CYCLE D'INGENIEURE EN GENIE INFORMATIQUE

Année universitaire : 2025-2026

Business Intelligence

Rapport Mini Projet : Détection de sites de phishing avec XGBoost

Réalisé par :

Wijdane Oubhat

Sous l'encadrement de :

Objectif du projet

L'objectif principal est de classer les sites web en deux catégories : légitime (safe) et phishing (malveillant). Pour cela, nous utiliserons les caractéristiques du site (features) comme la structure de l'URL, le contenu de la page, ou la présence du protocole HTTPS afin d'entraîner un modèle capable d'identifier les sites suspects avec une forte précision.

Algorithme utilisé : XGBoost

XGBoost (Extreme Gradient Boosting) est un algorithme de Machine Learning supervisé basé sur les arbres de décision. Son principe repose sur la création d'une série d'arbres, où chaque nouvel arbre vient corriger les erreurs du précédent. Grâce à cette approche, le modèle devient de plus en plus précis à chaque itération.

Avantages de XGBoost :

- Haute performance et rapidité
- Réduction du surapprentissage (overfitting)
- Gestion efficace des grandes bases de données

Résultats attendus

À l'issue du projet, nous obtiendrons un modèle fiable et précis capable d'identifier automatiquement les sites de phishing, de réduire les risques liés aux cyberattaques et de contribuer à la sécurité des utilisateurs sur Internet.

Dataset:

Descriptions de toutes les caractéristiques du dataset

1. Caractéristiques liées à l'URL

- **FILENAME** : Nom du fichier associé à l'URL, souvent utilisé comme identifiant technique.
- **URL** : L'adresse web complète.
- **URLLength** : Longueur totale de l'URL ; une URL très longue est souvent suspecte.
- **Domain** : Nom de domaine extrait de l'URL.
- **DomainLength** : Longueur totale du domaine ; les domaines trop longs sont souvent générés automatiquement.
- **IsDomainIP** : Indique si le domaine est une adresse IP (1) ou un nom de domaine normal (0). Les phishing utilisent souvent des IP pour masquer leur identité.
- **TLD** : Extension du domaine (ex : .com, .net, .xyz).
- **URLSimilarityIndex** : Mesure la similarité entre l'URL et une URL légitime connue, pour détecter les imitations.
- **CharContinuationRate** : Mesure la présence de répétitions de caractères (comme "aaa", "////"), signe d'obfuscation.
- **TLDLegitimateProb** : Probabilité que le TLD corresponde à un domaine fiable.
- **URLCharProb** : Score indiquant si les caractères de l'URL ressemblent à ceux que l'on retrouve dans une URL normale.
- **TLDLength** : Longueur de l'extension du domaine.
- **NoOfSubDomain** : Nombre de sous-domaines présents dans l'URL.
- **HasObfuscation** : Indique si l'URL contient des éléments d'obfuscation (encodage, caractères masqués…).
- **NoOfObfuscatedChar** : Nombre total de caractères obfusqués dans l'URL.
- **ObfuscationRatio** : Proportion des caractères obfusqués par rapport à la longueur totale de l'URL.
- **NoOfLettersInURL** : Nombre de lettres dans l'URL.
- **LetterRatioInURL** : Pourcentage de lettres dans l'URL.
- **NoOfDegitsInURL** : Nombre total de chiffres dans l'URL.
- **DegitRatioInURL** : Pourcentage de chiffres dans l'URL ; les URLs frauduleuses contiennent souvent beaucoup de chiffres.
- **NoOfEqualsInURL** : Nombre de symboles « = ».
- **NoOfQMarkInURL** : Nombre de symboles « ? ».
- **NoOfAmpersandInURL** : Nombre de symboles « & ».

- **NoOfOtherSpecialCharsInURL** : Nombre d'autres caractères spéciaux dans l'URL.
- **SpacialCharRatioInURL** : Proportion totale de caractères spéciaux.
- **IsHTTPS** : Indique si l'URL utilise le protocole HTTPS ou non.

2. Caractéristiques liées au code source HTML de la page

- **LineOfCode** : Nombre total de lignes dans le code source ; les pages phishing sont souvent très simples et limitées.
- **LargestLineLength** : Longueur de la plus grande ligne du code source.
- **HasTitle** : Indique la présence d'une balise <title>.
- **Title** : Contenu textuel du titre de la page.
- **DomainTitleMatchScore** : Score de similarité entre le domaine et le titre.
- **URLTitleMatchScore** : Similarité entre l'URL et le titre de la page.
- **HasFavicon** : Indique si un favicon est présent.
- **Robots** : Présence d'un fichier robots.txt.
- **IsResponsive** : Indique si la page est responsive ; les pages légitimes le sont généralement.
- **NoOfURLRedirect** : Nombre total de redirections effectuées par la page.
- **NoOfSelfRedirect** : Redirections vers la même URL.
- **HasDescription** : Présence d'une meta description.
- **NoOfPopup** : Nombre de pop-ups détectés sur la page.
- **NoOfiFrame** : Nombre de balises <iframe>, souvent utilisées en phishing pour masquer du contenu.
- **HasExternalFormSubmit** : Indique si un formulaire envoie les données vers un domaine externe (typique du vol d'identifiants).
- **HasSocialNet** : Présence de liens vers des réseaux sociaux.
- **HasSubmitButton** : Présence d'un bouton de soumission dans un formulaire.
- **HasHiddenFields** : Présence de champs cachés dans la page.
- **HasPasswordField** : Présence d'un champ « password » dans un formulaire.

3. Mots clés présents dans la page

- **Bank** : Présence de mots liés aux banques.
- **Pay** : Présence de mots liés aux paiements.
- **Crypto** : Termes liés aux cryptomonnaies.

- **HasCopyrightInfo** : Indique si la page contient une mention de copyright.

4. Statistiques d'éléments HTML

- **NoOfImage** : Nombre total d'images dans la page.
- **NoOfCSS** : Nombre de fichiers CSS utilisés.
- **NoOfJS** : Nombre de fichiers JavaScript.
- **NoOfSelfRef** : Nombre de liens internes vers le même domaine.
- **NoOfEmptyRef** : Liens vides (du type « # »).
- **NoOfExternalRef** : Liens vers des domaines externes.

5. Label (variable cible)

- **label** :
  - **0** = site légitime
  - **1** = site phishing