Αλγόριθμοι Μηχανικής μάθησης για τον εντοπισμό κακόβουλων ιστοσελίδων

1. Τι είναι το Phishing και πότε μια Ιστοσελίδα θεωρείται Phishing

Το **phishing** είναι μια κακόβουλη τεχνική όπου απατεώνες δημιουργούν ψεύτικες ιστοσελίδες που μοιάζουν με νόμιμες, με σκοπό να εξαπατήσουν τους χρήστες ώστε να παρέχουν προσωπικές πληροφορίες, όπως κωδικούς πρόσβασης, στοιχεία πιστωτικών καρτών ή άλλες ευαίσθητες πληροφορίες. Αυτές οι ιστοσελίδες συχνά μοιάζουν εξαιρετικά με τις αυθεντικές, καθιστώντας δύσκολη την αναγνώρισή τους από τους χρήστες.

Μια ιστοσελίδα θεωρείται **phishing** όταν χρησιμοποιεί παραπλανητικές τακτικές για να μιμηθεί μια έγκυρη ιστοσελίδα, με στόχο να παραπλανήσει τους χρήστες ώστε να δώσουν τα προσωπικά τους δεδομένα. Συνήθεις ενδείξεις phishing περιλαμβάνουν:

- **Παράξενες URL διευθύνσεις**: Οι URL μπορεί να είναι είτε υπερβολικά μεγάλες ή να περιέχουν περίεργους χαρακτήρες και σύμβολα που δεν ταιριάζουν με την κανονική διεύθυνση της αυθεντικής ιστοσελίδας.
- **Έλλειψη ασφαλούς σύνδεσης (HTTPS)**: Αν μια ιστοσελίδα που υποτίθεται ότι είναι ασφαλής δεν χρησιμοποιεί πρωτόκολλο HTTPS, μπορεί να είναι phishing.
- **Χρήση IP διεύθυνσης αντί για domain**: Οι phishing ιστοσελίδες συχνά χρησιμοποιούν απλές IP διευθύνσεις αντί για κανονικά domain.
- **Ανακατεύθυνση σε άλλες σελίδες**: Πολλαπλές ανακατευθύνσεις σε ύποπτες ιστοσελίδες είναι μια συνηθισμένη τακτική.
- **Αίτηση υποβολής προσωπικών πληροφοριών**: Οι phishing ιστοσελίδες συχνά ζητούν ευαίσθητα δεδομένα με ακατάλληλο τρόπο, όπως μέσω popup παραθύρων ή ανύπαρκτων φόρμας επαφής.

Ο τελικός στόχος του phishing είναι η κλοπή των δεδομένων και η χρήση τους για κακόβουλους σκοπούς, όπως η οικονομική απάτη ή η πρόσβαση σε άλλες υπηρεσίες των θυμάτων.

2. Περιγραφή της Άσκησης

Σε αυτήν την άσκηση, θα μελετήσετε τους αλγόριθμους Δένδρων Απόφασης και k-Κοντινότερου Γείτονα k-NN για την ταξινόμηση ιστοσελίδων ως phishing ή νόμιμες, χρησιμοποιώντας ένα πραγματικό dataset (https://archive.ics.uci.edu/dataset/327/phishing+websites) που περιέχει χαρακτηριστικά ιστοσελίδων. Ο στόχος είναι να εφαρμόσετε τους αλγόριθμους αυτούς και να συγκρίνετε την απόδοσή τους στην πρόβλεψη phishing ιστοσελίδων.

1. **Decision Trees:** Ο αλγόριθμος Decision Trees δημιουργεί ένα δέντρο αποφάσεων με βάση τα χαρακτηριστικά των δεδομένων. Κάθε κόμβος του δέντρου αντιπροσωπεύει μια απόφαση πάνω σε κάποιο χαρακτηριστικό, και τα φύλλα του δέντρου αντιπροσωπεύουν την τελική απόφαση (π.χ. αν η ιστοσελίδα είναι phishing ή όχι).
2. **k-NN:** Ο αλγόριθμος k-Nearest Neighbors βασίζεται στην εύρεση των "κ" πιο κοντινών γειτόνων μιας ιστοσελίδας στο χώρο των χαρακτηριστικών της, με βάση την ευκλείδια απόσταση. Η ταξινόμηση γίνεται με βάση την κατηγορία των περισσότερων γειτόνων.

Η εργασία σας θα πρέπει να περιλαμβάνει τα εξής βήματα:

1. Φορτώστε το dataset που παρέχεται.
2. Προετοιμάστε τα δεδομένα αφαιρώντας τυχόν ελλιπή σημεία και κωδικοποιήστε τα κατάλληλα για χρήση από τους αλγόριθμους.
3. Χρησιμοποιήστε έναν αλγόριθμο Decision Tree και τον αλγόριθμο k-NN για να ταξινομήσετε τις ιστοσελίδες.
4. Χρησιμοποιήστε κατάλληλη μέθοδο εκπαίδευσης για την ταξινόμηση των δεδομένων ώς κακόβουλο ή όχι
5. Αξιολογήστε την απόδοση των αλγορίθμων (π.χ. ακρίβεια, ανάκληση, F1 score).
6. Συγκρίνετε τα αποτελέσματα και συζητήστε πότε και γιατί είναι κατάλληλος κάθε αλγόριθμος.

3. Περιγραφή του Συνόλου Δεδομένων

Το σύνολο δεδομένων που θα χρησιμοποιήσετε περιέχει χαρακτηριστικά που σχετίζονται με το αν μια ιστοσελίδα είναι phishing ή νόμιμη. Αυτά τα χαρακτηριστικά αναλύουν διάφορα τεχνικά στοιχεία της ιστοσελίδας, όπως η δομή της URL, η χρήση του πρωτοκόλλου HTTPS, και άλλες τεχνικές λεπτομέρειες που μπορεί να βοηθήσουν στην ανίχνευση phishing ιστοσελίδων.

Κάποια χαρακτηριστικά περιλαμβάνουν:

- **Μήκος της URL**: Αν το μήκος της URL είναι μεγάλο, μπορεί να υποδηλώνει phishing.
- **Χρήση του IP αντί του domain**: Ιστοσελίδες που χρησιμοποιούν IP διεύθυνση αντί για domain είναι συχνά phishing.
- **Χρήση του "@" συμβόλου στην URL**: Αυτό μπορεί να δείχνει ότι η URL είναι phishing.
- **HTTPS και Πιστοποιητικά**: Έλεγχος αν η ιστοσελίδα χρησιμοποιεί έγκυρο HTTPS πιστοποιητικό.
- **Αριθμός subdomains**: Ιστοσελίδες με πολλαπλά subdomains είναι συχνά ύποπτες.
- **Favicon**: Αν το favicon φορτώνεται από διαφορετικό domain από αυτό της ιστοσελίδας, αυτό μπορεί να δείχνει phishing.
- **Ανακατευθύνσεις**: Ιστοσελίδες που κάνουν πολλαπλές ανακατευθύνσεις μπορεί να είναι phishing.

Το dataset αποτελείται από URLs με χαρακτηριστικά όπως τα παραπάνω και κάθε γραμμή αντιπροσωπεύει μια ιστοσελίδα. Κάθε ιστοσελίδα ταξινομείται είτε ως "phishing" είτε ως "legitimate", και αυτή η ετικέτα χρησιμοποιείται ως στόχος για την ταξινόμηση.

4. Τι θα πρέπει να περιλαμβάνει η λύση:

1. **Αναλυτικός κώδικας σε Python**:
   - Κάθε γραμμή κώδικα πρέπει να περιέχει επεξηγηματικά σχόλια για την κατανόηση της λειτουργίας της.
   - Ο κώδικας πρέπει να περιλαμβάνει τη διαδικασία φόρτωσης και προεπεξεργασίας των δεδομένων, την εφαρμογή των αλγορίθμων, και την αξιολόγηση των αποτελεσμάτων.
2. **Αναλυτικό περιγραφικό κείμενο (Αναφορά)**:
   - **Πλήθος δεδομένων και ισορροπία dataset**: Παρουσιάστε πίνακες ή σχήματα που δείχνουν το πλήθος των δεδομένων για κάθε κατηγορία (phishing/νόμιμο) και εξηγήστε αν το dataset είναι **unbalanced** ή όχι, παρέχοντας τεκμηρίωση.
   - **Κανονικοποίηση**: Αναφέρετε αν εφαρμόσατε κανονικοποίηση στα δεδομένα, αιτιολογώντας την απόφασή σας (ναι ή όχι).
   - **Μέθοδος εκπαίδευσης**: Εξηγήστε ποια μέθοδο εκπαίδευσης επιλέξατε και γιατί θεωρείτε ότι είναι η κατάλληλη για το πρόβλημα[1].
   - **Παράμετρος max_leaf_nodes για Decision Trees**: Σημειώστε ποιες τιμές για το **max_leaf_nodes** δοκιμάσατε για τη βελτίωση της ακρίβειας των Decision Trees.
   - **Παράμετρος k για k-NN**: Προσδιορίστε την παράμετρο **k** που βελτιστοποιεί την απόδοση του k-NN. Αναφέρετε πώς επιλέξατε το κατάλληλο k.
   - **Ακρίβεια και απόδοση**: Συγκρίνετε την απόδοση των δύο αλγορίθμων με βάση μετρικές όπως η ακρίβεια, η ανάκληση και το F1 score. Παρουσιάστε τα τελικά ποσοστά επιτυχίας των αλγορίθμων.
   - **Πίνακες Σύγχυσης**: Παρουσιάστε και σχολιάστε τους **πίνακες σύγχυσης** που προκύπτουν από την εφαρμογή των αλγορίθμων. Οι πίνακες αυτοί θα πρέπει να αναλύουν την ταξινόμηση μεταξύ phishing και νόμιμων ιστοσελίδων.

Η τελική αναφορά πρέπει να είναι τεκμηριωμένη, με γραφήματα και στατιστικά δεδομένα που εξηγούν τα αποτελέσματα κάθε αλγορίθμου.

5. Bonus: (To be set)

6. Χρήσιμο Software

- Προτείνεται η εργασία να υλοποιηθεί με Python
- Για την υλοποίηση της εργασίας θα χρειαστείτε το εργαλείο Scikit Learn.
- Για τον αλγόριθμο Δένδρα Αποφάσεων θα χρειαστείτε την κλάση DecisionTreeClassifier. Θέστε ***random_state****=1*

[1] Hint: Χρησιμοποιήστε cross validation με C=10

```
from sklearn.tree import DecisionTreeClassifier
```

- Για τον αλγόριθμο k-NN θα χρειαστείτε την κλάση KNeighborsClassifier.

```
from sklearn.neighbors import KNeighborsClassifier
```

7. Αναφορές

[1] Phishing Websites, https://archive.ics.uci.edu/dataset/327/phishing+websites

[2] Scikit Learn, https://scikit-learn.org/dev/index.html