Μάθημα: ΜΕΘΟΔΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ακαδημαϊκή περίοδος: Χειμερινό εξάμηνο 2021-2022

1η Εργασία

Το αρχείο market.csv περιλαμβάνει δεδομένα για 440 πελάτες χονδρικής, 298 από τη βιομηχανία τροφίμων και 142 από τη βιομηχανία του λιανικού εμπορίου. Οι πελάτες προέρχονται από δύο μεγάλες πόλεις (Λισαβόνα και Πόρτο) ή από κάποια άλλη περιοχή της Πορτογαλίας. Τα δεδομένα αφορούν στις ετήσιες δαπάνες σε νομισματικές μονάδες ανά κατηγορία προϊόντος: νωπά προϊόντα, γαλακτοκομικά προϊόντα, είδη παντοπωλείου, κατεψυγμένα προϊόντα, απορρυπαντικά / χαρτικά και είδη delicatessen. Οι παραπάνω πληροφορίες καταγράφονται στις εξής μεταβλητές:

- `Channel`: Βιομηχανία (1: βιομηχανία τροφίμων, 2: λιανικό εμπόριο)
- `Region`: Περιοχή (1: Λισαβόνα, 2: Πόρτο, 3: άλλη περιοχή)
- `Fresh`: Ετήσια δαπάνη σε νωπά προϊόντα
- `Milk`: Ετήσια δαπάνη σε γαλακτοκομικά προϊόντα
- `Grocery`: Ετήσια δαπάνη σε είδη παντοπωλείου
- `Frozen`: Ετήσια δαπάνη σε κατεψυγμένα προϊόντα
- `Detergents_Paper`: Ετήσια δαπάνη σε απορρυπαντικά/ χαρτικά
- `Delicatessen`: Ετήσια δαπάνη σε είδη delicatessen

Σκοπός είναι η ταξινόμηση των πελατών σε ομάδες με βάση την πληροφορία που είναι διαθέσιμη για τις ετήσιες δαπάνες τους ανά κατηγορία προϊόντος (όλες δηλαδή τις μεταβλητές εκτός από τις `Channel` και `Region`).

Εφαρμόστε διαφορετικές τεχνικές clustering (k-means, ιεραρχική ομαδοποίηση, model-based clustering) προκειμένου να απαντήσετε στα ακόλουθα ερωτήματα:

i. Ποια μέθοδος είναι προτιμότερη για τα δεδομένα μας και γιατί;
ii. Πόσο επιτυχημένη μπορεί να θεωρηθεί συνολικά η ομαδοποίηση στην οποία καταλήξατε;
iii. Υπάρχουν ομάδες που να αναγνωρίζονται με σαφήνεια από όλες τις μεθόδους;
iv. Υπάρχουν ομάδες που να μην είναι και τόσο ευδιάκριτες;
v. Ερμηνεύστε τα αποτελέσματα της ομαδοποίησης στην οποία καταλήξατε λαμβάνοντας υπόψη όλη τη διαθέσιμη πληροφορία (λάβετε δηλαδή υπόψη στην ερμηνεία και τις μεταβλητές `Channel` και `Region`).

Τα παραδοτέα της εργασίας περιλαμβάνουν:

- Το script file με τον κώδικά σας,
- Μία γραπτή αναφορά που θα περιλαμβάνει **αιτιολογημένες** απαντήσεις στα παραπάνω ερωτήματα και μία παράγραφο που να συνοψίζει τις τελικές σας σκέψεις και συμπεράσματα. Όπως σε κάθε ανάλυση, θα πρέπει προφανώς να ξεκινήσετε και να συμπεριλάβετε στην παρουσίαση των αποτελεσμάτων σας βασικά περιγραφικά μέτρα και γραφήματα για όλες τις μεταβλητές.

Η εργασία θα βαθμολογηθεί με άριστα το 10 και θα μετρήσει κατά 10% στον τελικό σας βαθμό.

Η εργασία θα πρέπει να αναρτηθεί στο eclass μέχρι την Παρασκευή 17 Δεκεμβρίου 2021 στις 24:00. Καμία εργασία δε θα γίνει δεκτή μετά από τη συγκεκριμένη ημερομηνία και ώρα.

Καλή επιτυχία!