Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2022-2023

Διδάσκοντες:
Καθηγητής Β. Μεγαλοοικονόμου,
Αναπληρωτής Καθηγητής Χ. Μακρής

Γλώσσα Υλοποίησης

Ως γλώσσα υλοποίησης της άσκησης ορίζεται η Python. Μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκη επιθυμείτε αρκεί να την συμπεριλάβετε στην αναφορά σας.

Σύνολο δεδομένων

Το αρχείο data.csv[1] περιέχει στατιστικά στοιχεία σχετικά με την καθημερινή εξέλιξη της νόσου COVID19 σε 104 χώρες από τον Ιανουάριο του 2020 έως και τον Φεβρουάριο του 2021. Επιπρόσθετα περιέχει μερικές βασικές πληροφορίες σχετικά με την τοποθεσία της κάθε χώρας, τον πληθυσμό της αλλά και το σύστημα υγείας της.

Ερώτημα 1

Πραγματοποιήστε μια πρώτη ανάλυση του συνόλου δεδομένων αλλά και κατάλληλες γραφικές παραστάσεις για αυτό έτσι ώστε να το κατανοήσετε καλύτερα. Πιο συγκεκριμένα, καλείστε να υπολογίστε τα βασικά συγκεντρωτικά στατιστικά μεγέθη για τις δοθέντες τιμές, να ανακαλύψετε αν η μορφή των γραφικών παραστάσεων ακολουθεί συγκεκριμένα μοτίβα αλλά και να προσπαθήσετε να εντοπίσετε συσχετίσεις μεταξύ των διάφορων στηλών του συνόλου δεδομένων αλλά και μεταξύ των στατιστικών στοιχείων που υπολογίσατε.

[1] https://www.kaggle.com/datasets/sambelkacem/covid19-algeria-and-world-dataset

Ερώτημα 2

Επιχειρήστε να χωρίστε τις χώρες σε συστάδες με βάση τις επιδόσεις τους στην αντιμετώπιση του ιού. Προσπαθήστε να χρησιμοποιήσετε για την ανάλυσή σας διάφορα κριτήρια που να αξιολογούν την επιτυχία ή μη της κάθε χώρας (ποσοστό θετικότητας, ποσοστό θνησιμότητας, συνολικός αριθμός κρουσμάτων σε σχέση με τον πληθυσμό της χώρας κ.ο.κ.). Υπάρχουν χώρες που να ξεχωρίζουν είτε αρνητικά είτε θετικά; Μοιράζονται αυτές κοινά χαρακτηριστικά;

Ερώτημα 3

Φανταστείτε πως την 1/1/2021 σας ζητείται η καθημερινή ανάλυση των δεδομένων έτσι ώστε να συνδράμετε στο έργο των υπεύθυνων του σχεδιασμού των πολιτικών για την αντιμετώπιση του ιού. Προσπαθήστε να εκπαιδεύσετε δύο παλινδρομητές: έναν βασισμένο σε RNNs και έναν σε SVMs, οι οποίοι να μαντεύουν το ποσοστό της θετικότητας στην Ελλάδα 3 ημέρες μετά από την ημερομηνία στην οποία γίνεται η ανάλυση. Αξιολογήστε και συγκρίνετε τα μοντέλο σας χρησιμοποιώντας τις γνωστές μετρικές για την παλινδρόμηση και λαμβάνοντας υπόψιν τις ιδιαιτερότητες του προβλήματος που επιλύετε.

Παραδοτέα

1. Τα αρχεία κώδικα που υλοποιούν τα ζητούμενα των ασκήσεων.
2. Μια αναφορά σε μορφή pdf η οποία θα πρέπει να περιέχει τα ακόλουθα:
    - Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
    - Σύντομη περιγραφή της διαδικασία υλοποίησης.
    - Σχολιασμό των τελικών αποτελεσμάτων.

Διαδικαστικά

1. Η άσκηση μπορεί να υλοποιηθεί είτε **ατομικά** είτε σε **ομάδες των δύο**.
2. Η άσκηση μπορεί να υποβληθεί έως και **τρεις ημέρες πριν την ημερομηνία της γραπτής εξέτασης** του μαθήματος στις **23:59**.
3. Η άσκηση θα εξεταστεί προφορικά σε ημερομηνία που θα ανακοινωθεί στο τέλος του εξαμήνου.
4. Η υποβολή της άσκησης πρέπει να γίνει μέσω του eclass του μαθήματος.
5. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.