ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Απαλλακτική Εργασία

1.ΣΚΟΠΟΣ

Η εργασία αυτή θα ασχοληθεί με την ανάπτυξη μοντέλων μηχανικής μάθησης για το classification πακέτων δικτύου με βάση την εφαρμογή από την οποία προέρχονται (π.χ. HTTP, DHCP, DNS, TLS, Google κλπ). Σκοπός είναι η κατηγοριοποίηση της κυκλοφορίας δικτύου για τη βελτίωση της ασφάλειας, την αποτελεσματική κατανομή πόρων ή οποιαδήποτε άλλη ανάγκη απαιτείται.
Ως dataset θα χρησιμοποιηθεί μια καταγραφή δικτύου από ένα πανεπιστήμιο της Κολομβίας[1] με συνολικά περίπου 3.5 εκατ. εγγραφές. Το dataset περιλαμβάνει 87 features με αρκετές λεπτομέρειες όπως μέγιστο/ελάχιστο/μέσο μέγεθος πακέτων, συνολικός αριθμός πακέτων προς κάθε κατεύθυνση κλπ.
Με βάση τα παραπάνω features το μοντέλο θα πρέπει να είναι ικανό να κατηγοριοποιεί παρόμοιες καταγραφές έτσι ώστε να υπάρχει ένα insight σχετικά με την κίνηση ενός δικτύου.

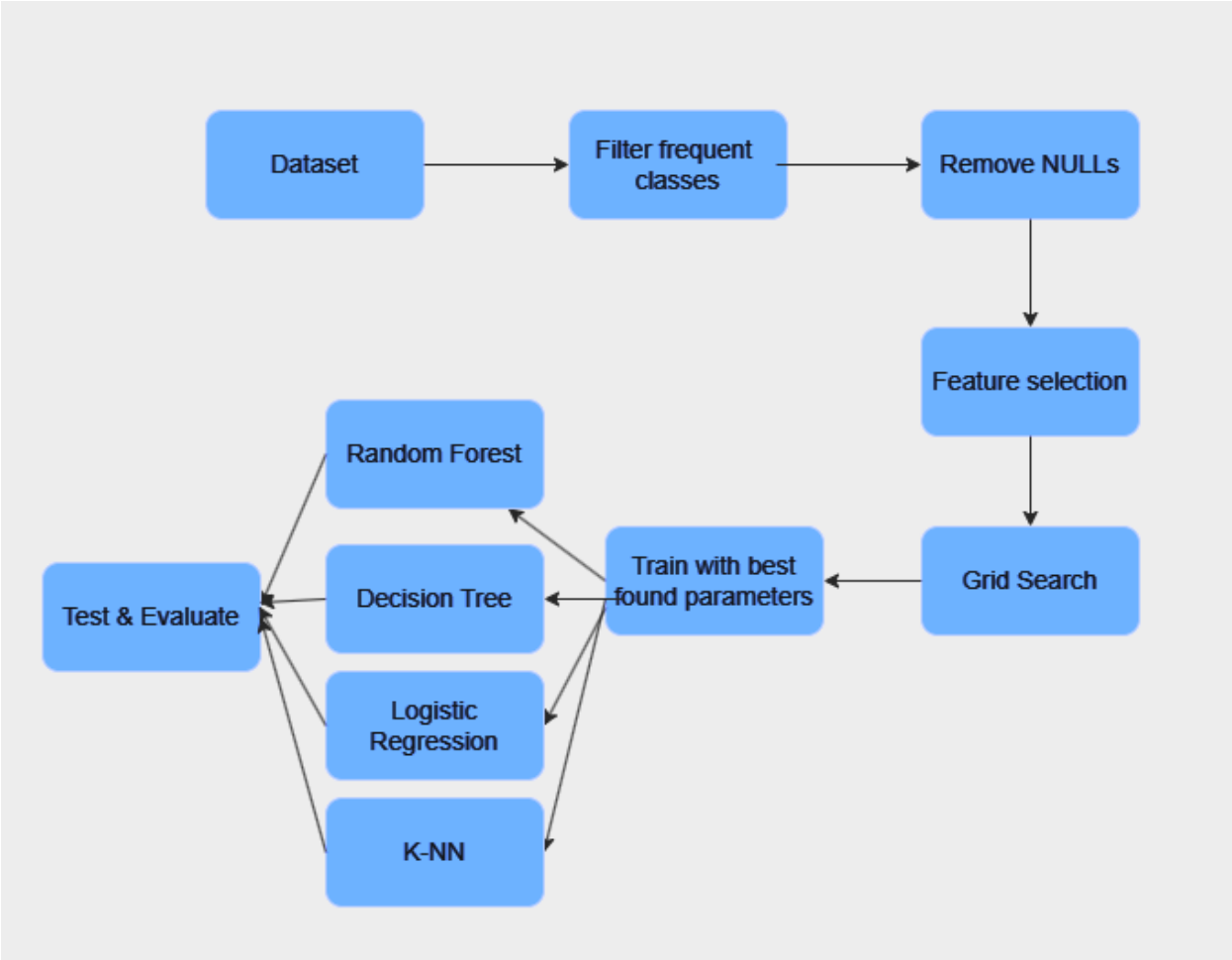
2.ΣΧΕΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Η εργασία είναι εμπνευσμένη από την έρευνα *Network Traffic Classification Using Machine Learning for Software Defined Networks*[2](Kuranage et al., 2019). Μια βασική διαφορά της εργασίας αυτής σε σχέση με την έρευνα είναι πως η εργασία βασίζεται σε supervised learning με έτοιμες κλάσεις. Επιπλέον, αξιοσημείωτο είναι ότι η έρευνα χρησιμοποίησε ακριβώς το ίδιο dataset όμως για διαφορετικούς σκοπούς και με διαφορετικό τρόπο. Συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος k-means για το clustering και την δημιουργία των κλάσεων και έπειτα το classification με βάση αυτές. Αντίθετα, σε αυτή την εργασία θα αξιοποιήσουμε τις έτοιμες κλάσεις του dataset για το classification χωρίς την ανάγκη για clustering.

3.ΜΕΘΟΔΟΛΟΓΙΑ

Για την υλοποίηση των μοντέλων ακολουθήθηκε η διαδικασία της *Εικόνας 1*.

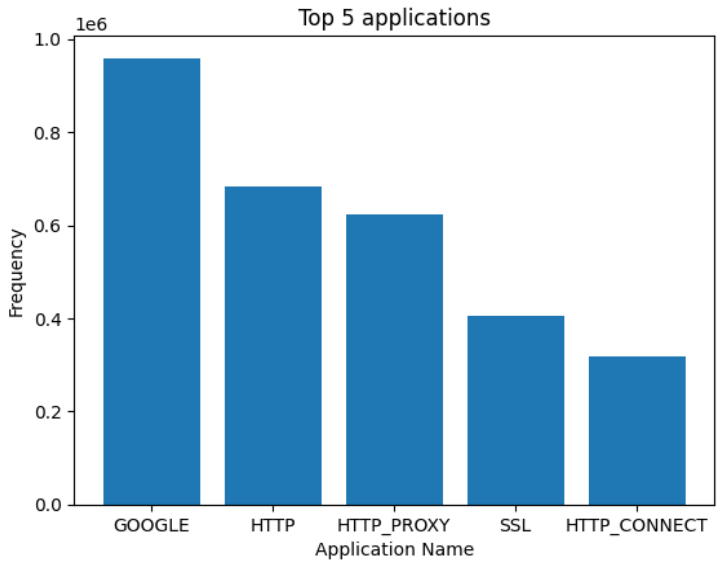
[1] https://www.kaggle.com/datasets/jsrojas/ip-network-traffic-flows-labeled-with-87-apps
[2] Kuranage, M. P. J., Piamrat, K., & Hamma, S. *Network Traffic Classification Using Machine Learning for Software Defined Networks*. 2nd International Conference on Machine Learning for Networking (MLN), Dec 2019, Paris, France.



Εικόνα 1: Flowchart

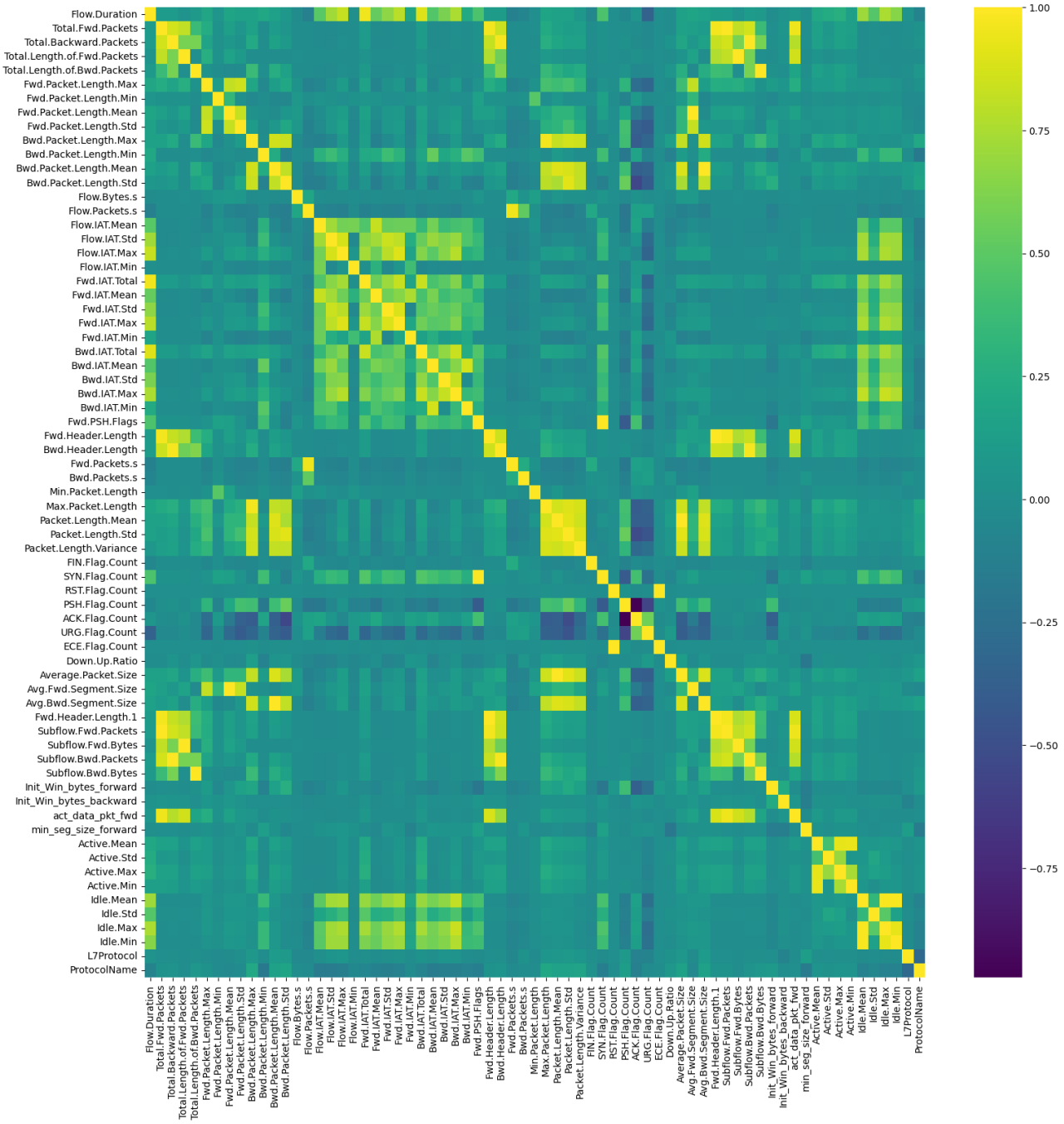
Προεπεξεργασία

Καθώς το μέγεθος του dataset ήταν αρκετά μεγάλο, επιλέχθηκαν μόνο κλάσεις με παραπάνω από 10.000 εγγραφές. Μετά από την αφαίρεση και των τιμών null το τελικό dataset περιείχε 170.000 εγγραφές με 87 features.



Εικόνα 2: Ιστόγραμμα κλάσεων

Για την επιλογή των features χρησιμοποιήθηκε η pearson correlation και παρέμειναν τα features με απόλυτη συσχέτιση μεγαλύτερη από 0,01. Εάν υπήρχαν παραπάνω από ένα features με την ίδια συσχέτιση επιλέχθηκε τυχαία ένα από αυτά. Επιπλέον, αφαιρέθηκαν και οι στήλες που δεν προσφέρουν χρήσιμες πληροφορίες για την συσχέτιση όπως Flow.ID, Source.IP, Timestamp, Destination.IP κλπ. Έτσι, το τελικό σχήμα του dataset ήταν (170.000, 29)



Εικόνα 3: Correlation Matrix

Training and Tuning

Τα μοντέλα μάθησης και οι παράμετροι που επιλέχθηκαν για την εκπαίδευση τους είναι τα εξής:

- **Random Forest**
  - n_estimators: 100, 150, 200

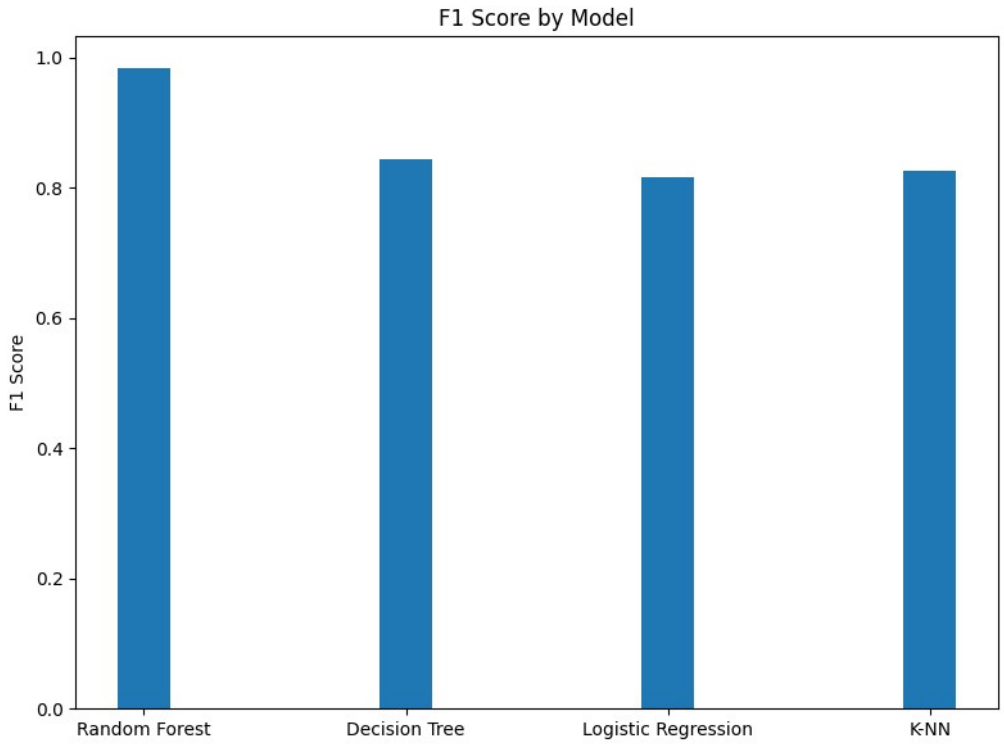
- **Decision Tree**
  - max_depth: 3, 4, 5, 6
  - min_samples_split: 2, 5, 10
  - min_samples_leaf: 1, 2, 4
- **Logistic Regression**
  - solver: lbfgs, sag
  - max_iter: 100, 200, 300
- **K-NN**
  - n_neighbors: 3, 5, 7, 9

Για την επιλογή των καλύτερων παραμέτρων χρησιμοποιήθηκε Grid Search με 5 folds στο training set. Έπειτα έγινε εκπαίδευση του μοντέλου με τις καλύτερες παραμέτρους και τέλος οι εκτιμήσεις των αποτελεσμάτων με βάση το test set.

4.ΑΠΟΤΕΛΕΣΜΑΤΑ

Παρακάτω φαίνονται συγκεντρωτικά οι αποδόσεις των τεσσάρων μοντέλων στο συγκεκριμένο πρόβλημα.

Model	**F1 Score**
Random Forest	98.3%
Decision Tree	84.2%
Logistic Regression	81.6%
K-NN	82.6%

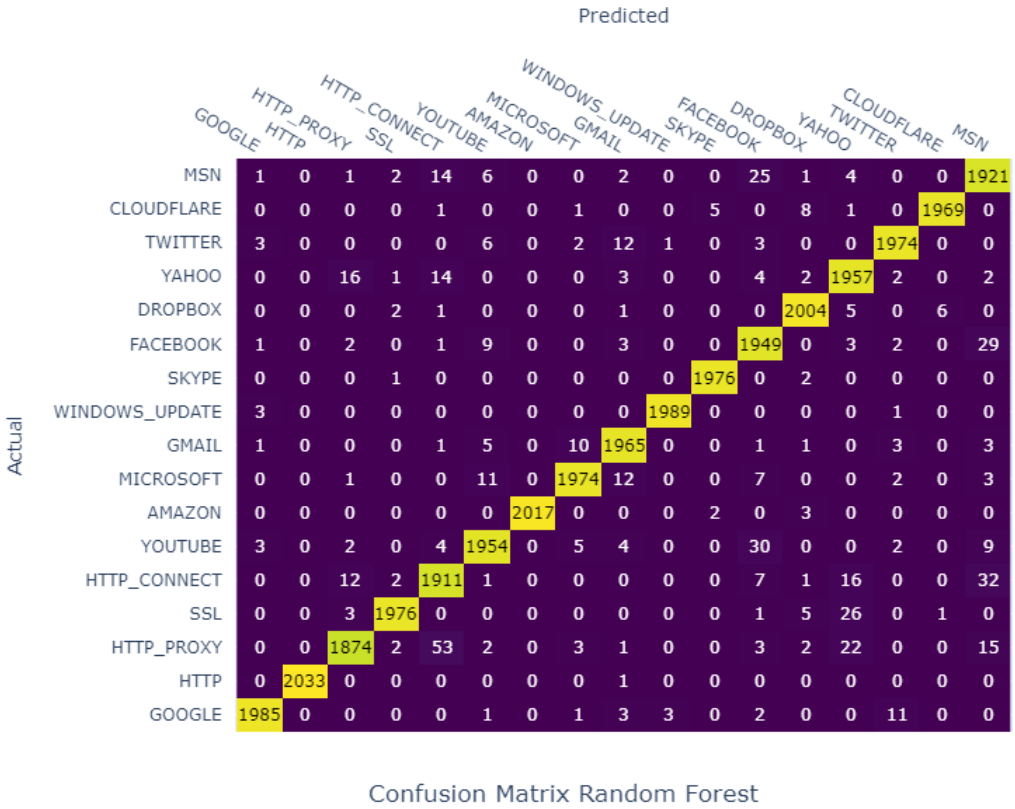


Εικόνα 4: Σύγκριση F1 Score

Random Forest

Best parameters

- n_estimators: 200



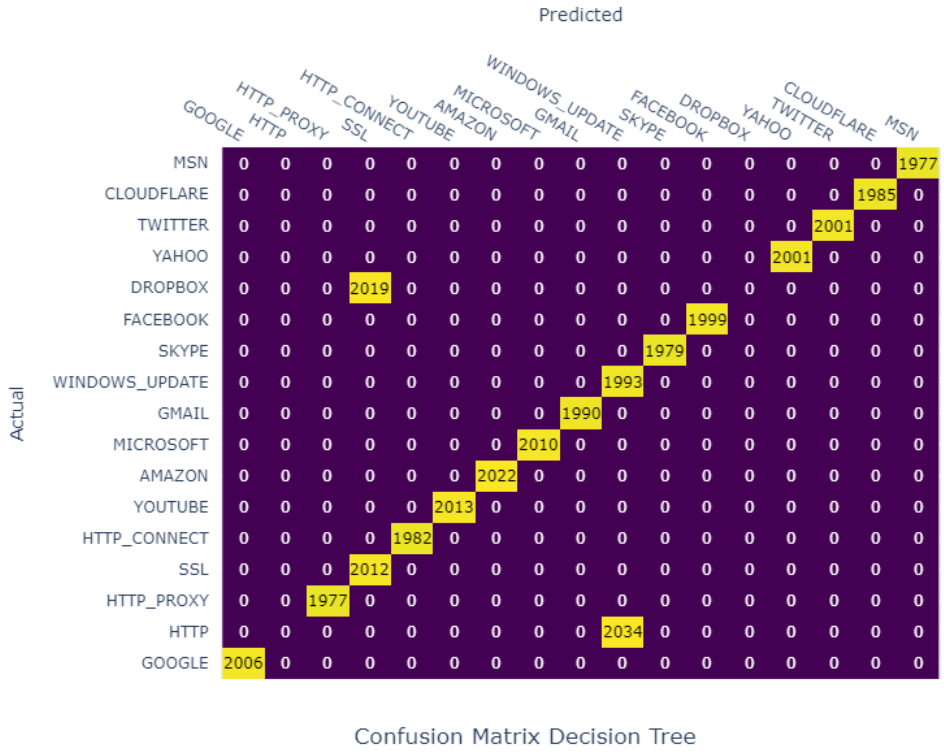
Εικόνα 5: Random Forest

Το μοντέλο Random Forest επιτυγχάνει υψηλή ακρίβεια και υπερτερεί των άλλων με ελάχιστο αριθμό misclassification. Άρα το συγκεκριμένο μοντέλο είναι καλύτερο στο να αναγνωρίζει διαφοροποιήσεις στα δεδομένα, οδηγώντας σε πιο ακριβείς προβλέψεις σε διαφορετικές κατηγορίες.

Decision Tree

Best parameters

- max_depth: 6
- min_samples_split: 2
- min_samples_leaf: 1



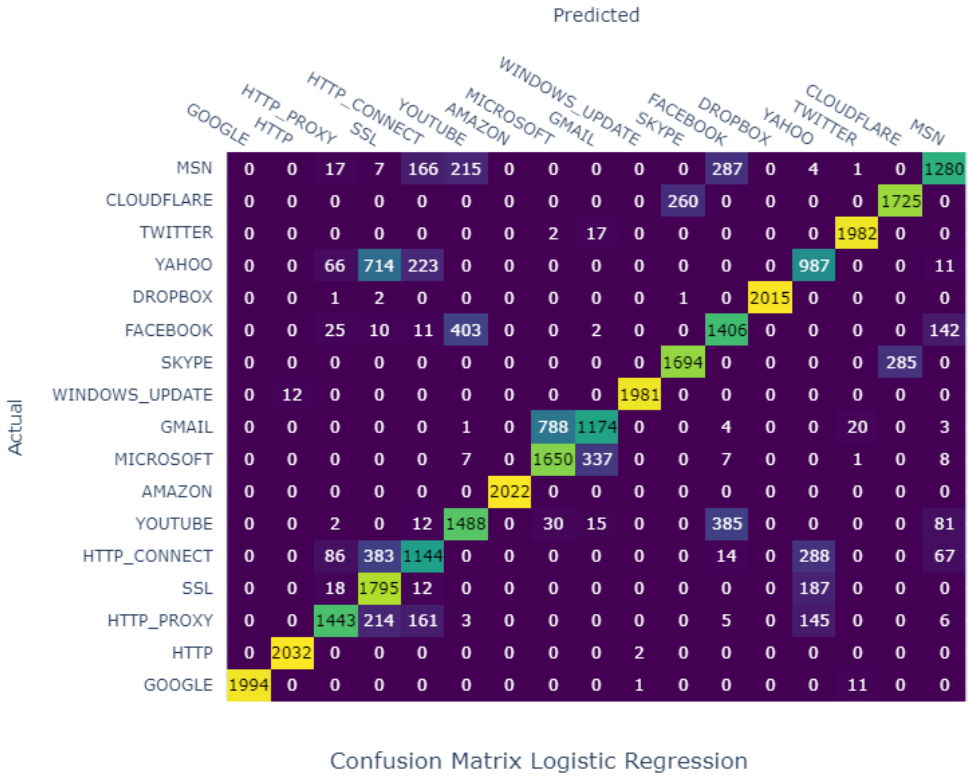
Εικόνα 6: Decision Tree

Το μοντέλο Decision Tree δείχνει υψηλή ακρίβεια με μια πρώτη ματιά. Έχει ελάχιστες αλλά σοβαρές εσφαλμένες ταξινομήσεις. Κάθε κατηγορία προβλέπεται κυρίως σωστά, αλλά παρουσιάζει overfit εφόσον η ακρίβεια φτάνει το 100% στις περισσότερες κλάσεις. Επομένως, παρόλο που είναι το αμέσως καλύτερο μοντέλο μετά το Random Forest, με βάση το F1 score, δεν είναι το καταλληλότερο για αυτή την περίπτωση.

Logistic Regression

Best parameters:

- solver: lbfgs
- max_iter: 200



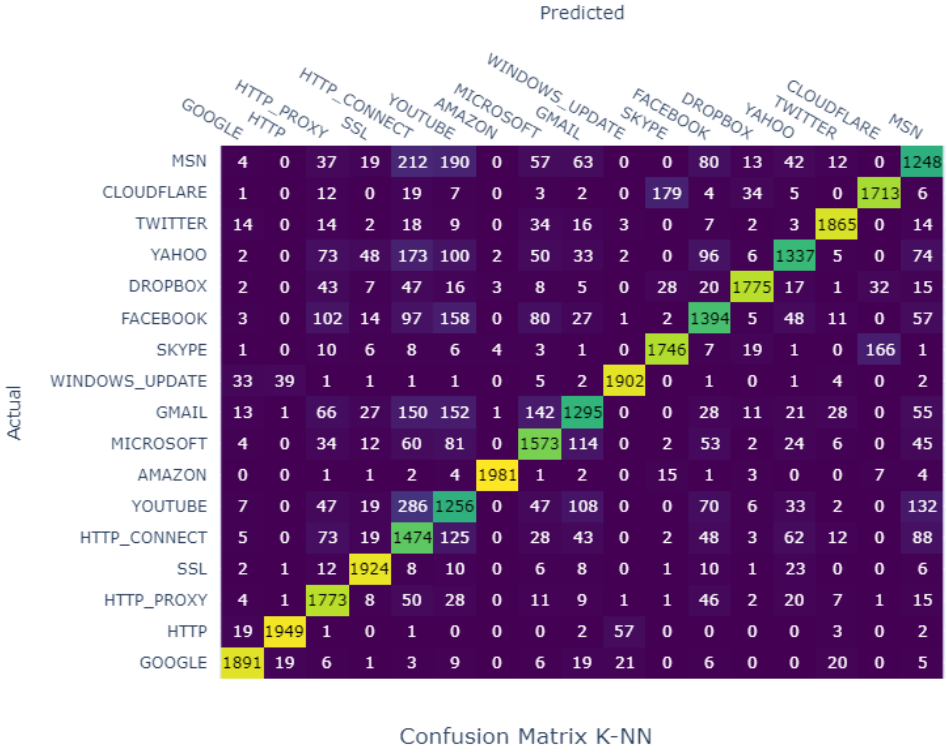
Εικόνα 7: Logistic Regression

Η Logistic Regression εμφανίζει ένα μέτριο επίπεδο ακρίβειας με αξιοσημείωτα misclassifications. Ενώ το μοντέλο ταξινομεί σωστά τις περισσότερες κατηγορίες, δυσκολεύεται με άλλες, εμφανίζοντας μεγάλες τιμές εκτός της διαγώνιου. Αυτό δείχνει ότι ενώ μπορεί να καταγράψει γραμμικές σχέσεις, δεν μπορεί να διαχειριστεί πιο περίπλοκα μοτίβα στα δεδομένα.

K-NN

Best parameters:

- n_neighbors: 3



Εικόνα 8: K-NN

Το μοντέλο K-NN δείχνει ένα πιο κατανεμημένο σύνολο προβλέψεων, με αρκετές τιμές εκτός διαγώνιου. Ενώ η διαγώνιος εξακολουθεί να έχει τις υψηλότερες τιμές, υπάρχουν διαφορές μεταξύ ορισμένων κλάσεων. Αυτό υποδηλώνει ότι ο K-NN δυσκολεύεται με κλάσεις που είναι παρόμοιες ή έχουν παρόμοια χαρακτηριστικά, οδηγώντας σε ένα λιγότερο ακριβές μοντέλο σε σύγκριση με το Random Forest.

Συμπεράσματα

Εν ολίγοις, το μοντέλο Random Forest είναι το καλύτερο για το συγκεκριμένο πρόβλημα και υπερτερεί σημαντικά έναντι των υπολοίπων. Αντίθετα το μοντέλο Decision Tree δεν είναι το καταλληλότερο ενώ τα υπόλοιπα δύο μοντέλα έχουν περιθώρια βελτίωσης με καλύτερες παραμέτρους ή καλύτερης ποιότητας δεδομένα.