Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής
Προπτυχιακό Πρόγραμμα Σπουδών
Ακαδ. έτος 2021-22 (εαρινό εξάμηνο)

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

Απαλλακτική Εργασία
(Ομάδες των 1-3 ατόμων)
Ημερομηνία παράδοσης: *Ημερομηνία εξέτασης του μαθήματος, 23:59μμ*

Στο πλαίσιο της εργασίας θα χρησιμοποιήσετε το σύνολο δεδομένων "US Accidents" από το Kaggle (https://www.kaggle.com/sobhanmoosavi/us-accidents). Αποτελείται από περίπου 2.8 εκατομμύρια εγγραφές, και περιέχει πληροφορίες σχετικά με αυτοκινητιστικά ατυχήματα στην Αμερική κατά το χρονικό διάστημα Feb. 2016 - Dec 2021 (ανανεώνεται σε ετήσια βάση).

Σκοπός της εργασίας είναι, δοθέντος του παραπάνω dataset, να κατασκευάσετε μια “ιστορία” γύρω από τα δεδομένα, της οποίας τα συμπεράσματα θα ενισχύσετε μέσω της χρήσης μοντέλων Μηχανικής Μάθησης.

Ερώτημα 1 - Εξοικείωση με τα δεδομένα

Το πρώτο (και βασικότερο) βήμα σε οποιοδήποτε πείραμα Αναλυτικής Δεδομένων είναι η εξοικείωση του ερευνητή με τα δεδομένα του. Αφού κατεβάσετε το παραπάνω σύνολο δεδομένων, προχωρήσετε σε όποια προπαρασκευαστική εργασία (επιλογή, οπτικοποίηση, καθαρισμό, μετασχηματισμό, δειγματοληψία, κλπ.) θεωρήσετε απαραίτητη ώστε: α) να «καθαρίσετε» τα δεδομένα από ελλιπείς ή εσφαλμένες τιμές, β) να κανονικοποιήσετε – διακριτοποιήσετε τα δεδομένα (π.χ. για αντιμετώπιση των συνεχών πεδίων τιμών), γ) να μειώσετε τον όγκο των δεδομένων (π.χ., μείωση διαστάσεων).

Ερώτημα 2 - Ομαδοποίηση

Έχοντας “καθαρίσει” το εν λόγω σύνολο δεδομένων, το επόμενο βήμα στην πειραματική μας διαδικασία είναι η χρήση τεχνικών ομαδοποίησης, προκειμένου να αποκαλύψουμε ιδιότητες του συνόλου δεδομένων, που δεν ήταν ορατές κατά το στάδιο της προεπεξεργασίας.

Επιπλέον, κάποιες τεχνικές ομαδοποίησης μπορούν να χρησιμοποιηθούν (με την κατάλληλη μοντελοποίηση) και για ταξινόμηση χωρίς επίβλεψη (unsupervised classification). Εφόσον μοντελοποιήσετε κατάλληλα (<Feature(s)>, <Label(s)>) το σύνολο δεδομένων, χρησιμοποιήστε τουλάχιστον τρεις διαφορετικές τεχνικές και συγκρίνετε τις ως προς την

ποιότητα/αποτελεσματικότητα της ταξινόμησης (π.χ., με scatter plots, confusion matrices, clustering metrics, etc.).

Ερώτημα 3: Classification/Regression/Storytelling

Έχοντας εξοικειωθεί με τα δεδομένα, το τελευταίο βήμα της πειραματικής μας διαδικασίας είναι να χρησιμοποιήσουμε μοντέλα Μηχανικής Μάθησης, η χρήση των οποίων θα αποτελέσει σημαντικό αρωγό στη δημιουργία της "ιστορίας" μας.

Σε αυτό το στάδιο, σημαντικό ρόλο παίζει η διατύπωση του προβλήματος (π.χ. εκτίμηση σοβαρότητας ατυχήματος), καθώς δυνητικά μπορεί να διαφέρει αρκετά από αυτή του προηγούμενου βήματος. Εφ' όσον καταλήξετε στην τελική δομή του συνόλου δεδομένων, δημιουργήστε (μέσω του εργαλείου TensorFlow/Keras) τουλάχιστον 3 μοντέλα Μηχανικής Μάθησης και (όπως και στο προηγούμενο βήμα) συγκρίνετέ τα ως προς την ποιότητα/ αποτελεσματικότητα της ταξινόμησης (π.χ. confusion matrix, ROC-AUC curves, etc.).

Για το καλύτερο μοντέλο που εκπαιδεύσατε, μέσα από απεκονίσεις (ή παρόμοιες διαδικασίες) καταγράψτε τις παρατηρήσεις και τα συμπεράσματά σας, τόσο ως προς το μοντέλο όσο και ως προς τα αποτελέσματα που προέκυψαν (data story).

Παραδοτέο εργασίας, προθεσμία και τρόπος παράδοσης

Το τελικό παραδοτέο θα αποτελείται από ένα αρχείο zip, το οποίο θα υποβληθεί ηλεκτρονικά στην ενότητα «Εργασίες» του μαθήματος στο gunet και θα περιέχει τα εξής:

- Τεχνική αναφορά (report) με αναλυτική περιγραφή των προσεγγίσεων που ακολουθήσατε σε καθένα από τα βήματα (π.χ. παράμετροι αλγορίθμων, προπαρασκευή δεδομένων, κτλ.) και ερμηνεία των αποτελεσμάτων που προέκυψαν. Στην τεκμηρίωση πρέπει να αναγράφονται τα στοιχεία των φοιτητών της ομάδας.
- Τα αρχεία πηγαίου κώδικα (source code) και τυχόν συμπληρωματικά αρχεία (που είναι απαραίτητα για την εκτέλεση του κώδικα), καθώς και τα αποτελέσματα που παρήχθησαν (π.χ. plots).

Ζητήματα δεοντολογίας

(α) σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται, (β) σε περίπτωση αμφιβολίας για το κατά πόσο η ομάδα που αναγράφεται ήταν εκείνη που ανέπτυξε την εργασία, ενδέχεται να της ζητηθεί να την παρουσιάσει για τυχόν διευκρινίσεις.