ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ - ΕΡΓΑΣΙΑ 2020-2021

Στην εργασία σας καλείστε να αξιοποιήσετε τα σύνολα δεδομένων του Airbnb που παρέχονται από την ίδια την υπηρεσία για διάφορες πόλεις του κόσμου (http://insideairbnb.com/get-the-data.html)
Προτείνεται να εστιάσετε στα δεδομένα των αρχείων listings.csv που περιέχουν αρκετά αριθμητικά γνωρίσματα, αλλά μπορείτε να χρησιμοποιήσετε βοηθητικά και τα υπόλοιπα δεδομένα, καθώς και εξωτερικά δεδομένα.
Θα ασχοληθούμε με τα δεδομένα για τις Ελληνικές περιοχές μόνο για τους μήνες Ιούνιο και Ιούλιο του 2020.

1ο μέρος - Επεξεργασία δεδομένων

Στο μέρος αυτό θα πρέπει να εστιάσετε στην προεπεξεργασία των δεδομένων σας ώστε να υποστηρίξετε την επεξεργασία που θα ακολουθήσει στα υπόλοιπα μέρη της εργασίας.

Στόχοι σας είναι:
α) Να εξάγετε επιπλέον γνωρίσματα που θα σας βοηθήσουν να βελτιώσετε την πρόβλεψη. Τα γνωρίσματα μπορούν να προκύπτουν από το ίδιο το σύνολο δεδομένων ή να αξιοποιούν εξωτερικά δεδομένα (π.χ. Δεδομένα για POIs, Δεδομένα για περιοχές
β) Να εντοπίσετε αν υπάρχουν γνωρίσματα με μεγάλη συσχέτιση μεταξύ τους.
γ) Να μετατρέψετε τα γνωρίσματα σε τύπους που απαιτεί η κάθε τεχνική ή ο αντίστοιχος αλγόριθμος.

2ο μέρος - Εξόρυξη γνώσης

Στο μέρος αυτό θα πρέπει να αξιολογήσετε διαφορετικές τεχνικές εξόρυξης γνώσης στα δεδομένα του ίδιου συνόλου. Θα χρησιμοποιήσετε για εκπαίδευση το σύνολο δεδομένων για Αθήνα και Θεσσαλονίκη για τον Ιούνιο και Ιούλιο του 2020 (δείτε τα archived data κάθε πόλης).
Θα πρέπει να εκπαιδεύσετε/επαληθεύσετε σε μια αναλογία 90/10 στα δεδομένα αυτά.

Α) Παλινδρόμηση (Πειράματα)
Να εκπαιδεύσετε μοντέλα που θα προβλέπουν την τιμή ενοικίασης ενός ακινήτου λαμβάνοντας υπόψη τα βασικά γνωρίσματα που έχετε στο αρχικό dataset ή και περισσότερα γνωρίσματα αν έχετε προσθέσει μετά από επεξεργασία.
Να δώσετε ένα συγκριτικό πίνακα με τα αποτελέσματα όλων των μοντέλων που θα δοκιμάσετε.
Παρουσιάστε τη διαφορά στις επιδόσεις σας αν εκπαιδεύσετε διαφορετικά μοντέλα για κάθε πόλη. Χρησιμοποιήστε αντίστοιχο πίνακα με πριν.

Β) Κατηγοριοποίηση (Πειράματα)
Θα πρέπει να εκπαιδεύσετε ένα κατηγοριοποιητή που θα προβλέπει την κατηγορία τιμής ενός ακινήτου. Τα όρια είναι 50, 100, 200 και 500.
Να δώσετε ένα συγκριτικό πίνακα με τα αποτελέσματα όλων των μοντέλων που θα δοκιμάσετε.

Επιλέξτε το καλύτερο μοντέλο σας και απαντήστε αν είναι σημαντικά καλύτερα από τα υπόλοιπα και σε ποιο διάστημα εμπιστοσύνης.

Γ) Γενίκευση μοντέλων
Εκπαιδεύστε το μοντέλο σας στην Αθήνα (σε όλο το dataset) και δοκιμάστε την απόδοσή του στη Θεσσαλονίκη.
Κάνε το το ίδιο για τον Ιούνιο στις 2 πόλεις και δοκιμάστε στον Ιούλιο (για τις δύο πόλεις).
Δείξτε αν το μοντέλο σας είναι αρκετά γενικό.

Δοκιμάστε το ίδιο μοντέλο για τα πιο πρόσφατα δεδομένα μιας άλλης πόλης.

Τι μπορείτε να κάνετε για να έχετε ένα πιο γενικό μοντέλο;

Πώς μπορείτε να αξιοποιήσετε το αρχικό μοντέλο (εκπαιδευμένο σε Αθήνα και Θεσσαλονίκη) για να μην ξαναρχίσετε από την αρχή, εκπαιδεύοντας ένα μοντέλο για μια άλλη πόλη;

<u>3ο μέρος - Παρουσίαση αποτελεσμάτων και Τεκμηρίωση</u>
Συγκεντρώστε τα συμπεράσματα τις δουλειάς σας σε 15 διαφάνειες που θα περιγράφουν i) την προεπεξεργασία που κάνατε στο σύνολο δεδομένων, ii) τη μεθοδολογία που χρησιμοποιήσατε σε regression και classification και τα αποτελέσματα που επιτύχατε, iii) τα αποτελέσματά σας στη γενίκευση των μοντέλων.

Αποτυπώστε επίσης όλα τα παραπάνω σε μια αναλυτική τεχνική αναφορά.

ΠΑΡΑΔΟΣΗ
Μια μέρα πριν την τελευταία διάλεξη (ενδεικτικά στις **17/1/2021**) στο eclass θα πρέπει να ανεβάσετε την παρουσίασή σας και στην τελευταία διάλεξη θα γίνει η παρουσίαση σε όλους. Λίγες ημέρες μετά και πριν την έναρξη της εξεταστικής θα πρέπει να ανεβάσετε και την τελική σας τεκμηρίωση και πάλι στο e-class.

Η εργασία είναι για 2 (το πολύ) άτομα.