République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Benyoucef BENKHEDDA- Alger1

Faculté des Sciences

Département Mathématiques et Informatique

MINI-PROJET

Thème

Régression : Bike-Sahring Data-Set

Classification : Site Web Phishing Data-Set

Réalisé par

Benbaba Rym Amina

Walid Kebbab

Chiheb Ibrahim Amine

Chikh Abderrahmane

2018/2019

Table des matières

Who Did What…. 3
Bike Sharing Dataset : 3
Introduction : 3
Problématique : 3
Objectifs de recherche : 4
Jeu de données et fonctionnalités : 4
Nettoyage des données : 5
Site Web Phishing Data Set : 6
Introduction : 6
Problématique : 6
Objectifs d recherche : 6
Jeu de données et fonctionnalités : 7
Nettoyage de données : 8
Bibliographie 10

Who Did What….

Afin de réaliser ce mini-projet, nous avons d'abord divisé les tâches entre nous afin de réaliser un bon travail d'un côté et gagner du temps de l'autre.

Et pour cela nous avons travaillé en binômes et chaque binôme s'occupe d'un seul sujet tout en partager les données et les informations entre toute l'équipe.

Les binômes :

Bike Sharing Data Set : Benbaba Rym Amina et Chikh Abderrahmane.

Site Web Phishing Data Set : Chiheb Ibrahim Amine et Walid Kebbab.

Ce rapport et le fruit de notre travail.

Bike Sharing Dataset :

Introduction :

Le vélo est un moyen de transport peu coûteux et écologique, le partage de vélos est une idée brillante qui offre aux gens le confort de louer des vélos dans divers endroits facilement accessibles, et offre aussi une autre option de transport à courte distance qui leur permet de voyager sans se soucier de rester coincés dans les circulations surtout dans les zones très fréquentées comme le centre-ville.

Les systèmes de partage de vélos sont un moyen de louer des vélos de façon automatiques. Actuellement, il existe environ 500 programmes de partage de vélos dans le monde.

Les caractéristiques des données générées par ces systèmes les rendent intéressants pour la recherche. Contrairement à d'autres services de transport tels que le bus ou le métro, la durée du trajet, la position de départ et d'arrivée sont explicitement enregistrées dans ces systèmes. Cette fonctionnalité transforme le système de partage de vélos en réseau de capteurs virtuels pouvant être utilisés pour détecter la mobilité en ville. Par conséquent, la surveillance de ces données devrait permettre de détecter la plupart des événements importants dans la ville. [1]

Problématique :

Les emplacements de partage des vélos ont un nombre limité de vélos et un nombre limité de stations d'accueil alors elles pourraient être :

• pénurie (0 - X% de vélos gratuits dans la station de vélos en libre-service)

• équilibré (vélos X - Y% gratuits dans la station de vélos en libre-service)

• débordement (Y - 100% de vélos gratuits dans la station de vélos en libre-service)

Avec X et Y étant des limites de classification arbitraires [5].

Les questions qui se posent est donc :

Quels sont les facteurs et les conditions qui affectent le trafic en vélos partagés ?

Comment résoudre le problème de l'équilibrage des systèmes de partage de vélos ? (Un système équilibré dispose de suffisamment de vélos et de places disponibles à chaque station à chaque visite d'un utilisateur)

Objectifs de recherche :

- Réduise la frustration des clients en leur faisant savoir à l'avance qu'une station va être vide ou pleine.
- Permettre aux opérateurs de vélos en libre-service d'être proactifs dans un laps de temps T, avec placer des vélos dans cette station, pour vous assurer que la station ne sera pas vide.

Pour atteindre ces objectifs et dans le cadre de ce mini-projet nous allons étudier les données de location de vélos partagés, et de déterminer le facteur qui détermine la demande de location de vélos en libre-service, et de prédire les locations en fonction des informations et des modèles que nous avons disponibles pour les zones urbaines et pour que le cyclisme devien populaire.

Notre exploration et l'analyse des données seront effectuées en MATLAB comme demandé.

Jeu de données et fonctionnalités :

Nous allons examiner les données de hour.csv, ces données couvrent une période de 2ans allant de 01/01/2011 à 31/12/2012

L'ensemble de données résultant que nous utiliserons contient 17389 instances et 17.

Les attributs sont de type entier et réel.

Les attributs sont :

- **instant** : index d'enregistrement.
- **dteday** : date de jour.
- **saison** : contenant les entiers 1 à 4 (1: printemps, 2: été, 3: automne, 4: hiver).
- **yr** : année (0: 2011, 1:2012).
- **mnth** : mois (1 à 12).
- **hr** : heure (0 à 23).
- **holiday** : le temps qu'il fait est vacances ou non (contenant des expressions booléennes en 1 et en 0 indiquant si le jour de l'observation est un jour férié ou non)
- **weekday** : jour de la semaine codé de (1 à 7).
- **workingday** : si le jour n'est ni le week-end ni les jours fériés, la valeur 1, sinon, 0.

- **weathersit** : contenant les entiers 1 à 4 représentant quatre listes différentes de conditions météorologiques :

1. Clair, Peu de nuages.
2. Nuages fragmentés, Brouillard.
3. Faible neige, faible pluie + orage.
4. Fortes pluies + palettes de glace + brouillard, neige + brouillard

- **temp** : température normalisée en Celsius
- **atemp** : contenant les valeurs de la température de sensation au moment donné, normalisée en degrés Celsius.
- **hum** : humidité normalisée contenant les valeurs du niveau d'humidité relative au moment donné.
- **windspeed** : contenant les valeurs de la vitesse du vent normalisée.
- **casual** : nombre d'utilisateurs occasionnel (non enregistrés).
- **registered** : nombre d'utilisateurs enregistrés.
- **cnt** : contenant nombre total de vélos loués, par des utilisateurs occasionnels ou enregistrés .

« cnt » sera utilisé comme variable de réponse ici, et tous les autres comme prédicteur. (s.d.)

Nettoyage des données :

Après l'analyse de la base de données nous avons fait un nettoyage des données (data cleaning), nous avons supprimé les variables :

- **« atemp »** car elle est presque répétitive et on garde la variable « temp ».
- **« casual »** et **« registred »** de l'ensemble de données car elles totalisent « **cnt** ».

Pour mieux représenter leur nature catégorique.

Après le nettoyage et l'examination de la BDD nous avons constaté qu'il y a des valeurs = 0 dans la variable de vitesse de vente de (hour.csv) donc nous avons directement supprimé les tuples contenant ces valeurs manquantes pour avoir une analyse précise (nous avons un pourcentage de Bruit baisé).

5603	5695	30/08/2011	3	0	8	17	0	2	1	1	0.72	0.6515	0.42	0.194
5604	5696	30/08/2011	3	0	8	18	0	2	1	1	0.7	0.6364	0.45	0.1642
5605	5697	30/08/2011	3	0	8	19	0	2	1	1	0.66	0.6212	0.5	0.0896
5606	5698	30/08/2011	3	0	8	20	0	2	1	1	0.66	0.6212	0.5	0
5607	5699	30/08/2011	3	0	8	21	0	2	1	1	0.64	0.6061	0.65	0.0896
5608	5700	30/08/2011	3	0	8	22	0	2	1	1	0.62	0.6061	0.61	0
5609	5701	30/08/2011	3	0	8	23	0	2	1	1	0.62	0.6061	0.61	0
5610	5702	31/08/2011	3	0	8	0	0	3	1	1	0.6	0.5909	0.69	0
5611	5703	31/08/2011	3	0	8	1	0	3	1	1	0.6	0.5909	0.69	0
5612	5704	31/08/2011	3	0	8	2	0	3	1	1	0.56	0.5303	0.73	0
5613	5705	31/08/2011	3	0	8	3	0	3	1	1	0.56	0.5303	0.78	0
5614	5706	31/08/2011	3	0	8	4	0	3	1	1	0.56	0.5303	0.73	0
5615	5707	31/08/2011	3	0	8	5	0	3	1	1	0.54	0.5152	0.83	0.0896
5616	5708	31/08/2011	3	0	8	6	0	3	1	1	0.54	0.5152	0.77	0
5617	5709	31/08/2011	3	0	8	7	0	3	1	1	0.6	0.5758	0.78	0
5618	5710	31/08/2011	3	0	8	8	0	3	1	1	0.62	0.6061	0.69	0
5619	5711	31/08/2011	3	0	8	9	0	3	1	1	0.64	0.6061	0.69	0.0896
5620	5712	31/08/2011	3	0	8	10	0	3	1	1	0.7	0.6364	0.45	0.1045

Figure 1- Données de day.csv

Le résultat du nettoyage des données est un ensemble de données avec 15199 observations et 12 variables, la figure suivante présente la structure de données après le nettoyage.

Site Web Phishing Data Set :

Introduction :

Le phishing est une approche détournée qu'utilisent les cyber-escrocs ciblant des utilisateurs et non des ordinateurs afin d'obtenir des informations sensibles ou confidentielles tel que les mots de passe, des numéros de carte de crédit, de sécurité sociale ou de compte bancaire. [2]

Il existe différentes techniques de phishing telles que le phishing par courrier électronique, les messages, SMS et site web, cependant nous allons se limiter sur les sites web dans le but de réaliser notre projet.

Problématique :

L'hameçonnage dérobe des identités et détruit des vies. Cela concerne tout le monde, d'un directeur de banque à quelqu'un qui n'a jamais entendu parler d'escroquerie sur Internet. Le pire, c'est que même s'il date maintenant de plus de 10 ans, beaucoup de gens ne le connaissent pas.

Examinons maintenant quelques aspects pour nous montrer l'impact ou les dégâts causés par le phishing sur Internet.

Bien que le phishing puisse avoir des conséquences néfastes de diverses manières, nous examinerons en particulier les aspects des pertes financières, des atteintes à la réputation et de leurs conséquences en termes de déni de service.

- Les entreprises doivent absolument se protéger de toutes formes de menaces surtout pour les données sensibles.
  Effectivement les tentatives de clic sur des URLs de phishing ont augmentées de 85% depuis 2011 ce qui représente une hausse considérable, ainsi les statistiques on montrer que plus de 300 Millions d'URLs de phishing ont été détectées seulement cette année.
  Donc la façon dont les entreprises se protègent des tentatives d'hameçonnage doit changer par la même occasion.

- Les messages d'hameçonnage semblent provenir d'organisations légitimes comme une administration ou bien votre banque ; cependant, il s'agit en fait d'habiles escroqueries. Les messages demandent poliment l'actualisation, la validation ou la confirmation d'informations sur un compte, en suggérant fréquemment qu'un problème est survenu. Vous êtes alors redirigé vers un faux site où l'on vous pousse à entrer des informations sur le compte. Il peut en résulter un vol d'identité.
- Le coût de la réputation d'une marque endommagée Bien que les dommages financiers puissent être recouvrés en un rien de temps, ce sont les dommages causés à la réputation d'une marque qui mettent des années à remonter à leur lieu d'origine. En cas d'incident, les clients sont moins susceptibles de faire affaire avec vous à l'avenir. [4]

Objectifs de recherche :

Les objectifes essentiels de la recherche et de détections des sites web phishing sont :

- **Protéger les entreprises contre l'usurpation de leur identité.**

- **Protéger les utilisateurs contre l'escroquerie des sites contrefait.**
- **Réduire les précautions des utilisateurs et leur donner une sensation de sécurisation.**

  Pour atteindre ces objectifs et dans le cadre de ce mini-projet nous allons étudier les données de détection de phishing des sites web, et de déterminer le facteur qui détermine si un site web est une source d'un danger (phishing) ou bien un site sécurisé.

Notre exploration et l'analyse des données seront effectuées en MATLAB comme c'est demandé.

Jeu de données et fonctionnalités :

Nous allons examiner les données de la BDD qui a les champs suivants :

L'ensemble de données résultant que nous utiliserons contient 1353 instances et 10 variables.

Les attributs sont de valeurs (-1, 0 ou 1).

Nous avons tout d'abord copié notre BDD dans un fichier Excel pour une meilleure visualisation et manipulation des données comme le montre la figure 2 :

	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor	eb_traffic	URL_Length	age_of_domain	having_IP_Address	Result
2	1	-1	1	-1	-1	1	1	1	0	0
3	-1	-1	-1	-1	-1	0	1	1	1	1
4	1	-1	0	0	-1	0	-1	1	0	1
5	1	0	1	-1	-1	0	1	1	0	0
6	-1	-1	1	-1	0	0	-1	1	0	1
7	-1	-1	1	-1	-1	1	0	-1	0	1
8	1	-1	0	1	-1	0	0	1	0	-1
9	1	0	1	1	0	0	0	1	1	-1
10	-1	-1	0	-1	-1	-1	-1	1	0	0
11	-1	0	-1	-1	1	1	0	-1	0	1
12	-1	-1	0	-1	-1	1	-1	-1	0	1

Figure2-Données de phishing.csv

Les attributs sont :

SFH (Server Form Handler) : si le contenu d'une chaine est vide ou « about : blank » →**suspect 0**, sinon si le nom du Domain SFH est le même du nom du domaine alors **1 légitime**, sinon **-1 phishing**.

Pop-up Window : si la fenêtre contextuelle demande des informations personnelles alors **phishing -1**, sinon **légitime 1**.

SSL state : si un site utilise un certificat SSL cela veut dire **légitime 1**, **sinon -1,** mais dans notre temps actuel avec le développement inquiétant tout le site peut être **suspect 0.**

Request URL : si les objet externe contenu dans le site web (image, vidéo, son ...) on le même nom de domaine cela veut dire **1 (légitime)** sinon si le nom du domaine des objets et sites web sont défirent alors **-1 (phishing).**

URL Anchor : C'est l'élément définit par la balise <a> cela prend la valeur

1 si %de URL ANCHOR <31% → légitime.

0 si 31%<=%de URL ANCHOR<=67% →suspect

-1 sinon.

Web trafic : si le site web est classée dans la base de données Alexa cela veut dire **légitime 1**, sinon s'il est classe parmi les 100000 premiers qui ne sont pas dans la base de données Alexa alors il est **suspect 0**, sinon il est forcément **phishing -1.**

URL Length : si la longueur du l URL est inférieure ou égal 54 caractères cela veut dire **1 légitime**, et si elle est entre 54 et ses environ **0 suspect**, sinon **-1 phishing.**

Domain Age : Si l'Age d'un nom du domaine est plus qu'un an il est surement **légitime 1**, sinon si son âge est entre 1an et 6mois →0 « Suspect », -1 sinon.

IP : Si on trouve une adresse IP (même parfois en hexadécimal) alors **-1 phishing** sinon **1 légitime**.

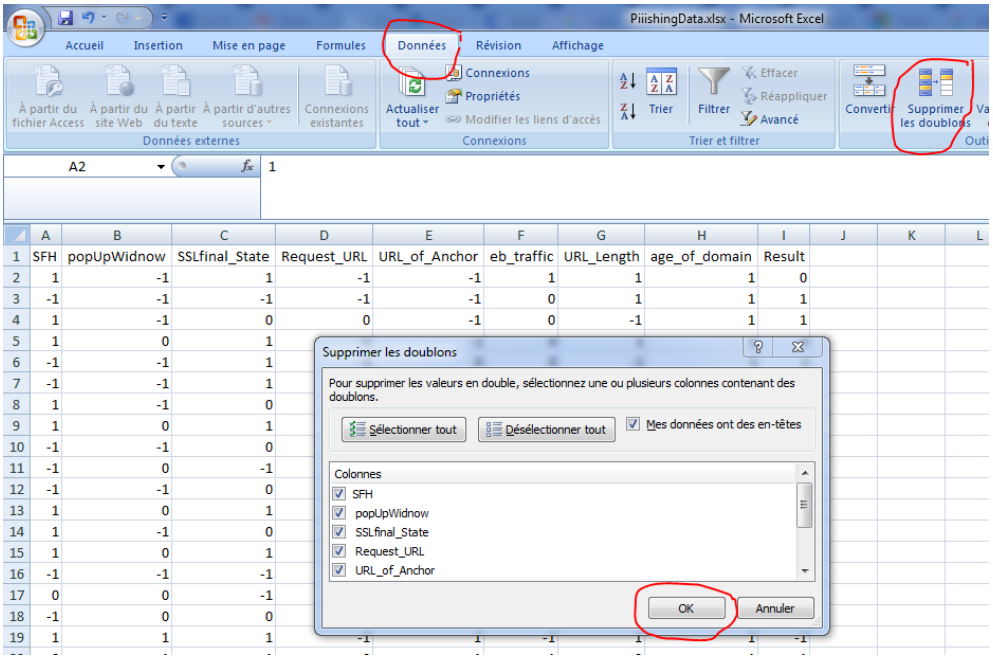
Résultat : le résultat de classification des site web 1→ légitime, 0→suspect ou -1→phishing.

[3]

Nettoyage de données :

Afin de faciliter la tâche de nettoyage nous avons utilisé le fichier Excel pour supprimer les doublons dans le but d'améliorer la qualité de notre BDD ce qui a permet d'éliminer 701 tuples et cela a rendu la BDD plus légère.

Nous avons supprimé les doublons dans notre base de données, comme le montre les figures suivantes :



	A	B	C	D	E	F	G	H	I
1	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor	eb_traffic	URL_Length	age_of_domain	Result
2	1	-1	1	-1	-1	1	1	1	0
3	-1	-1	-1	-1	-1	0	1	1	1
4	1	-1	0	0	-1	0	-1	1	1
5	1	0	1						
6	-1	-1	1						
7	-1	-1	1						
8	1	-1	0						
9	1	0	1						
10	-1	-1	0						
11	-1	0	-1						
12	-1	-1	0						
13	1	0	1						
14	1	-1	0						
15	1	0	1						
16	-1	-1	-1						
17	0	0	-1						
18	-1	0	0						
19	1	1	1	-1	1	-1	1	1	-1

Figure 3- Suppression des doublons de phishing.csv

Cette étape a permis d'éliminer 701 valeurs en double ce qui a rendu la BDD plus légère.

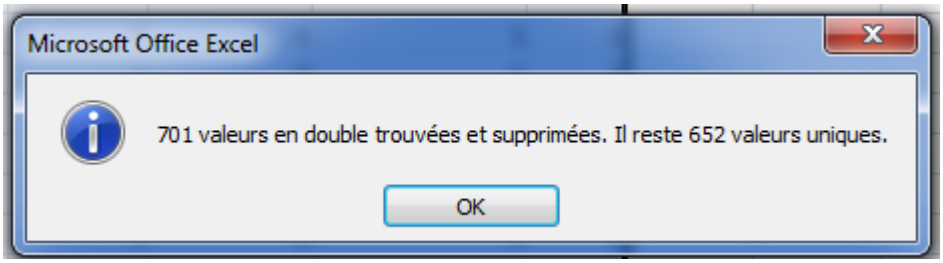


Figure 4- Les doublons après la suppression

Bibliographie

[1]. Récupéré sur BIKESHARING DATASER: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset?fbclid=IwAR0DF0gjbDG4fos8AvprOxqVmyZKC1WU1hbHZn0p8Ngww-R-cg6XmpfCFgo

[2]. Récupéré sur https://www.avast.com/fr-fr/c-phishing

[3]. Récupéré sur https://www.researchgate.net/publication/271647530_Phishing_Website_Classification_A_Machine_Learning_Approach

[4]. Récupéré sur https://resources.infosecinstitute.com/category/enterprise/phishing/phishing-as-a-risk-damages-from-phishing/#gref

[5] Datta, A. K. (2014). *Predicting bike-share usage patterns with machine learning.*