Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση

Εαρινό Εξάμηνο 2024-2025

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 σύνολο δεδομένων στο οποίο καλείστε να δουλέψετε είναι το *TII-SSRC-23* [1] που θα βρείτε στον σύνδεσμο https://www.kaggle.com/datasets/daniaherzalla/tii-ssrc-23 και περιέχει δεδομένα κυκλοφορίας δικτύου, συγκεντρωμένα για την υποστήριξη της ανάπτυξης και της έρευνας Συστημάτων Ανίχνευσης Εισβολών (IDS). Αποτελείται από δύο μέρη: το πρώτο μέρος περιέχει τα ακατέργαστα δεδομένα κυκλοφορίας δικτύου για κάθε τύπο κυκλοφορίας σε αρχεία PCAP, ενώ το δεύτερο μέρος είναι ένα αρχείο CSV το οποίο παρέχει σε μορφή πίνακα διάφορα χαρακτηριστικά που έχουν εξαχθεί από τα ακατέργαστα δεδομένα. Στην παρούσα εργασία θα εργαστείτε μόνο πάνω στο CSV αρχείο.

Ερώτημα 1

Πραγματοποιήστε μια πρώτη ανάλυση του συνόλου δεδομένων έτσι ώστε να το κατανοήσετε καλύτερα. Πιο συγκεκριμένα, καλείστε να **βρείτε πληροφορίες για τις στήλες** του και τι αυτές σημαίνουν, στην συνέχεια να υπολογίστε **τα βασικά συγκεντρωτικά στατιστικά μεγέθη** για την κάθε στήλη, να φτιάξετε **κατάλληλες γραφικές αναπαραστάσεις** για αυτές και να ανακαλύψετε αν η μορφή των γραφικών παραστάσεων ακολουθεί **συγκεκριμένα μοτίβα**. Στην συνέχεια καλείστε να προσπαθήσετε να εντοπίσετε **συσχετίσεις** μεταξύ των διαφόρων στηλών του συνόλου δεδομένων αλλά και μεταξύ των στατιστικών στοιχείων που υπολογίσατε.

Ερώτημα 2

Το σύνολο δεδομένων που έχετε να διαχειριστείτε είναι πολύ μεγάλο. Σε αυτό το ερώτημα καλείστε να το μειώσετε κρατώντας όσο **το δυνατόν περισσότερη πληροφορία**. Προσπαθήστε με βάση τα ευρήματα σας στο προηγούμενο ερώτημα να αφαιρέσετε ή να συνδυάσετε κάποιες στήλες ώστε να **μειώσετε τη διαστατικότητα του**.

Στην συνέχεια, θα πρέπει **να ελαττώσετε και τις γραμμές του συνόλου δεδομένων** σας με δύο διαφορετικούς τρόπους:

1. θα πρέπει να κατασκευάσετε ένα μικρότερο σύνολο μέσω δειγματοληψίας.
2. θα πρέπει να δοκιμάστε τεχνικές συσταδοποίησης, έτσι ώστε να συμπιέσετε τα δεδομένα σας. Επιλέξτε δύο τεχνικές που θεωρείτε πως είναι κατάλληλες και δικαιολογήστε τις επιλογές σας.

Στο τέλος του ερωτήματος θα πρέπει να δημιουργήσετε 3 νέα**, αντιπροσωπευτικά του αρχικού**, σύνολα δεδομένων πάνω στα οποία θα δουλέψετε για το επόμενο ερώτημα.

Ερώτημα 3

Χρησιμοποιώντας τα σύνολα δεδομένων **που παρήχθησαν στο προηγούμενο ερώτημα**, προσπαθήστε να εκπαιδεύσετε έναν κατηγοριοποιητή βασισμένο σε **Neural Networks** και έναν σε **SVM**, οι οποίοι να μαντεύουν αν **η κυκλοφορία ήταν κανονική ή κακόβουλη** (στήλη Label). Στην συνέχεια εφαρμόστε τις ίδιες μεθόδους για να προβλέψετε **το είδος της κυκλοφορίας** (στήλη Traffic Type). Αξιολογήστε και συγκρίνετε τα μοντέλα σας χρησιμοποιώντας **τις γνωστές μετρικές** για την ταξινόμηση. Ποιο μοντέλο είχε τα καλύτερα αποτελέσματα σε ποιο από τα σύνολα δεδομένων που κατασκευάσατε;

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 προς το τέλος του εξαμήνου. Κατά την εξέταση, θα πρέπει να είστε σε θέση να εξηγήσετε και να αιτιολογήσετε τις επιλογές σας, τόσο σε επίπεδο σχεδιασμού όσο και υλοποίησης της λύσης. Ελλιπής τεκμηρίωση ή αδυναμία αιτιολόγησης των αποφάσεων μπορεί να οδηγήσει σε μείωση της βαθμολογίας στο αντίστοιχο σκέλος.
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.

Βιβλιογραφία

[1] Herzalla, D., Lunardi, W. T., & Andreoni, M. (2023). TII-SSRC-23 Dataset: Typological Exploration of Diverse Traffic Patterns for Intrusion Detection. *IEEE Access*. doi:10.1109/ACCESS.2023.3319213