Εξόρυξη Δεδομένων 2023-2024

Στα πλαίσια της εργασίας καλείστε να επεξεργαστείτε ένα σύνολο δεδομένων με σκοπό να εξάγετε χρήσιμη γνώση κάνοντας χρήση βασικών αλγορίθμων εξόρυξης και αξιολογώντας τα αποτελέσματά σας.

ΔΕΔΟΜΕΝΑ

Για το σκοπό της εργασίας θα πρέπει να κατεβάσετε το σύνολο δεδομένων που θα βρείτε στο eclass: Το αρχείο movies.xlsx περιέχει ένα σύνολο από ταινίες του Hollywood που κυκλοφόρησαν από το 2007 μέχρι και το 2018. Κάποιες από τις ταινίες έλαβαν Oscar και κάποιες άλλες όχι.

Καλείστε

α) να εκπαιδεύσετε ένα μοντέλο που θα προβλέπει αν μια ταινία θα πάρει βραβείο Oscar λαμβάνοντας υπόψη το σύνολο των διαθέσιμων γνωρισμάτων.

β) να κάνετε μια ομαδοποίηση των ταινιών χρησιμοποιώντας τα διαθέσιμα γνωρίσματα και να περιγράψετε τις δύο πολυπληθέστερες ομάδες.

γ) να εντοπίσετε αν δημιουργείται μια ομάδα που περιλαμβάνει κυρίως τις ταινίες που πήραν βραβείο Oscar και να δώσετε μια περιγραφή γι' αυτή.

Προτείνεται να μελετήσετε περισσότερο το παράδειγμα που υπάρχει εδώ:
https://www.kaggle.com/code/farzadnekouei/customer-segmentation-recommendation-system

ΕΡΓΑΣΙΕΣ

Α) Προετοιμασία (10%)

Το αρχείο που σας δίνεται είναι σε μορφή xlsx οπότε μπορείτε να χρησιμοποιήσετε Colab ή όποια άλλη πλατφόρμα επιθυμείτε και να κάνετε τις απαραίτητες μετατροπές σε τύπους δεδομένων (π.χ. numeric σε nominal), και να απορρίψετε γνωρίσματα που αποφασίζετε ότι δεν χρειάζεστε. Επίσης μπορείτε να χρησιμοποιήσετε εξωτερική γνώση για να το συμπληρώσετε τιμές που λείπουν (π.χ imdb rating, distributor).

Β1) Κατηγοριοποίηση (50-60%)

Το μοντέλο που θα εκπαιδεύσετε θα πρέπει να μπορεί να κατηγοριοποιεί κάθε στιγμιότυπο σε μια από τις 2 κλάσεις (1-πήρε oscar ή 0-δεν πήρε oscar). Θα πρέπει να δοκιμάσετε αλγόριθμους κατηγοριοποίησης της επιλογής σας με στόχο να έχετε όσο το δυνατόν καλύτερες επιδόσεις και το μοντέλο σας να μπορεί να γενικεύσει αποτελεσματικά.

Βεβαιωθείτε ότι έχετε αποφύγει να υπερεκπαιδεύσετε το μοντέλο σας.

Β2) Πρόβλεψη σε άγνωστες ταινίες (10%)

Στις 23/12/2023 θα σας μοιραστεί από το eclass ένα επιπλέον σύνολο δεδομένων που θα περιέχει στοιχεία για άγνωστες ταινίες, για τις οποίες θα πρέπει να προβλέψετε αν πήραν oscar ή όχι. Οι ταινίες αφορούν τα έτη 2018 και μετά, και περιέχουν έναν μικρό αριθμό από νικήτριες.

Γ) Ομαδοποίηση (30%)

Εφαρμόζοντας 2-3 τεχνικές συσταδοποίησης καταλήξτε σε ομάδες ταινιών με κοινά χαρακτηριστικά. Δώστε χαρακτηρισμό για την κάθε ομάδα.

Αξιολογήστε την ποιότητα της συσταδοποίησης ως προς το αν κατάφερε να διαχωρίσει τις ταινίες που πήραν βραβείο από αυτές που δεν πήραν.

ΠΑΡΑΔΟΣΗ

- Στις 22/12/2023 θα πρέπει να ανεβάσετε στο eclass μια αρχική αναφορά των εργασιών (τουλάχιστον προετοιμασίας) που έχετε υλοποιήσει μέχρι τότε.
- Στις 12/1/2024 θα πρέπει να ανεβάσετε τις τελικές προβλέψεις σας για τις αγνωστες ταινίες που πήραν oscar.
- Στις 19/1/2024 θα πρέπει να ανεβάσετε το τελικό κείμενο με την τεκμηρίωση της εργασίας σας στο eclass.

Στην τελική τεκμηρίωση που θα παραδώσετε θα πρέπει
1) να αναλύσετε τη διαδικασία που ακολουθήσατε σε κάθε εργασία (προετοιμασία, εκπαίδευση ταξινομητή, αξιολόγηση ταξινομητή σε γνωστά και άγνωστα δείγματα) και αφορά: τους μετασχηματισμούς που κάνατε στο σύνολο δεδομένων, τον αλγόριθμο που χρησιμοποιήσατε, τις παραμέτρους που δοκιμάσατε και πως καταλήξατε σε αυτές, τις επιδόσεις που είχατε στα δεδομένα εκπαίδευσης αλλά και στα δεδομένα ελέγχου κλπ.
2) να δώσετε τις εκτιμήσεις σας για τις συστάδες που δημιουργούνται, τις περιγραφές αυτών και την ποιότητα της συσταδοποίησης που παράξατε, σε σχέση με το διαχωρισμό των οσκαρικών ταινιών από τις υπόλοιπες

Οι εργασίες θα βαθμολογηθούν:
ι) για την προετοιμασία: η περιγραφή των ενεργειών (10%)
ιι) για την κατηγοριοποίηση: α) η περιγραφή των ενεργειών (30%-40%), β) οι επιδόσεις που είχατε στα δεδομένα εκπαίδευσης και ελέγχου (20%), γ) η επίδοση του αλγορίθμου στα άγνωστα δεδομένα (10%).
ιιι) για τη μεθοδολογία που ακολουθήσατε για τη συσταδοποίηση (30%)

Η εργασία είναι για 3 (το πολύ) άτομα.