Analyse rapide du dataset

Dataset: *https://www.kaggle.com/datasets/datasetengineer/teenage-online-behavior-and-cybersecurity-risks/data*

résumé

- collecté depuis des logs provenants d'outils d'e-safety
- 2017-2024
- Texas, Californie
- urban / suburban

features intéressantes

- Device type: le type d'appareil utilisé lors de la session
- Malware détection (Yes/No): si un malware à été détecter sur le device
  - peut être corréler avec le type de device > système d'exploitation
- Phishing attempts: nombres de tentatives de phishing lors de la session
- Social Media Usage (Low, Medium, High): fréquence d'utilisation de média sociaux
  - définir média social (diff de réseau social ?)
- VPN Usage (Yes, No): utilisation d'un vpn lors de la session
  - peut être corréler avec divers facteurs de risque pour déterminer si l'utilisation d'un vpn réduit les risques cyber.
- Cyberbullying Reports: nombres d'incidents lié au cyberharcèlement sur cette session
- Parental Control Alers: Nombres d'alertes déclanchés par des programmes de controle parental installer sur la machine.
- Firewall Logs: nombre de connections bloquée par le firewall
  - le firewall est un programme permettant de bloquer les ports d'un pc ou l'accès internet de certains programmes. Certains firewall se basent sur une liste de logiciels malvenant à bloqué tendis que le plus basique se charge de bloqué certains ports / protocoles. Une bonne pratique sur un serveur / PC est de bloqué tout les ports par défaut et d'ouvrire uniquement ceux dont on à besoin (ex, le port 22 pour ssh.)
- Login attempts: nombres de tentatives de connections lors de la session
  - usage ?
- Download Risk (Low, Medium, High): Niveau de risque associé aux fichiers télécharger
  - Basé sur ?
- Password Strength (Weak, Moderate, Strong): Résistence des mots de passes utilisés
  - la qualification de la résistence d'un MDP se base sur sa capacité à résister à un bruteforce sur la durée. Pour cela, on vas regarder si ce dernier est bien aléatoire, qu'il ne comporte pas de pattern identifiable, pas de mots du ditionnaire, une variété de caractères, ect...
- Data Breach Notifications: Nombre d'altertes lié aux informations personelles de cet identifiant, ex: mot de passe, address, téléphone, ...
- Online Purchase Risk: Similaire au **download risk** mais pour les achats en ligne
- Education Content Usage: Fréquence d'engagement de l'utilisateur avec du conteneu éducationel
- Age Groupe (under 13, 13-16, 17-19): catégotie d'âge auquel appartient le sujet
- Geolocation: localisation de l'utilisateur
- Public Network Usage (Yes, No): Est ce que l'activité enregistrée était elle sur un réseau public ?
  - Permettrais de corréler le niveau de risques des différents types de réseau
- Network Type: Type de connection utilisée. ex: wifi, 5G, ect...
- Hours Online: Total d'heure passée sur la session
- Website Visits: Nombres de sites visités lors de la session
- Peer Interactions: Nombres d'interaction en *Peer-To-Peer* durant la session
  - Le Peer-To-Peer est une méthode de communication ou deux PC communiques directement entres eux sans intermédiaires. ex: le protocol bitorrent
- Risky Website (Yes, No): Est ce que des visites sur des sites à risque ont été enregistrés ?
  - Source ?
- Cloud Service Usage (Yes, No): Est ce que des services dans le *cloud* on été utilisé lors de cette session ?
  - Définir cloud
- Unencrypted Trafic (Yes, No): Est ce que du trafic réseau non chiffré à été utilisé lors de cette session.
  - La plupart du trafic sur internet est chiffré via le protocol *SSL* afin d'éviter qu'un acteur mal intentioné puisse avoir accès aux données de cette connection. Il est possible de piéger un utilisateur à utiliser une connection non chiffrée afin de lui subtiliser des données.
- Ad Clicks (Yes, No): Est ce que des publicités ont été cliqués lors de cette session ?
- Insecure Login Attempts: Nombre de tentatives de connections sur des réseaux non sécurisé.