαρμογές. Στην κατηγορία των εταιριών που πρωτοστάτησαν σε αυτή τη νέα τάξη πραγμάτων ήταν και η Amazon, η οποία ξεκίνησε από την ηλεκτρονική πώληση βιβλίων και άλλων στοιχείων, δημιουργώντας στη συνέχεια ένα πολύ φιλικό προς τον χρήστη σύστημα συστάσεων[2].

Η παραγωγή μεγάλων όγκων δεδομένων σε εικοσιτετράωρη βάση καλύπτει μια τεράστια γκάμα ανθρώπινων δραστηριοτήτων και όχι μόνο, όπως είναι τα δεδομένα από το καλάθι αγορών, τον ιατρικό φάκελο του ασθενούς, τις συζητήσεις ή και ανακοινώσεις στα κοινωνικά μέσα δικτύωσης, τις τραπεζικές ή και χρηματιστηριακές συναλλαγές, τα ίχνη κινούμενων οχημάτων, τα δεδομένα αισθητήρων από κινητήρες αεροσκαφών, η καταγραφή συνομιλιών σε κέντρα εξυπηρέτησης πελατών κ.λπ. Τα δεδομένα αυτά διαφέρουν πάρα πολύ μεταξύ τους τόσο σε μορφή (εικόνα, βίντεο, κείμενο, πολυδιάστατα ή πραγματικού χρόνου δεδομένα, ακολουθίες DNA και άλλα πολλά) όσο και στην ταχύτητα συλλογής. Εάν, μάλιστα, δεν υποστούν άμεση ανάλυση, ίσως να είναι ιδιαίτερα δύσκολο να αποθηκευτούν ή να τα επεξεργαστούν οι άνθρωποι, δημιουργώντας έτσι μία καινούρια ερευνητική περιοχή, γνωστή

με τον όρο Μεγάλα Δεδομένα. Η Εξόρυξη των Δεδομένων στοχεύει να καλύψει τις ανάγκες που δημιουργούνται από αυτόν τον νέο τομέα και να προσφέρει λύσεις για την κλιμακούμενη και αποτελεσματική επεξεργασία δεδομένων.

Οι δύο πρωταρχικοί στόχοι στην πρακτική της Επιστήμης των Δεδομένων είναι η δημιουργία μοντέλων, τα οποία να μπορούν να χρησιμοποιηθούν τόσο για την πρόβλεψη, όσο και για την περιγραφή των δεδομένων. Η πρόβλεψη αφορά στη χρήση κάποιων μεταβλητών ή πεδίων μίας βάση δεδομένων, μέσω των τιμών των οποίων μπορεί να εκτιμηθεί η άγνωστη ή μελλοντική τιμή ενός άλλου γνωρίσματος. Η περιγραφή των δεδομένων εστιάζει στην εύρεση κατανοητών από τον άνθρωπο προτύπων, τα οποία περιγράφουν τα δεδομένα, όπως, δηλαδή, γίνεται κατά την εύρεση συστάδων ή ομάδων αντικειμένων με παρόμοια χαρακτηριστικά.

Η Εξόρυξη Δεδομένων είναι ένα πολλά υποσχόμενο επιστημονικό πεδίο. Όπως σε όλα τα επιστημονικά πεδία, έτσι και σε αυτό, υπάρχουν αρκετές προκλήσεις που καλείται να αντιμετωπίσει ο ερευνητής. Οι προκλήσεις αυτές είναι κυρίως τεχνικές και κοινωνικές-ηθικές. Η βασικότερη τεχνική πρόκληση είναι ο ολοένα αυξανόμενος όγκος των δεδομένων, καθώς και ο πολυδιάστατος και πολύπλοκος χαρακτήρας τους. Η υλοποίηση των λύσεων πρέπει να διέπεται από κλιμάκωση (scalability) και ευελιξία. Με άλλα λόγια, οι τεχνικές και οι αλγόριθμοι της Εξόρυξης Δεδομένων θα πρέπει να μπορούν να χειριστούν τον μεγάλο όγκο δεδομένων εξίσου καλά με ένα μικρότερο σύνολο δεδομένων[3].

1.2 Αντικείμενο διπλωματικής

Αντικείμενο της διπλωματικής αποτελεί η παραγωγή συστάσεων από δεδομένα συνδυαστικών κανόνων (association rules). Οι συνδυαστικοί κανόνες προκύπτουν από επεξεργασία τύπου data mining συνόλων δεδομένων του Μοντέλου Καλαθιού αγοράς και μπορεί να χρησιμοποιηθεί τροφοδοτώντας το με δεδομένα από διαφορετικά σύνολά όπως από εξεταστικές περιόδους του προγράμματος σπουδών του Τμήματος. Στόχος της διπλωματικής εργασίας συνιστά η διαμόρφωση, η αξιολόγηση στην πράξη και η τεκμηριωμένη διατύπωση μίας νέας πρότασης για ένα νέα πλαίσιο ιεραρχικής ταξινόμησης των συστάσεων που προκύπτουν στην έξοδο ενός Συστήματος Παραγωγής Συστάσεων (Recommender System, RS) που βασίζεται στην Εξόρυξη Συνδυαστικών Κανόνων (Association Rule Mining, ARM). Το μοντέλο της νέας πρότασης βασίζεται στην ταξινόμηση των συστάσεων στην εφαρμογή σχήματος με το οποίο καταρτίζεται ο πίνακας κατάταξης των χωρών βάσει των μεταλίων που κατακτούν οι αθλητές τους στους Ολυμπιακούς Αγώνες (Olympic Medal Standings Table, OMST). Το νέο πλαίσιο ιεράρχησης συστάσεων ARM πρόκειται να υλοποιηθεί σε περιβάλλον R/RStudio και να αξιολογηθεί σε σύγκριση με παραλλαγές της κλασικής προσέγγισης όπου οι συστάσεις ταξινομούνται με τη χρήση των αριθμητικών τιμών των μέτρων αξίας των συνδυαστικών κανόνων: lift, conviction, confidence, και support. Η ποιότητα των συστάσεων

που προκύπτουν πρόκειται να αξιολογηθεί σε τρόπο ώστε να προκύψει το μοντέλο με τα μεγαλύτερα ποσοστά επιτυχίας στις προβλέψεις. Δημιουργείται το πρότυπο-πλαίσιο μοντέλου το οποίο πρόκειται να χρησιμοποιηθεί για τη δημιουργία συστάσεων σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένου του περιβάλλοντος ανάλυσης βαθμολογικών δεδομένων του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙΠΑΕ.

1.2.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται στα παρακάτω :

1. Δημιουργήθηκε και προτείνεται μοντέλο-πλαίσιο παραγωγής συστάσεων με μια μέθοδο ταξινόμησης εμπνευσμένη από την κατάταξη των χωρών με βάση τα μετάλλια που αποκτούν από τους Ολυμπιακούς Αγώνες.
2. Αναπτύχθηκε κώδικας ως πρότυπο με δυνατότητα τροφοδότησής του δεδομένα διαφόρων συνόλων για παραγωγή συστάσεων.
3. Αξιολογήθηκε η προτεινόμενη μέθοδος έναντι της κλασικής μεθόδου διαβάθμισης των συστάσεων στην έξοδο ενός συστήματος ARM (Association Rule Mining).

1.3 Οργάνωση κειμένου

Στο Κεφάλαιο 2 περιγράφεται η έννοια των συνδυαστικών κανόνων και ποια είναι τα μέτρα για να υπολογιστεί η ισχύς του κάθε κανόνα, περιγράφεται ο αλγόριθμός Apriori και η κλασική μέθοδος παραγωγής συστάσεων με τη χρήση των συνδυαστικών κανόνων.

Από το Κεφάλαιο 3 παρουσιάζεται της συμβολής της παρούσας διπλωματικής εργασίας καθώς αναφέρεται η κεντρική ιδέα στη οποία βασίστηκε η διπλωματική εργασία και ο ψευδοκώδικας της μεθοδολογίας που αναπτύχθηκε.

Στο Κεφάλαιο 4 περιγράφονται οι τεχνικές και το περιβάλλον επεξεργασίας των δεδομένων που χρησιμοποιήθηκε.

Στο Κεφάλαιο 5 παρουσιάζεται η επεξεργασία που έγινε στα δεδομένα, η παραγωγή συστάσεων με τη μέθοδο medal standings και γίνεται αξιολόγηση των αποτελεσμάτων της μεθόδου που προτάθηκε έναντι των υπολοίπων κλασικών μεθόδων ταξινόμησης κανόνων κατά την παραγωγή συστάσεων.

Στο Κεφάλαιο 6 σχολιάζονται τα ευρήματα της μεθόδου και παρατίθενται τα συμπεράσματα που εξήχθησαν.

Στο Κεφάλαιο 7 περιγράφονται προτάσεις για περαιτέρω έρευνα και η εφαρμογή του προτεινόμενου μοντέλου σε πραγματικές συνθήκες.

Στο Κεφάλαιο 8 (Επίλογος) γίνεται μια σύνοψη του περιεχομένου και της συμβολής της διπλωματικής εργασίας.

2 *Συνδυαστικοί Κανόνες και Παραγωγή Συστάσεων*

2.1 Συνδυαστικοί κανόνες

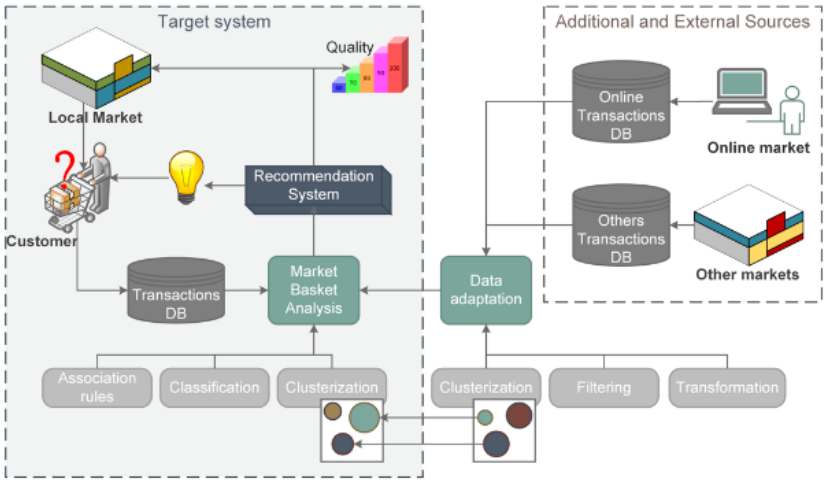
Το Μοντέλο του Καλαθιού Αγορών (Market Basket) αποτελεί κλασικό πεδίο εφαρμογής κανόνων που ανακαλύπτουν την κρυμμένη σύνδεση μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων, των επονομαζόμενων συνδυαστικών κανόνων (association rules). Τα δεδομένα του συνόλου που εξετάζει το μοντέλο του καλαθιού αγοράς μπορεί να είναι είτε κάποιο προϊόν είτε κατηγορίες προϊόντων, βιβλία, επιστημονικά άρθρα ακόμα και βαθμοί μαθημάτων. Μια από τις βασικές έννοιες που είναι σημαντικές για την κατανόηση του μοντέλου είναι η έννοια του συνδυασμού η οποία ορίζει την αλληλοσυσχέτιση μεταξύ των προϊόντων που αγοράζονται μαζί στην ίδια συναλλαγή. Με τον όρο συναλλαγή νοείται η αγορά κάποιων στοιχειοσύνολων μαζί. Ο κανόνας είναι ένας συστηματικός τρόπος προκειμένου να αναδειχθεί το είδος της πληροφορίας που διέπει αυτές τις αλληλοσυσχετίσεις [4].

Οι συνδυαστικοί κανόνες είναι μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Η έρευνα σχετικά με τους συνδυαστικούς κανόνες έχει γίνει πολύ δημοφιλής στους ερευνητές της εξόρυξης δεδομένων. Μπορούν να χρησιμοποιηθούν για την εξόρυξη χρήσιμης πληροφορίας από μεγάλους όγκους δεδομένων. Σήμερα, οι συνδυαστικοί κανόνες χρησιμοποιούνται στην εξόρυξη δεδομένων Ιστού, σε συστήματα συστάσεων, στην ανίχνευση εισβολών, στο μάρκετινγκ, στο ηλεκτρονικό εμπόριο, στην ανάλυση διαφόρων περιπτώσεων. Χρησιμοποιούνται ευρέως στη λήψη αποφάσεων που σχετίζονται με πολλούς τομείς όπως τα τηλεπικοινωνιακά δίκτυα, η αγορά και η διαχείριση κινδύνου, ο έλεγχος αποθεμάτων και σε πολλούς ακόμα.

Όπως αναφέρθηκε, μία από τις εφαρμογές των συνδυαστικών κανόνων είναι η χρήση τους στην ανάλυση του καλαθιού αγορών, μέσω της οποίας εντοπίζονται οι αγοραστικές συνήθειες των πελατών αναλύοντας τα προϊόντα που τοποθετούν στα καλάθια αγορών τους. Για παράδειγμα, αυτή η μέθοδος μπορεί να χρησιμοποιηθεί για την ανάλυση των προϊόντων που αγοράζουν οι

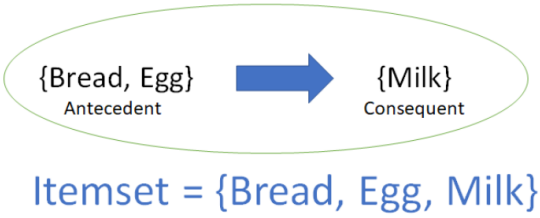
πελάτες και της αλληλουχίας τους σε μια χρονική περίοδο. Μπορεί δηλαδή να γίνει σαφές ότι όταν οι πελάτες αγοράζουν το προϊόν Α, πόσο πιθανό είναι να προσθέσουν το προϊόν Β στο καλάθι των αγορών τους.

Σε ένα κατάστημα, όλα τα λαχανικά τοποθετούνται στον ίδιο διάδρομο, όλα τα γαλακτοκομικά είδη τοποθετούνται μαζί και τα καλλυντικά αποτελούν ένα άλλο σύνολο τέτοιων ομάδων. Η επένδυση χρόνου και πόρων σε σκόπιμες τοποθετήσεις προϊόντων όπως αυτή, όχι μόνο μειώνει τον χρόνο αγορών ενός πελάτη, αλλά υπενθυμίζει επίσης στον πελάτη ποια σχετικά είδη μπορεί να ενδιαφέρεται να αγοράσει, βοηθώντας έτσι τα καταστήματα να πραγματοποιήσουν πολλαπλές πωλήσεις. Οι συνδυαστικοί κανόνες βοηθούν στην αποκάλυψη όλων αυτών των σχέσεων μεταξύ στοιχείων από τεράστιες βάσεις δεδομένων. Οι κανόνες δεν εξάγουν την προτίμηση ενός ατόμου, αλλά αναδεικνύουν σχέσεις μεταξύ στοιχειοσυνόλων κάθε ξεχωριστής συναλλαγής Αυτό είναι χρήσιμο για να προταθούν αντικείμενα σε ιστότοπους ηλεκτρονικού εμπορίου, τραγούδια σε εφαρμογές όπως το spotify και πολλά ακόμη[5].

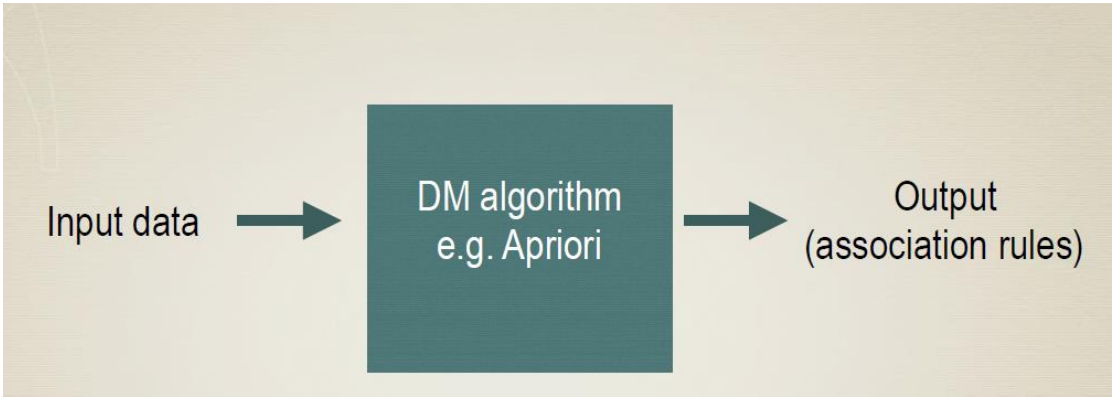


Εικόνα 2.1: Συνδυαστικοί Κανόνες και παραγωγή συστάσεων

Κατά τη διαδικασία εξόρυξης κανόνων, ο κανόνας ορίζεται ως η μορφή του $X \rightarrow Y$ με δύο περιορισμούς $X, Y \subset I$ και $X \cap Y = \emptyset$, όπου το I αντιπροσωπεύει το σύνολο των στοιχείων. Το X είναι ένα σύνολο αντικειμένων που ονομάζεται σώμα ή αριστερή πλευρά (LHS) και το Y υποδηλώνει ένα σύνολο στοιχείων που αναφέρονται ως η δεξιά πλευρά (RHS) ή κεφαλή του κανόνα[6]. Στην Εικόνα 2.2 φαίνεται η παρουσίαση ενός κανόνα που προκύπτει κατά την ανάλυση καλαθιών αγορών. Ο κανόνας ερμηνεύεται ως εξής: όταν κάποιος αγοράζει ψωμί και αυγά συνηθίζει να αγοράζει και γάλα.



Εικόνα 2.2: Μορφή κανόνα σε καλάθι αγοράς



Εικόνα 2.3: Εικόνα Επεξεργασίας

Η Εικόνα 2.3 απεικονίζει το διάγραμμα ροής του Μοντέλου. Στην είσοδό του τροφοδοτείται με δεδομένα εισόδου, τις συναλλαγές που πραγματοποιούνται. Με τη βοήθεια ενός αλγορίθμου εξόρυξης δεδομένων παράγονται στην έξοδο οι συνδυαστικοί κανόνες οι οποίοι στη συνέχεια ταξινομούνται ανάλογα με την έντασή τους. Για την παραγωγή συνδυαστικών κανόνων που πρεσβεύουν χρήσιμη και αξιοποιήσιμη πληροφορία είναι σημαντικό και σχεδόν απαραίτητο να τροφοδοτείται με σωστά και ποιοτικά δεδομένα ο αλγόριθμος. Επομένως, είναι σημαντικό να επιδέχονται τα δεδομενα την κατάλληλη επεξεργασία πριν αποτελέσουν στοιχεία εισόδου στον αλγόριθμο. Είναι πολύ σημαντικό για την σωστή επεξεργασία των δεδομένων η κατανόησή τους προκειμένου να γίνει σαφές ποια πληροφορία θα μπορούσε να εξαχθεί μετά την επεξεργασία και την είσοδό τους στον αλγόριθμο.

Η ένταση ενός συνδυαστικού κανόνα συνήθως ποσοτικοποιείται από τα μέτρα υποστήριξης (support), εμπιστοσύνης (confidence), και ανύψωσης (lift).

Το support δίνει μια ιδέα για το πόσο συχνά εμφανίζεται ένα στοιχειοσύνολο στο σύνολο των συναλλαγών. Μαθηματικά, το support ορίζεται ως το πηλίκο του συνολικού αριθμού συναλλαγών στις οποίες λαμβάνει χώρα ένα σύνολο στοιχείων προς το σύνολο όλων του συναλλαγών που πραγματοποιήθηκαν(Εξίσωση 1). Το support ενός κανόνα δείχνει το ποσοστό των συναλλαγών που περιέχουν την ένωση των συνόλων X και Y, και λαμβάνεται ως η πιθανότητα, εκφρασμένη ως P (X ∪ Y)[7]. Η τιμή του support βοηθά να προσδιοριστούν οι κανόνες εκείνοι που αξίζει να εξεταστούν για περαιτέρω ανάλυση. Για παράδειγμα, θα

μπορούσε κανείς να θελήσει να λάβει υπόψη μόνο τα σύνολα στοιχείων που εμφανίζονται τουλάχιστον 50 φορές από ένα σύνολο 10.000 συναλλαγών, δηλαδή η τιμή του support να είναι ίση με 0,005.

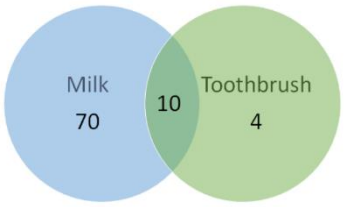
$$Support\ (\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ or\ transactions} \qquad (1)$$

Το μέτρο confidence ορίζεται μαθηματικά ως ένα κλάσμα όπου αριθμητής του είναι το σύνολο των συναλλαγών οι οποίες περιέχουν ένα συγκεκριμένο σύνολο στοιχείων, το σώμα και την κεφαλή του κανόνα, ενώ παρονομαστής του είναι το πλήθος των συναλλαγών που περιέχουν μόνο τα στοιχειοσύνολα του σώματος του κανόνα. Ένας άλλος τρόπος έκφρασης του confidence ενός κανόνα είναι το πηλίκο του Support (X,Y) προς το support (X). Δηλαδή το confidence είναι η πιθανότητα όταν μια συναλλαγή περιέχει το X να περιέχει και το Y [8]. Γίνεται αντιληπτό πως οι ακραίες περιπτώσεις είναι δύο, αυτή οπού το confidence μπορεί να είναι ίσο με μηδέν κάτι το οποίο σημαίνει πως καμία συναλλαγή δεν περιέχει ούτε το X ούτε το Y, και η άλλη ακραία περίπτωση είναι αυτή οπού το confidence μπορεί να είναι ίσο με 1 πράγμα που σημαίνει ότι όλες οι συναλλαγές που περιέχουν το X περιέχουν και το Y.

$$conf\ (\{X\} \rightarrow \{Y\}) = P(Y|X) = \frac{supp(X \cap Y)}{supp(X)} = \frac{number\ of\ transactions\ containing\ X\ and\ Y}{number\ of\ transactions\ containing\ X} \qquad (2)$$

Για το παράδειγμα της Εικόνας 2.4, για τον κανόνα {Toothbrush} → {Milk} θα είναι

Confidence: $\frac{10}{10+4} = 0,7$ και Support: $\frac{10}{70+10+4} = 0,0119$



Εικόνα 2.4: Απεικόνιση κανόνα

Παρόμοια με το Confidence και η τιμή του support βρίσκεται στο διάστημα [0,1]. Για τιμή ίση με το 0 τα X και Y δεν περιλαμβάνονται σε καμία συναλλαγή ενώ για τιμή ίση με 1 όλες οι συναλλαγές περιέχουν τα X και Y.

Για την επιλογή του ισχυρότερου κανόνα γίνεται ταξινόμηση πρώτα ως προς confidence και έπειτα ως προς support. Αυτό συμβαίνει διότι η confidence είναι ένα μέτρο το οποίο τείνει να δίνει επιπλέον πληροφορία σε σχέση με το support: τη διεύθυνση της αμοιβαίας εξάρτησης των X και Y.

Μια ακόμα σημαντική έννοια για τους συνδυαστικούς κανόνες είναι η έννοια του θορύβου. Θόρυβο μπορεί να προκαλούν δύο στοιχειοσύνολα που είναι στατιστικά ανεξάρτητα μεταξύ τους. Δύο στοιχειοσύνολα είναι στατιστικά ανεξάρτητα μεταξύ τους όταν το support του συνδυασμού τους είναι το γινόμενο του support του ενός με το support του άλλου στοιχειοσυνόλου [9]. Δηλαδή :

$$Support\ \{X, Y\} = Support\{X\} * Support\{Y\} \quad (3)$$

Ο θόρυβος ενός κανόνα μπορεί ποσοτικοποιηθεί και να υπολογιστεί από το support της κεφαλής του κανόνα.

Για να ληφθεί υπόψιν η ύπαρξη του θορύβου και να εξαλειφθεί ώστε να μην παράγονται κανόνες οι οποίοι δεν οδηγούν σε αξιοποιήσιμη πληροφορία εισάγεται το μέτρο lift [10]. Το μέτρο lift ενός κανόνα ισούται μαθηματικά με το πηλίκο του confidence κανόνα προς το support της κεφαλής του κανόνα, δηλαδή του θορύβου. Επομένως , οι κανόνες που η τιμή του Lift είναι ~1 είναι κυρίως θόρυβος.

$$lift(X \to Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)} \quad (4)$$

Δηλαδή το μέτρο lift ενός κανόνα όσο υψηλότερο είναι τόσο πιο ισχυρός είναι ο κανόνας και τόσο πιο αξιοποιήσιμη είναι η πληροφορία που μας παρέχει. Επίσης στο lift ισχύει η συμμετρική ιδιότητα σε αντίθεση με το μέτρο confidence, δηλαδή

$$lift\ (X \to Y) = lift(Y \to Y) \quad (5)$$

Το μέτρο conviction ενός κανόνα ορίζεται βάση της λογικής ισοδυναμίας του X=>Y με το Not(X=>not Y). Ισοδύναμα, όσο αυξάνεται η πιθανότητα του κανόνα P(X=>Y) τόσο η πιθανότητα P(X=>not Y) μικραίνει[11]. Στο conviction επίσης δεν ισχύει η συμμετρική ιδιότητα αλλά η τιμή του είναι μεγαλύτερη του 1.

$$conv(X \to Y) = \frac{1 - supp(Y)}{1 - conf\,(X \to Y)} \quad (6)$$

2.2 *Ο αλγόριθμος Apriori*

Ο αλγόριθμος Apriori [12] υπήρξε ο πρώτος προτεινόμενος κλασικός αλγόριθμος εξόρυξης συνδυαστικών κανόνων, του οποίου η διαδικασία υλοποίησης είναι σχετικά απλή. Βασίζεται στο ότι τα υπερσύνολα ενός μη συχνού στοιχειοσυνόλου είναι συχνά.

Apriori Pseudocode

$Apriori\ (T, \varepsilon)$

$L_1 \leftarrow \{large\ 1 - itemsets\ that\ appear\ in\ more\ than\ \varepsilon\ transactions\}$

$k \leftarrow 2$

$while\ L_{k-1} \neq \emptyset$

$C_k \leftarrow Generate\ (L_{k-1})$ — Join step and prune step

$for\ transactions\ t\ \in T$

$C_t \leftarrow Subset\ (c_k t)$

$for\ candidates\ c\ \in C_t$

$count\ |c| \leftarrow count|c| + 1$

$L_k \leftarrow \{c\ \in\ C_k | count|c| \geq \varepsilon\}$

$k\ \leftarrow k + 1$

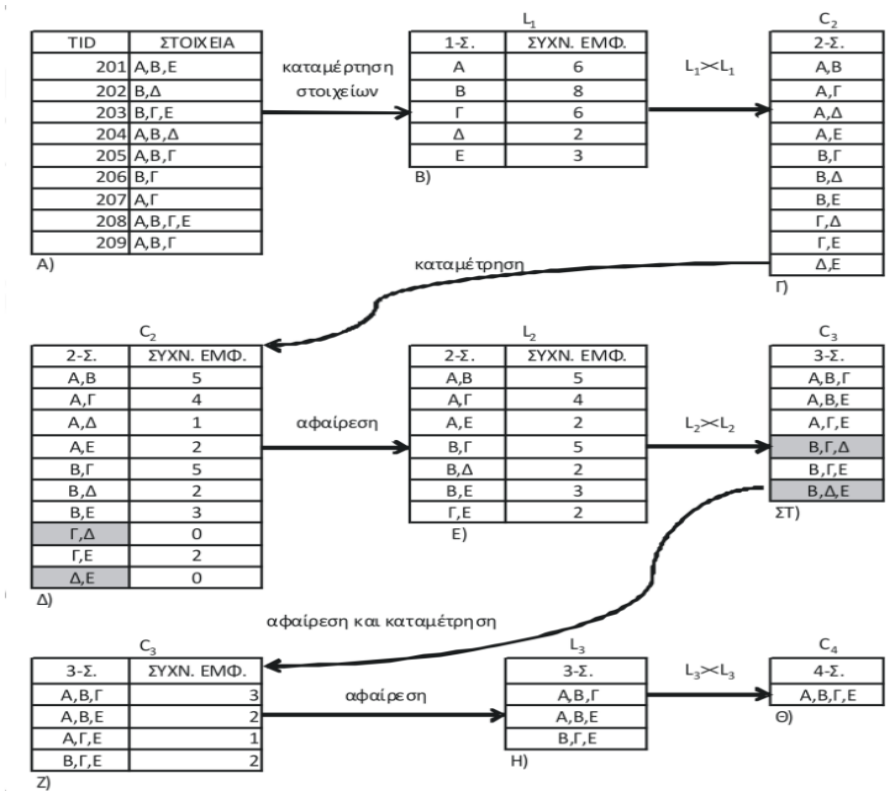
$return\ \cup L_k$

Για την καλύτερη κατανόηση του αλγορίθμου Apriori παραθέτουμε το παράδειγμα της Εικόνας 2.4. Στο τμήμα Α) του σχήματος απεικονίζονται τα δεδομένα. Η βάση δεδομένων περιέχει εννέα συνολικά συναλλαγές. Για κάθε συναλλαγή καταγράφονται τα εμπορεύματα που αγοράζονται ως Α, Β κλπ. Έστω ότι θέλουμε να βρούμε συχνά στοιχειοσύνολα με συχνότητα εμφάνισης ίση ή μεγαλύτερη από δύο. Η βάση δεδομένων σαρώνεται και υπολογίζεται η συχνότητα εμφάνισης κάθε δυνατού στοιχειοσυνόλο. Τα αποτελέσματα παρουσιάζονται στο τμήμα Β). Ξεκινούμε θεωρώντας το σύνολο των 1-στοιχειοσυνόλων C1.

Παρατηρούμε ότι η συχνότητα εμφάνισης κάθε εμπορεύματος είναι ίση ή μεγαλύτερη από δύο. Αυτό σημαίνει ότι όλα τα 1-στοιχειοσύνολα είναι συχνά και κανένα δεν πρέπει να διαγραφεί. Σε αυτήν την περίπτωση C1=L1. Στη συνέχεια, το L1 συνενώνεται με τον εαυτό του και προκύπτει το σύνολο 2-στοιχειοσυνόλων C2 το οποίο απεικονίζεται στο τμήμα Γ). Πραγματοποιείται καταμέτρηση των εμφανίσεων των μελών του C2 στη βάση δεδομένων, και τα αποτελέσματα απεικονίζονται στο τμήμα Δ). Παρατηρούμε ότι τρία στοιχειοσύνολα, το {Α,Δ}, το {Γ,Δ} και το {Δ,Ε}, έχουν συχνότητα εμφάνισης μικρότερη από δύο. Τα τρία αυτά στοιχειοσύνολα απομακρύνονται από το C21 και προκύπτει το σύνολο των συχνών 2-

στοιχειοσυνόλων L2, που παρουσιάζεται στο τμήμα Ε). Στη συνέχεια, το L2 συνενώνεται με τον εαυτό του. Για να μπορούν να συνδυαστούν δύο 2-στοιχειοσύνολα, μέλη του L2, πρέπει το πρώτο μέλος τους να είναι κοινό. Σύμφωνα με αυτόν τον κανόνα, το {Α,Β} συνδυάζεται με το {Α,Γ} και το {Α,Ε}. Επίσης, το {Α,Γ} συνδυάζεται με το {Α,Ε}. Με τον ίδιο τρόπο συνδυάζονται και τα μέλη του L2 των οποίων το πρώτο στοιχείο είναι το Β. Το {Γ,Ε} δεν μπορεί να συνδυαστεί με κανένα μέλος του L2. Το αποτέλεσμα της συνένωσης είναι το σύνολο C3 και απεικονίζεται στο τμήμα ΣΤ). Πραγματοποιείται έλεγχος στο C3 για να βρεθούν μέλη που έχουν μη συχνά υποσύνολα. Δύο μέλη του C3 έχουν μη συχνά υποσύνολα. Ειδικότερα, το {Β,Γ,Δ} έχει υποσύνολο το {Γ,Δ} το οποίο είναι μη συχνό καθώς δεν ανήκει στο L2. Επίσης, το {Β,Δ,Ε} έχει υποσύνολο το {Δ,Ε} το οποίο είναι μη συχνό. Συμπεραίνουμε ότι και τα {Β,Γ,Δ}, {Β,Δ,Ε} είναι μη συχνά. Για τον λόγο αυτό, τα δύο στοιχειοσύνολα απομακρύνονται από το C3, και για τα υπόλοιπα στοιχειοσύνολα του πραγματοποιείται καταμέτρηση στη βάση δεδομένων. Τα αποτελέσματα παρουσιάζονται στο τμήμα Ζ). Παρατηρούμε ότι ένα μέλος του C3, το {Α,Γ,Ε}, έχει συχνότητα εμφάνισης ίση με ένα.

Το στοιχειοσύνολο αυτό είναι μη συχνό και απομακρύνεται από το C3. Το αποτέλεσμα είναι το σύνολο συχνών 3-στοιχειοσυνόλων L3, το οποίο παρουσιάζεται στο τμήμα Η). Στη συνέχεια, το L3 συνενώνεται με τον εαυτό του. Τα μέλη του που μπορούν να συνδυαστούν είναι το {Α,Β,Γ} και το {Α,Β,Ε}, γιατί αυτά τα δύο στοιχειοσύνολα έχουν τα δυο πρώτα μέλη τους κοινά. Το αποτέλεσμα της συνένωσης είναι το C4, το οποίο απεικονίζεται στο τμήμα Θ). Το C4 έχει ένα μόνο μέλος, το {Α,Β,Γ,Ε}. Παρατηρούμε ότι το {Α,Β,Γ,Ε} έχει υποσύνολο το {Α,Γ,Ε} το οποίο είναι μη συχνό, οπότε συμπεραίνουμε ότι και το {Α,Β,Γ,Ε} είναι μη συχνό. Με την ολοκλήρωση της εκτέλεσης του αλγορίθμου έχουμε εντοπίσει τα συχνά στοιχειοσύνολα, τα οποία είναι τα μέλη του L2 και του L3.



Εικόνα 2.4:Εκτέλεση αλγορίθμου Apriori

Για την παραγωγή των κανόνων ορίζεται μια ελάχιστη τιμή του support (s) και ένα κατώφλι για το confidence (c) και καλείται ο αλγόριθμος να βρει όλους τους κανόνες της μορφής Α -> Β, όπου τα Α και Β είναι στοιχειοσύνολα, έτσι ώστε:

1. Τα Α και Β εμφανίζονται μαζί σε τουλάχιστον s% των συναλλαγών.

2. Το Β εμφανίζεται τουλάχιστον στο c% των συναλλαγών στις οποίες υπάρχει το Α.

Ένας συνδυαστικός κανόνας παρέχει αξιοποιήσιμη πληροφορία όταν ικανοποιεί τις ελάχιστες τιμές support και confidence που έχουμε ορίσει στον αλγόριθμο apriori για την παραγωγή του.

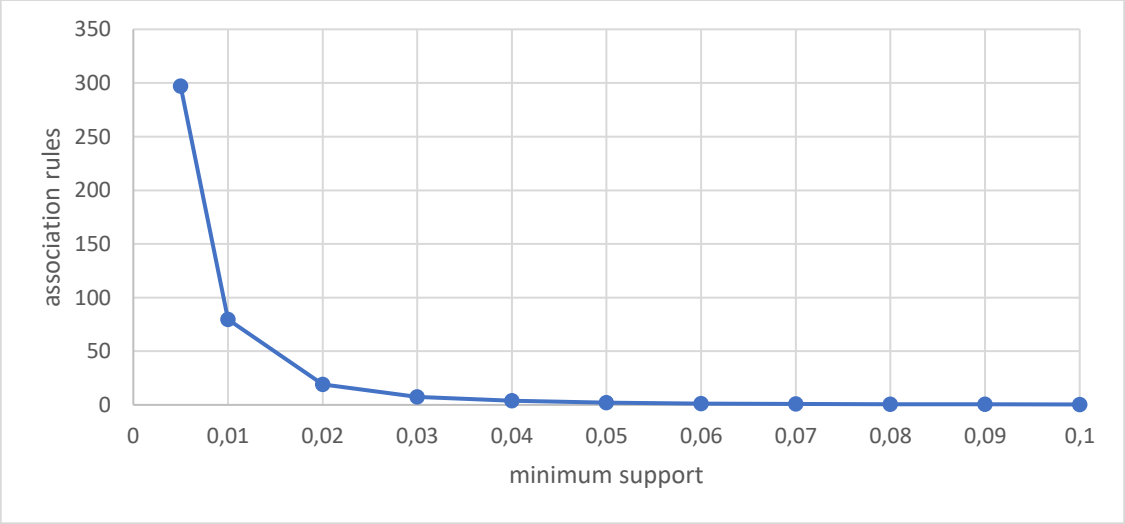
Μετά την εκτέλεση του αλγορίθμου έχουν εντοπιστεί όλοι οι ισχυροί κανόνες, οι οποίοι έχουν τιμή support και confidence μεγαλύτερη από τις ελάχιστες καθορισμένες τιμές.

Η διαδικασία εξόρυξης συνδυαστικών κανόνων μπορεί να αναδείξει ενδιαφέροντες κανόνες, ωστόσο συχνά παρουσιάζονται τα παρακάτω προβλήματα:

- Το πλήθος των κανόνων που προκύπτουν είναι υπερβολικά μεγάλο.
- Πολλοί από τους κανόνες δεν έχουν σχέση με το θέμα που μελετά ο αναλυτής και γι' αυτόν τον λόγο του είναι αδιάφοροι.
- Ο χρόνος εξόρυξης των κανόνων είναι υπερβολικά μεγάλος.

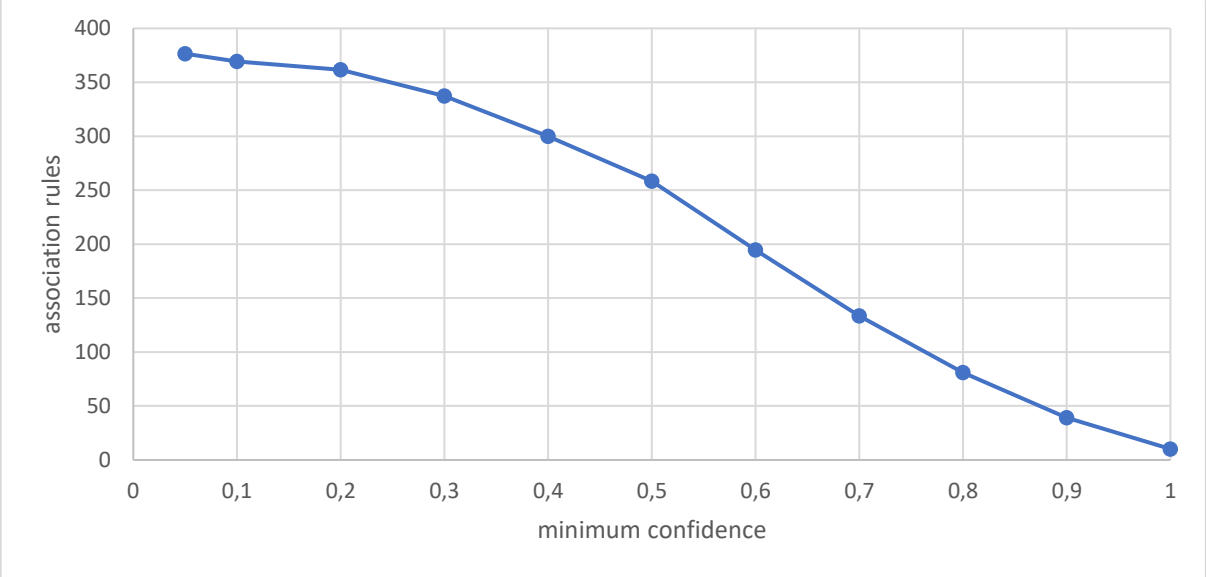
Δεν υπάρχει προφανής μέθοδος για την επιλογή της κατάλληλης τιμής support και confidence. Αν οι κατώτατες τιμές που επιλεγούν είναι υψηλές, τότε ενδιαφέρουσες συσχετίσεις μπορεί να χαθούν[13][1].

Για να γίνει κατανοητό το μέγεθος αυτού του προβλήματος και ιδιαίτερα η δυσκολία επιλογής των κατάλληλων ορίων των τιμών, παρατίθενται συνδυαστικοί κανόνες στα δεδομένα που χρησιμοποιήθηκαν προς ανάλυση στην παρούσα διπλωματική εργασία με ποικίλα επίπεδα υποστήριξης και εμπιστοσύνης. Η Εικόνα 2.5 δείχνει τον αριθμό των συνδυαστικών κανόνων που βρέθηκαν με confidence 0,1 καθώς το support κυμαίνεται από 0,005 έως 0.1. Ο αριθμός των κανόνων είναι αμελητέος πάνω από το 0,02 support αλλά αυξάνεται πολύ γρήγορα κάτω από 0,005. Το Εικόνα 2.6 δείχνει ένα παρόμοιο αποτέλεσμα, αυτή τη φορά διατηρώντας σταθερή την τιμή του support στο 0,001 και κυμαινόμενη η τιμή του confidence από 0.005 έως 1. Η αύξηση εμφανίζεται σημαντικά μικρότερη αλλά αυτό οφείλεται σε μεγάλο βαθμό σε έναν αριθμό περιττών κανόνων με πολλά στοιχεία με εξαιρετικά υψηλή την τιμή του confidence. Να σημειωθεί ότι από 10 έως 5%, ο αριθμός των κανόνων περισσότερο από διπλασιάζεται. Συνολικά, οι Εικόνες 2.5 και 2.6 δείχνουν αυτή τη συσχέτιση με τους κανόνες οι οποίοι μπορεί να είναι απίστευτα ευαίσθητοι στην επιλογή της κατώτατης τιμής των παραμέτρων support και confidence.



Εικόνα 2.5: Πλήθος συνδυαστικών κανόνων που παράγονται σε σχέση με το support (όταν confidence = 0.5)

[1] Cran.r-project.org. 2021. CRAN Repository Policy. [online] Available at: <https://cran.r-project.org/web/packages/policies.html> [Accessed 15 January 2021]



Εικόνα 2.6: Πλήθος συνδυαστικών κανόνων που παράγονται σε σχέση με το confidence (όταν support =0.003)

Ένα δεύτερο πρακτικό ζήτημα είναι ότι πολλοί κανόνες μπορεί να μην περιέχουν πληροφορία και να είναι είτε περιττοί είναι να περιέχουν θόρυβο. Ένας αριθμός διαφορετικών τεχνικών έχει αναπτυχθεί για την αντιμετώπιση αυτού του ζητήματος. Μια προσέγγιση για την καταπολέμηση της έκρηξης πληθώρας κανόνων και απαλοιφής των περιττών είναι ο υπολογισμός πρόσθετων μέτρων ενδιαφέροντος. Αυτά τα μέτρα μπορούν στη συνέχεια να χρησιμοποιηθούν είτε για την ταξινόμηση των κανόνων κατά σπουδαιότητα (και να παρουσιάζουν μια ταξινομημένη λίστα στον χρήστη) ή ως πρόσθετο κριτήριο κλαδέματος των κανόνων. Ένα από τα απλούστερα τέτοια μέτρα και αντικειμενικό κριτήριο αποτίμησης των κανόνων είναι το lift, το οποίο αποτελεί στατιστικό μέτρο. Επιπλέον ,τα μέτρα είναι κυρίως χρήσιμα όταν η εμπειρία ή η βασική γνώση είναι διαθέσιμη και βοηθά στην επιλογή του κατάλληλου μέτρου. Δεδομένου ότι από τα πρώτα βήμα στην διαδικασία αποτελεί η επιλογή της ελάχιστης τιμής στην παράμετρο support και confidence, μπορεί κανείς να φανταστεί αρκετές λογικές επιλογές. Για παράδειγμα, μπορεί κανείς να επιλέξει ένα αυθαίρετα υψηλό όριο και να το μειώσει σταδιακά και επαναλαμβανόμενα μέχρι ο αριθμός των κανόνων να πάψει να είναι μη διαχειρίσιμος.

2.3 Παραγωγή Συστάσεων

Τα Συστήματα Συστάσεων (RS: Recommender Systems) έχουν αναπτυχθεί κυρίως λόγω της αυξανόμενης χρήσης του διαδικτύου ως μέσου ψυχαγωγίας και ως μέσου ηλεκτρονικών και επιχειρηματικών συναλλαγών [14]. Καταλυτικό παράγοντα για την ανάπτυξη τους αποτελεί η ευκολία με την οποία το διαδίκτυο επιτρέπει στους χρήστες να αφήνουν ανατροφοδότηση σχετικά με τις προτιμήσεις τους. Άλλες μορφές ανατροφοδότησης, ακόμη ευκολότερες στη

συλλογή, είναι τα δεδομένα που παράγονται κατά την αγορά ενός προϊόντος από ένα χρήστη ή αυτά που παράγονται κατά την περιήγηση του σε ιστοσελίδες και κατά τις αναζητήσεις που πραγματοποιεί. Έτσι ο χρήστης, χωρίς να το αντιλαμβάνεται τροφοδοτεί συνεχώς Σύστηματα Συστάσεων με δεδομένα που αφορούν τις προτιμήσεις του. Τα μοντέλα των Συστημάτων Συστάσεων (ΣΣ) σε συνδυασμό με συνδυαστικούς κανόνες βασίζονται κατά κύριο λόγο στην παραγωγή συστάσεων χρησιμοποιώντας σύνθετη ταξινόμηση στους κανόνες πρώτα ως προς την τιμή του lift και έπειτα ως προς την τιμή του confidence [2].

[2] 'Market Basket Analysis using R', DataCamp Community, Αυγούστου 21, 2018. https://www.datacamp.com/community/tutorials/market-basket-analysis-r (ημερομηνία πρόσβασης 25 Αυγούστου 2020)

3 Προτεινόμενες Βελτιώσεις

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο (Συνδυαστικοί Κανόνες και Παραγωγή Συστάσεων) η κλασική μέθοδος που χρησιμοποιείται για την παραγωγή συστάσεων με τη χρήση των συνδυαστικών κανόνων, ταξινομεί την ισχύ του κανόνα αρχικά ως προς την τιμή του lift του κανόνα και έπειτα ως προς confidence. Στην παρούσα εργασία εξετάζεται η παραγωγή συστάσεων ταξινομώντας τους κανόνες πρώτα ως προς conviction και μετά ως προς την τιμή του support και επιπλέον προτείνεται μέθοδο ταξινόμησης των κανόνων και κατ' επέκταση παραγωγή συστάσεων βασιζόμενη στη διακριτοποίηση των κανόνων σε μετάλλια και την μετέπειτα ταξινόμησή τους με βάση τα μετάλλια που σχηματίστηκαν ως προς το conviction και στη συνέχεια ως προς το support. Στο παρόν κεφάλαιο αναλύεται η κεντρική ιδέα της προτεινόμενη μεθόδου. Τα αποτελέσματα της αξιολόγησης των συστάσεων που προκύπτουν με την εφαρμογή της προτεινόμενης μεθόδου έναντι της κλασικής αλλά και της «παραλλαγής» της κλασικής μεθόδου παρουσιάζονται σε επόμενο κεφάλαιο (Κεφάλαιο 5 : Επεξεργασία των Δεδομένων).

3.1 Η κεντρική ιδέα

Εκμεταλλευόμενοι όλες τις διαθέσιμες τεχνικές και με τη χρήση των εργαλείων του περιβάλλοντος της R, ερευνάται και παρουσιάζεται μια μέθοδος παραγωγής προβλέψεων η οποία βασίζεται στην ισχύ των συνδυαστικών κανόνων ταξινομώντας τους όχι κατά απόλυτη τιμή τις μετρικές του αλλά αφού πρώτα διακριτοποιηθούν σε διαστήματα που προκύπτουν βασιζόμενα στη μέθοδο της συχνότητας.

Η έρευνα διεξάγεται σε δεδομένα που αφορούν συναλλαγές καλαθιού σε κατάστημα υπεραγοράς. Από τα προϊόντα που εμπεριέχονται στο καλάθι μιας συναλλαγής παράγονται οι συνδυαστικοί κανόνες και αυτοί με τη σειρά τους αφού θα ταξινομηθούν θα προκύψει η σύσταση- πρόβλεψη. Για να γίνει κατανοητό έστω το καλάθι απαρτίζεται από προϊόντα Π1, Π2, Π3. Το ζητούμενο είναι να βρεθεί το ΠΧ. Για την αναζήτηση σχηματίζονται όλοι οι δυνατοί συνδυασμοί που προκύπτουν από τα Π1,Π2,Π3: Π1, Π2, Π3, Π1Π2, Π1Π3, Π2Π3, και Π1Π2Π3. Εφόσον σχηματιστούν όλοι οι δυνατοί συνδυασμοί αναζητούνται οι κεφαλές των

κανόνων που έχουν στο σώμα κάποιον από τους συνδυασμούς των προϊόντων. Οι κανόνες αυτοί ενδέχεται να είναι αρκετοί και ο καθένας τους να έχει διαφορετική κεφαλή. Για να οριστεί το πιο πιθανό προϊόν που βρίσκεται στο καλάθι με τα υπόλοιπα είναι αναγκαίο οι συστάσεις που προκύπτουν να ταξινομηθούν. Η κλασική μέθοδος χρησιμοποιεί ως κριτήριο ταξινόμησης την απόλυτη τιμή των μέτρων, δηλαδή του conviction, support και lift.

α)

	body	head	support	confidence	coverage	lift	count	conviction	rulenumber	medallift_freq	medalconviction_freq	medalsupport_freq
1	57,93,97	61	0.004787339	0.5450450	0.008783383	4.991309	121	1.957999	4640	A	C	B
2	57,93,97	101	0.004510386	0.5135135	0.008783383	4.839319	114	1.837435	4619	A	C	B
3	57,93,97	49	0.005024728	0.5720721	0.008783383	4.093749	127	2.010285	4641	A	C	B
4	57,97	49	0.010840752	0.5404339	0.020059347	3.867346	274	1.871890	952	B	C	A
5	57,93,97	53	0.005974283	0.6801802	0.008783383	3.682063	151	2.549160	4646	B	B	B
6	57,93,97	85	0.005657765	0.6441441	0.008783383	3.600341	143	2.307361	4644	B	B	B
7	57,93	53	0.013570722	0.6340111	0.021404550	3.432133	343	2.227587	1058	B	B	A
8	57,97	53	0.012621167	0.6291913	0.020059347	3.406042	319	2.198632	956	B	B	A
9	57,97	85	0.012106825	0.6035503	0.020059347	3.373448	306	2.071103	954	B	C	A
10	57,93,97	81	0.005103858	0.5810811	0.008783383	3.298186	129	1.966533	4642	B	C	B
11	57,93,97	33	0.007042532	0.8018018	0.008783383	3.281869	178	3.812787	4647	B	A	A
12	57,93	85	0.012067260	0.5637708	0.021404550	3.151107	305	1.882240	1057	B	C	A
13	57,97	33	0.015232443	0.7593688	0.020059347	3.108186	385	3.140439	957	B	B	A
14	93,97	53	0.016142433	0.5513514	0.029277943	2.984666	408	1.817173	944	B	C	A
15	57,93	33	0.015430267	0.7208872	0.021404550	2.950676	390	2.707463	1059	B	B	A
16	57,93	81	0.010919881	0.5101664	0.021404550	2.895678	276	1.681832	1056	B	C	A
17	57,97	81	0.010128586	0.5049310	0.020059347	2.865962	256	1.664047	953	B	C	A
18	93,97	85	0.014876360	0.5081081	0.029277943	2.839989	376	1.669245	943	B	C	A
19	57	53	0.048189911	0.5014409	0.096102868	2.714483	1218	1.635257	26	C	C	A
20	57,93,97	41	0.008269041	0.9414414	0.008783383	2.586125	209	10.860315	4649	C	A	A
21	57,97	41	0.018516320	0.9230769	0.020059347	2.535678	468	8.267537	959	C	A	A
22	57	33	0.058555885	0.6093042	0.096102868	2.493954	1480	1.934209	27	C	C	A
23	57,93,97	65	0.006686449	0.7612613	0.008783383	2.468045	169	2.896693	4648	C	B	A
24	57,93	41	0.019109792	0.8927911	0.021404550	2.452483	483	5.932013	1061	C	A	A
25	93,97	33	0.017368942	0.5932432	0.029277943	2.428214	439	1.857836	945	C	C	A
26	57,97	65	0.014757666	0.7357002	0.020059347	2.385175	373	2.616547	958	C	B	A
27	57,93	65	0.015113749	0.7060998	0.021404550	2.289209	382	2.353020	1060	C	B	A
28	57	41	0.076755687	0.7986826	0.096102868	2.193968	1940	3.159013	29	C	B	A
29	93,97	41	0.022710188	0.7756757	0.029277943	2.130769	574	2.835022	947	C	B	A
30	93,97	65	0.018595450	0.6351351	0.029277943	2.059138	470	1.895367	946	C	C	A
31	57,93,97	77	0.004589515	0.5225225	0.008783383	1.990768	116	1.544632	4645	C	C	B
32	57,97	77	0.010089021	0.5029586	0.020059347	1.916231	255	1.483834	955	C	C	A
33	57	65	0.055074184	0.5730753	0.096102868	1.857937	1392	1.619848	28	C	C	A
34	97	41	0.058911968	0.6638431	0.088743818	1.823566	1489	1.891868	20	C	C	A
35	97	65	0.048229476	0.5434686	0.088743818	1.761951	1219	1.514798	19	C	C	A
36	93	41	0.064451039	0.6410862	0.100534125	1.761054	1629	1.771914	23	C	C	A
37	93	65	0.051671612	0.5139709	0.100534125	1.666318	1306	1.422863	22	C	C	A

β)

	body	head	support	confidence	coverage	lift	count	conviction	rulenumber	medallift_freq	medalconviction_freq	medalsupport_freq
1	57,93,97	61	0.004787339	0.5450450	0.008783383	4.991309	121	1.957999	4640	A	C	B
2	57,93,97	101	0.004510386	0.5135135	0.008783383	4.839319	114	1.837435	4619	A	C	B
3	57,93,97	49	0.005024728	0.5720721	0.008783383	4.093749	127	2.010285	4641	A	C	B
4	57,97	49	0.010840752	0.5404339	0.020059347	3.867346	274	1.871890	952	B	C	A
5	57,93,97	53	0.005974283	0.6801802	0.008783383	3.682063	151	2.549160	4646	B	B	B
6	57,93,97	85	0.005657765	0.6441441	0.008783383	3.600341	143	2.307361	4644	B	B	B
7	57,93	53	0.013570722	0.6340111	0.021404550	3.432133	343	2.227587	1058	B	B	A
8	57,97	53	0.012621167	0.6291913	0.020059347	3.406042	319	2.198632	956	B	B	A
9	57,97	85	0.012106825	0.6035503	0.020059347	3.373448	306	2.071103	954	B	C	A
10	57,93,97	81	0.005103858	0.5810811	0.008783383	3.298186	129	1.966533	4642	B	C	B
11	57,93,97	33	0.007042532	0.8018018	0.008783383	3.281869	178	3.812787	4647	B	A	A
12	57,93	85	0.012067260	0.5637708	0.021404550	3.151107	305	1.882240	1057	B	C	A
13	57,97	33	0.015232443	0.7593688	0.020059347	3.108186	385	3.140439	957	B	B	A
14	93,97	53	0.016142433	0.5513514	0.029277943	2.984666	408	1.817173	944	B	C	A
15	57,93	33	0.015430267	0.7208872	0.021404550	2.950676	390	2.707463	1059	B	B	A
16	57,93	81	0.010919881	0.5101664	0.021404550	2.895678	276	1.681832	1056	B	C	A
17	57,97	81	0.010128586	0.5049310	0.020059347	2.865962	256	1.664047	953	B	C	A
18	93,97	85	0.014876360	0.5081081	0.029277943	2.839989	376	1.669245	943	B	C	A
19	57	53	0.048189911	0.5014409	0.096102868	2.714483	1218	1.635257	26	C	C	A
20	57,93,97	41	0.008269041	0.9414414	0.008783383	2.586125	209	10.860315	4649	C	A	A
21	57,97	41	0.018516320	0.9230769	0.020059347	2.535678	468	8.267537	959	C	A	A
22	57	33	0.058555885	0.6093042	0.096102868	2.493954	1480	1.934209	27	C	C	A
23	57,93,97	65	0.006686449	0.7612613	0.008783383	2.468045	169	2.896693	4648	C	B	A
24	57,93	41	0.019109792	0.8927911	0.021404550	2.452483	483	5.932013	1061	C	A	A
25	93,97	33	0.017368942	0.5932432	0.029277943	2.428214	439	1.857836	945	C	C	A
26	57,97	65	0.014757666	0.7357002	0.020059347	2.385175	373	2.616547	958	C	B	A
27	57,93	65	0.015113749	0.7060998	0.021404550	2.289209	382	2.353020	1060	C	B	A
28	57	41	0.076755687	0.7986826	0.096102868	2.193968	1940	3.159013	29	C	B	A
29	93,97	41	0.022710188	0.7756757	0.029277943	2.130769	574	2.835022	947	C	B	A
30	93,97	65	0.018595450	0.6351351	0.029277943	2.059138	470	1.895367	946	C	C	A
31	57,93,97	77	0.004589515	0.5225225	0.008783383	1.990768	116	1.544632	4645	C	C	B
32	57,97	77	0.010089021	0.5029586	0.020059347	1.916231	255	1.483834	955	C	C	A
33	57	65	0.055074184	0.5730753	0.096102868	1.857937	1392	1.619848	28	C	C	A
34	97	41	0.058911968	0.6638431	0.088743818	1.823566	1489	1.891868	20	C	C	A
35	97	65	0.048229476	0.5434686	0.088743818	1.761951	1219	1.514798	19	C	C	A
36	93	41	0.064451039	0.6410862	0.100534125	1.761054	1629	1.771914	23	C	C	A
37	93	65	0.051671612	0.5139709	0.100534125	1.666318	1306	1.422863	22	C	C	A

γ)

	body	head	support	confidence	coverage	lift	count	conviction	rulenumber	medallift_freq	medalconviction_freq	medalsupport_freq
1	57,93,97	41	0.008269041	0.9414414	0.008783383	2.586125	209	10.860315	4649	C	A	A
2	57,97	41	0.018516320	0.9230769	0.020059347	2.535678	468	8.267537	959	C	A	A
3	57,93	41	0.019109792	0.8927911	0.021404550	2.452483	483	5.932013	1061	C	A	A
4	57,93,97	33	0.007042532	0.8018018	0.008783383	3.281869	178	3.812787	4647	B	A	A
5	57	41	0.076755687	0.7986826	0.096102868	2.193968	1940	3.159013	29	C	B	A
6	57,97	33	0.015232443	0.7593688	0.020059347	3.108186	385	3.140439	957	B	B	A
7	57,93,97	65	0.006686449	0.7612613	0.008783383	2.468045	169	2.896693	4648	C	B	A
8	93,97	41	0.022710188	0.7756757	0.029277943	2.130769	574	2.835022	947	C	B	A
9	57,93	33	0.015430267	0.7208872	0.021404550	2.950676	390	2.707463	1059	B	B	A
10	57,97	65	0.014757666	0.7357002	0.020059347	2.385175	373	2.616547	958	C	B	A
11	57,93,97	53	0.005974283	0.6801802	0.008783383	3.682063	151	2.549160	4646	B	B	B
12	57,93	65	0.015113749	0.7060998	0.021404550	2.289209	382	2.353020	1060	C	B	A
13	57,93,97	85	0.005657765	0.6441441	0.008783383	3.600341	143	2.307361	4644	B	B	B
14	57,93	53	0.013570722	0.6340111	0.021404550	3.432133	343	2.227587	1058	B	B	A
15	57,97	53	0.012621167	0.6291913	0.020059347	3.406042	319	2.198632	956	B	B	A
16	57,97	85	0.012106825	0.6035503	0.020059347	3.373448	306	2.071103	954	B	C	A
17	57,93,97	49	0.005024728	0.5720721	0.008783383	4.093749	127	2.010285	4641	A	C	B
18	57,93,97	81	0.005103858	0.5810811	0.008783383	3.298186	129	1.966533	4642	B	C	B
19	57,93,97	61	0.004787339	0.5450450	0.008783383	4.991309	121	1.957999	4640	A	C	B
20	57	33	0.058555885	0.6093042	0.096102868	2.493954	1480	1.934209	27	C	C	A
21	93,97	65	0.018595450	0.6351351	0.029277943	2.059138	470	1.895367	946	C	C	A
22	97	41	0.058911968	0.6638431	0.088743818	1.823566	1489	1.891868	20	C	C	A
23	57,93	85	0.012067260	0.5637708	0.021404550	3.151107	305	1.882240	1057	B	C	A
24	57,97	49	0.010840752	0.5404339	0.020059347	3.867346	274	1.871890	952	B	C	A
25	93,97	33	0.017368942	0.5932432	0.029277943	2.428214	439	1.857836	945	C	C	A
26	57,93,97	101	0.004510386	0.5135135	0.008783383	4.839319	114	1.837435	4619	A	C	B
27	93,97	53	0.016142433	0.5513514	0.029277943	2.984666	408	1.817173	944	B	C	A
28	93	41	0.064451039	0.6410862	0.100534125	1.761054	1629	1.771914	23	C	C	A
29	57,93	81	0.010919881	0.5101664	0.021404550	2.895678	276	1.681832	1056	B	C	A
30	93,97	85	0.014876360	0.5081081	0.029277943	2.839989	376	1.669245	943	B	C	A
31	57,97	81	0.010128586	0.5049310	0.020059347	2.865962	256	1.664047	953	B	C	A
32	57	53	0.048189911	0.5014409	0.096102868	2.714483	1218	1.635257	26	C	C	A
33	57	65	0.055074184	0.5730753	0.096102868	1.857937	1392	1.619848	28	C	C	A
34	57,93,97	77	0.004589515	0.5225225	0.008783383	1.990768	116	1.544632	4645	C	C	B
35	97	65	0.048229476	0.5434686	0.088743818	1.761951	1219	1.514798	19	C	C	A
36	57,97	77	0.010089021	0.5029586	0.020059347	1.916231	255	1.483834	955	C	C	A
37	93	65	0.051671612	0.5139709	0.100534125	1.666318	1306	1.422863	22	C	C	A

Εικόνα 3.1: Ταξινόμηση α)ως προς lift -confidence β)lift - support γ)conviction-support

Στο παράδειγμα της Εικόνας 3.1 απεικονίζονται οι συστάσεις που προκύπτουν για την συναλλαγή {57,93,97}. Στο τμήμα (α) της εικόνας οι συστάσεις είναι ταξινομημένες πρώτα ως προς lift και μετά ως προς confidence. Αυτή είναι η κλασική μέθοδος ταξινόμησης της ισχύος τους. Στο τμήμα (β) οι συστάσεις ταξινομούνται πρώτα ως προς lift και μετά ως προς support και τέλος, στο τμήμα (γ) η ταξινόμηση γίνεται πρώτα ως προς το conviction και μετά ως προς

support. Παρατηρείται ότι στις (α) και (β) ο κανόνας που ενεργοποιεί την ισχυρότερη σύσταση είναι ο 4640 η ισχυρότερη σύσταση είναι η {61} ενώ στη (γ) η ισχυρότερη σύσταση είναι η {41} και προκύπτει από τον κανόνα 4649.

Σε αντίθεση με την κλασική μέθοδο ερευνάται η ταξινόμηση της ισχύος των κανόνων και η παραγωγή συστάσεων με τη μέθοδο των μεταλλίων. Σύμφωνα με τη μέθοδο αυτή τα μέτρα conviction, support και lift διακριτοποιούνται βάσει συχνότητας σε τρία τμήματα καθένα από τα οποία αντιστοιχούν σε μετάλλια A,B,C με φθίνουσα ταξινόμηση ακολουθώντας την λογική δημιουργίας του πίνακα κατάταξης των χωρών σύμφωνα με τα μετάλλια που κατέκτησαν στους Ολυμπιακούς αγώνες. Μια τέτοια κατάταξη παρουσιάζεται στην Εικόνα 3.2 η οποία απεικονίζει την κατάταξη των χωρών όπως διαμορφώθηκε από την κατάκτηση των μεταλλίων στους Ολυμπιακούς Αγώνες το 2020. Αξίζει να σημειωθεί ότι στην κατάταξη αυτή δεν έχει σημασία το πλήθος των μεταλλίων αλλά το είδος αυτών. Αν μια χώρα κατακτήσει 1 μόνο μετάλλιο αλλά χρυσό θα καταλάβει υψηλότερη θέση στον πίνακα κατάταξης από μια άλλη η οποία θα καταφέρει να συγκεντρώσει π.χ. 20 μετάλλια χάλκινα. Όπως φαίνεται και στην Εικόνα 3.2 η Ιαπωνία βρίσκεται στην τρίτη θέση με 58 στο σύνολο μετάλλια ενώ η οι Ρώσοι αθλητές έχουν καταλάβει την Πέμπτη θέση παρότι τα μετάλλια που έχουν συγκεντρώσει είναι στο σύνολο 71. Αυτό συμβαίνει διότι η Ιαπωνία έχει καταφέρει να συγκεντρώσει 27 χρυσά μετάλλια έναντι των 20 των Ρώσων Αθλητών. Παρόμοια, μια χώρα που συγκέντρωσε περισσότερα αργυρά βρίσκεται σε υψηλότερη θέση από μια άλλη που συγκέντρωσε περισσότερα χάλκινα, όπως για παράδειγμα συμβαίνει με τη Γαλλία και τη Γερμανία.

Πίνακας μεταλλίων Θερινών Ολυμπιακών Αγώνων 2020

Θέση	Χώρα	Χρυσά	Αργυρά	Χάλκινα	Σύνολο
1	Ηνωμένες Πολιτείες (USA)	39	41	33	113
2	Κίνα (CHN)	38	32	18	88
3	Ιαπωνία (JPN)*	27	14	17	58
4	Μεγάλη Βρετανία (GBR)	22	21	22	65
5	Ρώσοι αθλητές (ROC)	20	28	23	71
6	Αυστραλία (AUS)	17	7	22	46
7	Ολλανδία (NED)	10	12	14	36
8	Γαλλία (FRA)	10	12	11	33
9	Γερμανία (GER)	10	11	16	37
10	Ιταλία (ITA)	10	10	20	40
11	Καναδάς (CAN)	7	6	11	24
12	Βραζιλία (BRA)	7	6	8	21
13	Νέα Ζηλανδία (NZL)	7	6	7	20
14	Κούβα (CUB)	7	3	5	15
15	Ουγγαρία (HUN)	6	7	7	20
16	Νότια Κορέα (KOR)	6	4	10	20
17	Πολωνία (POL)	4	5	5	14
18	Τσεχία (CZE)	4	4	3	11
19	Κένυα (KEN)	4	4	2	10
20	Νορβηγία (NOR)	4	2	2	8

Εικόνα 3.2: Πίνακας μεταλλίων Θερινών Ολυμπιακών αγώνων 2020

3.2 Ψευδοκώδικας Υλοποίησης

Έχοντας σαν κεντρική ιδέα τα όσα περιεγράφηκαν παραπάνω αναπτύχθηκε υλοποίηση για την παραγωγή των συστάσεων αλλά και την αξιολόγηση αυτών που προέκυψαν έπειτα από την εφαρμογή της σε ένα μέρος δεδομένων που χρησιμοποιήθηκε για έλεγχο.

Μετά την επεξεργασία των δεδομένων και την παραγωγή των συνδυαστικών κανόνων, εφαρμόζοντας Εξόρυξης Δεδομένων και παράλληλα τα εργαλεία και τις τεχνικές που διαθέτει το RStudio, εισήχθησαν μερικές παράμετροι εισόδου. Αυτές είναι :

tts (test transactions size: 1,2,...tts): Αποτελεί το πλήθος των επαναλήψεων που θα πραγματοποιηθούν για την παραγωγή τυχαίων συναλλαγών κατά το στάδιο της αξιολόγησης του μοντέλου

k (body size): το πλήθος των στοιχείων που θα απαρτίζουν την κάθε συναλλαγή η οποία θα παραχθεί κατά το στάδιο αξιολόγησης που μοντέλου

d (depth): το πλήθος των ισχυρότερων συστάσεων που προκύπτουν ταξινομημένες σύμφωνα με τη μεθόδους των μεταλλίων αλλά και με τις μεθόδους των απολύτων τιμών.

ct (conviction threshold): το κατώφλι του μέτρου conviction που εφαρμόζεται στους κανόνες που προκύπτουν και θα περιορίσουν το σύνολό τους αποτελώντας ένα νέο υποσύνολο κανόνων.

st (support threshold): το κατώφλι του μέτρου support που εφαρμόζεται στους κανόνες που προκύπτουν και θα περιορίσουν το σύνολό τους αποτελώντας ένα νέο υποσύνολο κανόνων.

lt (lift threshold) : το κατώφλι του μέτρου lift που εφαρμόζεται στους κανόνες που προκύπτουν και θα περιορίσουν το σύνολό τους αποτελώντας ένα νέο υποσύνολο κανόνων.

ni (distinct items in testing set) : το σύνολο των μοναδικών στοιχειοσυνόλων που απαρτίζουν το σύνολο των συναλλαγών ελέγχου.

Εκτός από τις παραμέτρους εισόδου παρακάτω εξηγούνται επιπλέον παράμετροι που χρησιμοποιούνται στον κώδικα:

stripped_AR: Υποσύνολο κανόνων που προκύπτει από το αρχικό σύνολο που παράχθηκε από τα δεδομένα της εκπαίδευσης και αφού εφαρμόστηκαν σε αυτούς τα κατώφλια του conviction support, lift (ct,st,lt παράμετροι εισόδου)

test_set: Σύνολο δεδομένων ελέγχου. Το ένα τέταρτο των αρχικών συναλλαγών από τα προς ανάλυση δεδομένα. Η δημιουργία τους γίνεται με τυχαιοποιημένη επιλογή.

test_bag: Υποσύνολο από τα δεδομένα ελέγχου (test_bag). Χαρακτηριστικό του υποσυνόλου αυτού είναι ότι παράγεται με τυχαίο τρόπο από τις μοναδικές κατηγορίες προϊόντων που υπάρχουν στις συναλλαγές ελέγχου, προκύπτει δηλαδή με τυχαίο τρόπο από τα στοιχεία του **ni**. Επιπλέον πρέπει να υπάρχουν στο σώμα των κανόνων του stripped_AR.

MSR: Medal Standings Recommendations – Συστάσεις που προκύπτουν με τη μέθοδο των μεταλλίων.

CR: Classic Recommendations- Συστάσεις που προκύπτουν εφαρμόζοντας της μέθοδο ταξινόμησης των κανόνων πρώτα ως προς την τιμή του conviction και έπειτα ως προς την τιμή του support.

Mismatches: Πλήθος διαφορών ανάμεσα στις μεθόδους

Matches: Πλήθος ομοιοτήτων ανάμεσα στις μεθόδους

Hits: το σύνολο των συναλλαγών του test_set που συμπεριλαμβάνουν την υπό εξέταση συναλλαγή του test_bag και σύστασή της.

Επομένως, με τους κανόνες και τις παραμέτρους εισόδου αναπτύσσεται κώδικας. Ο κώδικας αποτελείται από δύο τμήματα. Το ένα μέρος αφορά τον υπολογισμό των συστάσεων με τη μέθοδο των μεταλλίων και το δεύτερο την αξιολόγηση των συστάσεων αυτών. Παρακάτω επισημαίνονται τα δύο αναφερόμενα τμήματα σε μορφή ψευτοκώδικα, το πρώτο από τη γραμμή 1 έως και τη γραμμή 16 και το δεύτερο από τη γραμμή 17 και έπειτα.

Ο ψευτοκώδικας συνίσταται από τα παρακάτω βήματα:

1. Μετά την δημιουργία των κανόνων και την εισαγωγή τιμών στις μεταβλητές εισόδου σχηματίζεται υποσύνολο κανόνων (stripped_AR) περιορίζοντας τις τιμές του conviction, support και lift από τις παραμέτρους των κατωφλίων. (γραμμή 1)

2. Για τόσες επαναλήψεις όσες η τιμή της μεταβλητής tts σχηματίζεται με τυχαίο τρόπο συναλλαγή με τα εξής χαρακτηριστικά (γραμμές 2-5) :

   - στοιχεία της είναι κάποια από τις μοναδικές κατηγορίες που απαρτίζουν τις συναλλαγές ελέγχου και έχουν αποθηκευτεί στο διάνυσμα [1,2,…,ni]
   - το πλήθος των στοιχείων της είναι ίσο με την τιμή της μεταβλητής εισόδου k

3. Για την παραπάνω συναλλαγή ελέγχεται αν τα στοιχεία που την σχημάτισαν συμμετέχουν στο σώμα του υποσυνόλου των κανόνων (stripped_AR). Οι συναλλαγές αυτές που συμμετέχουν στους κανόνες σχηματίζουν ένα νέο σύνολο συναλλαγών ελέγχου το test_bag.(γραμμές 6-12)

4. Παράγονται συστάσεις για τις συναλλαγές του test_bag με τη μέθοδο των μεταλλίων και τις κλασικές μεθόδους.(γραμμές 13-16)

5. Ορίζεται βάθος (d) των πιο ισχυρών συστάσεων όπως προέκυψαν με την ταξινόμηση των μεθόδων που εξετάζονται. Για το βάθος αυτό υπολογίζονται οι ομοιότητες και οι διαφορές μεταξύ των συστάσεων που προέκυψαν και αποθηκεύονται σε μεταβλητές matches και mismatches.(γραμμές 17-22)

6. Για κάθε μια συναλλαγή του test_bag υπολογίζονται (γραμμές 23-27)

   - το σύνολο των συναλλαγών του αρχικού συνόλου ελέγχου (test_set) που τις περιέχουν και αποθηκεύεται ως hits.
   - το σύνολο από τις συναλλαγές αυτές που περιέχουν επιπλέον και κάποια από τις συστάσεις που προέκυψαν ταξινομημένες έως βάθος d και αποθηκεύεται ως *μέθοδος*_recs

7. Υπολογίζεται για την κάθε μια συναλλαγή το ποσοστό του *μέθοδος*_recs/hits(γραμμές 29-30)

8. Υπολογίζεται ο μέσος όρος του ποσοστού του βήματος 7 ως ποσοστό επιτυχίας ανά μέθοδο (γραμμές 31-32).

```
Apply ct and st to AR, produce stripped_AR
counter=1
k_ids_counter=0
While counter < tts+1
  {k_ids: k random item ids in [1,...,ni] without replacement
   If exist appropriate rules in stripped_AR then
   {retain k_ids in test_bag;
     k_ids_counter++
    }
   Save the test_bag;
   Save the k_ids_counter;
}
For each member in test_bag
{
  create_medal_standings_recommendations (MSR)
  create_classical_recommendations (CR)
  Up to depth=d
  {
    Number of matches between MSR and CR
    Number of mismatches between MSR and CR
  }
}
For each member of the test_bag
{
  Calculate the percentage of hits by considering the top d of MSR in
test_set
  Calculate the percentage of hits by considering the top d of CR  in
test_set
}
Calculate and save the following:
Average Number of matches between MSR and CR
Average Number of mismatches between MSR and CR
AVG by considering the top d of MSR
AVG by considering the top d of CR
```

4

Υλοποίηση της Μεθοδολογίας Medal Standings

Το Κεφάλαιο αυτό πραγματεύεται τις τεχνικές και τα εργαλεία που χρησιμοποιήθηκαν προκειμένου να υλοποιηθεί το προτεινόμενο μοντέλο-πλαίσιο με τη βοήθεια του οποίου έγινε η επεξεργασία των δεδομένων, υπολογίστηκαν τα αντίστοιχα αποτελέσματα και απεικονίστηκαν γραφικά ώστε να είναι πιο κατανοητά στον αναγνώστη. Αναφέρονται βασικά στοιχεία της γλώσσας προγραμματισμού R που χρησιμοποιήθηκε, παρουσιάζονται τα πακέτα που χρησιμοποιήθηκαν για την ανάλυση καθώς και το περιβάλλον RStudio που αποτελεί το περιβάλλον ανάπτυξης της εφαρμογής. Τέλος, παρουσιάζεται και αναλύεται η μεθοδολογία που ακολουθήθηκε για την υλοποίηση των ανωτέρω.

4.1 Γλώσσα Προγραμματισμού R

Για την υλοποίηση του κώδικα η οποία αναλύεται παρακάτω χρησιμοποιήθηκε η γλώσσα προγραμματισμού R. Η R είναι μια γλώσσα με ποικίλες δυνατότητες και χρησιμοποιείται κατά κύριο λόγο για την υλοποίηση αλγορίθμων που αφορούν τη Μηχανική Μάθηση αλλά και σε κάθε μορφής έρευνα με στατιστικό περιεχόμενο και υπολογισμούς, για την παραγωγή γραφικών απεικονίσεων, για την επεξεργασία και ανάλυση των δεδομένων κατά την Εξόρυξη Δεδομένων. Επίσης είναι μια γλώσσα που χάρις στην ευκολία εκμάθησής της, τη συμβατότητά της με τα πιο διαδεδομένα και κοινά λειτουργικά συστήματα, την πληθώρα των έτοιμων πακέτων που διαθέτει αλλά και λόγω της δωρεάν διάθεσής την κατέστησαν ως μια γλώσσα δημοφιλή[3].

Ένα από τα πλεονεκτήματα της R είναι ότι δεν απαιτείται να δηλωθεί ρητά η κλάση στην οποία ανήκουν τα αντικείμενα αλλά ο τύπος τους καθορίζεται αυτόματα ανάλογα την τιμή που θα καταχωρηθεί στο εκάστοτε αντικείμενο. Διαθέτει πέντε βασικές κλάσεις:

[3] 'R: The R Project for Statistical Computing'. https://www.r-project.org/ (ημερομηνία Πρόσβασης 27 Αυγούστου 2020)

- Χαρακτήρας
- Αριθμός (αριθμητικός ή πραγματικός)
- Ακέραιος
- Σύνθετος
- Λογικός

Επιπλέον, στην R χρησιμοποιούνται διανύσματα τα οποία διευκολύνουν με τη χρήση τους υπολογισμούς και πράξεις χωρίς πολυπλοκότητα. Αξίζει επίσης να αναφερθεί η ύπαρξη ειδικών τιμών όπως είναι η τιμή NA η οποία αναπαριστά μη ορισμένη τιμή. Για τον έλεγχο αυτών των απόντων τιμών υπάρχουν συναρτήσεις όπως η is.na()[4]. Στην R υπάρχει η δυνατότητα αποθήκευσης δεδομένων σε μορφή πίνακα και για το σκοπό αυτό χρησιμοποιούνται τα πλαίσια δεδομένων (data frames) όπου σε κάθε στήλη μπορεί να της ανατεθεί ένα όνομα και με τη χρήση του να γίνεται αναφορά στη συγκεκριμένη στήλη.

Μια ακόμη από τις βασικές εργασίες που εκτελεί η R είναι η ανάγνωση και εξαγωγή δεδομένων από κάποιο αρχείο. Συνηθέστεροι τύποι αρχείων που χρησιμοποιούνται για αυτό το σκοπό είναι σε αρχεία .csv και xls ή xlsx. Ο πιο διαδεδομένος τρόπος ανάγνωσης των αρχείων .csv είναι με τη χρήση της συνάρτησης read.csv()[5], η οποία δέχεται ως βασικά ορίσματα το όνομα το αρχείου, ένα λογικό όρισμα, το οποίο σηματοδοτεί αν το αρχείο έχει γραμμή κεφαλίδας ή όχι, ένα αλφαριθμητικό, το οποίο ορίζει με τι διαχωρίζονται οι στήλες, π.χ. κόμμα, κενό κ.λπ. και ένα ακόμα λογικό όρισμα όπου η προεπιλεγμένη τιμή είναι TRUE. Παρακάτω αποτυπώνεται ο τρόπος σύνταξης της συνάρτησης:

```
pos_data<-read.csv("pos_data.csv",TRUE,";", stringsAsFactors = TRUE)
```

Η R διαθέτει επίσης δομές ελέγχου, όπως είναι η υπό συνθήκη εκτέλεση (if else) και οι βρόχοι επανάληψης (for). Αυτές οι δομές ελέγχου είναι πολύ απλές, αλλά και πολύ χρήσιμες. Οι επαναληπτικές δομές από την άλλη έχουν ως στόχο την εκτέλεση ενός κομματιού κώδικα για προκαθορισμένο ή μη αριθμό επαναλήψεων. Στην R, ο βρόχος επανάληψης for μπορεί να χρησιμοποιηθεί είτε με τον κλασσικό τρόπο, όπου μια μεταβλητή παίρνει τιμές από ένα εύρος τιμών που καθορίζεται εξαρχής σε κάθε επανάληψη, είτε μια μεταβλητή παίρνει τιμές από τα στοιχεία ενός συνόλου αντικειμένων, όπως για παράδειγμα από τα στοιχεία μια στήλης ενός πίνακα.

Η R διαθέτει πληθώρα συναρτήσεων οι οποίες είναι διαθέσιμες αφού ολοκληρωθεί η εγκατάσταση των απαιτούμενων πακέτων. Σε κάθε ένα από τα πολλά έτοιμα πακέτα που είναι διαθέσιμα, υπάρχουν διαφορετικές συναρτήσεις, οι οποίες χρησιμοποιούνται για διαφορετική

[4] 'Market basket introduction | R'. https://campus.datacamp.com/courses/market-basket-analysis-in-r/introduction-to-market-basket-analysis?ex=1 (ημερομηνία πρόσβασης 14 Μαΐου 2020)

[5] Readxl.tidyverse.org. 2021. Read Excel Files. [online] Available at: <https://readxl.tidyverse.org/> [Accessed 15 January 2021].

λειτουργία η καθεμιά. Εκτός όμως από τις έτοιμες συναρτήσεις της, η R επιτρέπει στον χρήστη να ορίσει τις δικές του συναρτήσεις. Η εγκατάσταση πακέτων γίνεται με τη συνάρτηση install.packages(), ενώ η εισάγεται στο περιβάλλον με τη συνάρτηση library(). Κάποια από τα πακέτα που χρησιμοποιήθηκαν στην παρούσα υλοποίηση του κώδικα είναι:

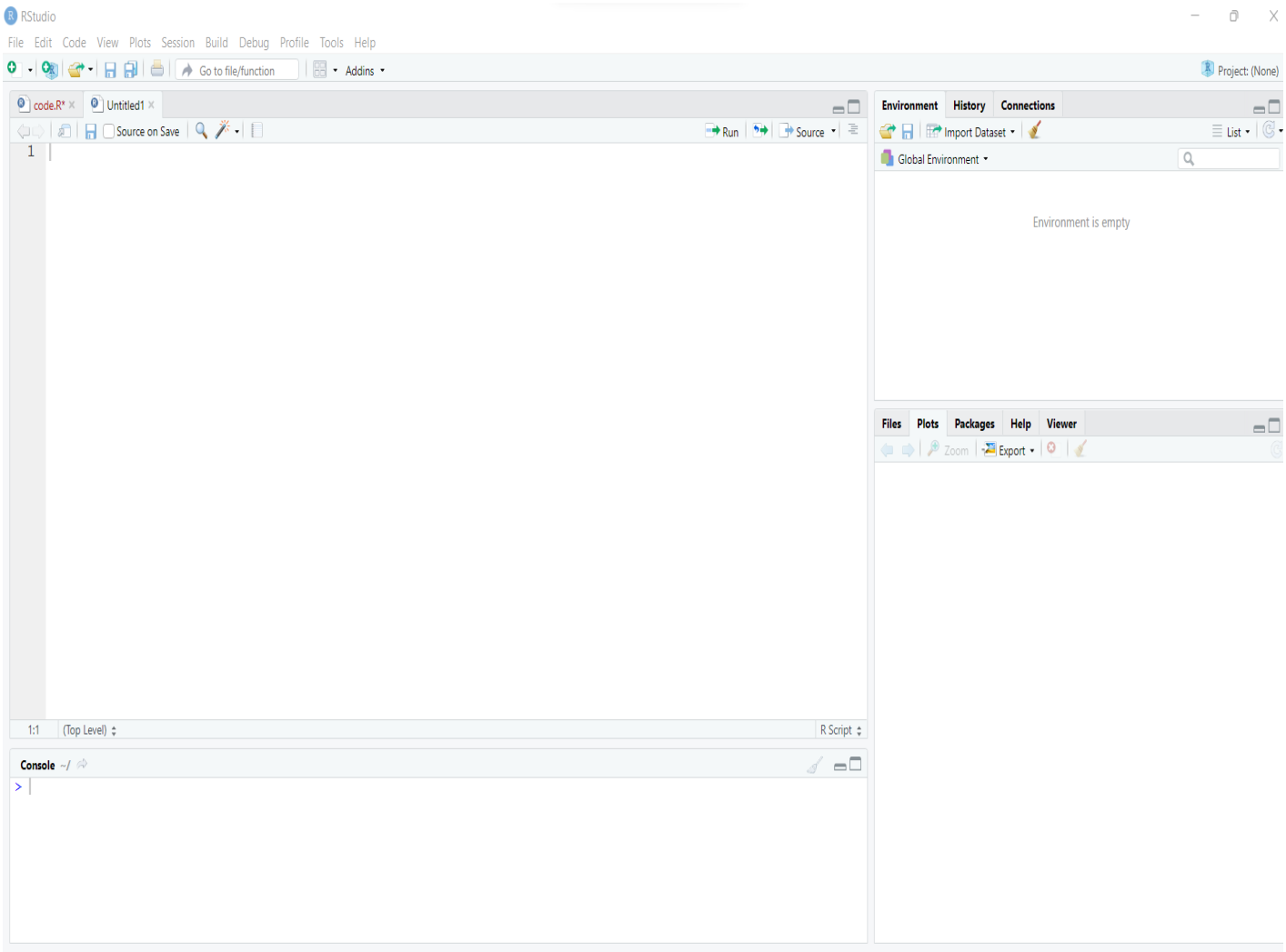
- Το πακέτο dplyr[6] χρησιμοποιείται για τον εύκολο χειρισμό των δεδομένων. Παρέχει 5 συναρτήσεις, οι οποίες καλύπτουν τις θεμελιώδεις εργασίες διαχείρισης δεδομένων. Αυτές είναι οι:
  i) select, για επιλογή-φιλτράρισμα στηλών του συνόλου δεδομένων,
  ii) filter, για επιλογή-φιλτράρισμα γραμμών του συνόλου δεδομένων,
  iii) arrange, για ταξινόμηση των γραμμών βάσει τιμών συγκεκριμένων στηλών,
  iv) mutate, για δημιουργία νέων μεταβλητών από τις ήδη υπάρχουσες,
  v) summarize, για συνάθροιση δεδομένων
- Το πακέτο plyr το οποίο παρέχει ένα σύνολο εργαλείων για διαχωρισμό δομών δεδομένων σε ομοιογενή κομμάτια, εφαρμογή συναρτήσεων σε κάθε κομμάτι και εν συνεχεία συνδυασμό όλων των αποτελεσμάτων μαζί
- Το πακέτο stringr παρέχει ένα συνεκτικό σύνολο λειτουργιών που έχουν σχεδιαστεί για να κάνουν την εργασία με συμβολοσειρές (strings) όσο το δυνατόν πιο εύκολη
- Το πακέτο arules [7] παρέχει στους χρήστες της R έτοιμες συναρτήσεις για την εξόρυξη συχνών στοιχειοσυνόλων και συνδυαστικών κανόνων. Η βασικότερη συνάρτηση του πακέτου αυτού είναι η apriori() , η οποία δέχεται 4 ορίσματα:
  i. data, τα δεδομένα από τα οποία θέλουμε να εξάγουμε τα στοιχειοσύνολα ή τους συνδυαστικούς κανόνες,
  ii. parameter, μια λίστα παραμέτρων, όπως το κατώφλι support και confidence,
  iii. appearance, μια λίστα παραμέτρων, που καθορίζει περιορισμούς πάνω σε στοιχεία ή και μέλη των κανόνων,
  iv. control, μια λίστα παραμέτρων, που αφορά περιορισμούς της επίδοσης του αλγορίθμου.

Το RStudio είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης για την R και το πλέον διαδεδομένο. Διατίθεται δωρεάν και διευκολύνει τον χρήστη με την απλή και εύχρηστη διεπαφή που έχει. Είναι σημαντικό να επιλεγεί η σωστή έκδοση ανάλογα με το διαθέσιμο λειτουργικό σύστημα στο οποίο θα εγκατασταθεί. Η αρχική οθόνη του Rstudio είναι όπως φαίνεται στην Εικόνα 4.1. Στο τμήμα πάνω αριστερά φαίνεται ο κώδικας των αρχείων που είναι ανοιγμένα. Κάθε καρτέλα αποτελεί και διαφορετικό αρχείο κώδικα. Στο πλαίσιο πάνω δεξιά εμφανίζονται οι μεταβλητές και οι συναρτήσεις. Στο επόμενο πλαίσιο κάτω δεξιά, στην καρτέλα “Plots”, εκτυπώνονται οι γραφικές παραστάσεις, ενώ μπορούμε να δούμε και ποια πακέτα έχουμε κατεβάσει ή χρειάζονται ενημέρωση μέσω της καρτέλας “Packages”. Επιπλέον, μέσω της καρτέλας “Help” μπορούμε να βρούμε πληροφορίες και βοήθεια για κάποια

[6] Dplyr.tidyverse.org. 2021. A Grammar Of Data Manipulation. [online] Available at: <https://dplyr.tidyverse.org/> [Accessed 15 January 2021].

[7] ‘Association Mining With R | arules’. http://r-statistics.co/Association-Mining-With-R.html (ημερομηνία πρόσβασης 25 Αυγούστου 2020).

συνάρτηση ή πακέτο. Τέλος, στο πλαίσιο αριστερά και κάτω βρίσκεται η κλασσική κονσόλα της R.



Εικόνα 4.1: Αρχική Οθόνη RSTUDIO

4.2 Τεχνικές

Οι κυριότερες τεχνικές στις οποίες βασίστηκε η μελέτη και αναπτύχθηκε ο κώδικας αφορούν στον μετασχηματισμό των δεδομένων, στην κανονικοποίηση τους και στη διακριτοποίηση. Ο μετασχηματισμός των δεδομένων χρησιμοποιείται κυρίως για την εξομάλυνση των δεδομένων και για την απομάκρυνση θορύβου, για τη συγκέντωση των δεδομένων, την τακτοποίηση των χαρακτηριστικών του συνόλου δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών με σκοπό τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα. Η πιο συχνή εφαρμογή του μετασχηματισμού δεδομένων είναι η τυποποίηση (standardization) και η δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα, η οποία είναι ιδιαίτερα χρήσιμη σε προβλήματα

κατηγοριοποίησης. Μια ακόμα ειδική μορφή μετασχηματισμού των δεδομένων αποτελεί η διακριτοποίηση (discretization)[15].

Επιπλέον, η τεχνική που αποτέλεσε την κεντρική ιδέα στην παρούσα διπλωματική εργασία είναι η κατηγοριοποίηση η οποία αποτελεί μια από τις βασικές εργασίες στο στάδιο της Εξόρυξης Δεδομένων. Βασίζεται στην εξέταση των τιμών μιας αριθμητικής μεταβλητής και με βάση αυτές την αντιστοίχιση σε ένα προκαθορισμένο σύνολο κατηγοριών (κλάσεων). Έχοντας ένα σύνολο από κατηγορίες (κλάσεις) και ένα σύνολο δεδομένων με δείγματα, για τα οποία ξέρουμε σε ποια κλάση ανήκουν, στόχος της κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να κατηγοριοποιήσει αυτόματα σε αυτές τις κατηγορίες νέα, άγνωστα και μη-κατηγοριοποιημένα δείγματα και έπεται η πρόβλεψη.

5 *Επεξεργασία των Δεδομένων - Αποτελέσματα*

Στο κεφάλαιο αυτό παρουσιάζεται ο τρόπος με τον οποίο έγινε η επεξεργασία των δεδομένων και τα αποτελέσματα τα οποία προέκυψαν από τον αλγόριθμο που αναπτύχθηκε,

5.1 Επεξεργασία Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν αφορούν συναλλαγές που πραγματοποίησαν πελάτες ενός καταστήματος υπεραγοράς, supermarket. Το σύνολό τους αποτελείται από τρία αρχεία μορφής CSV:

pos_data: περιγράφει την καθεμία συναλλαγή του πελάτη με κύρια χαρακτηριστικά τον κωδικό πελάτη (CUSTOMERID), κωδικό συναλλαγής (TRANDSID), κωδικό προϊόντος (ITEMID), ημερομηνία (DATE), ώρα (TIME), ποσότητα (QUANTITY), τιμή (PRICE).

articles: αντιστοιχίζει τον κωδικό συναλλαγής με την κατηγορία προϊόντων και απαρτίζεται από τον κωδικό συναλλαγής (ITEMID) και την κατηγορία των αγαθών (ARTICLE_CATEGORY).

article_categories: περιγράφει την κάθε κατηγορία προϊόντων και για το σκοπό αυτό έχει ως στοιχεία τον κωδικό της κατηγορίας (ID) και την αντίστοιχη περιγραφή της (DESC).

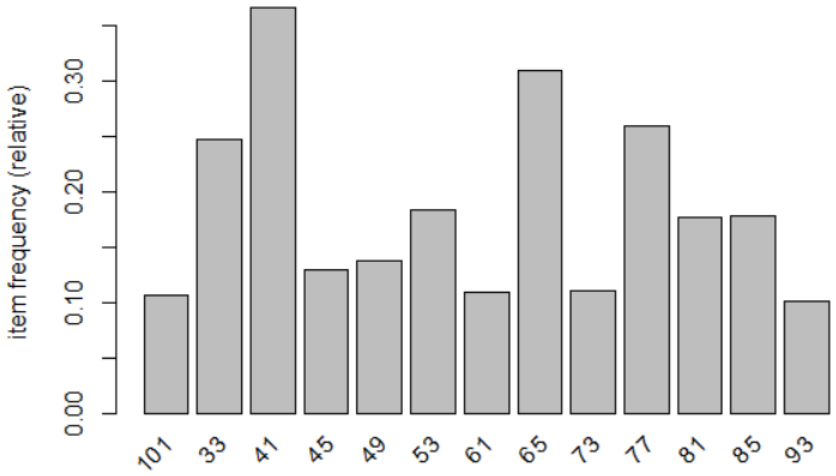
ΣΕ πρώτο στάδιο απαιτείται να γίνει επεξεργασία τους προκειμένου να είναι ολοκληρωμένη η πληροφορία «Συναλλαγή – Κατηγορία Προϊόντων. Για το σκοπό αυτό πραγματοποιήθηκε ενοποίηση των δεδομένων ώστε να προκύψει η μορφή που παρουσιάζονται στην Εικόνα 5.1.

	TRANSID	ITEM_CATEGORIES	id
1	22	65,45	1
2	24	65,33,81,53,49,169,265,281	2
3	160	41,197	3
4	283	77	4
5	291	77	5
6	293	41,77,249,73,145,101,45	6
7	294	73,33,57,81,85	7
8	299	61,281	8
9	302	81	9
10	306	341	10
11	552	77,33,85	11
12	555	33,193,57,225,341	12
13	560	41,33,77,53,93,145,205,417,61,85,97	13
14	562	209,65,309,77	14

Εικόνα 5.1:Τελική μορφή των προς επεξεργασία δεδομένων

Είναι σημαντικό πέραν της ανάλυσης και του υπολογισμού των αποτελεσμάτων να πραγματοποιηθεί και αξιολόγηση αυτών. Η αξιολόγηση θα επιτευχθεί με την χρήση δεδομένων ελέγχου. Για το σκοπό αυτό τα αρχικά δεδομένα-συναλλαγές χωρίζονται σε δύο ομάδες, την ομάδα που αποτελεί τα δεδομένα εκπαίδευσης και μια άλλη ομάδα την οποία αποτελούν τα δεδομένα ελέγχου.

Η επιλογή των στοιχείων για τον επιμερισμό τους στις δύο ομάδες γίνεται με τυχαίο τρόπο και η αναλογία που τηρήθηκε είχε ως εξής: τα τρία τέταρτα του συνόλου των συναλλαγών αποτελεί τα στοιχεία εκπαίδευσης και το υπόλοιπο ένα τέταρτο αποτελούν τα στοιχεία ελέγχου.



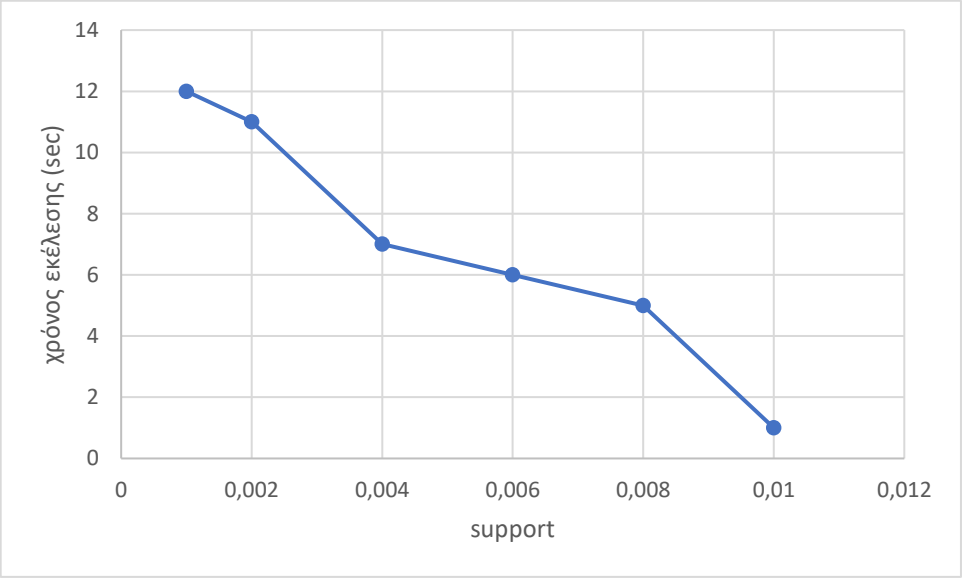
Εικόνα 5.2: Οι 10 πιο συχνές κατηγορίες προϊόντων των συναλλαγών εκπαίδευσης

Στην Εικόνα 5.2 απεικονίζονται οι δέκα πιο συχνές κατηγορίες προϊόντων που απαρτίζουν τις συναλλαγές εκπαίδευσης. Είναι οι κατηγορίες εκείνες που εμπλέκονται και στις περισσότερους συναλλαγές, όπου υπάρχει κάποιο προϊόν τους στα περισσότερα καλάθια αγορών [8].

```
> summary(pos_data_train)
transactions as itemMatrix in sparse format with
 25275 rows (elements/itemsets/transactions) and
 75 columns (items) and a density of 0.05148592

most frequent items:
     41      65      77      33      53 (Other)
   9263    7849    6558    6263    4662   63003

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   23   24
7918 4354 2944 2097 1705 1425 1125  884  756  621  427  327  238  173  102   84   46   30   10    3    3    2    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   3.000   3.861   5.000  24.000
```

Εικόνα 5.3: Το στατιστικό προφίλ των δεδομένων εισόδου

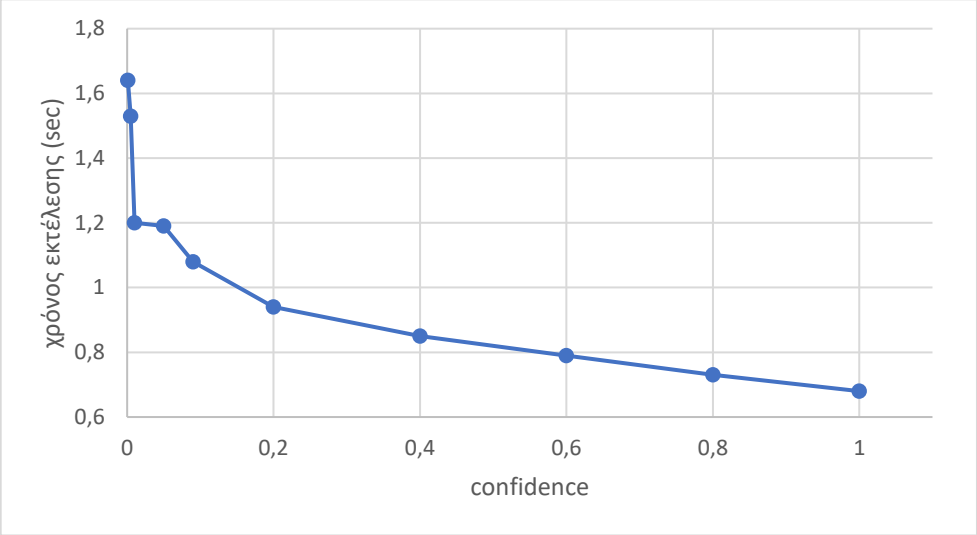
Από τα δεδομένα εκπαίδευσης και με τη χρήση του αλγορίθμου Apriori παράγονται οι συνδυαστικοί κανόνες. Κατά την εφαρμογή του αλγορίθμου επιλέχθηκαν οι τιμές για το support και confidence να είναι 0.003 και 0.5, αντίστοιχα. Η επιλογή των τιμών έγινε λαμβάνοντας υπόψιν παραμέτρους όπως ο χρόνος εκτέλεσης του αλγορίθμου και το επιθυμητό πλήθος των παραγόμενων κανόνων (ώστε να μην είναι υπερβολικά μεγάλο). Στην Εικόνα 5.4 παρουσιάζεται η μεταβολή του χρόνου εκτέλεσης του αλγορίθμου Apriori κρατώντας σταθερή την τιμή του confidence ίση με 0.5 και μεταβάλλοντας την τιμή του support. Όσο μεγαλύτερη είναι η τιμή του support τόσο μικρότερος είναι ο χρόνος εκτέλεσης του αλγορίθμου ενώ όταν η τιμή του ελαττώνεται ο χρόνος εκτέλεσης του αλγορίθμου αυξάνει.



Εικόνα 5.4:Χρόνος εκτέλεσης αλγορίθμου apriori σε σχέση με το support και σταθερό confidence =0.5

[8] 'summary function | R Documentation'. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary (ημερομηνία πρόσβασης 25 Αυγούστου 25, 2020).

Αντίστοιχα μελετήθηκε η μεταβολή του χρόνου εκτέλεσης του αλγορίθμου σε σχέση με τη μεταβολή του confidence κρατώντας σταθερή την τιμή του support ίση με 0.003. Τα αποτελέσματα απεικονίζονται στο διάγραμμα της Εικόνας 5.5.



Εικόνα 5.5:Χρόνος εκτέλεσης αλγορίθμου apriori σε σχέση με το confidence και σταθερό support = 0,003

Συγκρίνοντας το χρόνο εκτέλεσης του αλγορίθμου στις δυο περιπτώσεις παρατηρείται πως μεταβάλλοντας την τιμή του confidence ο χρόνος εκτέλεσης του αλγορίθμου παραμένει εξαιρετικά μικρός ενώ πιο αισθητή είναι η μεταβολή του σε σχέση με τη μεταβολή της τιμής του support. Επομένως η τιμή του support είναι εκείνη που θα καθορίσει την ταχύτητα εκτέλεσής σου και άρα και το χρόνο που θα απαιτηθεί για την ολοκλήρωσή του.

Εφόσον παραχθούν οι κανόνες, σημαντικό είναι να αφαιρεθούν από το σύνολό τους όσοι είναι πλεονάζοντες. Οι πλεονάζοντες κανόνες εντοπίζονται με τη συνάρτηση is.redundant() [9]. Ένας κανόνας μπορεί να οριστεί ως πλεονάζων εάν έχει την ίδια κεφαλή με κάποιον άλλο κανόνα ,το σώμα του ενός είναι υπερσύνολο του άλλου αλλά και η τιμή του confidence του κανόνα είναι μικρότερη από την τιμή του confidence του άλλου κανόνα. Δηλαδή αν ισχύουν οι τρείς συνθήκες που αναφέρονται παρακάτω:

$$L => R, L_1 => R$$

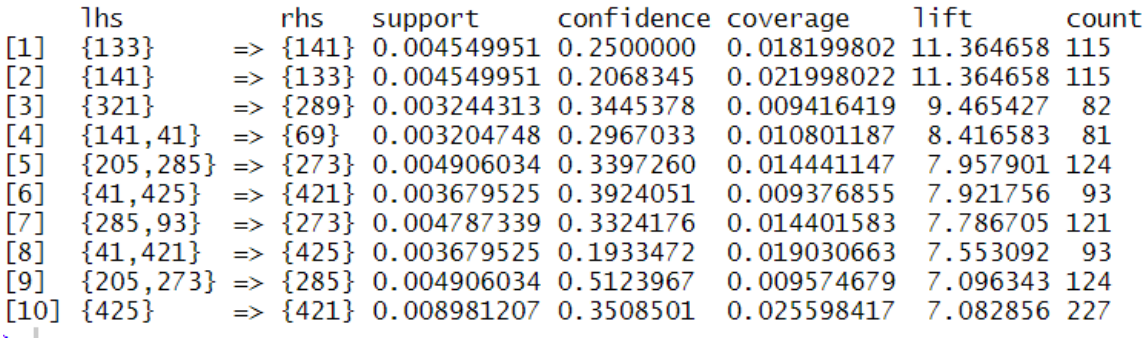
$$conf(L_1 => R) \leq conf(L => R) \qquad (1)$$

$$L \subset L_1$$

Με την επιλογή των παραπάνω αναφερόμενων τιμών για το support και confidence 0.003 και 0.5 αντίστοιχα παράγονται 72000 κανόνες περίπου και σχεδόν 20000 κανόνες χαρακτηρίστηκαν ως πλεονάζοντες και αφαιρέθηκαν από το σύνολο χωρίς να συμμετέχουν στη διαδικασία παραγωγής συστάσεων και στην ανάλυση που ακολουθεί.

[9] 'is.redundant function R Documentation'. https://www.rdocumentation.org/packages/arules/versions/1.6-6/topics/is.redundant

(ημερομηνία πρόσβασης 25 Αυγούστου 2020)



	lhs		rhs	support	confidence	coverage	lift	count
[1]	{133}	=>	{141}	0.004549951	0.2500000	0.018199802	11.364658	115
[2]	{141}	=>	{133}	0.004549951	0.2068345	0.021998022	11.364658	115
[3]	{321}	=>	{289}	0.003244313	0.3445378	0.009416419	9.465427	82
[4]	{141,41}	=>	{69}	0.003204748	0.2967033	0.010801187	8.416583	81
[5]	{205,285}	=>	{273}	0.004906034	0.3397260	0.014441147	7.957901	124
[6]	{41,425}	=>	{421}	0.003679525	0.3924051	0.009376855	7.921756	93
[7]	{285,93}	=>	{273}	0.004787339	0.3324176	0.014401583	7.786705	121
[8]	{41,421}	=>	{425}	0.003679525	0.1933472	0.019030663	7.553092	93
[9]	{205,273}	=>	{285}	0.004906034	0.5123967	0.009574679	7.096343	124
[10]	{425}	=>	{421}	0.008981207	0.3508501	0.025598417	7.082856	227

Εικόνα 5.6: Κανόνες με την υψηλότερη τιμή lift

Στην εικόνα 5.6 παρουσιάζονται από τους κανόνες που σχηματίστηκαν εκείνοι που έχουν την υψηλότερη τιμή του lift, με φθίνουσα ταξινόμηση.

Υπενθυμίζεται η ύπαρξη του μέτρου conviction, το οποίο όπως έχει αναφερθεί στο Κεφάλαιο 2 ποσοτικοποιεί την ισχύ του κανόνα. Μια υψηλή τιμή conviction σημαίνει ότι η κεφαλή του κανόνα εξαρτάται σε μεγάλο βαθμό από το σώμα του κανόνα. Η διαφορά του conviction και του lift είναι ότι το πρώτο κωδικοποιεί και το αιτιατό (causality) της συσχέτισης.

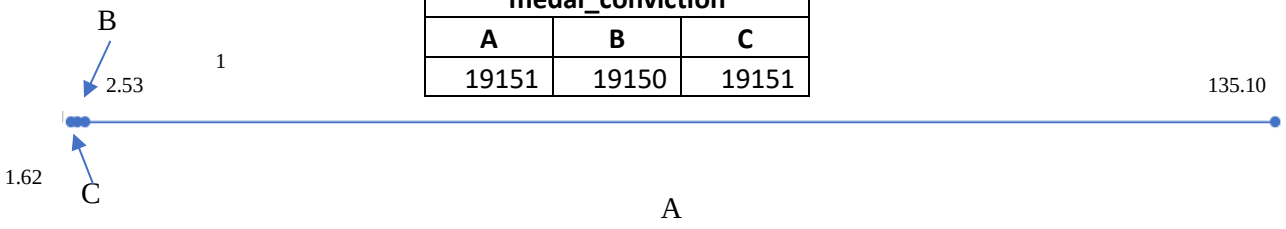
Για έναν κανόνα συσχέτισης , εάν η τιμή του lift είναι ίση με 1, σημαίνει ότι τα εμπλεκόμενα στοιχεία είναι ανεξάρτητα. Εάν η τιμή του lift είναι μεγαλύτερη από 1, σημαίνει ότι τα στοιχειοσύνολα συσχετίζονται θετικά και εάν η τιμή είναι χαμηλότερη από 1, σημαίνει ότι συσχετίζονται αρνητικά. Για το λόγο αυτό αφαιρούνται οι κανόνες που έχουν τιμή για lift μικρότερη του 1.

Η προτεινόμενη μέθοδος ανάλυσης των δεδομένων βασίζεται στη διακριτοποίηση. Η διακριτοποίηση είναι μια διαδικασία επεξεργασίας δεδομένων που μετατρέπει τα ποσοτικά δεδομένα σε ποιοτικά δεδομένα. Πολλοί είναι οι αλγόριθμοι διακριτοποίησης που έχουν προταθεί, κάποιοι χρησιμοποιούν τις πληροφορίες των ετικετών κλάσεων ενώ κάποιοι άλλοι όχι. Χρησιμοποιήθηκε η τεχνική των ίσων συχνοτήτων (frequency based). Κατά τη διακριτοποίηση ενός ποσοτικού χαρακτηριστικού, η διακριτοποίηση σταθερής συχνότητας προκαθορίζει μια επαρκή συχνότητα διαστήματος k και διαχωρίζει τις ταξινομημένες τιμές σε διαστήματα έτσι ώστε κάθε διάστημα να έχει περίπου τον ίδιο αριθμό k περιπτώσεων εκπαίδευσης. Μία άλλη τεχνική διακριτοποίησης είναι των ίσων διαστημάτων (Interval based), στην οποία διαχωρίζει τις ταξινομημένες τιμές σε k ίσα διαστήματα. Ακόμη μια άλλη τεχνική διακριτοποίησης είναι εκείνη όπου κάθε ένα διάστημα ορίζεται από τον χρήστη καθώς επιλέγει τα σημεία κοπής (user defined). Καθεμία από τις μεθόδους διακριτοποίησης παρέχει κάποιο είδος διακριτοποιημένων δεδομένων, για τα οποία υπάρχει το ερώτημα ποιο αποτελεί την καλύτερη επιλογή [13].

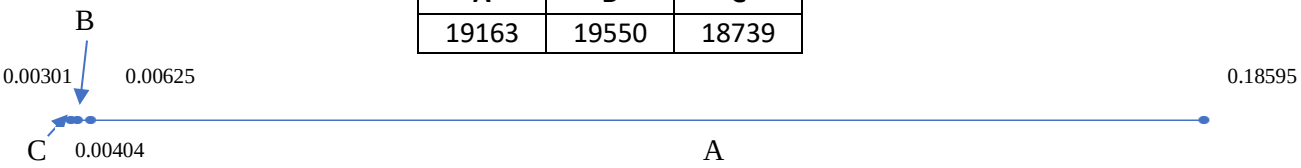
Διακριτοποιούμε τους κανόνες χρησιμοποιώντας μέθοδο ίσης συχνότητας με αριθμητικό k = 3. Υπολογίστηκαν τρία σημεία αποκοπής για κάθε χαρακτηριστικό και αξιολογήθηκαν οι κανόνες που προέκυψαν από τον αλγόριθμο Apriori βάσει των τριών διαστημάτων τιμών, καθένα από τα οποία πήρε σαν ετικέτα ένα από τα γράμματα A,B,C αντιπροσωπεύοντας μετάλλιο με ισχύς τέτοια όπου η συγκεκριμένη σειρά να δηλώνει φθίνουσα ταξινόμηση. Επομένως, οι πιο ισχυροί κανόνες για ένα μέτρο x ταξινομούνται στο μετάλλιο τύπου A και οι πιο ασθενείς στα μετάλλια τύπου C. Διακριτοποίηση πραγματοποιήθηκε για κάθε μια από τις παραμέτρους support, conviction και lift.

medal_lift		
A	**B**	**C**
19153	19148	19151

medal_conviction		
A	**B**	**C**
19151	19150	19151



medal_support		
A	**B**	**C**
19163	19550	18739



Εικόνα 5.7: Διαστήματα lift, conviction και support

Στην Εικόνα 5.7 απεικονίζονται τα διαστήματα των lift, conviction και support όπου έγινε η διακριτοποίησή τους καθώς και το πλήθος των κανόνων που αντιστοιχούν στο κάθε ένα διάστημα. Παρατηρείται ότι και για τις τρεις παραμέτρους οι κανόνες κατανέμονται σχεδόν ισομερώς στην κάθε μία κατηγορία μεταλλίων. Για τις παραμέτρους support και conviction η ανώτατη τιμή απέχει κατά πολύ από τις τιμές των επόμενων διαστημάτων ενώ το ίδιο δεν συμβαίνει στην παράμετρο lift. Λαμβάνοντας υπόψιν τα πλήθη των κανόνων που αντιστοιχίζονται σε κάθε διάστημα τα οποία είναι περίπου ίδια γίνεται αντιληπτό πως οι πιο αδύναμοι κανόνες συγκεντρώνονται σε πολύ χαμηλές τιμές σε σχέση με τους πιο ισχυρούς κανόνες.

body	head	support	confidence	coverage	lift	count	conviction	rulenumber	medallift_freq	medalconviction_freq	medalsupport_freq
101,33,49,61,65	41	0.008387735	0.9953052	0.008427300	2.721369	212	135.09804	47916	C	A	A
33,49,61,65,97	41	0.008348170	0.9952830	0.008387735	2.721309	211	134.46378	43108	C	A	A
33,45,49,53,65,77	41	0.008031652	0.9950980	0.008071217	2.720803	203	129.38967	57591	C	A	A
49,57,61,65,85	41	0.007952522	0.9950495	0.007992087	2.720670	201	128.12115	45879	C	A	A
33,53,61,81,97	41	0.007715134	0.9948980	0.007754698	2.720256	195	124.31557	43210	C	A	A
49,53,57,61,85	41	0.007636004	0.9948454	0.007675569	2.720112	193	123.04704	45865	C	A	A
33,45,49,65,81,85	41	0.007596439	0.9948187	0.007636004	2.720039	192	122.41278	57526	C	A	A
33,49,61,85,97	41	0.007556874	0.9947917	0.007596439	2.719965	191	121.77852	43050	C	A	A
53,61,65,81,97	41	0.007438180	0.9947090	0.007477745	2.719739	188	119.87573	43216	C	A	A
101,33,45,49,53	41	0.007438180	0.9947090	0.007477745	2.719739	188	119.87573	48297	C	A	A
101,33,53,61,65,85	41	0.007438180	0.9947090	0.007477745	2.719739	188	119.87573	56207	C	A	A
101,33,53,61,81	41	0.007359050	0.9946524	0.007398615	2.719585	186	118.60720	48024	C	A	A
57,61,65,77,85	41	0.007240356	0.9945652	0.007279921	2.719346	183	116.70441	46048	C	A	A
33,61,81,85,97	41	0.007200791	0.9945355	0.007240356	2.719265	182	116.07015	43160	C	A	A
101,33,53,61,77	41	0.006923838	0.9943182	0.006963403	2.718671	175	111.63031	48104	C	A	A
33,45,49,53,77,85	41	0.006923838	0.9943182	0.006963403	2.718671	175	111.63031	57567	C	A	A
45,53,61,77,85	41	0.006844708	0.9942529	0.006884273	2.718492	173	110.36178	50196	C	A	A
57,81,85,93	41	0.006805143	0.9942197	0.006844708	2.718401	172	109.72752	30215	C	A	A
101,33,53,81,97	41	0.006686449	0.9941176	0.006726014	2.718122	169	107.82473	41732	C	A	A

Εικόνα 5.8: Μετάλλια ανά κανόνα (παράδειγμα)

Στην Εικόνα 5.8 απεικονίζονται οι κανόνες που εξήχθησαν και πώς αυτοί αξιολογήθηκαν σύμφωνα με τη διαδικασία που περιγράφηκε. Στο αρχικό σύνολο κανόνων προστέθηκαν τρεις επιπλέον στήλες κάθε μία από τις οποίες αντιστοιχεί στα μετάλλια τύπου A,B,C ως προς lift (στήλη medallift_freq), ως προς conviction(στήλη medalconviction_freq) και ως προς support (στήλη medalsupport_freq). Σύμφωνα με την εικόνα ο κανόνας με αριθμό (rulenumber) 48565 έχει συγκεντρώσει ένα μετάλλιο τύπου A ως προς το conviction, ένα μετάλλιο τύπου A ως προς το support και ένα τύπου C ως προς το Lift . Επίσης διαπιστώνεται ότι, οι κανόνες με τις υψηλότερες τιμές ως προς τις προαναφερόμενες παραμέτρους αντιστοιχούν στα μετάλλια τύπου A ενώ κανόνες με τις χαμηλότερες τιμές αντιστοιχούν στα μετάλλια τύπου C.

Στην Εικόνα 5.9 απεικονίζονται οι συστάσεις του παραδείγματος της Εικόνας 3.1 για συναλλαγή {57,93,97} αυτή τη φορά ταξινομημένες με τη μέθοδο των μεταλλίων. Παρατηρείται η πιο ισχυρή σύσταση είναι η {41} με 3 μετάλλια τύπου A και 5 τύπου B ως προς conviction, 8 μετάλλια τύπου A ως προς support. Ακολουθεί η σύσταση {33}με 1 μετάλλιο τύπου A ως προς conviction και 1 μετάλλιο τύπου A ως προς support. Έπεται η σύσταση {53} με 4 μετάλλια τύπου B ως προς conviction και 4 μετάλλια τύπου A ως προς support. Η σύσταση {85} έχει 2 μετάλλια τύπου B και 7 τύπου C ως προς conviction και 7 μετάλλια τύπου A και 2 τύπου B ως προς support. Η σύσταση {65} έχει 10 μετάλλια τύπου C ως προς conviction και 10 μετάλλια τύπου A ως προς support. Η σύσταση {101} έχει 3 μετάλλια τύπου C ως προς conviction και 3 μετάλλια τύπου B ως προς support. Η σύσταση {81} και {77} έχει 1 μετάλλιο τύπου C ως προς conviction και 1 μετάλλιο τύπου B ως προς support.

	medalconviction_freq			medalsupport_freq		
head	A	B	C	A	B	C
41	3	5		8		
33	1			1		
53		4		4		
85		2	7	7	2	
65			10	10		
101			3		3	
81			1		1	
77			1		1	

Εικόνα 5.9: Ταξινόμηση με τη μέθοδο Medal Standing Recommendation

5.2 Αποτελέσματα

Η υλοποίηση περιλαμβάνει παραμέτρους εισόδου οι οποίες εισάγονται ελεύθερα από τον χρήστη και μελετάται ο βαθμός στον οποίο επηρεάζεται το αποτέλεσμα του αλγορίθμου από την επιλογή των τιμών που δίνονται στην κάθε παράμετρο.

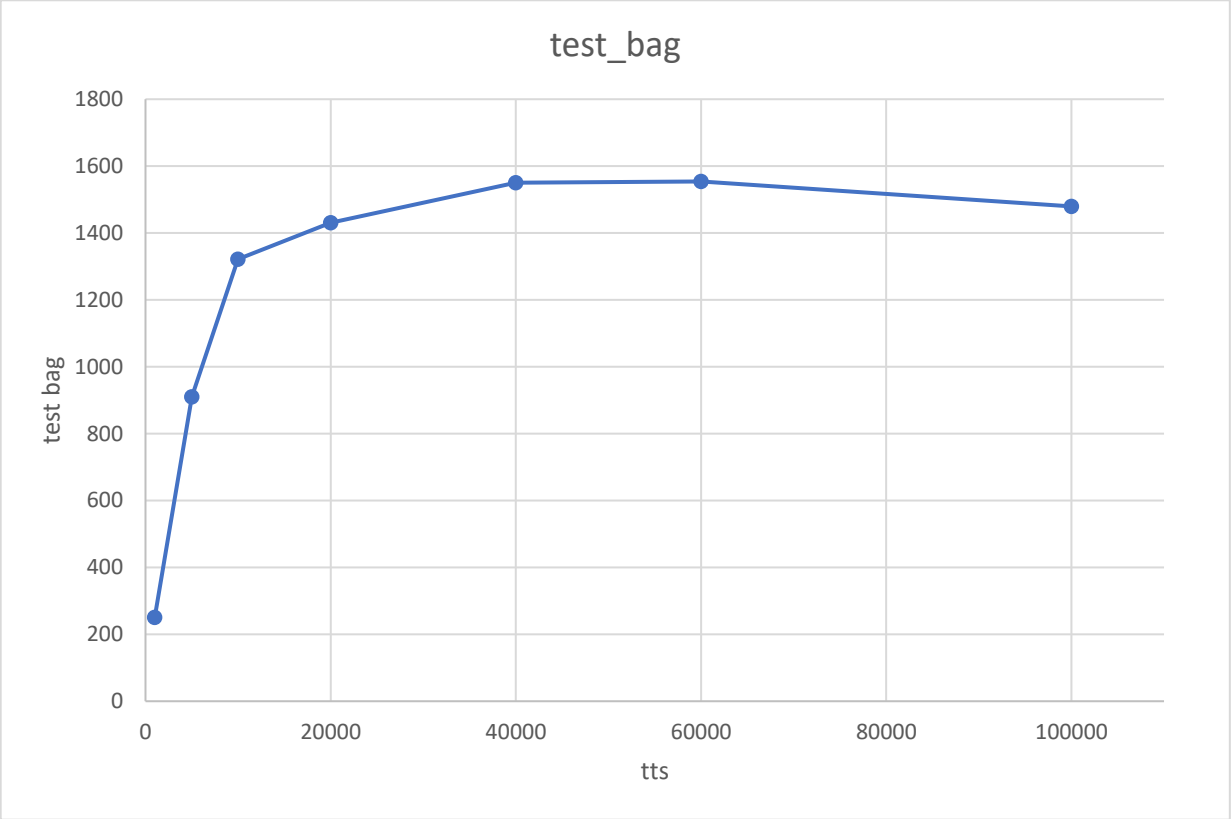
Ανάμεσα στις αναφερόμενες παραμέτρους βρίσκονται τα ct (conviction threshold), st (support threshold), lt (lift threshold). Καθεμία από τις τιμές των παραμέτρων περιορίζει το πλήθος των κανόνων και οδηγεί στον σχηματισμό μια υποομάδας κανόνων. Αρχικά παράγονται οι κανόνες με τη χρήση της συνάρτησης Apriori() από τις συναλλαγλες που αποτελούν τα δεδομένα της εκπαίδευσης. Στη συνέχεια επιβάλλονται επιπλέον κατώτατα όρια στις τιμές του conviction, support και lift, τα οποία κατώτατα όρια είναι οι τιμές του ct, st, lt αντίστοιχα. Θα πρέπει να γίνει σωστή επιλογή των τιμών των παραμέτρων και να μην δοθούν πολύ υψηλές τιμές λαμβάνοντας υπόψιν την απεικόνιση των διαστημάτων διακριτοποίησης των κανόνων όπως παρουσιάστηκαν στην Εικόνα 5.7. Από τη μία πλευρά οι υψηλές τιμές μπορεί να οδηγούν σε πιο ισχυρούς κανόνες από την άλλη όμως υπάρχουν και σχεδόν οι διπλάσιοι κανόνες που δεν θα συμπεριληφθούν στην υποομάδα των κανόνων και ενδεχομένως τα αποτελέσματα να χάσουν την αξιοπιστία τους.

Αρχικά, παρακολουθείται ο τρόπος μεταβολής του υποσυνόλου των συναλλαγών ελέγχου σε σχέση με την μεταβολή του μέγιστου αριθμού επαναλήψεων που επιλέγεται για το σχηματισμό του αυτό να σχηματιστεί. Όπως αναφέρθηκε στο Κεφάλαιο 3, επιλέγεται με τυχαίο τρόπο συνδυασμός μοναδικών κατηγοριών προϊόντων και σχηματίζονται συναλλαγές. Επιλέχθηκε να σχηματιστούν συναλλαγές με ζεύγη των μοναδικών κατηγοριών προϊόντων που υπάρχουν διαθέσιμοι.

Όπως φαίνεται στο διάγραμμα της Εικόνας 5.10, αυξάνοντας τον αριθμό επαναλήψεων και αναζητώντας τυχαία τα ζεύγη που παράγονται στο σώμα του υποσυνόλου κανόνων, μέχρι έναν αριθμό επαναλήψεων, ο ρυθμός αύξησης του πλήθους των συναλλαγών που απαρτίζουν το υποσύνολο ελέγχου είναι μεγάλος. Στη συνέχεια, όσο οι επαναλήψεις

αυξάνονται ο ρυθμός μειώνεται, φτάνει σε ένα μέγιστο αριθμό συναλλαγών και στη συνέχεια φθίνει συγκλίνοντας σε μια τιμή.

Επομένως, λαμβάνοντας υπόψιν το υποσύνολο ελέγχου που θα σχηματιστεί και για να έχουμε ένα μεγάλο αριθμό συναλλαγών για παραγωγή συστάσεων και κατ' επέκταση για αξιολόγηση της μεθόδου μεταλλίων επιλέγεται οι επαναλήψεις να είναι 15000. Η τιμή αυτή των επαναλήψεων παράγει μεγάλο αριθμό συναλλαγών και βρίσκεται κοντά στο όριο όπου ο ρυθμός αύξησης του περιεχομένου του test_bag αρχίζει να μειώνεται.



Εικόνα 5.10: Μεταβολή του μεγέθους του ελέγχου (test_bag) σε σχέση με το πλήθος των επαναλήψεων (tts)

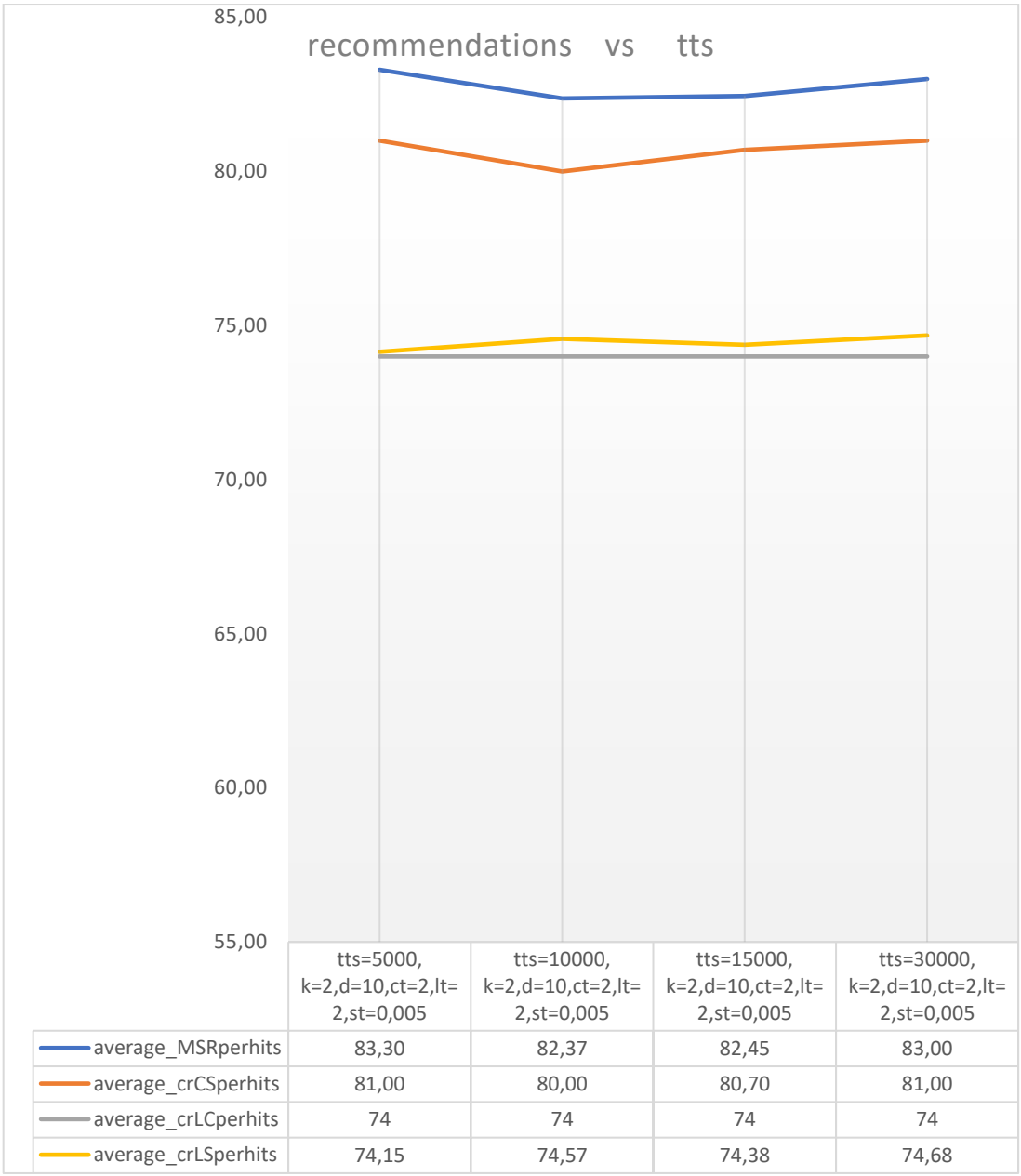
Στις Εικόνες που ακολουθούν απεικονίζονται οι γραφικές παραστάσεις μεταβολής του ποσοστού επιτυχίας για την κάθε μέθοδο σε σχέση με τη μεταβολή των παραμέτρων εισόδου όπως περιεγράφηκαν στο Κεφάλαιο 3, και χρησιμοποιούνται για την απεικόνιση οι εξής παράμετροι:

average_MSRperhits: μέσος όρος ποσοστού επιτυχίας της μεθόδου μεταλλίων

average_crCSperhits: μέσος όρος ποσοστού επιτυχίας της κλασικής μεθόδου ταξινόμησης Conviction - Support

average_crLSperhits: μέσος όρος ποσοστού επιτυχίας της κλασικής μεθόδου ταξινόμησης Lift - Support

average_crLCperhits: μέσος όρος ποσοστού επιτυχίας της κλασικής μεθόδου ταξινόμησης Lift - Confidence



	tts=5000, k=2,d=10,ct=2,lt=2,st=0,005	tts=10000, k=2,d=10,ct=2,lt=2,st=0,005	tts=15000, k=2,d=10,ct=2,lt=2,st=0,005	tts=30000, k=2,d=10,ct=2,lt=2,st=0,005
average_MSRperhits	83,30	82,37	82,45	83,00
average_crCSperhits	81,00	80,00	80,70	81,00
average_crLCperhits	74	74	74	74
average_crLSperhits	74,15	74,57	74,38	74,68

Εικόνα 5.11: Μεταβολή του ποσοστού των επιτυχών συστάσεων σε σχέση με τη μεταβολή του αριθμού επαναλήψεων tts.

Εκτός από το πλήθος του υποσυνόλου ελέγχου που θα σχηματιστεί ανάλογα με την επιλογή του αριθμού των επαναλήψεων, για την σωστή επιλογή τους, μελετάται η μεταβολή του ποσοστού επιτυχίας των συστάσεων μεταβάλλοντας τις επαναλήψεις αυτές. Επιλέγεται για τις υπόλοιπες παραμέτρους (k,d, ct, lt, st) , οι οποίες επεξηγήθηκαν στο Κεφάλαιο 3, που συμμετέχουν στην μελέτη τυχαίες αλλά σταθερές τιμές. Οι συναλλαγές είναι και πάλι ζεύγη κατηγοριών προϊόντων για την επίτευξη του μέγιστου δυνατού συνδυασμού τους. Εξετάζεται το ποσοστό επιτυχίας σταθερά για τις πρώτες δέκα συστάσεις όπως αυτές ταξινομούνται κατά φθίνουσα τάξη. Σταθερές παραμένουν και οι τιμές των κατωφλίων του lift, support και conviction.

Στο διάγραμμα της Εικόνας 5.11 απεικονίζει πως παρά την μεταβολή των επαναλήψεων του ποσοστού της επιτυχίας των συστάσεων που προκύπτουν από την μέθοδο των μεταλλίων και τις υπόλοιπες προκύπτει ότι η μέθοδος των μεταλλίων υπερτερεί με σταθερή διαφορά έναντι

των υπολοίπων μεθόδων. Μάλιστα το μέγιστο ποσοστό επιτυχίας εμφανίζεται για λίγες σχετικά επαναλήψεις, στις 5000, ενώ για επαναλήψεις 10000 με 15000 το ποσοστό επιτυχίας σχεδόν παραμένει σταθερό. Όσο οι επαναλήψεις αυξάνονται στις 20000 το ποσοστό επιτυχίας της μεθόδου των μεταλλίων έχει μια μικρή αύξηση.

Παρόμοια πορεία, όμως με σταθερά ποσοστά επιτυχίας ακολουθεί η μέθοδος CLperhits όπου προκύπτουν η ταξινόμηση της ισχύς του κανόνα γίνεται πρώτα ως προς το conviction και έπειτα ως προς το support.

Αρκετά χαμηλά βρίσκεται το ποσοστό επιτυχίας των συστάσεων που προέκυψαν ταξινομώντας τους κανόνες πρώτα ως προς το lift και έπειτα ως προς το support (μέθοδος LSperhits), ενώ ακολουθεί με μικρή διαφορά από αυτή το ποσοστό επιτυχίας προβλέψεων ταξινομώντας τους κανόνες πρώτα ως προς lift και στη συνέχεια ως προς confidence (μέθοδος LCperhits).

Αξίζει να σημειωθεί ότι και στις τέσσερις περιπτώσεις μετά τις 10000 επαναλήψεις το ποσοστό επιτυχίας είναι σχεδόν σταθερό καθώς οι μεταβολές με την αύξηση των επαναλήψεων είναι πολύ μικρές και σχεδόν δυσδιάκριτες ιδιαίτερα για τις συστάσεις που προκύπτουν από την ταξινόμηση του ζεύγους παραμέτρων lift - confidence.

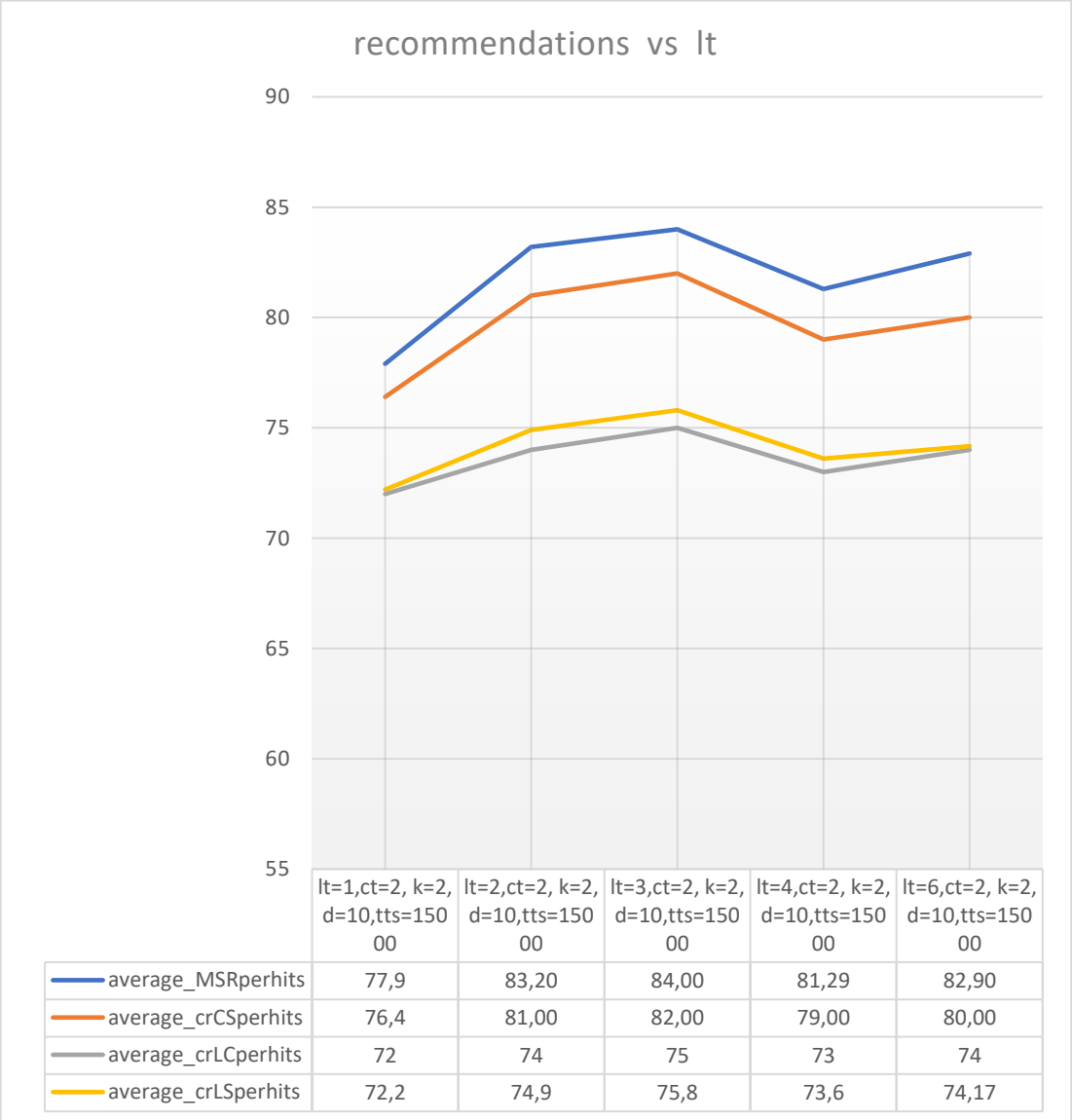
Με δεδομένο την μελέτη που πραγματοποιήθηκε για την παρακολούθηση των ποσοστών επιτυχίας ανά μέθοδο από την οποία προκύπτουν επιλέγεται ο αριθμός των επαναλήψεων που θα γίνουν προκειμένου να παραχθεί το υποσύνολο ελέγχου να είναι 15.000. Είναι μια τιμή όπου τα ποσοστά επιτυχίας για όλες τις μεθόδους σταθεροποιούνται ώστε να μπορεί να μελετηθεί στη συνέχεια ο τρόπος με τον οποίο οι τιμές των λοιπών παραμέτρων επηρεάζουν την επιτυχία πρόβλεψης και σε ποιο βαθμό.

Στο διάγραμμα της Εικόνας 5.12 φαίνεται ο τρόπος που επηρεάζεται το ποσοστό των επιτυχών συστάσεων σε σχέση με την μεταβολή του κατωφλίου lift ενώ οι τιμές των υπολοίπων παραμέτρων παραμένουν σταθερές. Η παρακολούθηση των μεταβολών γίνεται για ένα μέγιστο σύνολο επαναλήψεων έως 15000, για συναλλαγές που απαρτίζονται από δύο διαφορετικές κατηγορίες προϊόντων (k=2) και εξετάζεται για τις πρώτες δέκα συστάσεις (βάθος d=10) ταξινομημένες κατά φθίνουσα τάξη ως προς την ισχύ τους. Παρατηρείται ότι το ποσοστό της επιτυχίας των συστάσεων όπως προκύπτουν λαμβάνοντας υπόψιν την μέθοδο των μεταλλίων που προτείνεται, είναι σταθερά υψηλότερο σε σχέση με το ποσοστό επιτυχίας των συστάσεων που προέκυψαν ταξινομώντας την ισχύ τους βάσει των τιμών των παραμέτρων lift, support και confidence.

Για όλες τις μεθόδους παραγωγής συστάσεων η μεταβολή του ποσοστού επιτυχίας τους έχει την ίδια απεικόνιση μεταβάλλοντας μόνο το κατώφλι του lift(παράμετρος lt) όταν οι άλλες παράμετροι παραμένουν σταθερές. Αυτό σημαίνει ότι το Lift επηρεάζει με τον ίδιο τρόπο κάθε μέθοδο και δεν βοηθάει ούτε δυσχεραίνει την επιτυχία κάποιας. Δεν λειτουργεί ως βοηθητικός ούτε ως ανασταλτικός παράγοντας σε κάποια από τις μεθόδου έναντι κάποιας άλλης. Μεταβάλλοντας την τιμή του αυτό που συμβαίνει είναι να αλλάξει το ποσοστό της επιτυχίας αλλά κατά τη σύγκριση μεταξύ των μεθόδων τα συμπεράσματα δεν επηρεάζονται.

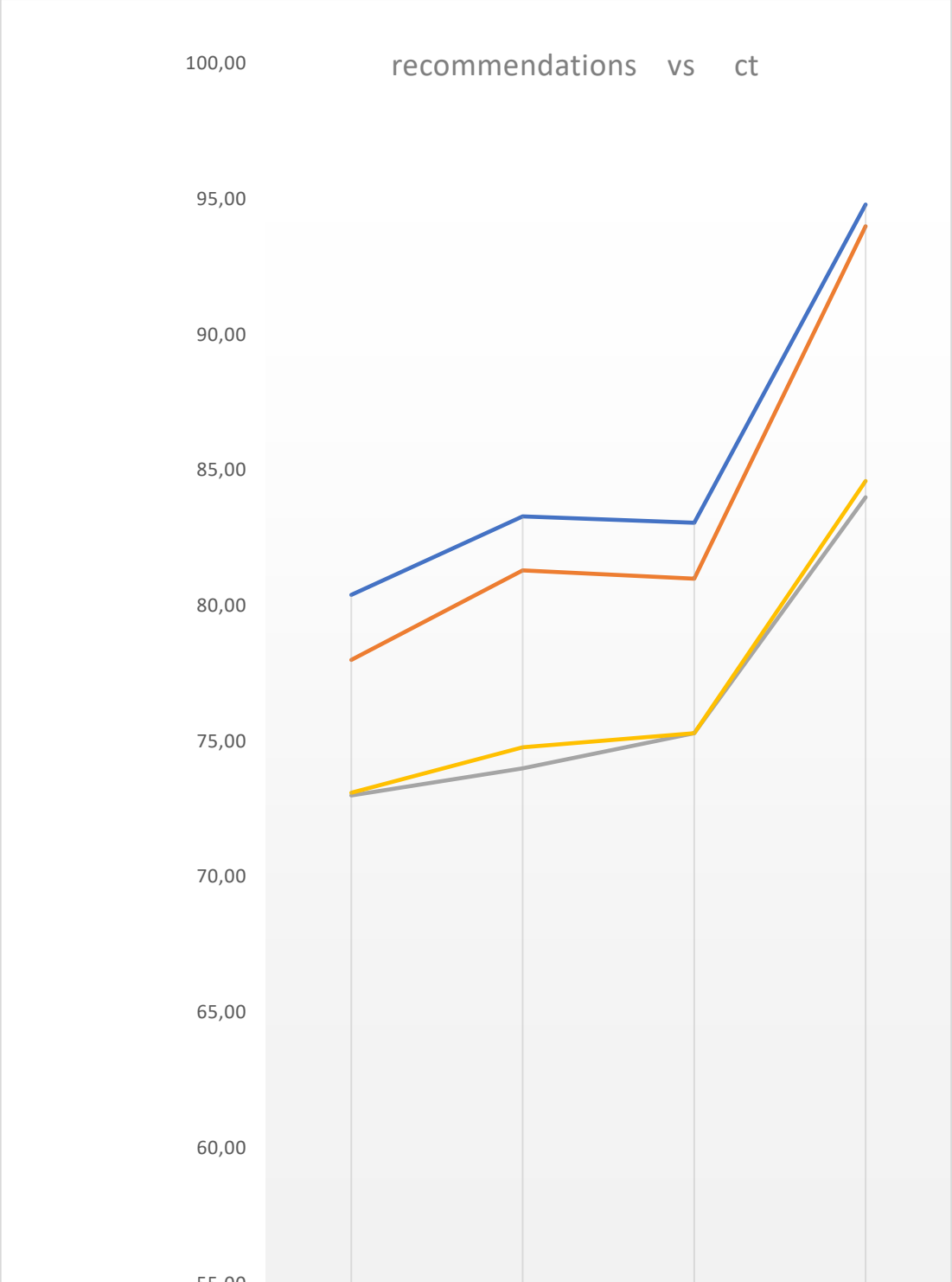
Παρατηρώντας μεμονωμένα τα τμήματα της γραφικής παράστασης που απεικονίζουν την μετάβαση από τη μια τιμή στην άλλη παρατηρούνται αλλαγές που αφορούν στο ρυθμό μεταβολής των ποσοστών επιτυχίας. Όταν το κατώφλι του lift αυξάνεται από 2 σε 3 το ποσοστό επιτυχίας των μεθόδων αυξάνεται, έχει θετικό αριθμό μεταβολής. Αυτό συμβαίνει καθώς όπως φάνηκε νωρίτερα στην Εικόνα 5.6, από τα σημεία τομής των Lift, το διάστημα 1

έως 3 περίπου αντιστοιχεί σε κανόνες με χαμηλή τιμή του lift , άρα αφαιρώντας τους κανόνες αυτούς αυξάνεται και η επιτυχία της πρόβλεψης. Θα περίμενε κανείς ότι ο ρυθμός μεταβολής θα είχε ανάλογη συμπεριφορά και αυξάνοντας την τιμή του κατωφλίου του lift από 3 σε 4. Σε αυτό το σημείο είναι που αφαιρούνται και οι κανόνες με ενδιάμεση ισχύ και λαμβάνονται υπόψιν μόνο οι πολύ ισχυροί κανόνες. Γίνεται αντιληπτό ότι οι κανόνες με ενδιάμεση ισχύ αποτελούν πηγή πληροφορίας και δεν συνιστούν ομάδα πλεοναζόντων κανόνων. Για το διάστημα μεταβολής του κατωφλίου του lift από 4 σε 6 όπου στο σημείο αυτό υπάρχουν μόνο πολύ ισχυροί κανόνες η μεταβολή του ποσοστού επιτυχίας σχεδόν σταθεροποιείται.



	lt=1,ct=2, k=2, d=10,tts=15000	lt=2,ct=2, k=2, d=10,tts=15000	lt=3,ct=2, k=2, d=10,tts=15000	lt=4,ct=2, k=2, d=10,tts=15000	lt=6,ct=2, k=2, d=10,tts=15000
average_MSRperhits	77,9	83,20	84,00	81,29	82,90
average_crCSperhits	76,4	81,00	82,00	79,00	80,00
average_crLCperhits	72	74	75	73	74
average_crLSperhits	72,2	74,9	75,8	73,6	74,17

Εικόνα 5.12: Μεταβολή του ποσοστού των επιτυχών συστάσεων σε σχέση με τη μεταβολή του κατωφλίου lift



	ct=1, d=10,k=2,tts=15000,st=0,005,lt=2	ct=2, d=10,k=2,tts=15000,st=0,005,lt=2	ct=3, d=10,k=2,tts=15000,st=0,005,lt=2	ct=8, d=10,k=2,tts=15000,st=0,005,lt=2
average_MSRperhits	80,400	83,300	83,060	94,800
average_crCSperhits	78,000	81,300	81,000	94,000
average_crLCperhits	73,000	74,000	75,300	84,000
average_crLSperhits	73,100	74,780	75,300	84,600

Εικόνα 5.13: Μεταβολή του ποσοστού επιτυχας των συστάσεων σε σχέση με τη μεταβολή του κατωφλίου conviction

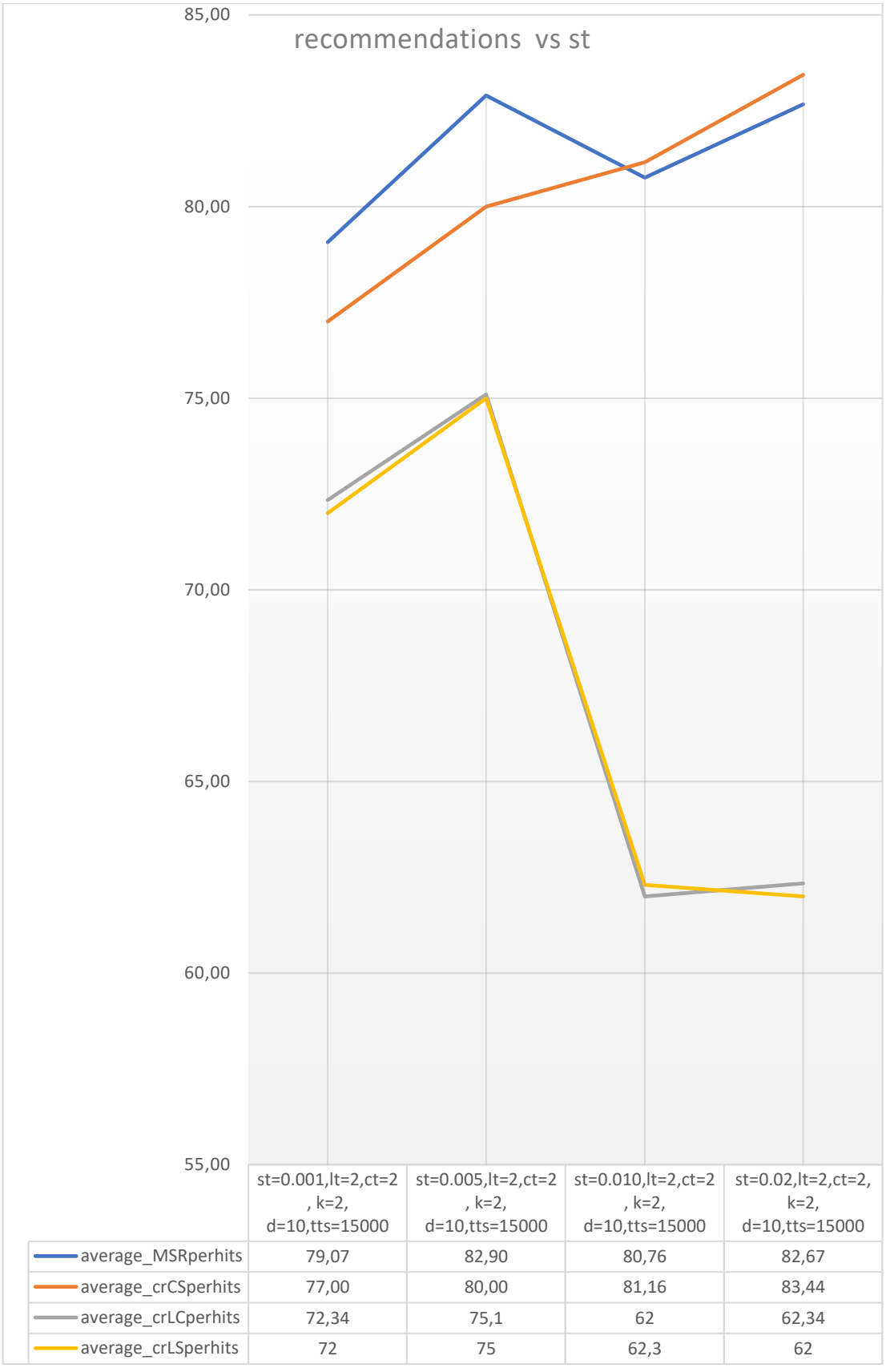
Στην Εικόνα 5.13 απεικονίζεται η μεταβολή του ποσοστού επιτυχίας μεταβάλλοντας μόνο το κατώφλι του conviction και διατηρώντας σταθερές οι τιμές των λοιπών παραμέτρων. Και εδώ ο ανώτατος αριθμός επαναλήψεων είναι οι 15000, μελετώνται ζεύγη κατηγοριών προϊόντων και το κατώφλι του lift έχει την τιμή 2 ώστε να συμπεριλαμβάνονται κανόνες με διαφορετική δυναμική και εμπλεκομένων όλων των μεταλλίων σύμφωνα με όσα αναφέρθηκαν κατά την ανάλυση του προηγούμενου διαγράμματος (Εικόνα 11).

Όπως και κατά την μεταβολή του κατωφλίου του lift έτσι και εδώ η μεταβολή του κατωφλίου του conviction επηρεάζει με παρόμοιο τρόπο τα ποσοστά επιτυχίας των προβλέψεων ανεξάρτητα από την μέθοδο την οποία προέκυψαν. Έτσι το κατώφλι στην τιμή του conviction δεν λειτουργεί υπέρ ή κατά κάποιας από τις μεθόδους παραγωγής συστάσεων. Η μέθοδος η οποία εμφανίζει υψηλότερα ποσοστά επιτυχίας κατά τη μεταβολή του κατωφλίου του conviction είναι προβλέψεις που προέκυψαν από τη μέθοδο των μεταλλίων.

Ξεκινώντας τον υπολογισμό των ποσοστών επιτυχίας ανά μέθοδο και ορίζοντας ως κατώφλι του conviction την τιμή 1 λαμβάνονται υπόψιν όλοι οι κανόνες χωρίς να αφαιρεθούν οι πιο αδύναμοι. Εξάλλου όπως αναφέρθηκε και στην μεθοδολογία ανάλυσης δεδομένων και παραγωγής κανόνων οι κανόνες με conviction μικρότερο του 1 έχουν αφαιρεθεί.

Μεταβάλλοντας το κατώφλι από 1 σε 2 και θυμίζοντας τα σημεία αποκοπής του conviction όπως απεικονίστηκαν στην Εικόνα 5.7 , τα ποσοστά επιτυχίας των μεθόδων αυξάνονται, καθώς χάνουν τον ενεργό ρόλο τους κανόνες με χαμηλότερη ισχύ.

Μεταβάλλοντας την τιμή του κατωφλίου από 2 σε 3 ο ρυθμός μεταβολής του ποσοστού επιτυχίας της εκάστοτε μεθόδου παραμένει σχεδόν αμετάβλητο. Αυξάνοντας όμως το κατώφλι σε τιμές μεγαλύτερες του 3 παρατηρείται με αλματώδη αύξηση των ποσοστών επιτυχίας καθώς στο σημείο αυτό υπερτερούν οι πολύ ισχυροί κανόνες. Υπενθυμίζεται ότι σχεδόν σε όλες τις μεθόδους το μέτρο του conviction συντιστά καθοριστική τιμή στην ταξινόμηση ισχύος των κανόνων έναντι των υπολοίπων μέτρων (lift, confidence και support).



	st=0.001,lt=2,ct=2, k=2, d=10,tts=15000	st=0.005,lt=2,ct=2, k=2, d=10,tts=15000	st=0.010,lt=2,ct=2, k=2, d=10,tts=15000	st=0.02,lt=2,ct=2, k=2, d=10,tts=15000
average_MSRperhits	79,07	82,90	80,76	82,67
average_crCSperhits	77,00	80,00	81,16	83,44
average_crLCperhits	72,34	75,1	62	62,34
average_crLSperhits	72	75	62,3	62

Εικόνα 5.14: Μεταβολή του ποσοστού επιτυχίας των συστάσεων σε σχέση με τη μεταβολή του κατωφλίου support

Η Εικόνα 5.14 δείχνει διαγραμματικά τη μεταβολή των ποσοστών επιτυχίας των συστάσεων αφήνοντας σταθερές τις τιμές όλων των παραμέτρων πλην του κατωφλίου του support.

Σε αντίθεση με τη μεταβολή των υπολοίπων παραμέτρων όπου την εικόνα της γενικής συμπεριφοράς όλων των κανόνων δεν την επηρέαζε τη μεταβολή των κατωφλίων, δεν συμβαίνει το ίδιο και για την μεταβολή του κατωφλίου του support.

Μέχρι την τιμή 0,005 του support τα ποσοστά επιτυχίας ακολουθούν ανοδική πορεία. Όσο όμως η τιμή του support αυξάνεται τα ποσοστά επιτυχίας των μεθόδων φθίνουν πλην της κλασικής μεθόδου παραγωγής συστάσεων Conviction-Support η οποία συνεχίζει την ανοδική της πορεία με μικρότερο όμως ρυθμό.

Για τις μεθόδους παραγωγής συστάσεων που η ταξινόμηση ισχύος του κανόνα γίνεται πρώτα ως προς το lift μειώνεται το ποσοστού επιτυχίας. Η αύξηση του support φαίνεται να οδηγεί σε μείωση του lift μέχρι κάποια τιμή όπου μετά σχεδόν σταθεροποιείται.

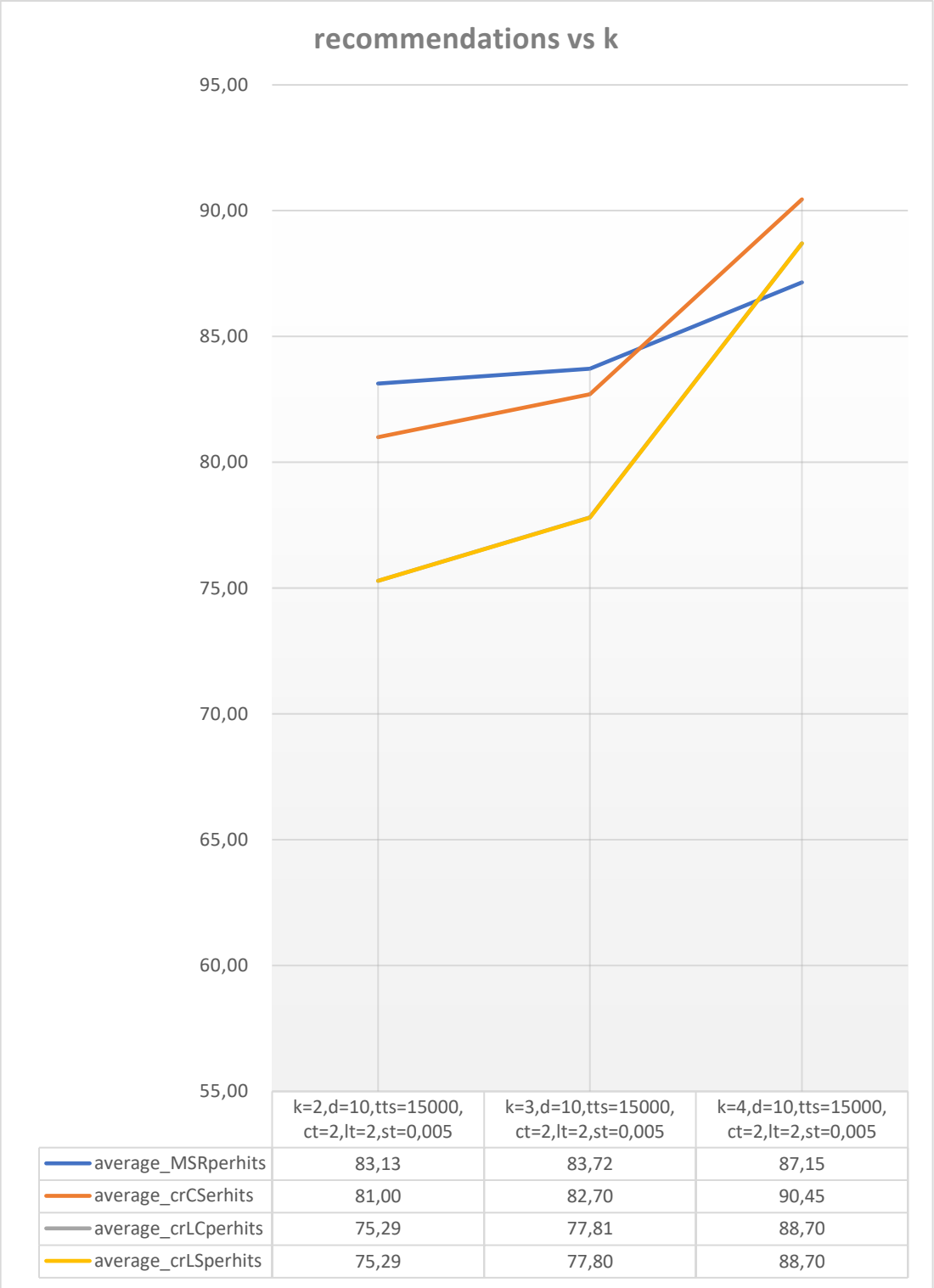
Σε ότι αφορά στην κλασική μέθοδο όπου οι κανόνες ταξινομούνται ως προς την ισχύ πρώτα με βάση το conviction και έπειτα με βάση το support δεν επηρεάζεται ο ρυθμός των ποσοστών επιτυχίας ιδιαίτερα καθώς στη μέθοδο αυτή κυριαρχεί έντονα η ταξινόμηση του conviction και αφήνει σχεδόν ανεπηρέαστη η μεταβολή του support. Στην μέθοδο των μεταλλίων όσο το support αυξάνει τόσο αυξάνει η συγκέντρωση μεταλλίων τύπου Α ως προς το support.

Σημαντικό είναι να αναφερθεί το διάστημα μεταξύ 0,005 και 0,01 του support όπου στο διάστημα αυτό και περίπου για τιμή support κάπου κοντά στο 0,009 οι δυο μέθοδοι MSR και Conviction-Support τέμνονται και τα ποσοστά επιτυχίας της κλασικής μεθόδου Conviction-Support αυξάνονται έναντι αυτών της μεθόδου των μεταλλίων. Να σημειωθεί και πάλι τα σημεία αποκοπής του support, όπου η τιμή του 0,005 έχει αποδυναμώσει τους κανόνες με μετάλλια τύπου C και συμμετέχουν πιο ενεργά ένα μέρος από κανόνες με ενδιάμεση ισχύ ως προς το support και οι περισσότεροι ισχυροί.

Μετά την μελέτη του τρόπου με τον οποίο η επιλογή των κατώτατων τιμών των παραμέτρων που εμπλέκονται με τους κανόνες μπορούν να επηρεάσουν την επιτυχία των προβλέψεων των μεθόδων που αναφέρθηκαν, είναι σημαντικό να εξεταστεί και ο τρόπος επιρροής στα ποσοστά επιτυχίας και του πλήθους των διαφορετικών κατηγοριών που υπάρχει σε κάθε συναλλαγή. Είναι σημαντικό να εξεταστεί η δυναμική των κανόνων σε ότι αφορά το «μήκος» της συναλλαγής.

Στην Εικόνα 5.15 απεικονίζονται οι μεταβολές στα ποσοστά επιτυχίας μεταβάλλοντας μόνο το πλήθος των μοναδικών κατηγοριών (k) και διατηρώντας όλες οι υπόλοιπες παραμέτρους εισόδου σταθερές. Παρατηρείται ότι τα ποσοστά επιτυχίας για τη μέθοδο των μεταλλίων παραμένουν υψηλότερα για τιμή του k έως 3. Όσο το k αυξάνει τότε τα ποσοστά επιτυχίας της κλασικής μεθόδου παραγωγής συστάσεων αυξάνουν και ξεπερνούν τα ποσοστά επιτυχίας της μεθόδου των μεταλλίων.

Στο σημείο αυτό αξίζει να σημειωθεί πως στις συναλλαγές με λίγα στοιχεία, ο βαθμός δυσκολίας της ορθής πρόβλεψης αυξάνει καθώς η πληροφορία που έρχεται ως είσοδος δεν επαρκεί για διαμόρφωση του προφίλ του χρήστη και άρα και επιτυχία στην πρόβλεψη των επόμενων αγορών για αυτό και τα ποσοστά επιτυχίας όλων των μεθόδων είναι αρκετά χαμηλά.



Εικόνα 5.15: Μεταβολή του ποσοστού επιτυχίας των συστάσεων σε σχέση με τη μεταβολή του πλήθους των συναλλαγών

	d=5,k=2,tts=15000,ct=2,lt=2,st=0,005	d=10,k=2,tts=15000,ct=2,lt=2,st=0,005	d=15,k=2,tts=15000,ct=2,lt=2,st=0,005	d=20,k=2,tts=15000,ct=2,lt=2,st=0,005
average_MSRperhits	74,600	83,200	87,12754	89,000
average_crCSperhits	75,000	81,000	84,23903	86,000
average_crLCperhits	59,600	74,000	81	84,000
average_crLSperhits	59,600	74,000	81,6831	84,700

Εικόνα 5.16: Μεταβολή των επιτυχών συστάσεων σε σχέση με τη μεταβολή του βάθους.

Επιπλέον, μελετήθηκαν τα ποσοστά επιτυχίας των μεθόδων παραγωγής συστάσεων σε συνάρτηση με το πλήθος των πιο ισχυρών συστάσεων που θα ληφθούν υπόψιν για την εξαγωγή συμπερασμάτων και την αξιολόγηση, το επονομαζόμενο βάθος. Παρατηρήθηκε ότι για μικρά βάθη έως d ίσο με 5, λαμβάνοντας δηλαδή υπόψιν τις πρώτες 5 πιο ισχυρές συστάσεις που προέκυψαν ανάλογα με τον τρόπο ταξινόμησης της κάθε μεθόδου, τα ποσοστά επιτυχίας της μεθόδου των μεταλλίων είναι ελάχιστα χαμηλότερα από τα ποσοστά επιτυχίας των τριών παραλλαγών της κλασικής μεθόδου.

Για βάθη μεγαλύτερα του 5 τα ποσοστά επιτυχίας της μεθόδου των μεταλλίων αυξάνονται και αυξάνει και ο ρυθμός με τον οποίο μεταβάλλονται. Μάλιστα για βάθος ίσο με 20 η διαφορά με τα ποσοστά επιτυχίας των λοιπών μεθόδων έχει αυξηθεί εμφανώς και πλησιάζει το 10%.

Κατόπιν της μελέτης που πραγματοποιήθηκε για την μεταβολή των ποσοστών επιτυχίας των συστάσεων έναντι της τιμής που λαμβάνει η κάθε μεταβλητή η βέλτιστη επιλογή τιμών για

τις παραμέτρους εισόδου για το συγκεκριμένο τύπο προβλήματος (καλάθι αγορών) έχει ως εξής:

- Πλήθος επαναλήψεων (tts)=15000
- Κατώφλι Lift (lt)=3
- Κατώφλι Conviction (ct) = 2
- Κατώφλι support (st) = 0.005
- Πλήθος συναλλαγών (k) =2
- Βάθος συστάσεων (d) = 10

V1	V2	id	hits	msr_recs	cr_recs	cr_recs_ls	cr_recs_lc	msrperhits	crperhits	crlsperhits	crlcperhits
101	61	1	901	758	730	666	666	84.12875	81.02109	73.91787	73.91787
285	49	2	615	473	440	387	387	76.91057	71.54472	62.92683	62.92683
49	85	3	1156	1089	1086	984	984	94.20415	93.94464	85.12111	85.12111
45	57	4	1116	998	998	940	940	89.42652	89.42652	84.22939	84.22939
285	61	5	615	454	429	406	406	73.82114	69.75610	66.01626	66.01626
205	85	6	636	521	507	475	475	81.91824	79.71698	74.68553	74.68553
205	85	7	636	521	507	475	475	81.91824	79.71698	74.68553	74.68553
417	85	8	489	434	428	412	412	88.75256	87.52556	84.25358	84.25358
417	57	9	489	421	418	418	418	86.09407	85.48057	85.48057	85.48057
417	61	10	489	418	418	411	411	85.48057	85.48057	84.04908	84.04908
57	93	11	835	788	788	721	721	94.37126	94.37126	86.34731	86.34731
101	49	12	901	749	723	686	686	83.12986	80.24417	76.13762	76.13762
101	33	13	901	785	772	698	698	87.12542	85.68257	77.46948	77.46948
285	49	14	615	473	440	387	387	76.91057	71.54472	62.92683	62.92683

ıg 1 to 14 of 138 entries, 12 total columns

Εικόνα 5.17: Τιμές μεταβλητών μετά της εκτέλεση του αλγορίθμου

Έχοντας ως είσοδο τις παραπάνω τιμές προκύπτει το υποσύνολο ελέγχου test_bag όπως φαίνεται στη Εικόνα 5.17. Στις δύο πρώτες στήλες φαίνονται τα ζεύγη συναλλαγών που παρήχθησαν με τυχαίο τρόπο και στις επόμενες στήλες τα αποτελέσματα για κάθε ένα μέτρο αξιολόγησης. Για παράδειγμα στη γραμμή 1 (id = 1) υπάρχει η συναλλαγή που αποτελείται από την κατηγορία προϊόντων με αρίθμηση 101 και την κατηγορία προϊόντων με αρίθμηση 61. Αυτή η συναλλαγή που παράχθηκε τυχαία εμφανίζεται στο αρχικό σύνολο ελέγχου σε 901 συναλλαγές, αυτό δηλώνεται με τη τιμή της μετρικής hits δίπλα σε κάθε συναλλαγή. Από τις συναλλαγές αυτές οι πρώτες 10 (d=10) συστάσεις που προκύπτουν για την συναλλαγή αυτή εφαρμόζοντας την μέθοδο των μεταλλίων υπάρχουν στις 758 από τις 901 συναλλαγές μέτρου msr_recs. Ομοίως αποτυπώνεται το πλήθος των συστάσεων που προέκυψαν από τις υπόλοιπες μεθόδους στις μεταβλητές cr_recs, cr_recs_ls, cr_recs_lc. Μετατρέπεται σε Τα αντίστοιχα ποσοστά επι τοις εκατό και καταχωρούνται στις αντίστοιχες perhits μεταβλητές. Τελικά υπολογίζεται ο συνολικός μέσος όρος του ποσοστού επιτυχίας για την κάθε μια μέθοδο (βλ. Εικόνα 5.18).

```
> average_MSRperhits
[1] 89.23326
> average_crCSperhits
[1] 88.60211
> average_crLSperhits
[1] 81.02789
> average_crLCperhits
[1] 81.02789
```

Εικόνα 5.18: Αποτελέσματα ποσοστού επιτυχιών για τις tts=15000,lt=3, ct=2,st=0.005, k=2, d=10

6 *Σχόλια και Συμπεράσματα*

6.1 Παράμετροι αξιολόγησης

Όπως αναφέρθηκε στο Κεφάλαιο 3 για την αξιολόγηση των μεθόδων παραγωγής συστάσεων χρησιμοποιήθηκε, ανά μέθοδο, παράμετρος μέτρησης ποσοστού επιτυχίας.

Από τις συστάσεις που προέκυψαν υπολογίστηκε το ποσοστό επιτυχίας για την κάθε μέθοδο ως εξής:

- Ταξινόμηση πρώτα ως προς lift και μετά ως προς confidence και αντιστοιχεί στην παράμετρο crLCperhits
- Ταξινόμηση πρώτα ως προς lift και μετά ως προς support και αντιστοιχεί στην παράμετρο crLSperhits
- Ταξινόμηση πρώτα ως προς conviction και μετά ως προς support και αντιστοιχεί στην παράμετρο CRperhits
- Ταξινόμηση με τη χρήση μεταλλίων και αντιστοιχεί στην παράμετρο MSRperhits .

6.2 Συμπεράσματα

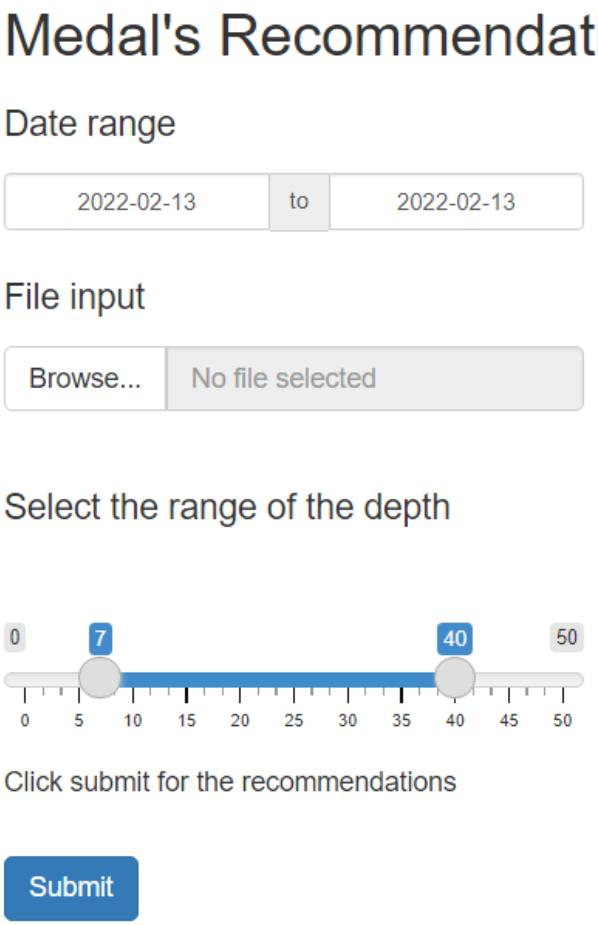
Ο κώδικας που αναπτύχθηκε αποτελεί ένα πρότυπο αλγορίθμου παραγωγής συστάσεων και μπορεί να τροφοδοτηθεί στην είσοδό του δεδομένα που προέρχονται από κάθε είδους συναλλαγή. Μπορεί οι συναλλαγές να αφορούν συναλλαγές καλαθιού αγορών από σούπερ μάρκετ, βιβλίων από μια βιβλιοθήκη, ταινιών αλλά και σύνολα μαθημάτων-βαθμοί φοιτητών ενός τμήματος. Ανεξάρτητα από τα δεδομένα εισόδου παράγονται συστάσεις βασιζόμενες στη μέθοδο των μεταλλίων, προκύπτουν από την διακριτοποίηση των κανόνων με βάση τη συχνότητα σε 3 μετάλλια. Η συχνότερη εφαρμοζόμενη μέθοδος παραγωγής συστάσεων είναι εκείνη που βασίζεται στην ταξινόμηση των συνδυαστικών κανόνων πρώτα ως προς lift και έπειτα ως προς confidence. Στην παρούσα διπλωματική εργασία τα αποτελέσματα των συστάσεων της προτεινόμενης μεθόδου αξιολογήθηκαν και συγκρίθηκαν έναντι της κλασικής μεθόδου αλλά και άλλων δύο ακόμα και παρατηρήθηκε τα ποσοστά επιτυχίας που υπερτερούν είναι αυτά της μεθόδου των μεταλλίων και ακολουθούν αυτά που προκύπτουν από την ταξινόμηση conviction-support.

7 *Μελλοντικές Επεκτάσεις*

Σε συνέχεια της διερευνητικής ανάλυσης των δεδομένων με τη μέθοδο παραγωγής συστάσεων με τη χρήση μεταλλίων παρατίθενται δύο (2) προτάσεις για την βελτίωση ή/και την επέκταση του προτεινόμενου μοντέλου.

Η πρώτη πρόταση αφορά στην παρουσίαση των αποτελεσμάτων σε ένα περιβάλλον διαδραστικό με τη χρήση του πακέτου shiny της R[10] στο περιβάλλον RStudio. Το πακέτο αυτό βοηθά στην ανάπτυξη διαδραστικών και αισθητικά ευχάριστων εφαρμογών ιστού χρησιμοποιώντας την R. Ένα ακόμα πακέτο της R είναι το flexdashboard [16] που επιτρέπει να δημιουργηθούν εύκολα ευέλικτοι, ελκυστικοί, διαδραστικοί πίνακες εργαλείων με την R. Σε συνδυασμό με το πακέτο shiny παρέχουν πρόσθετη διαδραστικότητα στα γραφήματα που δημιουργούνται. Ο χρήστης μπορεί να αλλάζει τις παραμέτρους στα εικονιζόμενα στοιχεία και να παρακολουθεί τη μεταβολή της γραφικής απεικόνισης. Στην Εικόνα 7.1 φαίνεται ένα ενδεικτικό διαδραστικό περιβάλλον όπου ο χρήστης έχει τη δυνατότητα να ανεβάσει με μορφή αρχείου τα δεδομένα του για να παραχθούν οι συστάσεις. Μπορεί επίσης να διαλέξει ένα εύρος ημερομηνιών και να εμφανίσει στην οθόνη του τις συστάσεις που προκύπτουν από ένα σύνολο εν δυνάμει σύνολο δεδομένων εκπαίδευσης τα οποία έχουν ήδη δημιουργηθεί, μπορεί να επιλέξει το εύρος των επιθυμητών συστάσεων και τιμές παραμέτρων εισόδου. Ενώ στην Εικόνα 7.2 απεικονίζεται ένα διαδραστικό διάγραμμα αποτελεσμάτων όπου το αποτέλεσμα διαμορφώνεται ανάλογα με την επιλογή του χρήστη από τη λίστα «Features».

[10] Shiny.rstudio.com. 2021. Shiny. [online] Available at: <https://shiny.rstudio.com/> [Accessed 15 January 2021].



Εικόνα 7.1: Διαδραστικό περιβάλλον για τον καθορισμό παραμέτρων εισόδου στην κλήση της μεθόδου MSR

Εικόνα 7.2: Διαδραστικό διάγραμμα απεικόνισης συστάσεων με τα μετάλλια που συγκεντρώνει η κάθε μια (διακριτοποίηση βάσει συχνότητας του conviction σε 5 μετάλλια)

Στόχο αποτελεί να προκληθεί το ενδιαφέρον των χρηστών και με το τρόπο αυτό να αυξηθει το κίνητρό του ώστε να χρησιμοποιηθεί η εφαρμογή πιο συστηματικά καταχωρώντας τα δεδομένα σε κάποια βάση δεδομένων και να αξιοποιηθούν για την βελτίωση των αποτελεσμάτων.

Η δεύτερη πρόταση αφορά διερεύνηση της παραγωγής συστάσεων με τη μέθοδο των μεταλλίων όταν το πλήθος μεταλλίων είναι μεγαλύτερο των τριών. Θα αποτελούσε ιδιαίτερο ενδιαφέρον αν η παρουσιαζόμενη στην παρούσα εργασία υλοποίηση εξεταζόταν με περισσότερα μετάλλια από αυτά που χρησιμοποιήθηκαν. Θα μπορούσε για παράδειγμα να γίνει διακριτοποίηση σε πέντε διαστήματα και παρακολουθώντας διαγραμματικά τις μεταβολές της επιτυχίας της μεθόδου σε σχέση με τις υπόλοιπες παραμέτρους εισόδου. Ο πίνακας συστάσεων όπως προκύπτει με διακριτοποίηση σε πέντε διαστήματα και ταξινομημένες ως προς τα πέντε πλέον μετάλλια και βάθος δέκα φαίνεται στην Εικόνα 7.3. Μπορεί να μελετηθεί ο τρόπος με

τον οποίο τα περισσότερα μετάλλια που σχηματίστηκαν επηρεάζουν τα αποτελέσματα των συστάσεων που προκύπτουν.

body	head	support	confidence	coverage	lift	count	conviction	rulenumber	medalconviction_freq	medalsupport_freq	medallift_freq	V1	V2	flag
417,57	41	0.011434224	0.9059561	0.012621167	2.491355	289	6.766628	773	A	A	E	417	57	1
417,57	33	0.009614243	0.7617555	0.012621167	3.089933	243	3.162599	771	B	A	D	417	57	2
57	41	0.078259149	0.7956557	0.098358061	2.188031	1978	3.114156	27	B	A	E	417	57	3
417,57	65	0.008862512	0.7021944	0.012621167	2.286519	224	2.326679	772	C	A	E	417	57	4
417,57	53	0.008150346	0.6457680	0.012621167	3.494281	206	2.301297	770	D	A	C	417	57	5
57	33	0.060138477	0.6114240	0.098358061	2.480138	1520	1.939059	25	D	A	E	417	57	6
417,57	85	0.007398615	0.5862069	0.012621167	3.266398	187	1.982957	769	D	B	C	417	57	7
417	65	0.029119683	0.5249643	0.055469832	1.709414	736	1.458623	13	E	A	E	417	57	8
417	41	0.037072206	0.6683310	0.055469832	1.837892	937	1.918660	14	E	A	E	417	57	9
57	53	0.049812067	0.5064360	0.098358061	2.740349	1259	1.651646	24	E	A	D	417	57	10
417,57	41	0.011434224	0.9059561	0.012621167	2.491355	289	6.766628	773	A	A	E	417	57	1
417,57	33	0.009614243	0.7617555	0.012621167	3.089933	243	3.162599	771	B	A	D	417	57	2
57	41	0.078259149	0.7956557	0.098358061	2.1880	1978	3.114156	27	B	A	E	417	57	3
417,57	65	0.008862512	0.7021944	0.012621167	2.286519	224	2.326679	772	C	A	E	417	57	4
417,57	53	0.008150346	0.6457680	0.012621167	3.494281	206	2.301297	770	D	A	C	417	57	5
57	33	0.060138477	0.6114240	0.098358061	2.480138	1520	1.939059	25	D	A	E	417	57	6

3.089933

Εικόνα 7.3: Οι πρώτες 10 συστάσεις με διακριτικοποίηση 5 μεταλλίων

8 *Επίλογος*

Στο κεφάλαιο αυτό ακολουθεί σύνοψη της διπλωματικής εργασίας.

8.1 Σύνοψη

Η παρούσα εργασία αποτελεί παράδειγμα της συμβολής της ακαδημαϊκής έρευνας στους φορείς ανάπτυξης και παραγωγής αλλά και σε κάθε είδους επιχείρηση. Θα μπορούσε να αποτελέσει ένα εργαλείο σε πολλούς κλάδους του επαγγελματικού κόσμου και να δώσει απαντήσεις σε ερωτήματα όπως

- √ Ανάλυση αποχωρήσεων – γιατί φεύγουν οι πελάτες, τι θα τους κρατήσει;
- √ Τί άλλο θα αγόραζε ο πελάτης;
- √ Ποιές περιπτώσεις μπορεί να εμπεριέχουν δόλο;
- √ Τι κίνδυνο εμπεριέχει μια επιχειρηματική απόφαση;
- √ Τι κοινά χαρακτηριστικά έχουν οι πελάτες;
- √ Τι διαφημίσεις να μπουν στο web ανάλογα με τις στρατηγικές πλοήγησης και αγορών των πελατών;
- √ Τι θα πουληθεί ανά χρονική περίοδο στο μέλλον;

Αυτά είναι μόνο μερικά από τα χιλιάδες επιχειρηματικά προβλήματα που απασχολούν καθημερινά τα εμπλεκόμενα μέρη των επιχειρήσεων.

Πέραν των επιχειρήσεων όμως μπορεί να απαντηθούν και αρκετά ακόμα ακαδημαϊκά ερωτήματα [17]

- √ Σε ποιο μάθημα οι φοιτητές τα πάνε καλύτερα;
- √ Σε ποιο μάθημα τα πάνε χειρότερα;
- √ Γιατί τους δυσκολεύει αυτό το μάθημα;
- √ Μήπως πρέπει να μετακινηθεί σε άλλο μικρότερο εξάμηνο;

Πολλά τα ερωτήματα που μπορεί να προβληματίζουν και να χρήζουν απάντησης, αρκεί να συγκεντρωθούν τα δεδομένα, να επεξεργαστούν καταλληλά και να τροφοδοτήσουν τον αλγόριθμο ο οποίος θα υπολογίσει με τα κατάλληλα αποτελέσματα.

.

8.2 Σημασία της διπλωματικής εργασίας

Η παρούσα εργασία αποτελεί σημαντικό σταθμό των μεταπτυχιακών μου σπουδών. Έχοντας μια πρώτη επαφή με το μάθημα «Αποθήκες Δεδομένων - Εξόρυξη Πληροφορίας» κατάφερα να ανακαλύψω πως ανάμεσα και στα φαινομενικά ασυσχέτιστα γεγονότα μπορεί με ανάλυση και με τα κατάλληλα μέσα να αποκαλυφθεί η σχέση ανάμεσά τους. Και αυτό με τη σειρά του μπορεί να αποτελέσει σημαντικό εργαλείο μιας επιχείρησης προκειμένου να γίνει πιο ανταγωνιστική και ευέλικτη στην αγορά. Γνώρισα επίσης την γλώσσα προγραμματισμού R που αποτέλεσε σημαντικό εργαλείο για την έρευνα και την εξαγωγή των συμπερασμάτων αλλά και για την απεικόνισή τους σε γραφήματα.

Αν και εκ πρώτης όψεως ως τραπεζικός υπάλληλος το αντικείμενο της εργασίας μου ίσως να φαίνεται όχι και τόσο σχετικό με την δουλειά μου, στην πραγματικότητα αποτέλεσε σημαντικό εργαλείο για αυτή καθώς προσέφερε μια νέα δυνατότητα επεξεργασίας και διαχείρισης του τεράστιου όγκου δεδομένων που προκύπτουν από συναλλαγές που εκτελούν καθημερινά οι πελάτες μας. Οι συναλλαγές αυτές εκτός από μια απλή αναφορά για την πορεία της τράπεζας συνιστούν πρωτογενή πληροφορία αρχικά για μελέτη και ανάλυση και στη συνέχεια μετασχηματίζεται σε δεδομένα εισόδου για τον αλγόριθμο που αναπτύχθηκε. Με αυτήν την επεξεργασία μπορούν να εξαχθούν συμπεράσματα σχετικά με το προφίλ των πελατών αλλά κυρίως και συστάσεις νέων προϊόντων και υπηρεσιών σε αυτούς ανάλογα με τις συνήθειες και τις ανάγκες τους όπως αυτές προκύπτουν από την έρευνα. Συνέπεια όλων αυτών θα είναιο εκσυγχρονισμός και η αυξημένη ανταγωνιστικότητα στην αγορά για την επιχείτηση που εργάζομαι.

Βέβαια, χρειάστηκε να κτίσω γερές βάσεις και να βελτιώσω σημαντικά τις γνώσεις σε θέματα τεχνολογίας που έπρεπε να χρησιμοποιήσω στην διπλωματική μου εργασία. Έμαθα όμως πολλά και ωφελήθηκα αποκτώντας γνώσεις.

Κατά την εκπόνηση της εργασίας και με την ολοκλήρωσή της, διεξήχθησαν δύο (2) παρουσιάσεις στο χώρο εργασίας μου όπου προτάθηκαν τρόποι για τη βελτίωση και την επέκταση της επιχείρησης και ομολογώ ότι οι αντιδράσεις και τα σχόλια από την διεύθυνση όσο και από την διοίκηση ήταν εξαιρετικά σε βαθμό που οι προτάσεις μου ενσωματώθηκαν στο στρατηγικό σχέδιο ανάπτυξης της επιχείρησης.

9 *Βιβλιογραφία*

[1] P. J. Azevedo and A. M. Jorge. ‘Comparing rule measures for predictive association rules’, in *ECML 2007: Proceedings of the 18th European conference on Machine Learning*, Springer, vol 4701, pp. 510-517, 2007

[2] Β. Σ. Βερύκιος, Β. Καγκλής, & Η.Κ. Σταυρόπουλος, ‘Η επιστήμη των δεδομένων μέσα από την γλώσσα R’, 2015 Αθήνα: Δράση-Κάλλιπος.

[3] R. Troy, C.V. Nitesh, ‘Market basket analysis with networks’, *Social Network Analysis and Mining volume*, vol.1, pp. 97–113, April 2011

[4] I. Benouaret, S. Amer-Yahia, S. B. Roy, C. Kamdem-Kengne and J. Chagraoui, "Enabling decision support through ranking and summarization of association rules for total customers" in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLIV*, Springer, vol.12380, pp. 160-193, 2020

[5] M. Sarnovsky, P. Butka and A. Huzvarova, "Twitter data analysis and visualizations using the R language on top of the Hadoop platform", in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI*), pp. 327-332, 2017

[6] J. B. Chandrasekar, S. Murugesh and V. R. Prasadula, "Deriving Big Data insights using Data Visualization Techniques," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 724-731, 2019.

[7] W. -L. Chang, T. -T. Ren, J. Luo, M. Feng, M. Guo and K. W. Lin, "Quantum Algorithms for Biomolecular Solutions of the Satisfiability Problem on a Quantum Machine," in *IEEE Transactions on NanoBioscience*, vol. 7, no. 3, pp. 215-222, Sept. 2008.

[8] J. Lee and H. Oh, "YouTube aware Personalized Ranking System for Future ICT Education," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 776-779, 2018.

[9] L. Shi, C. Jianping and X. Jie, "Prospecting Information Extraction by Text Mining Based on Convolutional Neural Networks–A Case Study of the Lala Copper Deposit, China," in IEEE Access, vol. 6, pp. 52286-52297, 2018

[10] H. I. Bulbul and Ö. Unsal, "Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data," *2011 10th International Conference on Machine Learning and Applications and Workshops*, pp. 298-301, 2011.

[11] M. Li, H. Wang and J. Li, "Mining conditional functional dependency rules on big data," in Big Data Mining and Analytics, vol. 3, no. 1, pp. 68-84, March 2020

[12] G. Ali, L. Bagheriye, H. Manhaeve and H. G. Kerkhoff, "On-chip EOL Prognostics Using Data-Fusion of Embedded Instruments for Dependable MP-SoCs", in *2020 IEEE 29th Asian Test Symposium (ATS)*, pp. 1-6, 2020.

[13] Y. Xun, J. Zhang and X. Qin, "FiDoop: Parallel Mining of Frequent Itemsets Using MapReduce," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 3, pp. 313-325, March 2016

[14] Q. Ding, Q. Ding and W. Perrizo, "PARM—An Efficient Algorithm to Mine Association Rules from Spatial Data," in *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), vol. 38, no. 6, pp. 1513-1524, Dec. 2008

[15] W. C. Braga de Sousa and R. Melo e Silva de Oliveira, "Coulomb's Law Discretization Method: a New Methodology of Spatial Discretization for the Radial Point Interpolation Method," in *IEEE Antennas and Propagation Magazine*, vol. 57, no. 2, pp. 277-293, April 2015

[16] S. Mallik, A. Mukhopadhyay and U. Maulik, "RANWAR: Rank-Based Weighted Association Rule Mining from Gene Expression and Methylation Data," in *IEEE Transactions on NanoBioscience*, vol. 14, no. 1, pp. 59-66, Jan. 2015

[17] Κ. Κελεσίδης, 'Προτυποποίηση αναλυτικής επεξεργασίας δεδομένων βαθμολογιών', Μεταπτυχιακή Διπλωματική Εργασία, ΠΜΣ: "Ευφυείς Τεχνολογίες Διαδικτύου, Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων, ΔΙ.ΠΑ.Ε Σεπτέμβριος 2020.