ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΑΘΗΜΑ: ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ-ΑΚΑΔ. ΕΤΟΣ 2023-2024

Στόχος της εργασίας

Στόχος της εργασίας είναι η εφαρμογή τεχνικών εξόρυξης δεδομένων σε δεδομένα με τη χρήση του λογισμικού Weka (https://www.cs.waikato.ac.nz/ml/weka/).

Δραστηριότητες (Tasks)

Δραστηριότητα 1: Συλλογή Δεδομένων
Μπορείτε να επιλέξετε να χρησιμοποιήσετε για την εργασία σας ένα ήδη υπάρχον σετ δεδομένων ή να συλλέξετε δικά σας πρωτογενή δεδομένα. Ενδεικτικές πηγές για σετ δεδομένων θα βρείτε εδώ:
http://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list
http://www.cs.ubc.ca/labs/beta/Projects/autoweka/datasets/
https://www.kaggle.com/
https://datasetsearch.research.google.com/
https://www.stats.govt.nz/large-datasets/csv-files-for-download/
https://www.bl.uk/collection-metadata/downloads

Φροντίστε το σετ δεδομένων σας να έχει τουλάχιστον 10 χαρακτηριστικά και τουλάχιστον τριψήφιο αριθμό παραδειγμάτων.
Επίσης, φροντίστε να καταλαβαίνετε τον μικρόκοσμο του σετ δεδομένων. Τι σημαίνει το κάθε χαρακτηριστικό εισόδου, οι τιμές του, η έξοδος. Διαφορετικά δεν θα μπορέσετε να κάνετε αξιολόγηση των αποτελεσμάτων σας.

Δραστηριότητα 2: Μετασχηματισμός των δεδομένων
Σε περίπτωση που συλλέγετε πρωτογενή δεδομένα, αυτά θα πρέπει

- να καθαριστούν από θόρυβο, διπλότυπα κλπ
- να μετασχηματιστούν σε διανύσματα χαρακτηριστικών-τιμών (feature value vectors, με την επιλογή κατάλληλων χαρακτηριστικών μάθησης).

Δραστηριότητα 3: Προεπεξεργασία των δεδομένων
Μελετήστε την προοπτική εφαρμογής διαφόρων τεχνικών προεπεξεργασίας των παραδειγμάτων μάθησης, εφόσον είναι εφαρμόσιμες στα δεδομένα σας:

- Διακριτοποίηση αριθμητικών μεταβλητών
- Κανονικοποίηση (Normalization of attribute values, …)
- Φιλτραρίσματα (Επιλογές σετ χαρακτηριστικών, επιλογές σετ παραδειγμάτων, feature selection, …)
- Εξισορρόπηση (Imbalanced datasets, SMOTE, resampling, one-sided sampling, Tomek links, …)
- Μετατροπή του αρχείου δεδομένων σε csv format

Δραστηριότητα 4: Μηχανική Μάθηση
Μελέτη τεχνικών μηχανικής μάθησης, και εκτενείς πειραματισμοί με εφαρμογή τους στα δεδομένα. Πειραματισμοί με διάφορα σετ παραμέτρων των αλγορίθμων, με διάφορα validation σχήματα κλπ.
Κανένας αλγόριθμος δεν εφαρμόζεται σαν μαύρο κουτί – πρέπει να είναι κατανοητή η λογική του, οι βασικές του παράμετροι και ο τρόπος λειτουργίας του.

Δραστηριότητα 5: Απεικόνιση, ανάλυση και αξιολόγηση των αποτελεσμάτων
Απεικόνιση των αποτελεσμάτων με πίνακες, γραφήματα, ιστογράμματα κλπ,

Ανάλυση και ποιοτική αξιολόγηση των αποτελεσμάτων
Σύγκριση των διαφόρων μεθόδων προεπεξεργασίας
Σύγκριση των διαφόρων τεχνικών μηχανικής μάθησης

Άρθρο εργασίας (μέχρι την ημερομηνία των εξετάσεων)
Συγγραφή άρθρου τουλάχιστον 6 σελίδων, σε 12 pt γραμματοσειρά, single line-spacing, που θα περιλαμβάνει

- Περίληψη (150 λέξεις)
- Εισαγωγή
    - Εισαγωγή στον ερευνητικό χώρο της εργασίας
    - Βασικές προσεγγίσεις επίλυσής του
    - Συνεισφορά της εργασίας
    - Σχεδιάγραμμα του άρθρου
- Ερευνητικός χώρος (με βιβλιογραφική επισκόπηση)
- Αλγόριθμοι μηχανικής μάθησης
- Μεθοδολογική διαδικασία
    - Περιγραφή δεδομένων
    - Περιγραφή προεπεξεργασίας
    - Πειράματα μάθησης
    - Ανάλυση και αξιολόγηση
- Συμπεράσματα και προτάσεις για μελλοντικές βελτιώσεις
- Βιβλιογραφία

Το άρθρο της εργασίας θα πρέπει να έχει υποβληθεί στην αντίστοιχη εργασία που θα ανοίξει στο opencourses του μαθήματος (https://opencourses.ionio.gr/courses/DDI200/**) μέχρι την ημερομηνία των εξετάσεων του μαθήματος.**

Την ημέρα των εξετάσεων του μαθήματος, κάθε φοιτητής

- **θα πρέπει να έχει ετοιμάσει 10λεπτη παρουσίαση της εργασίας του με διαφάνειες**
- **θα πρέπει να παρουσιαστεί την ώρα των εξετάσεων στην αίθουσα που αναγράφεται στο πρόγραμμα των εξετάσεων, προκειμένου να παρουσιάσει την εργασία του και να εξεταστεί πάνω σε αυτή. Όποιος φοιτητής δεν παρουσιαστεί στις εξετάσεις δεν θα βαθμολογηθεί για το μάθημα, ανεξάρτητα αν έχει υποβάλει ηλεκτρονικά το άρθρο της εργασίας του.**