Εξόρυξη Δεδομένων & Αλγόριθμοι Μάθησης

Αντωνίου Σωτήριος 1067512

Ερώτημα 1:

Κατανόηση των στηλών των δεδομένων.

1. Στήλες ταυτοποίησης ροής & δικτύου

Χρησιμότητα:
Εξυπηρετούν στην αναγνώριση κάθε σύνδεσης, εντοπισμό προέλευσης/προορισμού και φιλτράρισμα δεδομένων. Βοηθούν στον διαχωρισμό ή ομαδοποίηση των ροών.

Τυπικές στήλες:

- `Flow ID`
- `Src IP`
- `Src Port`
- `Dst IP`
- `Dst Port`
- `Protocol`
- `Timestamp`

---

2. Μετρικές διάρκειας και ρυθμού

Χρησιμότητα:
Αναδεικνύουν το χρονικό εύρος και τον όγκο μεταφοράς δεδομένων της κάθε ροής. Ιδιαίτερα χρήσιμες για ανίχνευση ανωμαλιών (ύποπτα μεγάλα διαστήματα, εξαιρετικά υψηλοί ρυθμοί μεταφοράς).

Τυπικές στήλες:

- `Flow Duration`
- `Flow Bytes/s`
- `Flow Packets/s`
- `Flow IAT Mean`, `Flow IAT Std`, `Flow IAT Max`, `Flow IAT Min`
- `Fwd Packets/s`
- `Bwd Packets/s`

---

3. Πληθυσμιακές μετρικές πακέτων

Χρησιμότητα:
Δείχνουν το μέγεθος και την κατανομή πακέτων ανά ροή, διευκολύνοντας τη διάκριση πρότυπων συμπεριφορών (π.χ. μεγάλα downloads, ασύμμετρες ή "σπαστές" ροές).

Τυπικές στήλες:

- `Total Fwd Packets`
- `Total Bwd Packets`
- `Fwd Packet Length Max`, `Min`, `Mean`, `Std`
- `Bwd Packet Length Max`, `Min`, `Mean`, `Std`
- `Total Length of Fwd Packets`
- `Total Length of Bwd Packets`
- `Packet Length Min`, `Max`, `Mean`, `Std`, `Variance`

---

4. Σημαίες TCP/ενεργειών πακέτων

Χρησιμότητα:
Οι flags αποκαλύπτουν τον τρόπο που λειτουργεί το πρωτόκολλο (π.χ. σύναψη or τερματισμός συνδέσεων, επανεκκίνηση). Εντοπίζουν επιθέσεις ή ασυνήθιστη δραστηριότητα.

Τυπικές στήλες:

- `FIN Flag Count`
- `SYN Flag Count`
- `RST Flag Count`
- `PSH Flag Count`
- `ACK Flag Count`
- `URG Flag Count`
- `CWR Flag Count`
- `ECE Flag Count`
- `Fwd PSH Flags`, `Bwd PSH Flags`
- `Fwd URG Flags`, `Bwd URG Flags`

---

5. Συγκεντρωτικά στατιστικά και διαστήματα

Χρησιμότητα:
Τα συγκεκριμένα πεδία δίνουν αίσθηση για το αν η δραστηριότητα είναι συνεχής ή διακεκομμένη, κάτι που διαφοροποιεί τη συνήθη χρήση από ως επί το πλείστον αυτοματοποιημένη δραστηριότητα.

Τυπικές στήλες:

- `Active Mean`, `Std`, `Max`, `Min`
- `Idle Mean`, `Std`, `Max`, `Min`

---

6. Μετρήσεις Bulk, Subflow & Παράθυρα

Χρησιμότητα:
Δείκτες για τη βελτιστοποίηση TCP (bulk μεταφορές, subflows, παράθυρα), χρήσιμοι για διαχωρισμό υπηρεσιών και εύκολη ανίχνευση ασυνήθιστης συμπεριφοράς.

Τυπικές στήλες:

- `Fwd Bulk Avg`, `Bwd Bulk Avg`
- `Fwd Bulk Rate Avg`, `Bwd Bulk Rate Avg`
- `Subflow Fwd Packets`, `Subflow Bwd Packets`
- `Subflow Fwd Bytes`, `Subflow Bwd Bytes`
- `Fwd Init Win Bytes`, `Bwd Init Win Bytes`

---

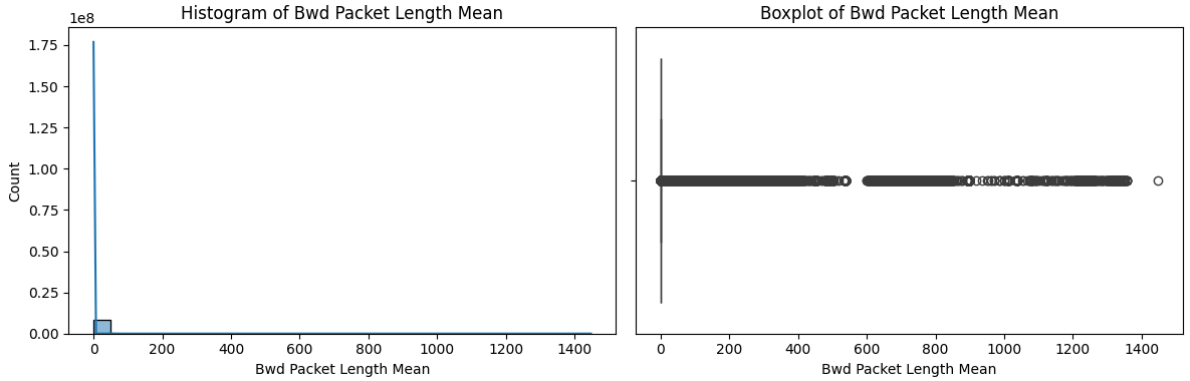
7. Κατηγοριοποιήσεις (Label/Traffic Type)

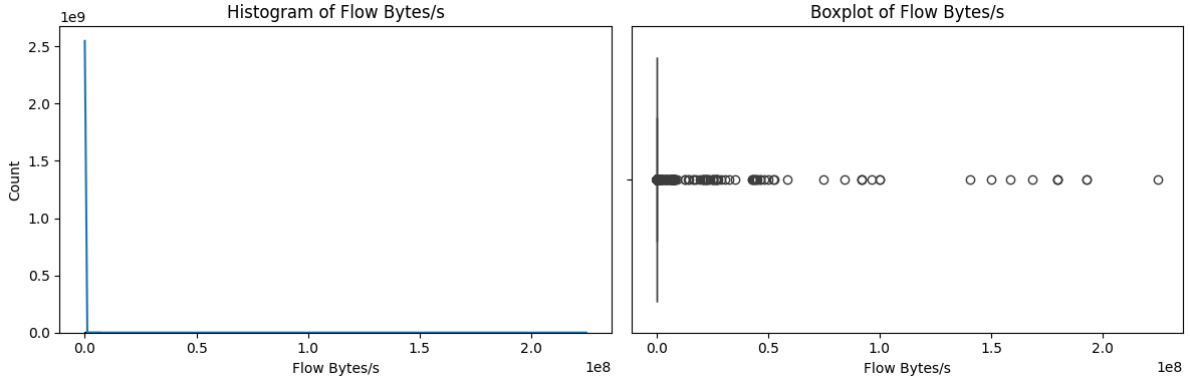
Χρησιμότητα:
Αφορούν άμεσα το σκοπό της εργασίας: κατανόηση, εκπαίδευση και αξιολόγηση ταξινομητών μέσω των ετικετών (benign οι βλαβερές ροές).

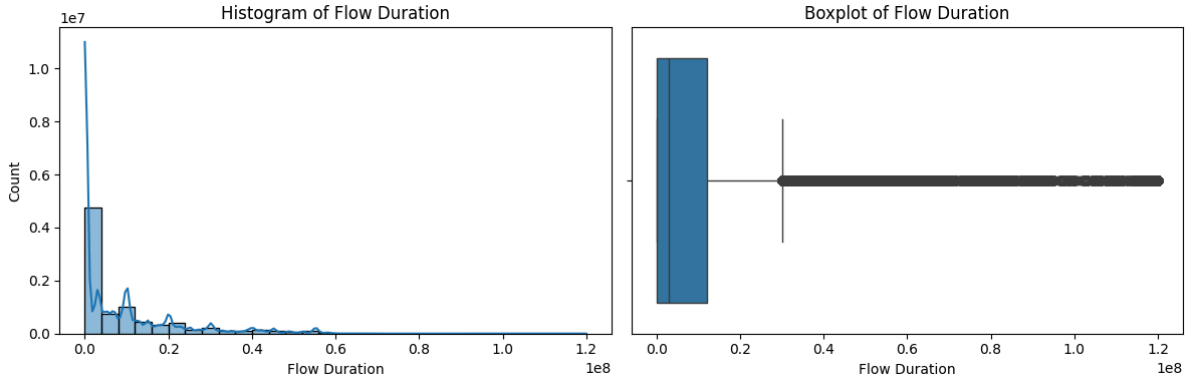
Τυπικές στήλες:

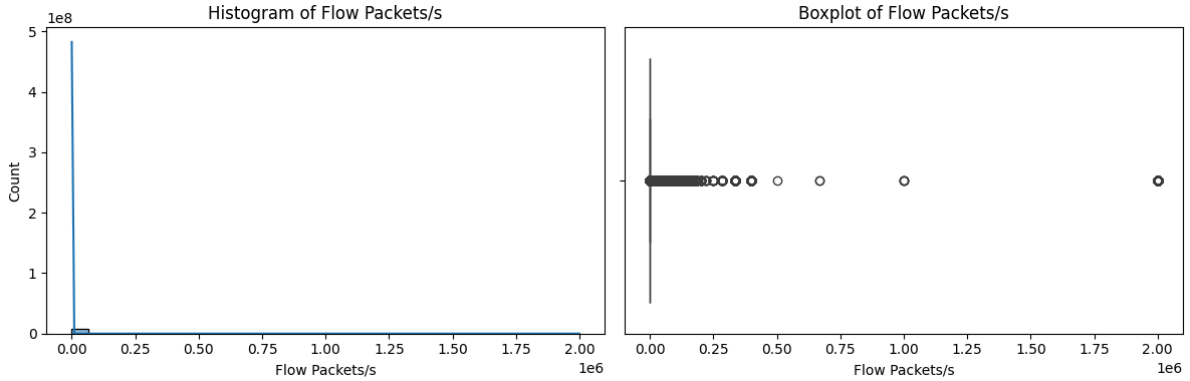
- `Label`
- `Traffic Type`
- `Traffic Subtype`

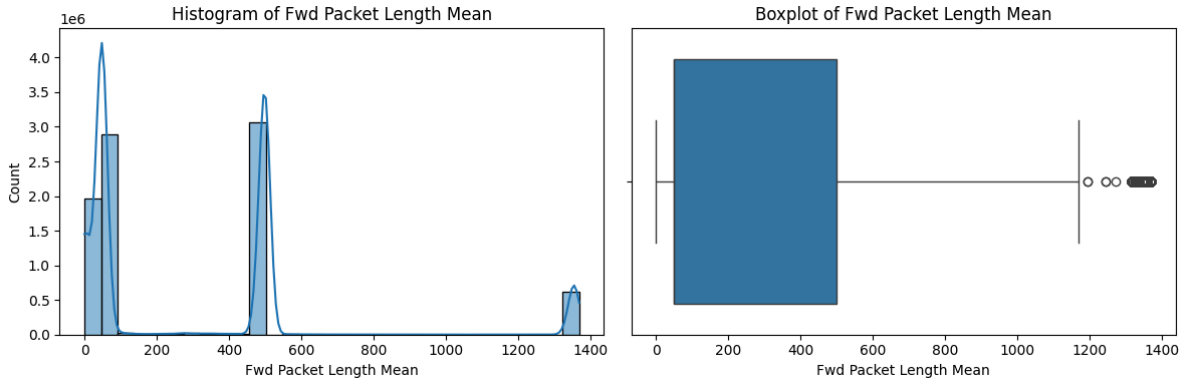
Δημιουργία γραφημάτων (παραθέτω μερικά παραδείγματα απ’τις “σημαντικότερες” στήλες αφού δεν μπορώ να παραθέσω γραφήματα για όλες τις στύλες):

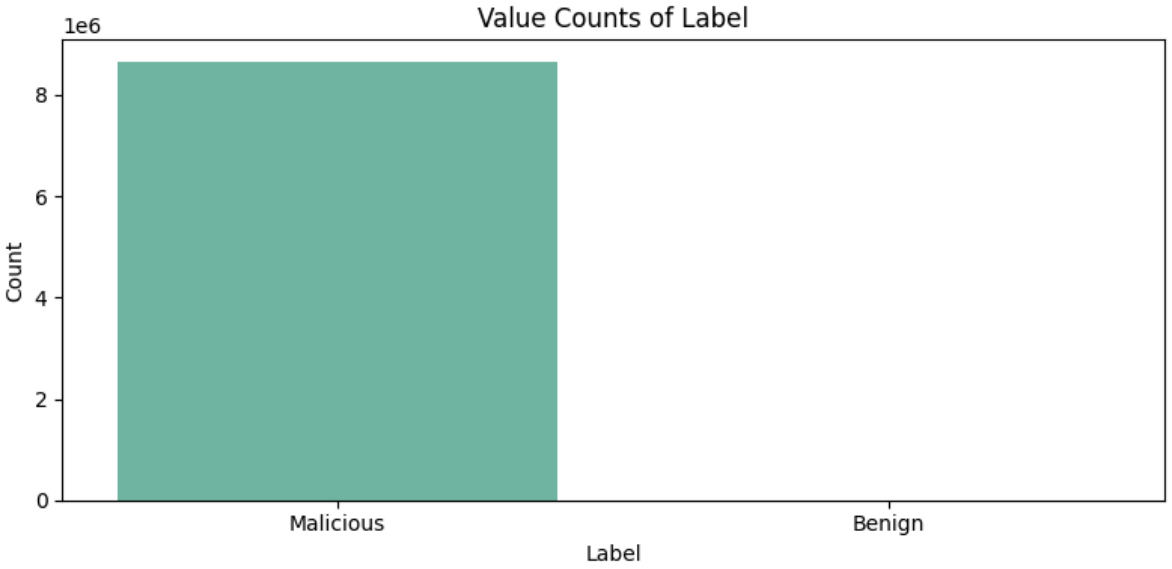


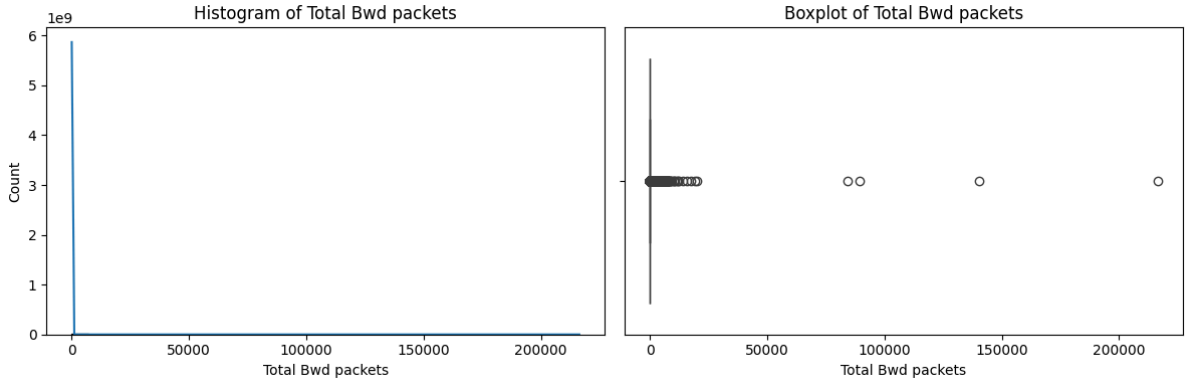


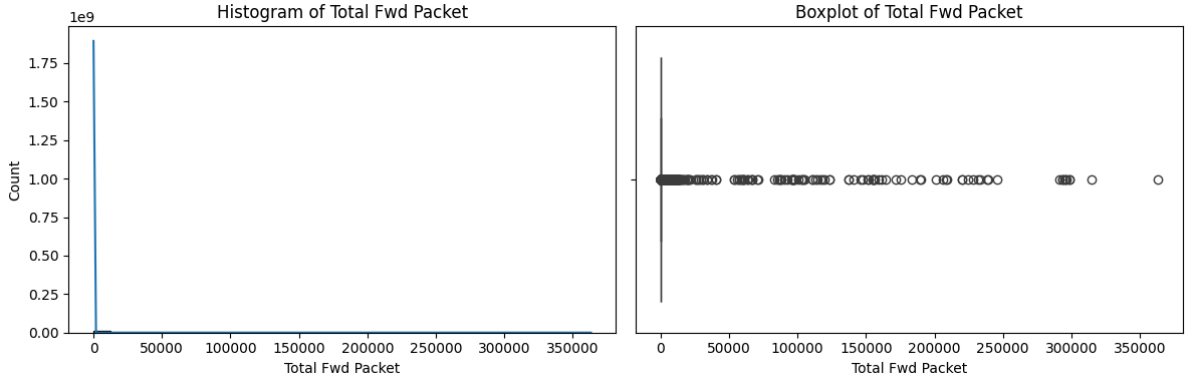


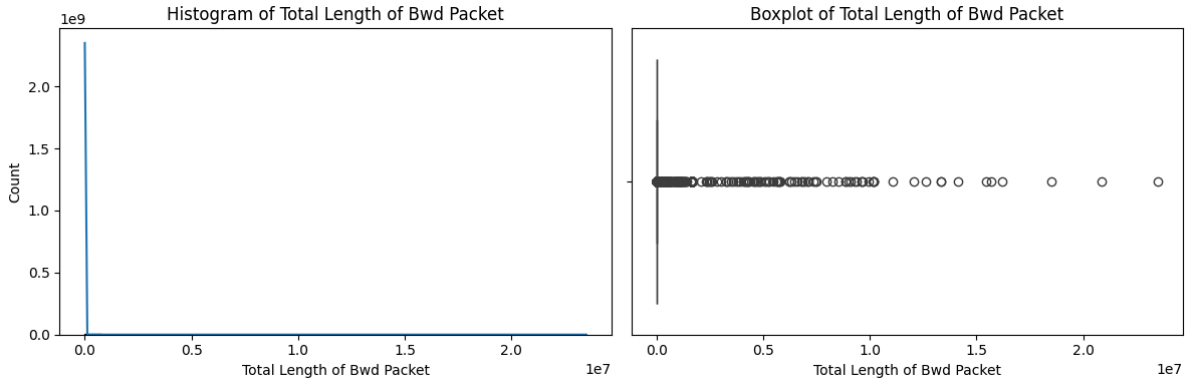


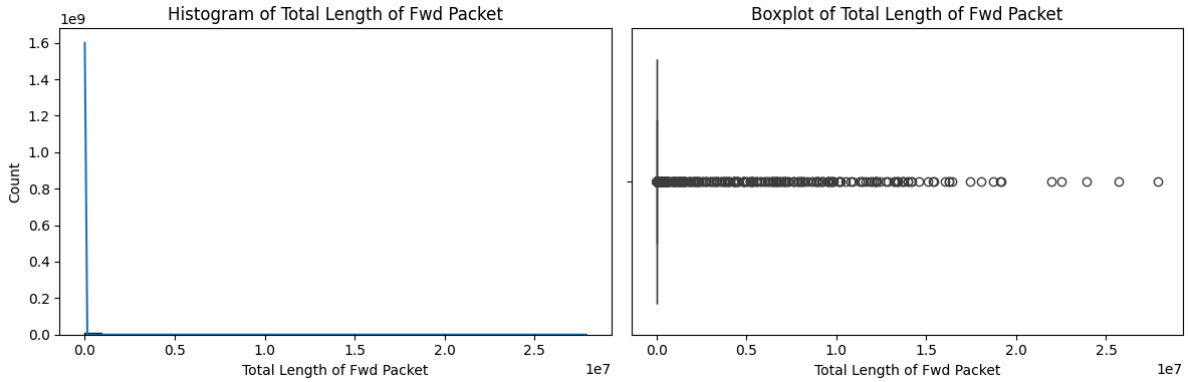

Histogram of Fwd Packet Length Mean
1e6
Count
Fwd Packet Length Mean
Boxplot of Fwd Packet Length Mean
Fwd Packet Length Mean

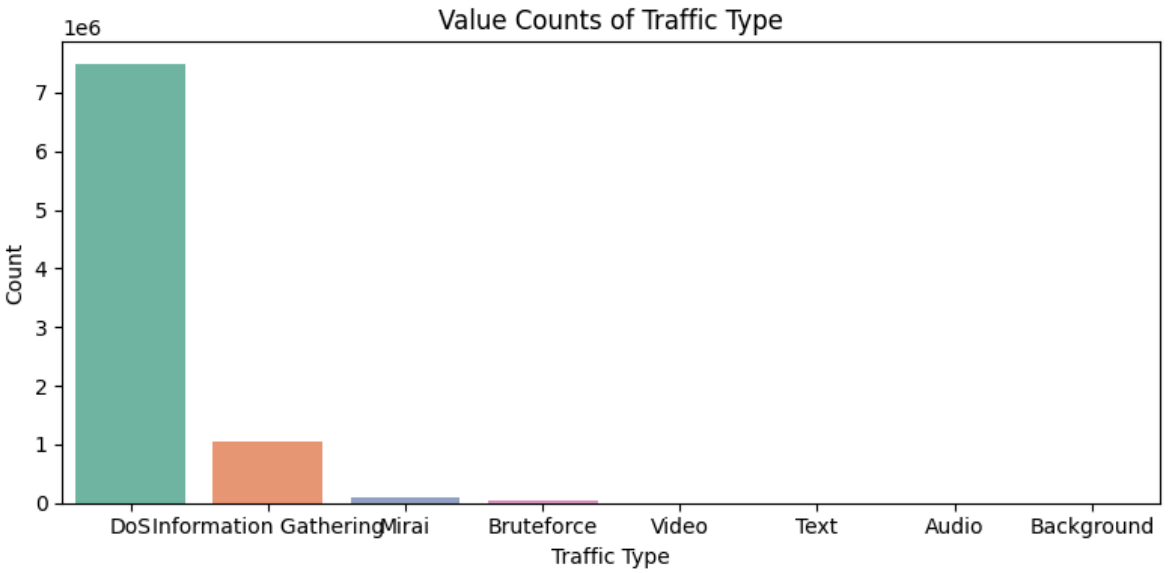

Value Counts of Label
1e6
Count
Malicious
Benign
Label

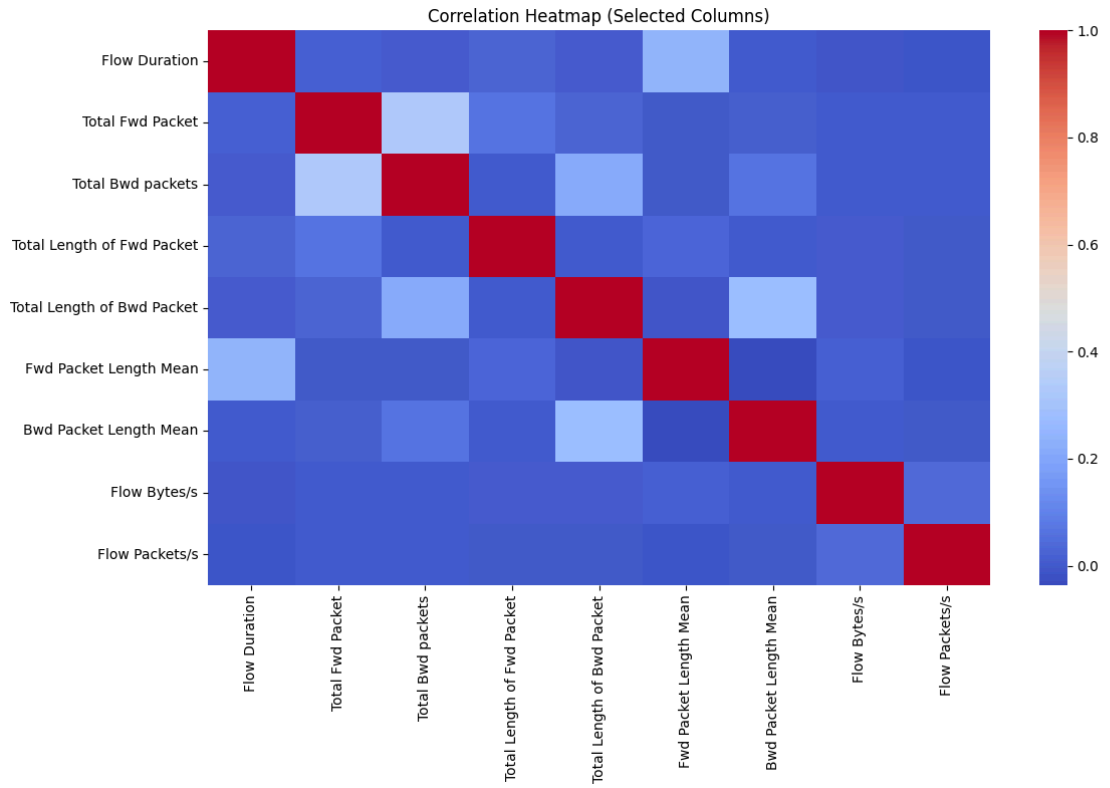

Histogram of Total Bwd packets
1e9
Count
Total Bwd packets
Boxplot of Total Bwd packets
Total Bwd packets


Histogram of Total Fwd Packet
1e9
Count
Total Fwd Packet
Boxplot of Total Fwd Packet
Total Fwd Packet


Histogram of Total Length of Bwd Packet
1e9
Count
Total Length of Bwd Packet
1e7
Boxplot of Total Length of Bwd Packet
Total Length of Bwd Packet
1e7


Histogram of Total Length of Fwd Packet
1e9
Count
Total Length of Fwd Packet
1e7
Boxplot of Total Length of Fwd Packet
Total Length of Fwd Packet
1e7


Value Counts of Traffic Type
1e6
Count
DoS
Information Gathering
Mirai
Bruteforce
Video
Text
Audio
Background
Traffic Type



Ερώτημα 2:

Επιλογή Συγκεκριμένων Στηλών

Για τη μείωση της πολυπλοκότητας του συνόλου δεδομένων και την αποτελεσματικότερη επεξεργασία του στις επόμενες φάσεις, επέλεξα να διατηρήσουμε μόνο ένα υποσύνολο των αρχικών χαρακτηριστικών. Τα κριτήρια επιλογής των στηλών βασίστηκαν:

- **Στη συσχέτιση των χαρακτηριστικών με τις ετικέτες («Label», «Traffic Type»)** από την ανάλυση του πρώτου ερωτήματος.
- **Στην απομάκρυνση πλεοναζόντων ή έντονα συσχετισμένων πεδίων** που δεν προσέφεραν σημαντική επιπλέον πληροφορία και θα μπορούσαν να οδηγήσουν σε υπερεκπαίδευση.
- **Στην εστίαση σε βασικούς δείκτες ροής (flow-related features)** που σύμφωνα με τη βιβλιογραφία συμβάλλουν καθοριστικά στη διάκριση μεταξύ κανονικής και κακόβουλης κυκλοφορίας.

Οι τελικές στήλες που επιλέχθηκαν είναι οι εξής:

- `Flow Duration`: Η συνολική διάρκεια κάθε ροής.
- `Total Fwd Packet` & `Total Bwd packets`: Ο συνολικός αριθμός προς-τα-εμπρός και προς τα πίσω πακέτων (χρήσιμος για την ανίχνευση ασύμμετρων ροών).
- `Fwd Packet Length Max` & `Bwd Packet Length Max`: Μέγιστο μήκος πακέτου ανά κατεύθυνση (αναδεικνύει ενδεχόμενα πρότυπα μεγάλου όγκου σε επιθέσεις).

- `Fwd Packet Length Mean` & `Bwd Packet Length Mean`: Μέσο μήκος πακέτων ανά κατεύθυνση (καταγραφή του γενικού μοτίβου της ροής).
- `Flow IAT Mean`: Μέσος χρόνος μεταξύ διαδοχικών πακέτων (καταγράφει τη «συμπεριφορά» της ροής στο χρόνο).
- `Label`, `Traffic Type`: Οι ετικέτες στόχου για εκπαίδευση και ταξινόμηση.

Η διατήρηση αυτών των χαρακτηριστικών μας επιτρέπει να εργαστούμε πάνω σε ένα περισσότερο διαχειρίσιμο και αντιπροσωπευτικό σύνολο δεδομένων.

Μείωση Δείγματος (Sampling)

Το αρχικό σύνολο δεδομένων είναι εξαιρετικά μεγάλο, με αποτέλεσμα να καθίσταται δύσκολη η επεξεργασία και η ανάλυσή του. Για το λόγο αυτό πραγματοποιήθηκε δειγματοληψία, ώστε να δημιουργηθεί ένα μικρότερο, αντιπροσωπευτικό υποσύνολο. Η δειγματοληψία έγινε με **στρωματοποίηση με βάση το πεδίο Label**, ώστε να διατηρηθούν σε μεγάλο βαθμό οι αναλογίες κανονικής και κακόβουλης κίνησης, αποτρέποντας πιθανή προκατάληψη του τελικού μοντέλου υπέρ της υπερεκπροσωπούμενης κλάσης.

Χρησιμοποιήθηκε ένα δείγμα της τάξης του 1% του αρχικού συνόλου, το οποίο επαρκεί τόσο για γρήγορη πειραματική αξιολόγηση όσο και για διατήρηση της στατιστικής εγκυρότητας.

Αποτελέσματα (data_sampled file and this terminal output):

(venv) ➜ ejorijidedomenwn python3 SampledAnalysis.py

Sampled dataset shape: (8570199, 10)

Class distribution in the sampled dataset:

Label

Malicious 0.99985

Benign 0.00015

Name: proportion, dtype: float64

Μείωση γραμμών μέσω συσταδοποίησης (Clustering)

Για περαιτέρω συμπίεση του συνόλου δεδομένων πέρα από τη δειγματοληψία, εφάρμοσα τεχνικές συσταδοποίησης (clustering), με στόχο να διατηρηθούν τα κύρια πρότυπα και διαφοροποιήσεις των ροών δικτύου, ενώ ταυτόχρονα να μειωθεί σημαντικά το πλήθος των δείγματων προς ανάλυση.

επιλογή τεχνικών clustering και κριτήρια

Επέλεξα να εφαρμόσω δύο διαδεδομένες τεχνικές συσταδοποίησης:

1. **KMeans Clustering**
   Η μέθοδος KMeans είναι από τις πιο γνωστές και αποτελεσματικές τεχνικές για

ανάλυση μεγάλων συνόλων δεδομένων με αριθμητικά χαρακτηριστικά. Η βασική της αρχή στηρίζεται στην ομαδοποίηση των δειγμάτων σε K ομάδες (clusters), με τέτοιο τρόπο ώστε κάθε δείγμα να ανήκει στο cluster του οποίου η μέση τιμή είναι η πλησιέστερη ως προς το δείγμα.
Επιλέχθηκε για την ταχύτητα και την κλιμακωσιμότητά της σε μεγάλα δεδομένα, καθώς και για την ικανότητά της να προσδιορίζει κέντρα (centroids) που συνοψίζουν όμοιες ροές δικτύου.

2. **Agglomerative (Ιεραρχική) Clustering**
Η δεύτερη τεχνική που επιλέχθηκε είναι η αθροιστική ιεραρχική συσταδοποίηση (Agglomerative Clustering). Σε αντίθεση με το KMeans, αυτή η μέθοδος δεν απαιτεί απαραίτητα σφαιρικές ή ισομεγέθεις ομάδες, ενώ μπορεί να αποτυπώσει πιο περίπλοκες εσωτερικές δομές των δεδομένων.
Παρόλο που δεν είναι τόσο αποδοτική υπολογιστικά για πολύ μεγάλα σύνολα, η ιεραρχική συσταδοποίηση μπορεί να αναδείξει εναλλακτικές ομαδοποιήσεις, δημιουργώντας πιθανά συμπληρωματικά αντιπροσωπευτικά δείγματα.

Και στις δύο περιπτώσεις, ως αντιπροσωπευτικά δείγματα κάθε cluster επιλέχθηκαν οι παρατηρήσεις που βρίσκονται πλησιέστερα στο κέντρο του κάθε cluster.

Η επιλογή των δύο διαφορετικών τεχνικών έγινε αφενός για να συγκριθούν τα αποτελέσματα ως προς την αντιπροσωπευτικότητα και την πιθανή ποιότητα των παραγόμενων συνόλων, και αφετέρου για να αξιολογηθεί η συμπεριφορά των αλγορίθμων σε δεδομένα IDS με υψηλή ανισορροπία κλάσεων.

Μέγεθος συνόλων και αποτελέσματα clustering

Για τη διαδικασία συσταδοποίησης, χρησιμοποιήθηκε υποσύνολο των δεδομένων λόγω περιορισμών μνήμης. Επιλέχθηκε να παραχθούν δύο νέα υποσύνολα με χρήση KMeans και Agglomerative Clustering, με αριθμό clusters 5000 και 100 (συμβατό μέγεθος τόσο με τις απαιτήσεις της υπολογιστικής ισχύος, όσο και με τη διατήρηση επαρκούς διακύμανσης στα δεδομένα).

Η ανάλυση κατανομής ετικετών στα παραγόμενα clusters έδειξε παρόμοια συμπίεση πληροφορίας με το sampled set, καθώς διατηρείται η αρχική ανισορροπία μεταξύ κανονικών και κακόβουλων ροών. Αυτό είναι σύνηθες σε σετ δικτυακής κυκλοφορίας, όπου το πλήθος των κακόβουλων ροών υπερτερεί σημαντικά.

Συνολικά παρήχθησαν 2 νέα, σαφώς μικρότερα αλλά αντιπροσωπευτικά σύνολα:

- Ένα μέσω KMeans Clustering (data_clustered.csv),
- Ένα μέσω Agglomerative Clustering (data_clustered_agglomerative.csv).

Κάθε ένα από αυτά θα χρησιμοποιηθεί για εκπαίδευση και σύγκριση ταξινομητών στα επόμενα βήματα της εργασίας.

Σχολιασμός αποτελεσμάτων

Η συσταδοποίηση επέτρεψε σημαντική μείωση του όγκου των δεδομένων, χωρίς να χαθεί η βασική πληροφορία που απαιτείται για την διάκριση μεταξύ των κατηγοριών. Τα σύνολα που

δημιουργήθηκαν παρουσιάζουν παρόμοια στατιστικά χαρακτηριστικά με το αρχικό, ειδικά ως προς τη βασική διακύμανση των κυριοτέρων χαρακτηριστικών και την παρουσία των δύο ετικετών (Label, Traffic Type). Το επίπεδο ανισορροπίας των κλάσεων διατηρήθηκε, γεγονός που θα πρέπει να ληφθεί υπόψιν στην εκπαιδευτική διαδικασία και στην ερμηνεία των αποτελεσμάτων.

Ερώτημα 3:

Για την εκπαίδευση και αξιολόγηση των κατηγοριοποιητών (SVM και Neural Network), χρησιμοποιήθηκαν αποκλειστικά τα σύνολα δεδομένων που παρήχθησαν από τη διαδικασία του kClustering στο προηγούμενο ερώτημα.

Αξιολόγηση και Σύγκριση Μοντέλων

1. Εκτίμηση σε Label (Κανονική ή Κακόβουλη Κυκλοφορία)

SVM

- **Accuracy**: 0.96
- **Benign (Κανονική):**
  - Precision: 0.99
  - Recall: 0.63
  - F1-score: 0.77
- **Malicious (Κακόβουλη):**
  - Precision: 0.95
  - Recall: 1.00
  - F1-score: 0.98

Neural Network

- **Accuracy**: 0.96
- **Benign (Κανονική):**
  - Precision: 0.96
  - Recall: 0.67
  - F1-score: 0.79
- **Malicious (Κακόβουλη):**
  - Precision: 0.96
  - Recall: 1.00
  - F1-score: 0.98

Συμπέρασμα:
Και τα δύο μοντέλα δίνουν εξαιρετικά αποτελέσματα, όμως εμφανίζουν δυσκολία στην ανίχνευση των κανονικών δειγμάτων (χαμηλότερο recall). Τα αποτελέσματά τους είναι σχεδόν ταυτόσημα, με ελαφρώς καλύτερο recall για το neural network στην κανονική κίνηση.

---

2. Εκτίμηση σε Traffic Type (Είδος Κυκλοφορίας)

SVM

- **Accuracy:** 0.92
- **Macro Avg F1-score:** 0.58
- **Σημαντικά Χαμηλοί Δείκτες** για ορισμένες κλάσεις όπως Background και Mirai (f1-score = 0.00).
- **Καλύτερα σε κλάσεις με πολλά δείγματα (DoS, Info Gathering).**
- **Περιορισμοί:** Υπάρχει πρόβλημα με τη διαχείριση μη ισορροπημένων δεδομένων.

Neural Network

- **Accuracy:** 0.93
- **Macro Avg F1-score:** 0.64
- **Βελτιωμένη Ανάκληση/Precision** στις περισσότερες κατηγορίες (ex: Audio, Bruteforce, Text).
- **Μικρό πρόβλημα σε κλάσεις με λίγα δείγματα (πχ. Background, Mirai), αλλά ελαφρώς καλύτερα από SVM.**

Συμπέρασμα:
Το νευρωνικό δίκτυο έχει καλύτερη συνολική απόδοση (ιδίως σε macro και weighted metrics) και αντιμετωπίζει καλύτερα την ταξινόμηση μικρότερων κλάσεων σε σύγκριση με το SVM.

Γενική Σύγκριση και Τελικό Συμπέρασμα

- **Και τα δύο μοντέλα αποδίδουν πολύ καλά στην ανίχνευση κακόβουλης/κανονικής κυκλοφορίας** (label), με υψηλό accuracy και recall για τα κακόβουλα δείγματα.
- **Για την ταξινόμηση είδους κυκλοφορίας (traffic type):**
    - **Το Νευρωνικό Δίκτυο αποδίδει καλύτερα συνολικά**, με υψηλότερα macro/weighted f1-scores.
    - Το SVM τα καταφέρνει καλά στις “ισχυρές” κλάσεις, αλλά τα νευρωνικά δίκτυα είναι πιο ανθεκτικά σε ανισορροπημένα δεδομένα και καλύτερα σε μικρότερες κατηγορίες.
- **Οι κύριες μετρικές που χρησιμοποιήθηκαν**: precision, recall, f1-score, accuracy, macro avg, weighted avg.

Ποιο μοντέλο είχε τα καλύτερα αποτελέσματα;

- **Για την προσέγγιση "Label":** Ισοδύναμα (slight edge to NN for recall).

- **Για "Traffic Type":** Το Νευρωνικό Δίκτυο είχε την καλύτερη συνολικά επίδοση.

Εργαλεία:

1) για την συγγραφή της παρούσας αναφοράς χρησιμοποιήθηκε το Google Docs.
2) για την υλοποίηση του project χρησιμοποιήθηκε το Pycharm.