**AVANCEMENT PROJET**

**Détection d’intrusions sur un réseau**

**21/04 – 28/04 : Début**

* Etude du projet, analyse des différentes features dans le dataset grâce aux fichiers donnés.
* La dataset : CIC-IDS-2018 de AWS
* Téléchargement de la dataset puis conversion en fichier csv afin de les exploiter.
* Résultat : 10 fichiers
* Découverte de GitHub

**04/05 : Point avancement 1**

* Chacun a travaillé sur 3 fichiers (nous avons laissé le 10ème car il faisait plus de 3Go)
  + Charles
    - Friday-02-03-2018\_TrafficForML\_CICFlowMeter
    - Friday-16-02-2018\_TrafficForML\_CICFlowMeter
    - Friday-23-02-2018\_TrafficForML\_CICFlowMeter
  + Théodore
    - Thursday-15-02-2018\_TrafficForML\_CICFlowMeter
    - Thursday-22-02-2018\_TrafficForML\_CICFlowMeter
    - Thursday-01-03-2018\_TrafficForML\_CICFlowMeter
  + Cécile
    - Wednesday-14-02-2018\_TrafficForML\_CICFlowMeter
    - Wednesday-21-02-2018\_TrafficForML\_CICFlowMeter
    - Wednesday-28-02-2018\_TrafficForML\_CICFlowMeter
* Analyse du fichier 🡪 génération d’un Report Profiling pour chaque dataset grâce à la librairie Pandas Profiling
  + On a pu voir les différents éléments présents dans chaque features et leur nombre d’apparition
  + On a remarqué que pour trois fichiers, le type de toutes les colonnes était ‘objet’, ainsi nous n’avons pu en tirer aucun résultat. Voici les trois fichiers à mettre de côté pour le moment :
    - Wednesday-28-02-2018\_TrafficForML\_CICFlowMeter
    - Thursday-01-03-2018\_TrafficForML\_CICFlowMeter
    - Friday-16-02-2018\_TrafficForML\_CICFlowMeter
  + On a donc pu en tirer comme conclusion quelle features retirer car l’ensemble de la feature a des valeurs nulles
* Voici les 10 features à supprimer pour les 6 datasets restantes :
  + Bwd PSH Flags
  + Fwd URG Flags
  + Bwd URG Flags
  + CWE Flag Count
  + Fwd Byts/b Avg
  + Fwd Pkts/b Avg
  + Fwd Blk Rate Avg
  + Bwd Byts/b Avg
  + Bwd Pkts/b Avg
  + Bwd Blk Rate Avg
* Intrusions présentes pour chaque fichier avec leur nombre d’apparition, et aussi les intrusions bénignes :
  + Wednesday-21-02-2018\_TrafficForML\_CICFlowMeter
    - DDOS attack-HOIC 686012
    - Benign 360833
    - DDOS attack-LOIC-UDP 1730
  + Wednesday-14-02-2018\_TrafficForML\_CICFlowMeter
    - Benign 667626
    - FTP-BruteForce 193360
    - SSH-Bruteforce 187589
  + Thursday-22-02-2018\_TrafficForML\_CICFlowMeter
    - Benign 1048213
    - Brute Force -Web 249
    - Brute Force -XSS 79
    - SQL Injection 34
  + Thursday-15-02-2018\_TrafficForML\_CICFlowMeter
    - Benign 996077
    - DoS attacks-GoldenEye 41508
    - DoS attacks-Slowloris 10990
  + Friday-02-03-2018\_TrafficForML\_CICFlowMeter
    - Benign 762384
    - Bot 286191
  + Friday-23-02-2018\_TrafficForML\_CICFlowMeter
    - Benign 1048009
    - Brute Force -Web 362
    - Brute Force -XSS 151
    - SQL Injection 53

**28/05 : Point avancement 2**

* Travail sur un dataset : Wednesday-14-02-2018\_TrafficForML\_CICFlowMeter
* Choix de quelques features afin d’entraîner un modèle de Machine Learning
  + Liste : 'Protocol', '**Flow Duration**', '**TotLen Fwd Pkts**', '**TotLen Bwd Pkts**', 'Fwd PSH Flags','Pkt Size Avg','Intrusion'
  + On a du enlever des features car elles contenaient des inf : comme par exemple *'Flow Byts/s','Flow Pkts/s'*; Nous allons par la suite regarder pour chaque feature le minimum et le maximum pour trouver celles qui contiennent des valeurs inf et donc les enlever
  + Normalisation de quelques features (celles en gras)
  + Création d’un fichier .csv avec ces features
* Pour ce nouveau fichier .csv, on a remplacé les valeurs NaN par 0, et arrondi les valeurs avec un nombre de décimal après la virgule
* Test du modèle **RandomForestRegression**: ça fonctionne ! car nous avions tous eu l’erreur   
  *ValueError: Input contains NaN, infinity or a value too large for dtype('float32').*

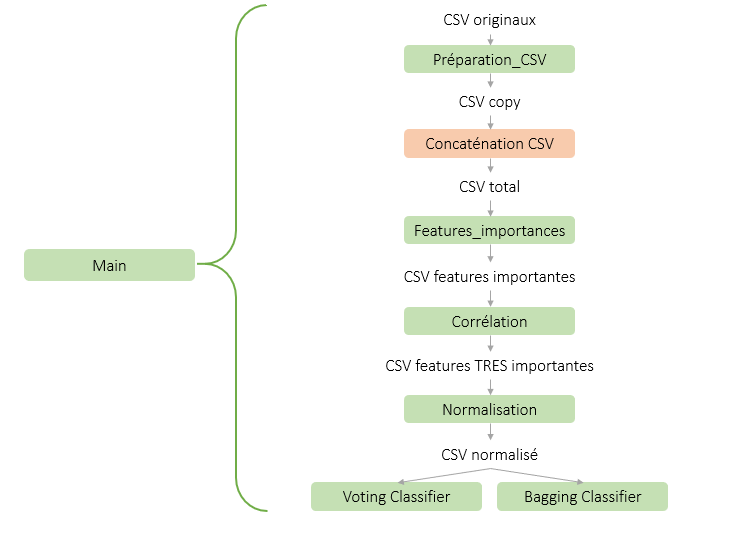
**17/06 : Point avancement 3**

* Equilibrage des datasets pour avoir autant de Begnin que d’intrusion, ainsi nous avons pu concaténer nos fichiers csv.
* Algorithme de l’importance des features, on a gardé que quelques features : 14 au total en comptant la colonne Index et Intrusion  
  Voici donc les features que nous avons retenus, avec un seuil de 0.001

Index ; Dst Port ; Tot Bwd Pkts ; Fwd IAT Std ; Fwd PSH Flags ; Fwd Header Len ; Fwd Pkts/s ; Bwd Pkts/s ; Pkt Len Max ; SYN Flag Cnt ; Init Fwd Win Byts ; Init Bwd Win Byts ; Fwd Seg Size Min ; Intrusion

* Et enfin corrélation sur ces features pour encore en éliminer : reste 8 features  
  Voici les features qui ont été retenu avec un seuil < -0.2 et > 0.2

Index ; Dst Port ; Fwd Pkts/s ; Bwd Pkts/s ; Pkt Len Max ; Init Bwd Win Byts ; Fwd Seg Size Min ; Intrusion

Schéma pour résumer notre travail jusqu’à présent :