ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΑΘΗΜΑ: **Επιχειρηματική Ευφυΐα και Επιχειρησιακή Έρευνα**
ΕΡΓΑΣΙΑ: **Εξόρυξη γνώσης από δεδομένα συναλλαγών καταστήματος λιανικής**
ΠΡΟΑΙΡΕΤΙΚΗ: ΝΑΙ
ΒΑΘΜΟΛΟΓΙΑ: Μέγιστο 1,5 μονάδα, εφόσον ο βαθμός των εξετάσεων είναι >= 5
ΟΜΑΔΙΚΗ: ΟΧΙ
ΗΜΕΡΟΜΗΝΙΑ ΑΝΑΚΟΙΝΩΣΗΣ: **21/4/2020**
ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ: **31/5/2020**
ΤΡΟΠΟΣ ΠΑΡΑΔΟΣΗΣ: **elearning.auth.gr**

Εξόρυξη γνώσης από δεδομένα συναλλαγών καταστήματος λιανικής

Πρόσφατα ανακαλύφθηκε το ταλέντο σας στην ανάλυση δεδομένων από γνωστή πολυεθνική αλυσίδα καταστημάτων λιανικής. Καλείστε επειγόντως να βοηθήσετε τον υπεύθυνο Marketing στην ανάλυση 7537 συναλλαγών (καλάθια) που έγιναν σε μια περίοδο 75 ημερών και αφορούν 170 κωδικούς προϊόντων. Αποφασίσατε να χρησιμοποιήσετε R για την συγκεκριμένη ανάλυση και μπορείτε να βοηθηθείτε και από τον κώδικα των *Εργαστηριακών Σημειώσεων R* που βρίσκονται στο elearning.

Άσκηση 1. Μετασχηματισμός πρωτογενών δεδομένων

Το αρχικό σύνολο δεδομένων που λαμβάνετε βρίσκετε στο παρακάτω σύνδεσμο:

https://www.icloud.com/iclouddrive/0I8P4KDIh0hyt_LgDeARtupTA#GroceriesInitial

Κάθε γραμμή αντιπροσωπεύει μια μεμονωμένη συναλλαγή. Κάθε στήλη αντιπροσωπεύει ένα χαρακτηριστικό (μεταβλητή) μια συναλλαγής. Τα χαρακτηριστικά αυτά είναι: η συνολική αξία της συναλλαγής (basket_value), το πλήθος ημερών που περάσαν από τη μέρα της συναλλαγής - αντί για ημερομηνία συναλλαγής (recency_days), ένας μοναδικός κωδικός συναλλαγής (id) και τέλος αναλυτικά τα προϊόντα που αγοράστηκαν στη συγκεκριμένη συναλλαγή.

Μετασχηματίστε τα παραπάνω πρωτογενή δεδομένα σε μορφή κατάλληλη για την εφαρμογή των μεθόδων κανόνων συσχέτισης της R (δυαδική μορφή συναλλαγών) και εισάγετε τα σε αυτή.

Λάβετε υπόψη ότι το τμήμα μάρκετινγκ σας ενημέρωσε ότι ενδιαφέρεται αποκλειστικά για τα 13 από τα 170 προϊόντα (αρα πρέπει να κρατήσετε μόνο αυτές τις στήλες από τη δυαδική μορφή συναλλαγών) και συγκεκριμένα για τα:

citrus fruit, tropical fruit, whole milk, other vegetables, rolls/buns, chocolate, bottled water, yogurt, sausage, root vegetables, pastry, soda, cream

Επίσης, για να μπορείτε να χρησιμοποιείστε και την αξία συναλλαγής (basket_value) στους κανόνες συσχέτισης στην (επόμενη) Ασκηση 2, διακριτοποίηστε την σε τρείς (περίπου) ισοπληθείς κατηγορίες:

low_value_basket, medium_value_basket, high_value_basket

Καταγράψτε εν συντομία το σύνολο της διαδικασίας που ακολουθήσατε.

Άσκηση 2. Μάθηση κανόνων συσχέτισης με την μέθοδο Apriori

Σημείωση: Πέρα απ όσα κάναμε στο μάθημα θα σας βοηθήσει και το παρακάτω tutorial για το arules, π.χ στην ταξινόμηση των κανόνων με βάσει το support: http://www.rdatamining.com/examples/association-rules

Χρησιμοποιείστε το επεξεργασμένο σύνολο δεδομένων που δημιουργήσατε στην Άσκηση 1 για τη μάθηση κανόνων συσχέτισης στην R αποκλειστικά για τα χαρακτηριστικά των **προϊόντων και τη διακριτοποιημένη αξία καλαθιού.**

Αν δεν ολοκληρώσατε επιτυχώς την Άσκηση 1 μπορείτε να κατεβάσετε τα δεδομένα σε επεξεργασμένη μορφή – έτοιμη για εισαγωγή - από εδώ:

https://www.icloud.com/iclouddrive/09wa4_rs1EkZSyysmYqt7fJiQ#GroceriesProcessed

α) Δοκιμάστε την εκτέλεση της μεθόδου Apriori με διάφορες παραμέτρους για το ελάχιστο Support

β) Βρείτε τους 20 κανόνες με το υψηλότερο confidence αποκλειστικά για τα προϊόντα. Καταγράψτε τους και ερμηνεύστε το αποτέλεσμα αναφερόμενοι π.χ στο συνδυασμό που σας έκανε τη μεγαλύτερη εντύπωση και γιατί.

γ) Βρείτε του 20 κανόνες με το υψηλότερο confidence για τα προϊόντα **και** την διακριτοποιημένη αξία καλαθιού. (ο αλγόριθμος να χρησιμοποιήσει τώρα και αυτές τις μεταβλητές). Καταγράψτε το αποτέλεσμα. Ποιο είναι πιθανών το ακριβότερο προϊόν και γιατί;

Άσκηση 3. Ομαδοποίηση συναλλαγών με χρήση μεθόδου k-means

Σημείωση: στην R θα χρησιμοποιήσετε την συνάρτηση kmeans(), περισσότερα στο http://www.statmethods.net/advstats/cluster.html

Στο επεξεργασμένο σύνολο δεδομένων καλείστε να ανακαλύψετε ομάδες συναλλαγών που μπορεί να έχουν ιδιαίτερο ενδιαφέρον για το τμήμα Μαρκετινγκ. π.χ συναλλαγές μεγάλης αξία που γινόταν παλαιότερα αλλά δεν γίνονται σήμερα. Στη συνέχεια, περιγράψετε το προφίλ των ομάδων που ανακαλύφθηκαν. Συγκεκριμένα:

α) Εφαρμόστε τη μέθοδο clustering k-means στα δύο συνεχή χαρακτηριστικά ***basket_value*** και ***recency_days*** για να εξάγετε 5 ομάδες συναλλαγών. Καταγράψτε συνοπτικά τη διαδικασία που ακολουθήσατε και την έξοδο που πήρατε από την R – ακριβώς όπως την πήρατε.

β) Για την ομαδοποίηση 5 ομάδων στην οποία καταλήξατε, αναφέρατε τη μέση τιμή των κέντρων των ομάδων που βγήκαν και τη τυπική τους απόκλιση. Ερμηνεύστε τις ομάδες μέσων αυτών. π.χ Ομάδα 1 --> “Ομάδα πρόσφατων συναλλαγών μικρής αξίας που αντιπροσωπεύει το 10% του συνόλου των συναλλαγών”. Αυτό είναι το αριθμητικό προφίλ της κάθε ομάδας.

- Υπάρχει κάποια ανησυχητική ομάδα συναλλαγών με την οποία θα έπρεπε να ασχοληθεί το τμήμα Μάρκετινγκ π.χ μεγάλης αξίας που γινόταν παλαιότερα;

γ) Εξάγετε τις αναθέσεις της κάθε μιας συναλλαγής σε ομάδα σε μια νέα ποιοτική μεταβλητή (στήλη) έτσι ώστε να είναι εφικτή η μάθηση κανόνων συσχέτισης και σε αυτό το νέο χαρακτηριστικό. Η ονομασία αυτού του νέου χαρακτηριστικού μπορεί να είναι “Cluster”. Στη συνέχεια όμως, θα πρέπει να μετατρέψτε και αυτήν την μεταβλητή για να έχει την κατάλληλη μορφή για εφαρμογή κανόνων συσχέτισης. Δηλαδή, θα πρέπει να αποθηκεύστε την ομάδα της κάθε συναλλαγής, με τη χρήση 5 χαρακτηριστικών-μεταβλητών (Cluster1, Cluster2 κλπ) για να είναι εφικτή η εφαρμογή των κανόνων συσχέτισης, πρέπει δηλαδή να παράγετε και πάλι την δυαδική μορφή των συναλλαγών.

Άσκηση 4. Συνδυαστική αξιοποίηση μεθόδων: περιγραφή προιοντικού προφίλ ομάδων με χρήση κανόνων συσχέτισης

Σημείωση: Αν δεν ολοκληρώσατε επιτυχώς την Άσκηση 3, ένα σύνολο με ενδεικτική ομαδοποίηση βρίσκεται εδώ: https://www.icloud.com/iclouddrive/0x_3z2rnFz4xteXa3tBQpp4DA#GroceriesClustered

Στο σύνολο δεδομένων που προέκυψε από την Άσκηση 3 προσπαθήστε να περιγράψτε το **προιοντικό προφίλ** της κάθε ομάδας συναλλαγών με χρήση κανόνων συσχέτισης. Συγκεκριμένα:

Με τη μέθοδο Apriori βρείτε τους 20 κανόνες με το υψηλότερο confidence αποκλειστικά **για τα προϊόντα και τις ομάδες.** Καταγράψτε τους και ερμηνεύστε το αποτέλεσμα.

- Ποια προϊόντα ή συνδυασμοί τους παρατηρείτε ότι αγοράζονται κατά κύριο λόγο από την κάθε ομάδα
- Αν εντοπίσατε από την Άσκηση 3 κάποια ανησυχητική ομάδα συναλλαγών, με ποιο προϊόν συνήθως αυτή σχετίζεται; Δώστε την ερμηνεία σας σχετικά με το τι μπορεί να έχεις συμβεί σχετικά με αυτό το προϊόν πιθανών συνδυαστικά και με όποια άλλη πληροφορία έχετε

Άσκηση 5. Συνδυαστική εφευρετικότητα: εφαρμογή μεθόδων Γραμμικού Προγραμματισμού σε αποτελέσματα ανάλυσης δεδομένων

Σε αυτή την άσκηση καλείστε να εντοπίσετε ένα πρόβλημα βελτιστοποίησης με περιορισμούς για μια διαφημιστική-προωθητική ενέργεια του τμήματος Μαρκετινγκ, χρησιμοποιώντας τα δεδομένα αλλά και τα αποτελέσματα των προηγούμενων ασκήσεων.

Ακολουθεί ενδεικτικό παράδειγμα το οποίο αν θέλετε μπορείτε να επεκτείνετε (ή εναλλακτικά να κάνετε κάτι εντελώς διαφορετικό):

“Για κάθε κανόνα μπορείτε να υπολογίσετε το μέσο basket value των συναλλαγών που τον επιβεβαιώνουν. Έστω ότι θα επιλέγονται τυχαία άτομα στο ταμείο για εκπτωτική προσφορά όταν επιβεβαιώνουν τουλάχιστον ένα από τους 5 κανόνες με το μεγαλύτερο support. Ποια η αναλογία ανά κανόνα των ατόμων που θα επιλέγονται έτσι ώστε να μεγιστοποιείται το μέσο basket value τους. Π.χ για κάθε 3 άτομα που δέχονται προσφορά για τον κανονα 1, 1 άτομο θα δέχεται προσφορά για τον κανόνα 2. Ο βασικός περιορισμός είναι ότι σε κανένα κανόνα δεν θα ανατεθούν αναλογικά περισσότεροι πελάτες από το support του κανόνα“

Τέλος, αφού παρουσιάσετε το πρόβλημα βελτιστοποίησης και την τυπική του μορφή (Γραμμικού Προγραμματισμού) επιχειρήστε να το λύσετε στην R με την βοήθεια των σχετικών πακέτων που αναφέρονται στις Εργαστηριακές Σημειώσεις R στο elearning.