ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΡΓΑΣΤΗΡΙΟ ΓΛΩΣΣΩΝ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ & ΤΕΧΝΟΛΟΓΙΑΣ ΛΟΓΙΣΜΙΚΟΥ

ΜΑΘΗΜΑ: Επιχειρηματική Ευφυία και Επιχειρησιακή Έρευνα
ΕΡΓΑΣΙΑ: Εξόρυξη γνώσης από δεδομένα συναλλαγών καταστήματος λιανικής
ΒΑΘΜΟΛΟΓΙΑ:
ΟΜΑΔΙΚΗ:
ΗΜΕΡΟΜΗΝΙΑ ΑΝΑΚΟΙΝΩΣΗΣ:
ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ:
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 Python για την συγκεκριμένη ανάλυση και μπορείτε να βοηθηθείτε και από τον κώδικα των Notebooks που κάναμε στο σχετικό μάθημα.

Άσκηση 1. Μετασχηματισμός και ανάλυση πρωτογενών δεδομένων

Το αρχικό σύνολο δεδομένων που λαμβάνετε βρίσκεται στο παρακάτω σύνδεσμο:

https://drive.google.com/file/d/1qklC6HBqynRmhn5q-zY3Ov2a0jcCwT9s/view?usp=sharing

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 πέρασαν 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.

Ξεκινήστε με ανάλυση των δεδομένων και επιλέξτε τις κατάλληλες οπτικοποιήσεις ώστε να αναδείξετε τυχόν γνώση ή ενδεικτικές τάσεις που κρύβονται στα δεδομένα.

Μετασχηματίστε τα παραπάνω πρωτογενή δεδομένα σε μορφή κατάλληλη για την εφαρμογή των μεθόδων κανόνων συσχέτισης της Python (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 χρησιμοποιήσετε και την αξία συναλλαγής (basket_value) στους κανόνες συσχέτισης στην Ασκηση 2, διακριτοποίηστε την σε τρεις (περίπου) ισοπληθείς κατηγορίες:

low_value_basket, medium_value_basket, high_value_basket

Καταγράψτε εν συντομία το σύνολο της διαδικασίας που ακολουθήσατε. Δικαιολογήστε τις αποφάσεις που πήρατε στην πορεία και οπτικοποιήστε κατάλληλα τα τελικά δεδομένα. Τι συμπεράσματα μπορούμε να εξάγουμε μετά την επεξεργασία;

Άσκηση 2. Μάθηση κανόνων συσχέτισης με την μέθοδο Apriori

Σημείωση: Ταξινόμηση κανόνων με βάση το Support και των άλλων μετρικών που παρουσιάστηκαν στο μάθημα.

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

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;

δ) Σε τι συμπεράσματα μπορείτε να καταλήξετε με βάση τους κανόνες; Είναι συναφή με την αρχική σας ανάλυση; Πως διαφέρουν οι τάσεις και γιατι;

Άσκηση 3. Ομαδοποίηση συναλλαγών με χρήση μεθόδου k-means

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 Python – ακριβώς όπως την πήρατε. Επιπλέον, παρουσιάστε τα αποτελέσματα με γραφήματα.

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 μέσω 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 σε ομάδα της κάθε μιας συναλλαγής σε μια νεα ποιοτική μεταβλητή (στήλη) έτσι ώστε να είναι εφικτή η μάθηση κανόνων συσχέτισης και σε αυτό το νέο χαρακτηριστικό. Για να έχει την κατάλληλη μορφή θα πρέπει να αποθηκεύεστε την ομάδα της κάθε συναλλαγής με τη χρήση 5 χαρακτηριστικών-μεταβλητών (Cluster1, Cluster2 κλπ) για να είναι εφικτή η εφαρμογή των κανόνων συσχέτισης, δηλαδή πρέπει να παράγετε και παλι την δυαδική μορφή των συναλλαγών. Οπτικοποιήστε τα δεδομένα σας ανάλογα ώστε να είναι δυνατή η εξαγωγή συμπερασμάτων.

Άσκηση 4. Συνδυαστική αξιοποίηση μεθόδων: περιγραφή προϊοντικού και γενικού προφίλ ομάδων με χρήση κανόνων συσχέτισης

Σημείωση: Αν δεν ολοκληρώσατε επιτυχώς την Άσκηση 3, δημιουργήστε ένα σύνολο δεδομένων με τυχαία ενδεικτική ομαδοποίηση και περιγράψτε τη διαδικασία που το δημιουργήσατε.

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

α) Με τη μέθοδο Apriori βρείτε τους 20 κανόνες με το υψηλότερο confidence αποκλειστικά **για τα προϊόντα και τις ομάδες των συναλλαγών.** 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

β) Βρείτε τους 20 κανόνες με το υψηλότερο confidence **για τα προϊόντα, τις ομάδες των συναλλαγών και την διακριτοποιημένη αξία καλαθιού**. (ο αλγόριθμος να χρησιμοποιήσει τώρα και αυτές τις μεταβλητές). Καταγράψτε το αποτέλεσμα. Προκύπτουν ενδιαφέροντες νέοι κανόνες; Αν ναι, καταγράψτε τους.

Κατάθεση εργασίας

Στο τέλος της εργασίας θα πρέπει να κατατεθεί μέσω elearning από κάθε φοιτητή/φοιτήτρια ένα αρχείο μορφής **ZIP** (όχι άλλης μορφής συμπίεσης όπως .tar, .7zip, κοκ) με όνομα ΑΕΜ_Ονοματεπώνυμο_ΕΕ1.zip (π.χ. 1234_ΓιώργοςΠαπαδόπουλος_ΕΕ1.zip). Το κάθε αρχείο θα πρέπει να περιέχει:

1. Ένα Python Notebook με τον κώδικα / λύση των ασκήσεων και τις απαντήσεις στις ασκήσεις - όπως αναφέρονται σε κάθε άσκηση. Να αξιοποιήσετε κελιά κώδικα και κειμένου αναλόγως ώστε να περιγράφονται τα αποτελέσματα. Το Notebook θα

πρέπει να μπορεί να εκτελεστεί χωρίς σφάλματα και να αναπαράγονται τα αποτελέσματα καθώς και τα γραφήματα.

Υποβολές που δεν έχουν τη σωστή ονομασία, κωδικοποίηση και δεν περιέχουν όλα τα αρχεία θα θεωρούνται άκυρες.