RAPPORT DE PROJET — Analyse et Détection de Profils Vulnérables au Phishing

1. Introduction

Dans le cadre de ce projet, nous avons travaillé sur un jeu de données fictif issu d'une campagne de phishing pédagogique.
L'objectif était double :

1. **Analyser les données** pour identifier quels profils sont les plus susceptibles de cliquer sur une attaque de phishing.
2. **Construire une attaque fictive et pédagogique** ciblant le profil le plus vulnérable identifié.

Pour cela, un travail de nettoyage, exploration, visualisation et interprétation des données a été mené avant de proposer un scénario de phishing réaliste.

2. Importation et exploration des données

Le dataset result.csv est constitué de 519 lignes et 8 colonnes :

- age
- gaming_interest_score
- insta_design_interest_score
- football_interest_score
- recommended_product
- canal_recommande
- campaign_success
- Id

Un premier aperçu (head(), info(), describe()) a permis de détecter plusieurs incohérences :

- **Âges impossibles** (entre 2 et 15 ans).
- **Scores anormaux** (de -100 à 740 alors que l'échelle logique est 0–100).

- **Variantes d’écriture** pour les canaux et produits (Mail, MAIL, insta, fornite, etc.).
- **Valeurs inutilisables** ou bruitées (test, nan).

Ces incohérences imposaient un nettoyage approfondi.

3. Nettoyage des données

3.1 Normalisation

- Passage en minuscules
- Suppression des espaces
- Uniformisation des catégories (mail, instagram, facebook, non_defini)
- Correction des fautes (fornite → fortnite)
- Transformation de campaign_success en booléen (True / False)

3.2 Détection et suppression des valeurs aberrantes

Les règles suivantes ont été appliquées :

- **Age** accepté uniquement entre **16 et 60 ans**
- Scores d’intérêt acceptés uniquement entre **0 et 100**

Résultat :

- **15 lignes anormales supprimées**
- Dataset final : **499 lignes propres**

Ce dataset propre est appelé **df_model**.

4. Analyse descriptive

4.1 Taux de réussite global

69.74%

Ce taux très élevé confirme la présence de patterns exploitables.

4.2 Taux de réussite par canal

Canal	Taux
Facebook	**85.16%**
Mail	66.34%
Instagram	63.06%
Non défini	44.44%

➡ **Facebook** est de loin le canal le plus vulnérable.
C'est un signal extrêmement fort.

4.3 Taux de réussite par produit

Produit	Taux
FIFA	**72.25%**
Fortnite	70.22%
Instagram Pack	66.22%

➡ Le thème **football/FIFA** est le plus efficace.
Dans un cadre réaliste : FIFA ≈ *contenu lié au football*.

4.4 Taux de réussite par tranche d'âge

Tranche	Taux
50–60	**78.57%**
40–49	71.43%
30–39	70.63%
25–29	68.60%
20–24	66.07%
16–19	57.78%

➡ Les **50–60 ans** sont la tranche la plus vulnérable.

4.5 Visualisations

Les graphiques réalisés (barplots + heatmap) montrent clairement :

- La domination de Facebook comme canal vulnérable
- Le rôle du thème "football"

- Une vulnérabilité croissante avec l’âge
- Une corrélation légère mais réelle entre centres d’intérêt et taux de clic

5. Analyse corrélative

La heatmap montre des corrélations faibles mais cohérentes :

- **football_interest_score → +0.10**
- **insta_design_interest_score → +0.13**
- **gaming_interest_score → +0.06**
- **âge → +0.06**

Analyse par niveaux d’intérêt :

- FOOT élevé : **76.44%**
- GAMING élevé : **74.42%**
- INSTAGRAM élevé : **79%**

➡ Les centres d’intérêt (notamment foot) **augmentent significativement la probabilité de clic.**

6. Définition du profil le plus vulnérable

En croisant **toutes** les analyses (canal + produit + âge + intérêts), on obtient :

Profil vulnérable final :

> **Utilisateur de 50–60 ans, actif sur Facebook, ayant un fort intérêt pour le football.**

Explication :

- Facebook est leur canal principal
- Le thème “football” est très engageant
- La tranche d’âge 50–60 est la plus sensible aux campagnes
- Les scores d’intérêt foot augmentent la probabilité de clic

- Dans la réalité, cette tranche est très fan de sport / actualités sportives

Ce profil est cohérent **statistiquement** ET **réaliste socialement**.

7. Scénario d’attaque fictive (datatelling)

Nous imaginons maintenant une attaque **fictive et pédagogique**.

Cible :

Personne de 50–60 ans, fan de football, présente quotidiennement sur Facebook.

- Message d’attaque :

" 🎉 Billets VIP Ligue des Champions à -80% !
Offre exclusive réservée aux abonnés Facebook.
Cliquez ici pour vérifier votre identité et réserver vos places.”

- Pourquoi cette attaque est crédible

- Le football est leur centre d’intérêt majeur
- L’offre “exclusivité + réduction” crée un fort effet FOMO
- Facebook est leur plateforme la plus utilisée
- Le design du message imite les pubs sponsorisées classiques
- La cible appartient à la tranche la plus vulnérable du dataset

8. Conclusion

Grâce à un nettoyage rigoureux, une analyse détaillée et des visualisations claires, nous avons identifié un profil de cible particulièrement vulnérable :

Utilisateur Facebook de 50–60 ans, grand fan de football.

Ces résultats ont permis de construire une attaque fictive, réaliste et basée sur les données, conformément aux objectifs du projet.

Ce travail montre à quel point la data peut révéler des comportements à risque et aider à concevoir des campagnes pédagogiques pour sensibiliser les utilisateurs aux dangers du phishing.