Επιχειρησιακή Έρευνα και Επιχειρηματική Ευφυΐα

1η Υποχρεωτική Εργασία

Στέφανος Καραμπέρας - ΑΕΜ 2910

Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η εξόρυξη γνώσης από δεδομένα συναλλαγών ενός καταστήματος λιανικής.

Συγκεκριμένα, ως δεδομένο προσφέρεται η ανάλυση 7537 συναλλαγών (καλάθια) που πραγματοποιήθηκαν στην επιχείρηση μέσα σε μια περίοδο 75 ημερών και αφορούν 170 κωδικούς προϊόντων.

Τα ερωτήματα ανάλυσης που καλούμαστε να απαντήσουμε αφορούν ποιοτικά και ποσοτικά χαρακτηριστικά των αγορών, κατηγοριοποίηση των πελατών σύμφωνα με τα μοτίβα συμπεριφοράς που μπορούν να εντοπιστούν από τις αγορές τους αλλά και προσδιορισμό κανόνων συσχετίσεως που παρουσιάζουν ιδιαίτερο ενδιαφέρον για τις μελλοντικές αποφάσεις της επιχείρησης.

Για τον σκοπό της ανάλυσης των δεδομένων θα χρησιμοποιηθεί η γλώσσα προγραμματισμού R.

Δομή κώδικα

Ο κώδικας που έχει συγγραφεί για τους σκοπούς της εξόρυξης γνώσης βρίσκεται στο αρχείο main.R.

Για ευκολία ανάγνωσης, κατανόησης, συντήρησης αλλά και ευκολία αξιοποίησης του κώδικα στα διαφορετικά σενάρια χρήσης που προκύπτουν από το δοθέν καθήκον, έγινε εκτενής διαχωρισμός του υλοποιημένου κώδικα σε συναρτήσεις (functions), ενώ έχει προστεθεί επαρκής σχολιασμός σε όλη την έκτασή του.

Οι συναρτήσεις που έχουν δημιουργηθεί παρουσιάζονται συνοπτικά παρακάτω, σύμφωνα με τα ερωτήματα της εργασίας στα οποία αντιστοιχούν. Αξίζει να σημειωθεί ότι ορισμένες συναρτήσεις είναι αναγκαίες σε περισσότερα από ένα ερωτήματα, συνεπώς θα κατηγοριοποιηθούν σύμφωνα με το ερώτημα βασικής αξιοποίησής τους.

Συναρτήσεις:

* Συνάρτηση εκτέλεσης: execute()
* Άσκηση 1:
  + prepareData()
* Άσκηση 2:
  + testAssociationRules(groceriesDiscrete)
  + generateAssociationRules(groceriesDiscrete)
* Άσκηση 3:
  + filterNormalizeCostRecency(groceriesDiscrete)
  + performClustering(normalizedCostAndRecency)
  + printClusteringCharts(groceriesDiscrete)
  + generateGroceriesWithBinaryClusterData(groceriesDiscrete, kmeansFit)
* Άσκηση 4:
  + clusterProductProfile(groceriesWithClusters)

Αναλυτική περιγραφή επίλυσης των ερωτημάτων

Άσκηση 1

1. Μετασχηματισμός των πρωτογενών δεδομένων σε δυαδική μορφή συναλλαγών

Αρχικά, γίνεται ανάγνωση των δεδομένων από το δοθέν αρχείο (GroceriesInitial.csv) με χρήση της εντολής:

groceries <- *read.csv*("GroceriesInitial.csv",header=TRUE,sep=",", stringsAsFactors=TRUE)

Ακολούθως, εφαρμόζεται η εντολή:

productNames <- *levels*(*unlist*(groceries[,4:35]))

με την οποία συλλέγονται τα διακριτά ονόματα των προϊόντων από το σύνολο δεδομένων **groceries**. Αυτό επιτυγχάνεται με την ενοποίηση των περιεχομένων των στηλών προϊόντων (Item\_1, Item\_2, …, Item\_32) σε μία ενιαία λίστα (χρήση έτοιμης συνάρτησης **unlist(groceries[,4:35])** ) και στη συνέχεια με την επιλογή των διακριτών ονομάτων προϊόντων από την προαναφερθείσα ενοποιημένη λίστα (χρήση έτοιμης συνάρτησης **levels()** ).

Με δεδομένο ότι στο αρχικό σύνολο δεδομένων υπάρχουν στήλες «Item\_Χ» που δεν περιέχουν όνομα προϊόντος, είναι προφανές ότι μεταξύ των διακριτών ονομάτων προϊόντων που έχουν εντοπιστεί συμπεριλαμβάνεται και η κενή συμβολοσειρά (“”). Για τον εντοπισμό και την αφαίρεση της κενής συμβολοσειράς από τα διακριτά ονόματα προϊόντων, γίνεται χρήση των εντολών:

blank <- *which*(productNames == "")   
productNames <- productNames[-blank]

Στη συνέχεια, πραγματοποιείται μετατροπή των πρωτογενών δεδομένων συναλλαγών σε δυαδική μορφή συναλλαγών (οι ονομασίες των διακριτών προϊόντων γίνονται ιδιότητες (στήλες) και κάθε καταχώρηση (σειρά) μπορεί να έχει είτε **«TRUE»** είτε **«FALSE»** κάτω από κάθε στήλη προϊόντος, ανάλογα με το αν το προϊόν περιλαμβάνεται ή όχι στην συναλλαγή).

Για τον σκοπό αυτό, χρησιμοποιείται η εντολή:

productsBinary <- *as.data.frame*(*t*(*apply*(groceries[,4:35],1, function(x)  
(productNames) %in% *as.character*(*unlist*(x)))))

Πιο αναλυτικά:

* Εφαρμόζεται επαναληπτικά η ανώνυμη συνάρτηση **function(x)** για τις ιδιότητες των προϊόντων (στήλες 4:35) κάθε καταχώρησης (σειράς) (χρήση έτοιμης συνάρτησης **apply()** ).
* Μέσω της **function(x)**, γιακάθε καταχώρηση συνδυάζονται τα δεδομένα όλων των ιδιοτήτων προϊόντων (“Item\_X”) σε μία ενιαία λίστα με χρήση της εντολής **unlist()**.

Στη συνέχεια, για κάθε όνομα προϊόντος της **productNames**, ελέγχεται αν το όνομα του προϊόντος εμφανίζεται μέσα στην προαναφερθείσα ενιαία λίστα, με αποτέλεσμα την επιστροφή της τιμής **«TRUE»** αν υπάρχει εμφάνιση ή της τιμής **«FALSE»** αν δεν υπάρχει.

* Ως αποτέλεσμα, επιστρέφεται μια λίστα που περιέχει τις διακριτές τιμές **«TRUE»** και **«FALSE»** με σειρά ανάλογη της σειράς των ονομάτων προϊόντων της **productNames**. Το αποτέλεσμα αποθηκεύεται στη μεταβλητή **productsBinary**.
* Η έτοιμη συνάρτηση **t()** καλείται για την εφαρμογή ενέργειας αντιμετάθεσης των γραμμών με τις στήλες του πίνακα που προκύπτει παραπάνω (transpose of matrix) προκειμένου να καταλήξουμε στην επιθυμητή μορφή των δεδομένων αποτελέσματος.

1. Φιλτράρισμα των 13 προϊόντων ενδιαφέροντος από το σύνολο δεδομένων

Προκειμένου να είμαστε σε θέση να επιλέξουμε τα 13/170 προϊόντα που μας έχουν γνωστοποιηθεί ως προϊόντα ενδιαφέροντος, ξεκινάμε αντιστοιχίζοντας στη μεταβλητή **productsBinary** τις ονομασίες των προϊόντων με αντιστοίχιση ένα προς ένα. Τα ονόματα των προϊόντων αποθηκεύονται ως τίτλοι των στηλών της **productsBinary**:

*names*(productsBinary) <- productNames

Αμέσως μετά, με την εντολή:

filteredProductsBinary <- productsBinary[, *c*("citrus fruit", "tropical fruit", "whole milk", "other vegetables","rolls/buns", "chocolate", "bottled water", "yogurt", "sausage", "root vegetables", "pastry", "soda", "cream")]

πραγματοποιείται «φιλτράρισμα» της **productsBinary**, διατηρώντας μόνο τις στήλες με τις ονομασίες που καθορίζονται εντός της λίστας που δημιουργείται από την έτοιμη συνάρτηση **c().**

Ακολούθως, προκειμένου να ενοποιήσουμε τη διακριτοποιημένη λίστα ονομάτων προιόντων αξίας με τις 3 πρώτες στήλες (id, basket\_value, recency\_days) της αρχικής συλλογής δεδομένων, εκτελούμε την εντολή:

groceriesBinary <- *cbind*(groceries[,1:3], filteredProductsBinary)

1. Διακριτοποίηση ιδιότητας αξίας συναλλαγής

Σε αυτό το σημείο πραγματοποιείται διακριτοποίηση της ιδιότητας αξίας συναλλαγής (basket\_value) των αρχικών δεδομένων σε τρεις (περίπου) ισοπληθείς κατηγορίες: **low\_value\_basket**, **medium\_value\_basket**, **high\_value\_basket**.

Για τον σκοπό αυτό, χρησιμοποιούνται οι εντολές:

groceriesDiscrete <- groceriesBinary  
cutPoints <- *quantile*(groceriesDiscrete$basket\_value, probs = *seq*(0, 1, 1/3), na.rm = TRUE, names = FALSE)

ώστε σύμφωνα με τις τιμές της ιδιότητας basket\_value να οριστούν τα 3 περίπου ισοπληθή διαστήματα.

Στη συνέχεια, τα διαστήματα αυτά τροφοδοτούνται στην εντολή:

groceriesDiscrete$basket\_value\_dis <- *cut*(groceriesDiscrete$basket\_value, breaks = cutPoints,  
 labels=*c*("Low","Medium","High"), include.lowest = TRUE)

υπό τη μορφή του ορίσματος σημείων διαχωρισμού (**breaks**). Η έτοιμη συνάρτηση **cut()** κατηγοριοποιεί για κάθε καταχώρηση την τιμή της ιδιότητας «**basket\_value**» ανάλογα με το διάστημα στο οποίο ανήκει ως «**Low**», «**Medium**» ή «**High**». Το αποτέλεσμα της κατηγοριοποίησης αποθηκεύεται στη νέα ιδιότητα (στήλη) «**basket\_value\_dis**» της **groceriesDiscrete.**

Τέλος, η συνάρτηση **prepareData()** επιστρέφει ως αποτέλεσμα τη **groceriesDiscrete**, η οποία περιέχει τις αρχικές ιδιότητες **id**, **basket\_value**, **recency\_days**, τις **διακριτοποιημένες** ιδιότητες με τις **ονομασίες των 13 προϊόντων ενδιαφέροντος**, καθώς και την **διακριτοποιημένη** ιδιότητα κατηγοριοποίησης σύμφωνα με την αξία συναλλαγής «**basket\_value\_dis**».

Άσκηση 2

Στόχος της άσκησης είναι η μάθηση κανόνων συσχέτισης μέσα από το επεξεργασμένο σύνολο των αρχικών δεδομένων (**groceriesDiscrete**), αποκλειστικά για τα χαρακτηριστικά των προϊόντων και τη διακριτοποιημένη αξία καλαθιού.

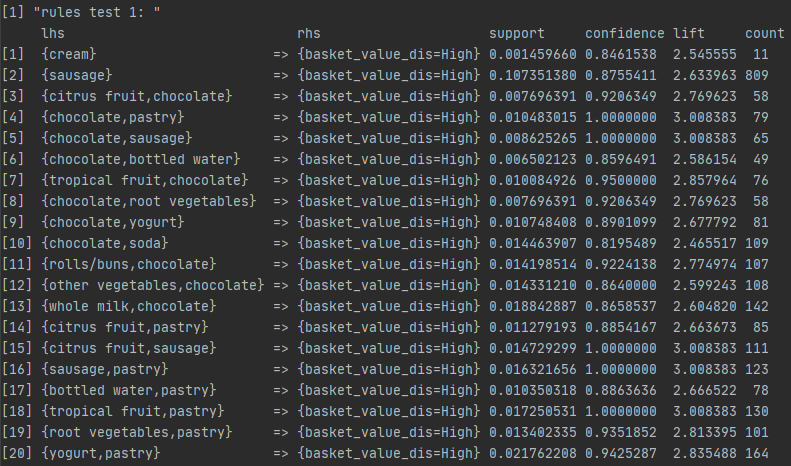
1. Πειραματισμός με διαφορετικές τιμές για το ελάχιστο Support

Στο πλαίσιο του πειραματισμού, πραγματοποιούμε αρχικά δοκιμαστικές εκτελέσεις της μεθόδου **Apriori** για την εξαγωγή κανόνων συσχέτισης, χρησιμοποιώντας διαφορετικές τιμές στο όρισμα του **ελάχιστου** **Support**.

Για την εκτέλεση των δοκιμών, δημιουργήθηκε η συνάρτηση **testAssociationRules(groceriesDiscrete)**, εντός της οποίας πραγματοποιούνται 4 συνολικά εκτελέσεις της **apriori()**, με ορίσματα για το ελάχιστο Support τις τιμές **0.001**, **0.02**, **0.03** και **0.04**.

Τα αποτελέσματα που λαμβάνουμε για κάθε μία από τις τιμές παρατίθενται παρακάτω (για λόγους ευκολίας απεικόνισης και μελέτης, σε κάθε εκτέλεση εμφανίζονται τα πρώτα 20 αποτελέσματα των κανόνων συσχέτισης που προκύπτουν):

1. Για **supp = 0.001**:



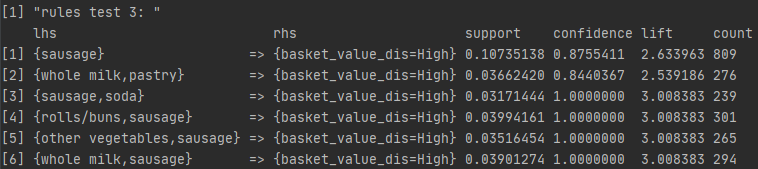
Αξίζει να σημειωθεί ότι για την συγκεκριμένη ελάχιστη τιμή του support, λαμβάνονται συνολικά 750 κανόνες συσχέτισης, από τους οποίους, όπως αναφέρθηκε και πριν, εξετάζονται οι 20.

1. Για **supp = 0.02**:



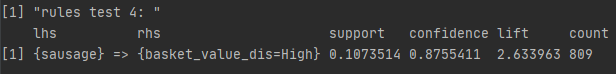
Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγονται μόλις 14 κανόνες συσχέτισης.

1. Για **supp = 0.03**:



Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγονται μόλις 6 κανόνες συσχέτισης.

1. Για **supp = 0.04**:



Παρατηρούμε ότι για την συγκεκριμένη ελάχιστη τιμή του support, εξάγεται μόλις 1 κανόνας συσχέτισης.

1. Κανόνες με το υψηλότερο confidence αποκλειστικά για τα προϊόντα

Σε συνέχεια της διεκπεραίωσης των καθηκόντων της άσκησης, καλούμαστε να βρούμε τους 20 κανόνες με το υψηλότερο confidence αποκλειστικά για τα προϊόντα.

Θα εφαρμοστεί η ίδια λογική με την διαδικασία πειραματισμού που πραγματοποιήθηκε στην υποενότητα 1 της παρούσας άσκησης, ωστόσο αυτή τη φορά η επιλογή της τιμής του ελάχιστου Support παρουσιάζει αυξημένο ενδιαφέρον: Σκοπός μας είναι η επιλογή της μέγιστης δυνατής τιμής του **Support**, για την οποία λαμβάνουμε ως αποτέλεσμα τουλάχιστον 20 κανόνες συσχέτισης προϊόντων (λόγω των ζητούμενων της άσκησης). Ο λόγος πίσω από την επιλογή της μέγιστης δυνατής τιμής **support** για το όρισμα του **ελάχιστου** **support** έχει να κάνει με το ότι οι κανόνες συσχέτισης με **υψηλότερο** **support** αφορούν εξ ορισμού προϊόντα που εμφανίζονται συχνότερα στο σύνολο δεδομένων που έχουμε, συνεπώς αποτελούν σημεία υψηλού ενδιαφέροντος για την ανάλυσή μας.

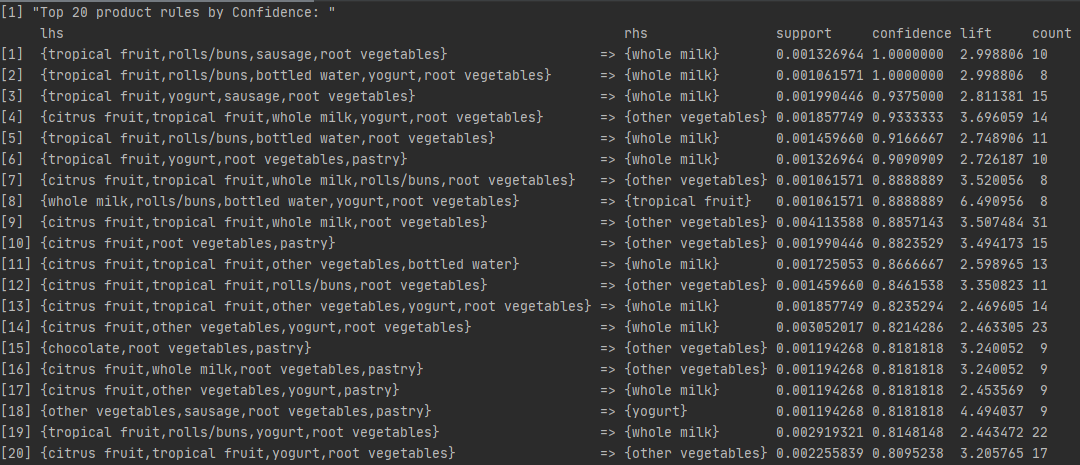
Δοκιμάζοντας διαφορετικές τιμές ελάχιστου **support**, καταλήγουμε στο συμπέρασμα ότι μία πολύ καλή τιμή για το όρισμα είναι η **supp = 0.001**, μιας και επιστρέφει ακριβώς 20 κανόνες συσχέτισης ως αποτέλεσμα, οι οποίοι εμφανίζουν επαρκή διαφοροποίηση μεταξύ τους στην τιμή του **confidence**.

* Για τιμές μικρότερες του **0.001**, λαμβάνουμε περισσότερους από 20 κανόνες συσχέτισης, οι οποίοι ωστόσο παρουσιάζουν μειωμένη διαφοροποίηση μεταξύ τους ως προς την τιμή του **confidence**.

Παράλληλα, το γεγονός ότι οι κανόνες συσχέτισης προέκυψαν από χαμηλότερο ελάχιστο **support** υποδηλώνει ότι στο μεταξύ των αποτελεσμάτων υπάρχουν προϊόντα/τιμές που εμφανίζονται σπανιότερα στο σύνολο δεδομένων που επεξεργαζόμαστε. Αυτό μπορεί να αποσπάσει την προσοχή μας από δεδομένα που παρουσιάζουν μεγαλύτερο ενδιαφέρον ανάλυσης λόγω υψηλότερης συχνότητας εμφάνισης στο σύνολο δεδομένων (**υψηλότερο** **support**).

* Για τιμές μεγαλύτερες του **0.001**, λαμβάνουμε λιγότερους από 20 κανόνες συσχέτισης ως αποτέλεσμα. Είναι προφανές λοιπόν πως λόγω των περιορισμών που δίνονται από την άσκηση, τιμές μεγαλύτερες του **0.001** στο όρισμα του **ελάχιστου** **support** αποκλείονται.

Τελικά, το αποτέλεσμα της εκτέλεσης είναι το εξής:



Διερμηνεύοντας το αποτέλεσμα, εντύπωση προκαλούν οι κανόνες συσχέτισης [1] και [2]:

* [1]: **tropical fruit**, **rolls/buns**, **sausage**, **root vegetables** => **whole milk**
* [2]: **tropical fruit**, **rolls/buns**, **bottled water**, **yogurt**, **root vegetable**s => **whole milk**

Παρατηρούμε δηλαδή ένα μοτίβο καταναλωτικής συμπεριφοράς, που υποδεικνύει ότι το προϊόν «**whole milk**» που βρίσκεται στο δεξί μέρος των 2 κανόνων συσχέτισης αγοράζεται με απόλυτη βεβαιότητα (**confidence = 1**) από καταναλωτές που συνδυάζουν την αγορά των προϊόντων «**tropical fruit**», «**rolls/buns**» και «**root vegetables**» με το προϊόν «**sausage**» ή το ζεύγος προϊόντων «**bottled water**» και «**yogurt**».

1. Κανόνες με το υψηλότερο confidence για τα προϊόντα και την διακριτοποιημένη αξία καλαθιού

Σε αυτό το στάδιο της ανάλυσης μας, καλούμαστε να βρούμε τους 20 κανόνες με το υψηλότερο confidence για τα προϊόντα **και** την διακριτοποιημένη αξία καλαθιού.

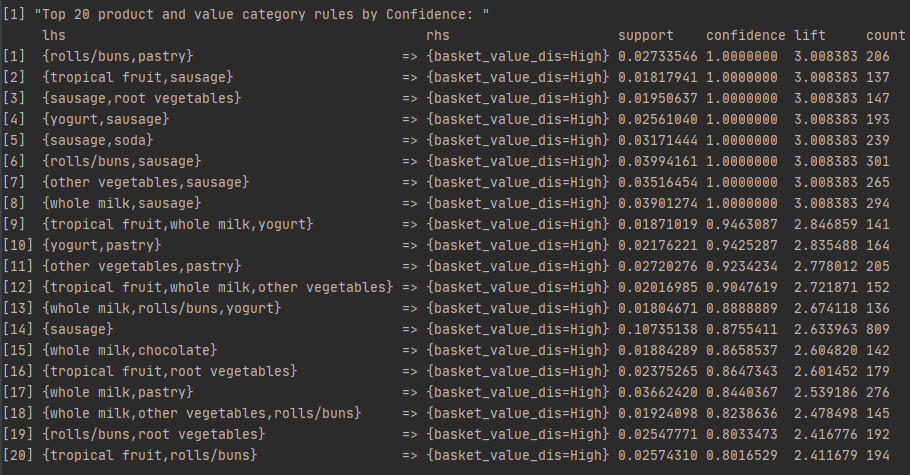
Θα εφαρμοστεί η ίδια λογική με την διαδικασία πειραματισμού που πραγματοποιήθηκε στην υποενότητα 1 της παρούσας άσκησης, ωστόσο αυτή τη φορά η επιλογή της τιμής του ελάχιστου Support παρουσιάζει αυξημένο ενδιαφέρον: Σκοπός μας είναι η επιλογή της μέγιστης δυνατής τιμής του **Support**, για την οποία λαμβάνουμε ως αποτέλεσμα τουλάχιστον 20 κανόνες συσχέτισης προϊόντων (λόγω των ζητούμενων της άσκησης). Ο λόγος πίσω από την επιλογή της μέγιστης δυνατής τιμής **support** για το όρισμα του **ελάχιστου** **support** έχει να κάνει με το ότι οι κανόνες συσχέτισης με **υψηλότερο** **support** αφορούν εξ ορισμού προϊόντα/τιμές που εμφανίζονται συχνότερα στο σύνολο δεδομένων που έχουμε, συνεπώς αποτελούν σημεία υψηλού ενδιαφέροντος για την ανάλυσή μας.

Δοκιμάζοντας διαφορετικές τιμές ελάχιστου **support**, καταλήγουμε στο συμπέρασμα ότι μία πολύ καλή τιμή για το όρισμα είναι η **supp = 0.018**, μιας και επιστρέφει ακριβώς 20 κανόνες συσχέτισης ως αποτέλεσμα, οι οποίοι εμφανίζουν επαρκή διαφοροποίηση μεταξύ τους στην τιμή του **confidence**.

* Για τιμές μικρότερες του **0.018**, λαμβάνουμε περισσότερους από 20 κανόνες συσχέτισης, οι οποίοι ωστόσο παρουσιάζουν μειωμένη διαφοροποίηση μεταξύ τους ως προς την τιμή του **confidence**.

Παράλληλα, το γεγονός ότι οι κανόνες συσχέτισης προέκυψαν από χαμηλότερο ελάχιστο **support** υποδηλώνει ότι στο μεταξύ των αποτελεσμάτων υπάρχουν προϊόντα/τιμές που εμφανίζονται σπανιότερα στο σύνολο δεδομένων που επεξεργαζόμαστε. Αυτό μπορεί να αποσπάσει την προσοχή μας από δεδομένα που παρουσιάζουν μεγαλύτερο ενδιαφέρον ανάλυσης λόγω υψηλότερης συχνότητας εμφάνισης στο σύνολο δεδομένων (**υψηλότερο** **support**).

* Για τιμές μεγαλύτερες του **0.018**, λαμβάνουμε λιγότερους από 20 κανόνες συσχέτισης ως αποτέλεσμα. Είναι προφανές λοιπόν πως λόγω των περιορισμών που δίνονται από την άσκηση, τιμές μεγαλύτερες του **0.018** στο όρισμα του **ελάχιστου** **support** αποκλείονται.

Τελικά, το αποτέλεσμα της εκτέλεσης είναι το εξής:  


Διερμηνεύοντας το αποτέλεσμα, παρατηρούμε αρχικά ότι στο δεξί μέρος των 20 κανόνων συσχέτισης με το υψηλότερο confidence έχουμε πάντα την ιδιότητα basket\_value\_dis = High.

Με πιο προσεκτική μελέτη των περιεχομένων του αριστερού μέρους των κανόνων συσχέτισης [2] => [8], παρατηρούμε ότι οι συγκεκριμένοι κανόνες συσχέτισης έχουν confidence = 1, συνεπώς το δεξί μέρος του κανόνα, δηλαδή η υψηλή διακριτοποιημένη αξία του καλαθιού, είναι απόλυτη συνέπεια για τα περιεχόμενα καλαθιού που παρουσιάζονται στο αριστερό μέρος των κανόνων [2] => [8].

Προχωρώντας σε ακόμα μεγαλύτερο επίπεδο ανάλυσης, παρατηρούμε ότι στους κανόνες [2] => [8] το προϊόν «**sausage**» συμπεριλαμβάνεται σε κάθε περίπτωση μεταξύ των προϊόντων του αριστερού μέρους των κανόνων συσχέτισης.

Αυτό μας οδηγεί στη διαπίστωση ότι με απόλυτη βεβαιότητα (**confidence = 1**), οι συναλλαγές στις οποίες συμπεριλαμβάνεται το προϊόν «**sausage**» παρουσιάζουν υψηλή διακριτοποιημένη αξία καλαθιού (**basket\_value\_dis = High**).

Συνεπώς, λαμβάνοντας υπόψιν τα παραπάνω, είναι πιθανόν ότι το ακριβότερο προϊόν είναι το «**sausage**».