



# **RAPPORT DE PROJET**

## Développement de l'outil **WinSecure**

Nom de la société :  
**UnixSecure**

Nom du projet :  
**Projet GN – MSI P9 S2**

Encadrant :  
**Guénaél RENAULT**

Chef de projet :  
**Lotfi DERRI**

Membres de l'équipe :  
**Eneda DEVOLLI**  
**Mohamed LAYADI**  
**Yousspha JAITEH**  
**Jeremy AHMADI**

# SOMMAIRE

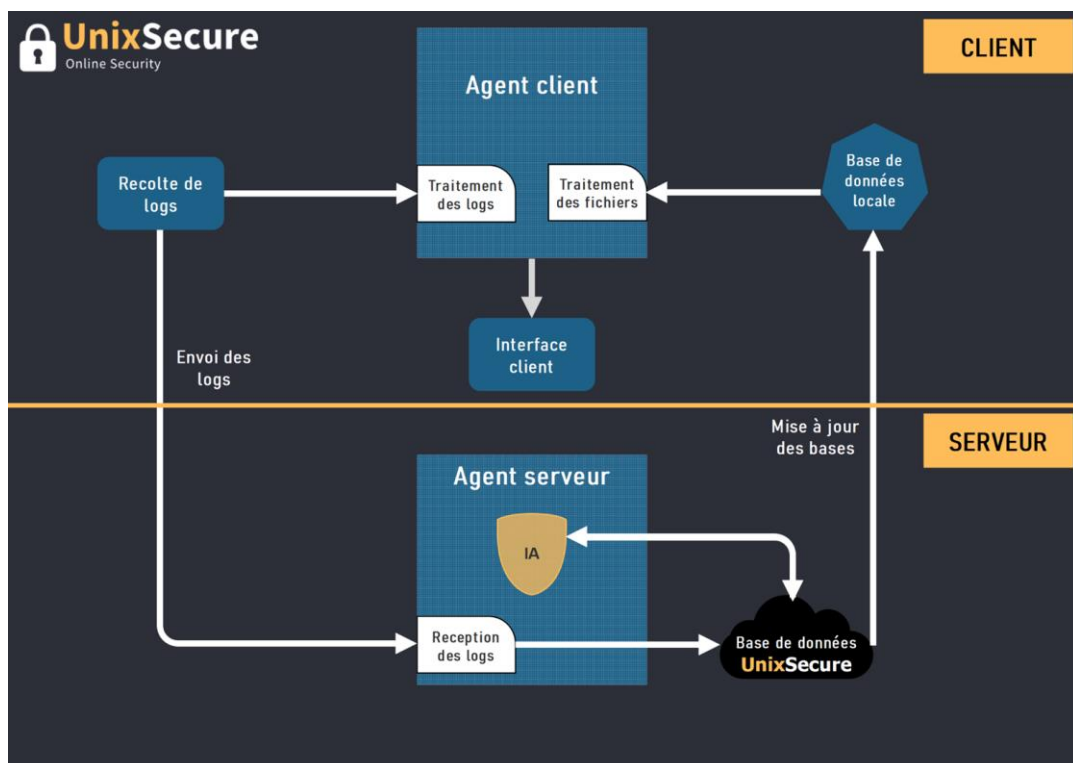
INTRODUCTION .....	2
MACHINE LEARNING .....	2
EXTRACTION ET PREDICTION .....	2
TESTS ET INTERPRETATION.....	2
POC.....	2
CONCLUSION.....	2

# INTRODUCTION

## 1. Premier objectif

Notre premier objectif consistait à mettre en place une solution de détection de comportement malicieux sur des systèmes d'exploitation Linux, en utilisant des bases de données contenant scenarios d'attaque et de l'intelligence artificielle pour corréler ces données et créer des règles de détection.

Voici l'ancien schéma fonctionnel de cette solution :



Comme nous le voyons sur ce schéma, le but était d'avoir un serveur central qui récolte et analyse les logs depuis les machines des utilisateurs, travaillant avec de l'intelligence artificielle, et alimentant la base de données du serveur en créant des règles de détection, et des Endpoint sur laquelle la détection se fera, avec une mise à jours régulières leurs bases de données locales pour pouvoir appliquer une sécurisation fiable sur tous les outils clients.

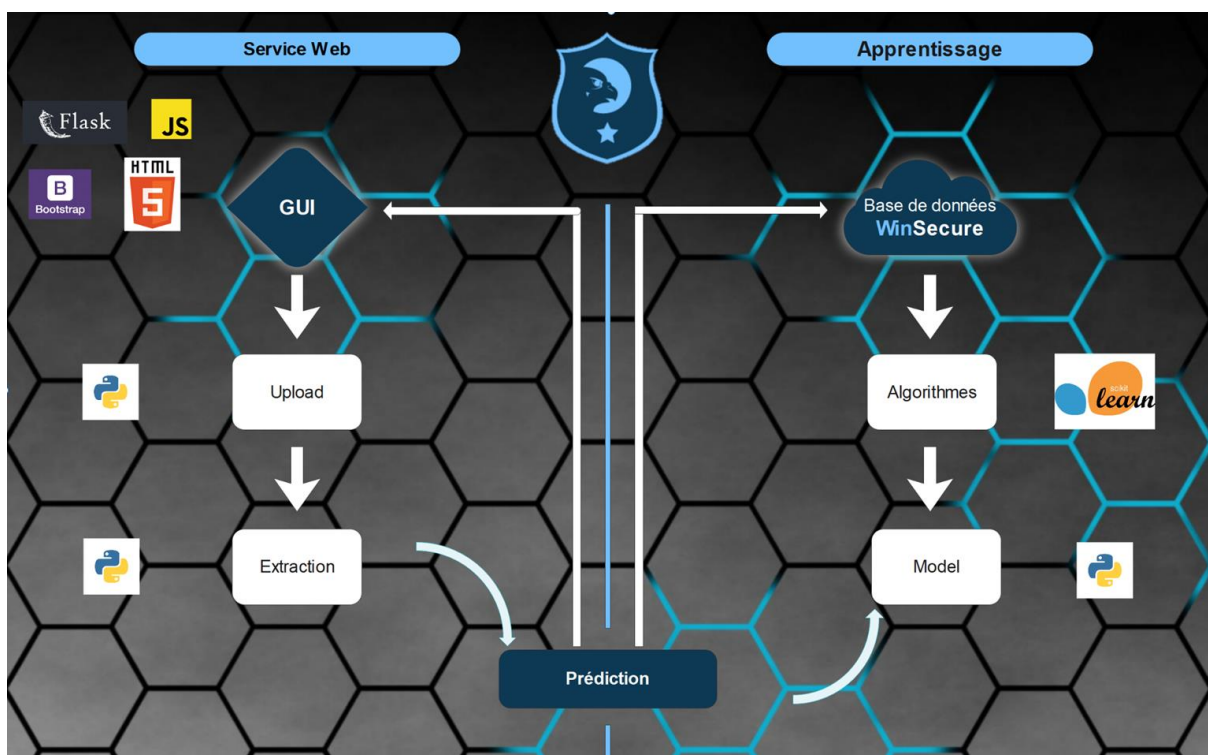
Nous nous sommes vite senti limités à cause du manque de disponibilité des données par rapport à ce que nous avons voulu réaliser, et comme l'IA a besoin de données pour apprendre, nous avons réévalué la faisabilité et décidé de partir sur une autre idée de projet.

## 2. Objectif final

Notre nouvel objectif était de réaliser une plateforme Web qui analyse des fichiers exécutables & systèmes Windows via de l'intelligence artificielle.

Après une étude de faisabilité et quelques recherches, nous avons trouvé assez de données pour pouvoir obtenir un résultat de l'apprentissage de l'IA, nous nous sommes alors fixé un objectif : Avoir un résultat proche de 100% sur la détection de malware en utilisant des algorithmes fiables.

Voici le nouveau schéma fonctionnel de la nouvelle solution :



Comme nous le pouvons voir sur ce schéma, nous avons deux parties fonctionnelles pour la solution WinSecure, une partie apprentissage, et une partie service Web.

Nous démarrons avec une base de données contenant 138.048 données, sur lesquelles nous exécutons de différents algorithmes, un algorithme de classification, puis des algorithmes d'apprentissage générant ainsi des modèles.

La partie service Web contient une interface graphique où les utilisateurs peuvent déposer un fichier que l'IA peut analyser (.exe & .dll). Lorsque le fichier, une fonction d'upload s'exécute pour le stocker sur le serveur, puis une fonction

d'extraction pour récupérer les informations que nous aurons besoin de passer à l'IA.

Une fois les informations extraites, nous exécutons la fonction de prédiction qui va s'appuyer sur les modèles générés par la partie d'apprentissage, alimenter la base de données pour avoir plus de données à la prochaine exécution de l'apprentissage et envoyer le résultat de l'analyse vers l'interface graphique.

# MACHINE LEARNING

## 1. Definitions

Le **Machine Learning** n'est entre autre que la rencontre des statistiques avec la puissance de calcul disponible. C'est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Pour apprendre et se développer, les ordinateurs ont toutefois besoin de données à analyser et sur lesquelles s'entraîner.

Il se découpe en **quatre étapes distinctes** mais **dépendantes** entre elles :

- Sélection et préparation des données
- Sélection des algorithmes
- Phase d'entraînement
- Génération et utilisation des modèles pour prédiction

Pour débiter l'implémentation de l'algorithme d'apprentissage, il est nécessaire de décider du type de l'apprentissage, à savoir **supervisé** ou **non supervisé**, et des algorithmes qui seront eux aussi nécessaire au bon développement de la **phase d'apprentissage**.

Dans notre cas, nous avons besoin d'un algorithme de classification, pour attribuer une catégorie à chaque observation d'un jeu de données. Nous en ont distinguons deux, **0 – Malicieux** et **1 – Légitime**. L'IA au cours de son apprentissage associera les statistiques des fichiers malicieux contenu dans la base de données avec la bonne catégorie, et va faire pareil pour les statistiques des fichiers légitimes. C'est le **principe de la classification**.

Comme nous pouvons le déduire, nous affectons une étiquette à nos données et c'est donc un **apprentissage supervisé** que nous allons réaliser.

Voici une représentation des données pour un apprentissage supervisé :

	Colonne 1	Colonne 2	Colonne 3	...	Colonne n	Etiquette
Ligne 1	data 1,1	data 1,2	data 1,3	...	data 1,n	result 1
Ligne 2	data 2,1	data 2,2	data 2,3	...	data 2,n	result 2
...	...	...	...	...	...	...
Ligne m	data n,1	data n,2	data n,3	...	data m,n	result m

Les algorithmes que nous utilisons établissent, à partir d'un jeu de données étiquetées, des **arbres de décisions**, servant à classier des observations futures lors de la **phase de prédiction**.

L'arbre de décision commence par une racine, puis une série de branches dont les intersections s'appellent des nœuds et termine par des feuilles qui correspondent chacun à une des classes à prédire.

## 2. Infrastructure

Pour comprendre notre réflexion à la manière dont nous avons construit les choses, il est nécessaire de comprendre l'organisation que nous avons mis en place.

Voici un capture d'écran reflétant cette dernière :

```
WinSecure@WinSecure:/var/www/WinSecure/static/IA$ ls -l
total 32
-rwxr-xr-x 1 root root 2032 Jun  3 11:56 apprentissage.py
drwxr-xr-x 2 root root 4096 Jun  3 12:07 data
drwxr-xr-x 2 root root 4096 Jun  3 13:38 extract
drwxr-xr-x 2 root root 4096 Jun  3 11:56 model
-rwxr-xr-x 1 root root 8120 Jun  4 01:12 prediction.py
drwxr-xr-x 2 root root 4096 Jun  3 11:58 stats
drwxr-xr-x 3 root root 4096 Jun  3 11:56 training_func
```

Ce dossier se trouve sur le chemin suivant depuis l'archive zip mise à votre disposition pour l'évaluation : « **WinSecure/static/IA** ».

Nous avons deux scripts sur la racine de l'espace de travail, un programme d'apprentissage "**apprentissage.py**", et un programme de prédiction "**prediction.py**".

Le programme d'apprentissage va récupérer les base de données **.csv** dans le répertoire "**data**", et va y stocker la référence des champs qu'il trouve les plus intéressant à analyser pendant la phase de classification. C'est aussi ces champs qu'il faudra extraire et passer en paramètre au modèle pour faire la prédiction, d'où le besoin de stockage afin de garder l'information.

Ce programme « **apprentissage.py** » appelle les algorithmes d'apprentissage sauvegardés dans le répertoire "**training\_func**", nous donnant chacun des informations sur leurs statistiques d'apprentissage, à savoir le score d'entraînement et le score de test, les faux positives etc..., nous pourrons les retrouver dans le dossier "**stats**".

Lorsque l'apprentissage est terminé, des modèles sont générées pour chaque algorithmes d'apprentissage exécutés, ils seront stockés dans le répertoire "**model**".

Le programme “prediction.py” va donc être appelé par le service web, pour traiter les fichiers des utilisateurs passés via l’interface graphique. Il appellera par la suite les fonctions d’extraction contenues dans le répertoire “**extract/**”, pour récupérer tous les champs dont nous aurons besoin par la suite, sélectionner les champs que le model attendra pour la prédiction, charger le model, exécuter la prédiction, alimenter la base de données et retourner un résultat au service Web.

### 3. Algorithme d’initialisation

L’algorithme d’initialisation “apprentissage.py” va, comme précisé plus haut, lire dans la base de données et exécuter l’algorithme de classification comme suit :

```
extratrees = ExtraTreesClassifier().fit(data_in, labels)
select = SelectFromModel(extratrees, prefit=True)
data_in_new = select.transform(data_in)
```

Nous lui donnons les données qui ont préalablement été mises dans la variable “**data\_in**”, et les étiquettes associés dans “**labels**”. Nous appelons par la suite la fonction servant à sélectionner les champs les plus intéressants à analyser, puis transformer les données de “**data\_in**” dans “**data\_in\_new**”.

Une fois les données transformées, nous les découpons en quatre variables pour pouvoir les passer en paramètre de l’IA :

```
X_train, X_test, y_train, y_test = train_test_split(data_in_new, labels, test_size=0.2)
```

**X\_train** : Les données sur lesquelles l’algorithme va apprendre.

**X\_test** : Les étiquettes des données dans X-train pour pouvoir classifier.

**y\_train** : Les données sur lesquelles l’algorithme va faire des tests pour s’auto-évaluer après l’apprentissage.

**y\_test** : Les étiquettes des données dans y\_train pour pouvoir comparer avec la prédiction dans la phase de test.

X\_train et X\_test contiennent 80% des données de la BDD, y\_train et y\_test 20%, c’est ce que signifie “**test\_size=0.2**” dans la fonction “**train\_test\_split**”.



Voici une représentation de données après l'appel à cette fonction :

		data_in_new					labels
		Colonne 1	Colonne 2	Colonne 3	...	Colonne n	Etiquette
80%	Ligne 1	data 1,1	data 1,2	data 1,3	...	data 1,n	result 1
	Ligne 2	data 2,1	data 2,2	data 2,3	...	data 2,n	result 2
	Ligne 3	data 3,1	data 3,2	data 3,3	...	data 3,n	result 3
	Ligne 4	data 4,1	data 4,2	data 4,3	...	data 4,n	result 4
	Ligne 5	data 5,1	data 5,2	data 5,3	...	data 5,n	result 5
	Ligne 6	data 6,1	data 6,2	data 6,3	...	data 6,n	result 6
	Ligne 7	data 7,1	data 7,2	data 7,3	...	data 7,n	result 7
	Ligne 8	data 8,1	data 8,2	data 8,3	...	data 8,n	result 8
20%	Ligne 9	data 9,1	data 9,2	data 9,3	...	data 9,n	result 9
	Ligne 10	data 10,1	data 10,2	data 10,3	...	data 10,n	result 10

## 4. Apprentissage

Avec ce découpage, le programme “**apprentissage.py**” va faire appel aux algorithmes d'apprentissage en passant ces données la en paramètre :

```
rfc.RandomForestClassifier_Training(X_train, X_test, y_train, y_test)
gbc.GradientBoostingClassifier_Training(X_train, X_test, y_train, y_test)
```

Nous avons l'algorithme **Random Forest Classifier** avec une efficacité de **99.36%** sur les 20% de données de test.

Les **forêts de décision aléatoire** est une méthode d'apprentissage d'ensemble pour la classification, la régression et d'autres tâches qui fonctionnent en construisant une multitude d'arbres de décision au moment de l'apprentissage. Pour les tâches de classification, la sortie de la forêt aléatoire est la classe sélectionnée par la plupart des arbres.

Voici les principales étapes de cette algorithme :

- Prendre le nombre X d'observations du jeu de données de départ
- Prendre le nombre Y des Z variables disponibles (colonne / features)
- Entraîne l'arbre de décision sur ce jeu de données
- Répète ces étapes N fois de sorte à obtenir N arbres de décisions

D'autre part, nous avons l'algorithme **Gradient Boosting Classifier** avec une efficacité de **98.74%** sur les 20% de données de test également.

L'**amplification de gradient** est une technique d'apprentissage automatique pour la régression, la classification et d'autres tâches, qui produit un modèle de prédiction sous la forme d'un ensemble de modèles de prédiction faibles, généralement des arbres de décision.

Pour la partie code voila comme sont appelées, respectivement, les algorithmes d'apprentissage.

Random Forest Classifier

```
classif = RandomForestClassifier(n_estimators=50)
classif.fit(X_train, y_train)
```

Gradient Boosting Classifier

```
classif = GradientBoostingClassifier(n_estimators=50)
classif.fit(X_train,y_train)
```

Suite à cette phase d'apprentissage, les deux algorithmes vont générer des modèles en **.bin**, grâce aux ligne suivantes :

```
with open('model/RandomForestClassifier_model.bin', 'wb') as f:
    pickle.dump(classif, f)
with open('model/GradientBoostingClassifier_model.bin', 'wb') as f:
    pickle.dump(classif, f)
```

## 5. DataSet

Le DataSet correspond à la base de données d'apprentissage de l'intelligence artificielle.

Elle prend le format suivant :

Name	Md5	Machine	SizeOfOpti onalheader	Characteri stics	....	Legitimat e
memtest.e xe	631ea355665f 28d4707448e 442fbf5b8	332	224	258	....	1

ose.exe	9d10f99a6712e 28f8acd5641e 3a7ea6b	332	224	3330	....	1
---------	--	-----	-----	------	------	---

Dans cette base de données, une ligne correspond à un fichier et chaque colonnes à une information extraite du fichier.

Pour chaque fichier, on dispose de 56 informations donc 56 colonnes. On y trouve le MD5, le type de fichier, la signature, la taille de ses options dans le header etc...

L'IA suit un apprentissage supervisé, la dernière colonne permet de donc différencier les fichiers légitimes des malwares.

La base de données comprend 138048 fichiers, dont 41323 légitimes ainsi que 96725 malwares.

# EXTRACTION ET PREDICTION

## 1. Extraction

Une fois que l'utilisateur dispose ses fichiers à scanner sur le site web, ils sont téléchargés sur le serveur web pour commencer leurs analyses.

La première étape est d'extraire les 56 informations vues précédemment dans la base de données de l'IA sans la colonne "legitimate" qu'on ne connaît pas à ce stade.

Pour cela on va utiliser PEFILE, c'est un module python multiplateforme développé par Ero Carerra qui nous permet de manipuler un fichier Portable Executable. il suffit de faire : `pefile.PE("fichier.exe")` puis il ne reste plus qu'à aller récupérer les informations qui nous intéressent sur les fichiers.

Par exemple les informations de la catégorie "FIXEDFILEINFO" sont récupérées de la manière suivantes :

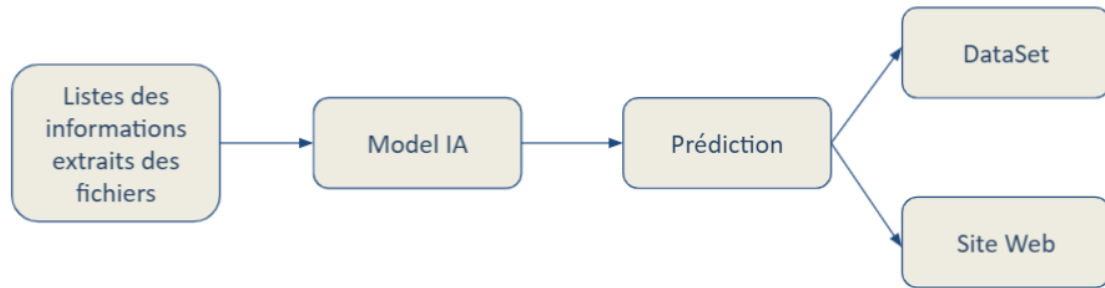
```
res['flags'] = pe.VS_FIXEDFILEINFO.FileFlags
res['os'] = pe.VS_FIXEDFILEINFO.FileOS
res['type'] = pe.VS_FIXEDFILEINFO.FileType
res['file_version'] = pe.VS_FIXEDFILEINFO.FileVersionLS
res['product_version'] = pe.VS_FIXEDFILEINFO.ProductVersionLS
ect...
```

## 2. Prédiction

Une fois qu'on dispose des 55 informations nécessaires pour chaque fichier téléchargés sur le serveur, il ne reste plus qu'à les envoyer au modèle IA pour qu'il puisse commencer les prédictions.

```
prediction = model.predict(ListForAllPrediction)
```

ListForAllPrediction correspond à une liste de liste, comportant les 55 informations de chaque fichier. Cela ressemble à la matrice de la base de données sans la colonne "legitimate".



Une fois les prédictions terminées, premièrement on va envoyer au site web le résultat ainsi que des informations qu'on souhaite affichées en relation avec le fichier analysé. Ces informations sont des détails que l'utilisateur peut consulter pour en savoir plus sur les fichiers qu'il a téléchargés.

Deuxièmement, tous les fichiers qui passent par la prédiction sont ensuite envoyés dans la base de données de l'apprentissage. Notre objectif est d'augmenter sans cesse le volume de la base de données pour améliorer le score de de prédiction.

# TESTS ET INTERPRETATION

## 1. Tests et interprétation des résultats :

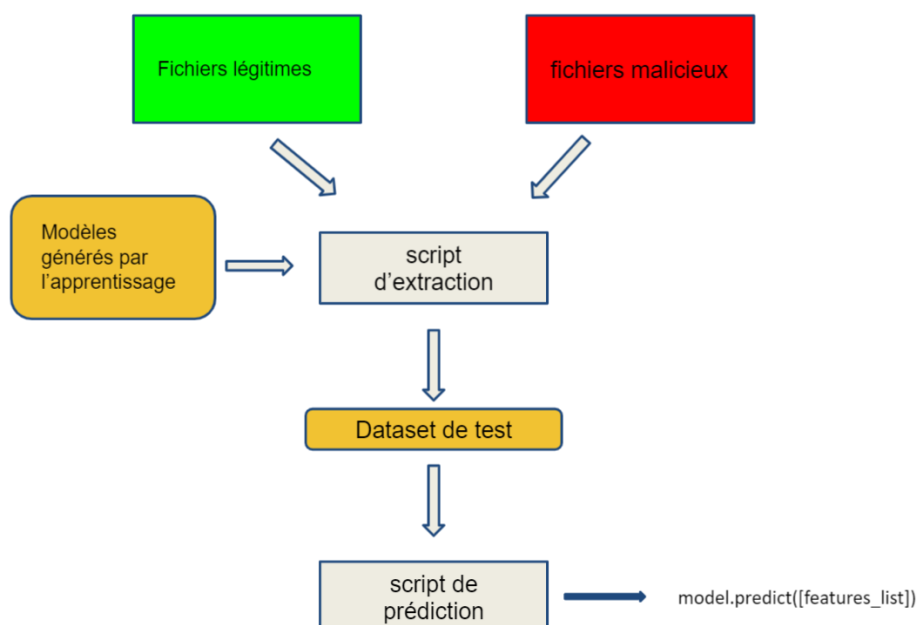
Afin de tester l'efficacité du model généré par l'IA, nous avons inclus dans notre projet une phase de test réalisée à l'aide de script python.

L'idée est dans un premier temps de recueillir un certain nombre de fichiers (.exe ou .dll) légitimes et malicieux.

Les fichiers légitimes recueillis provenaient en intégralité de nos propre machines tant que les fichiers malicieux ont été obtenus sur divers sites internet (<https://dasmalwerk.eu> et <https://bazaar.abuse.ch>).

Dans un second temps il fut question d'extraire de ces fichiers les données (pefile) qui nous intéressaient à savoir ceux présents dans le dataset d'apprentissage et plus spécifiquement celles que l'IA à considérer pertinents parmi l'ensemble du dataset d'entraînement.

Ce script d'extraction nous permet ainsi de créer notre propre dataset de test comprenant les données des fichiers recueillis et mis en forme correctement. Nous pouvons ainsi lancer notre script dédiée à la prédiction dont le rôle sera d'appliquer le model sur le dataset de test préalablement constitué.



Nous obtenons ainsi les résultats suivants :

```
590 fichiers malicieux provenant de: https://dasmalwerk.eu/
341 fichiers légitimes provenant de nos machines personnelles
WinSecure@WinSecure:~/Documents/IA$ sudo python3 prediction.py
[2, 3, 4, 13, 16, 20, 25, 26, 27, 35, 36, 49, 50, 55]

RandomForestClassifier

NUMBER OF ERRORS : 7/931
PERCENTAGE OF ERRORS: 0.7518796992481203 %

DecisionTreeClassifier

NUMBER OF ERRORS : 99/931
PERCENTAGE OF ERRORS: 10.633727175080558 %

GradientBoostingClassifier

NUMBER OF ERRORS : 9/931
PERCENTAGE OF ERRORS: 0.966702470461869 %
```

```
1100 fichiers malicieux provenant de: https://bazaar.abuse.ch/browse/
341 fichiers légitimes provenant de nos machines personnelles
WinSecure@WinSecure:~/Documents/IA$ sudo python3 prediction.py
[2, 3, 4, 13, 16, 20, 25, 26, 27, 35, 36, 49, 50, 55]

RandomForestClassifier

NUMBER OF ERRORS : 72/1431
PERCENTAGE OF ERRORS: 5.031446540880503 %

DecisionTreeClassifier

NUMBER OF ERRORS : 139/1431
PERCENTAGE OF ERRORS: 9.713487071977639 %

GradientBoostingClassifier

NUMBER OF ERRORS : 69/1431
PERCENTAGE OF ERRORS: 4.821802935010482 %
```

Il nous a semblé important de mettre en évidence les différences entre plusieurs algorithmes de classification.

On peut donc constater les pourcentages d'erreurs ont tendance à varier, raison pour laquelle afin de garder un esprit critique sur les résultats, l'intégration de plusieurs algorithmes sur le site web nous à sembler important.

De manière globale nous obtenons de bon score assez encourageant, cependant ce résultat est à relativiser car il nous aurait fallu une base de données bien plus conséquente pour affirmer que nos résultats sont fiables à 100%.

De même cet outil est plutôt recommandé pour être utilisé avec différents outils d'aide à la détection.

## 2. Sélection des données pertinentes :

L'une des étapes les plus importantes de l'apprentissage a été la sélection des features du dataset d'apprentissage dont l'IA a considéré qu'elles valaient la peine d'être considéré comme importante et jouant un rôle important dans la décision finale de considérer un fichier malicieux ou légitime.

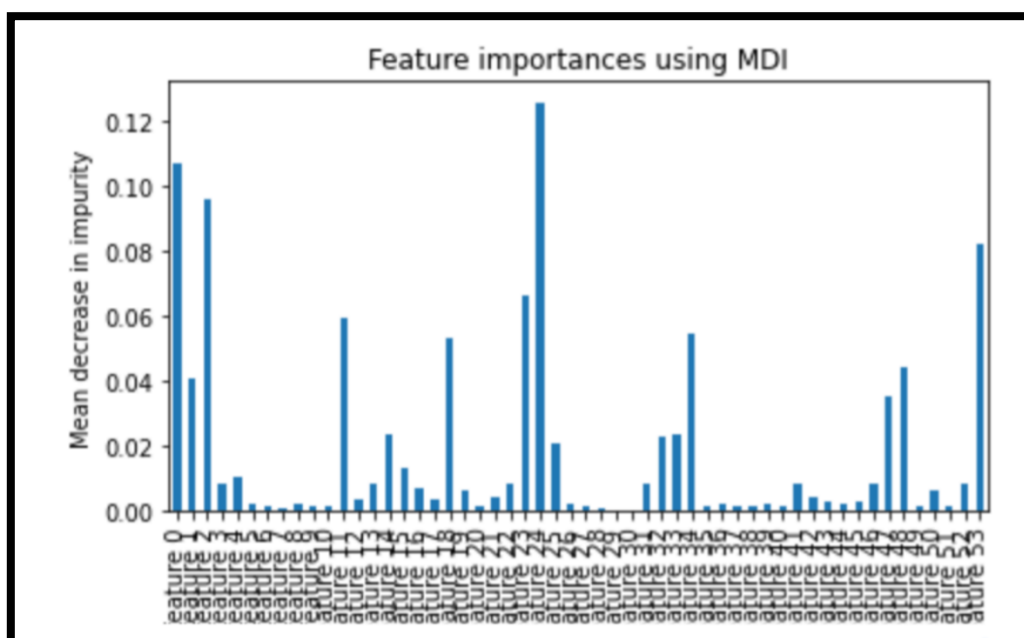
Pour ce faire la bibliothèque sklearn nous propose un outil très utile qui est la fonction `feature_importances_`.

Après avoir créé un arbre de décision, l'IA va donc analyser chaque colonne de notre dataset en vue de lui attribuer un score de pertinence et dont seuls les meilleurs scores seront retenus.

Il s'agit principalement d'un outil statistique basée sur l'indice (ou coefficient) de Gini dont le rôle est de permettre de rendre compte de la répartition d'une variable. Plus la valeur est élevée et plus la caractéristique est importante.

Comme on peut le voir ci-dessous, il s'agit des features sélectionnées par notre algorithme d'apprentissage, les résultats peuvent très légèrement variés d'une exécution à une autre dû à la présence de l'aléatoire dans la fonction `feature_importances_`.

```
1DllCharacteristics 0.12621938313213654
2Machine 0.10701699954782894
3Characteristics 0.09589584936454233
4VersionInformationSize 0.08204036188711286
5Subsystem 0.06625305527614885
6ImageBase 0.05935559681992614
7SectionsMaxEntropy 0.05478347126950434
8MajorSubsystemVersion 0.053228966395969426
9ResourcesMaxEntropy 0.044320321941247846
10SizeOfOptionalHeader 0.04086309322264022
11ResourcesMinEntropy 0.03511604283840767
12MajorOperatingSystemVersion 0.023255668691979335
13SectionsMinEntropy 0.023223397277293625
14SectionsMeanEntropy 0.022737669882002258
15SizeOfStackReserve 0.020293123644982324
[26, 2, 4, 55, 25, 13, 36, 20, 50, 3, 49, 16, 35, 34, 27]
```






### 3. Que nous apprend l'IA ?

Dans cette dernière partie il est question de se pencher sur ce que nous apprend finalement l'IA sur sa manière de fonctionner et ainsi de pouvoir justifier nos résultats plus aisément.

Démarche :

On s'est donc proposé la démarche suivante dans laquelle nous avons décidé d'isoler les fichiers malicieux et les fichiers légitimes dans des datasets distincts et ce pour le dataset d'apprentissage et le dataset de test. Ce qui nous donne un total de quatre datasets qu'on peut désormais analyser individuellement et en profondeur à l'aide du logiciel Excel.

<div>2 Machine</div> <div>The size of the optional header, which is required for executable files but not for object files. This value should be zero for an object file</div>	<div>Compte de « Machine »</div> <table><thead><tr><th>Étiquettes de lignes</th><th>Nombre sur M...</th></tr></thead><tbody><tr><td>332</td><td>25477</td></tr><tr><td>34404</td><td>15843</td></tr><tr><td>512</td><td>3</td></tr><tr><td>Total général</td><td>41323</td></tr></tbody></table>	Étiquettes de lignes	Nombre sur M...	332	25477	34404	15843	512	3	Total général	41323	<div>Étiquettes de lignes -&gt; Nombre de Machine</div> <table><tbody><tr><td>332</td><td>96656</td></tr><tr><td>34404</td><td>68</td></tr><tr><td>Total général</td><td>96724</td></tr></tbody></table>	332	96656	34404	68	Total général	96724	<div>Compte de « Machine »</div> <table><thead><tr><th>Étiquettes de lignes</th><th>Nombre sur Machine</th></tr></thead><tbody><tr><td>34404</td><td>309</td></tr><tr><td>332</td><td>23</td></tr><tr><td>Total général</td><td>332</td></tr></tbody></table>	Étiquettes de lignes	Nombre sur Machine	34404	309	332	23	Total général	332	<div>Compte de « Machine »</div> <table><thead><tr><th>Étiquettes de lignes</th><th>Nombre sur Machine</th></tr></thead><tbody><tr><td>332</td><td>594</td></tr><tr><td>34404</td><td>4</td></tr><tr><td>Total général</td><td>598</td></tr></tbody></table>	Étiquettes de lignes	Nombre sur Machine	332	594	34404	4	Total général	598	<div>Potentiellement pertinent pour détecter Legit</div> <div>34404 = Legit (majoritairement)</div>																												
Étiquettes de lignes	Nombre sur M...																																																																
332	25477																																																																
34404	15843																																																																
512	3																																																																
Total général	41323																																																																
332	96656																																																																
34404	68																																																																
Total général	96724																																																																
Étiquettes de lignes	Nombre sur Machine																																																																
34404	309																																																																
332	23																																																																
Total général	332																																																																
Étiquettes de lignes	Nombre sur Machine																																																																
332	594																																																																
34404	4																																																																
Total général	598																																																																
<div>3 SizeOfOptionalHeader</div> <div>The size of the optional header, which is required for executable files but not for object files. This value should be zero for an object file</div>	<div>Compte de « SizeOfOptionalHeader »</div> <table><thead><tr><th>Étiquettes de lignes</th><th>Nombre sur Si...</th></tr></thead><tbody><tr><td>224</td><td>25477</td></tr><tr><td>240</td><td>15846</td></tr><tr><td>Total général</td><td>41323</td></tr></tbody></table>	Étiquettes de lignes	Nombre sur Si...	224	25477	240	15846	Total général	41323	<div>Étiquettes de lignes -&gt; Nombre de SizeOfOptionalHeader</div> <table><tbody><tr><td>224</td><td>96653</td></tr><tr><td>240</td><td>68</td></tr><tr><td>232</td><td>1</td></tr><tr><td>352</td><td>1</td></tr><tr><td>248</td><td>1</td></tr><tr><td>Total général</td><td>96724</td></tr></tbody></table>	224	96653	240	68	232	1	352	1	248	1	Total général	96724	<div>Compte de « SizeOfOptionalHeader »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur SizeOfO...</th></tr></thead><tbody><tr><td>240</td><td>309</td></tr><tr><td>224</td><td>23</td></tr><tr><td>Total général</td><td>332</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur SizeOfO...	240	309	224	23	Total général	332	<div>Compte de « SizeOfOptionalHeader »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur SizeOfO...</th></tr></thead><tbody><tr><td>224</td><td>593</td></tr><tr><td>240</td><td>5</td></tr><tr><td>Total général</td><td>598</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur SizeOfO...	224	593	240	5	Total général	598	<div>Potentiellement pertinent pour détecter Malware</div> <div>240 != Malware (très souvent)</div>																								
Étiquettes de lignes	Nombre sur Si...																																																																
224	25477																																																																
240	15846																																																																
Total général	41323																																																																
224	96653																																																																
240	68																																																																
232	1																																																																
352	1																																																																
248	1																																																																
Total général	96724																																																																
Étiquettes de ligne...	Nombre sur SizeOfO...																																																																
240	309																																																																
224	23																																																																
Total général	332																																																																
Étiquettes de ligne...	Nombre sur SizeOfO...																																																																
224	593																																																																
240	5																																																																
Total général	598																																																																
<div>4 Characteristics</div> <div>The Characteristics field contains flags that indicate attributes of the object or image file.</div>	<div>Compte de « Characteristics »</div> <table><thead><tr><th>Étiquettes de lignes</th><th>Nombre sur Characteristics</th></tr></thead><tbody><tr><td>8450</td><td>10227</td></tr><tr><td>8226</td><td>13417</td></tr><tr><td>8462</td><td>1341</td></tr><tr><td>34</td><td>2176</td></tr><tr><td>258</td><td>2051</td></tr><tr><td>271</td><td>929</td></tr><tr><td>8462</td><td>363</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de lignes	Nombre sur Characteristics	8450	10227	8226	13417	8462	1341	34	2176	258	2051	271	929	8462	363	...	...	<div>Compte de « Characteristics »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur Characte...</th></tr></thead><tbody><tr><td>258</td><td>67151</td></tr><tr><td>271</td><td>6270</td></tr><tr><td>33167</td><td>4684</td></tr><tr><td>259</td><td>6146</td></tr><tr><td>783</td><td>4395</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur Characte...	258	67151	271	6270	33167	4684	259	6146	783	4395	...	...	<div>Compte de « Characteristics »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur Characte...</th></tr></thead><tbody><tr><td>34</td><td>185</td></tr><tr><td>8226</td><td>134</td></tr><tr><td>8450</td><td>5</td></tr><tr><td>258</td><td>4</td></tr><tr><td>3106</td><td>1</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur Characte...	34	185	8226	134	8450	5	258	4	3106	1	...	...	<div>Compte de « Characteristics »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur Characte...</th></tr></thead><tbody><tr><td>271</td><td>189</td></tr><tr><td>258</td><td>173</td></tr><tr><td>259</td><td>60</td></tr><tr><td>33166</td><td>47</td></tr><tr><td>33167</td><td>41</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur Characte...	271	189	258	173	259	60	33166	47	33167	41	...	...	<div>Potentiellement pertinent pour détecter Legit et Malware</div> <div>8826,34 = Legit(majoritairement)</div> <div>258,271 = Malware(majoritairement)</div>
Étiquettes de lignes	Nombre sur Characteristics																																																																
8450	10227																																																																
8226	13417																																																																
8462	1341																																																																
34	2176																																																																
258	2051																																																																
271	929																																																																
8462	363																																																																
...	...																																																																
Étiquettes de ligne...	Nombre sur Characte...																																																																
258	67151																																																																
271	6270																																																																
33167	4684																																																																
259	6146																																																																
783	4395																																																																
...	...																																																																
Étiquettes de ligne...	Nombre sur Characte...																																																																
34	185																																																																
8226	134																																																																
8450	5																																																																
258	4																																																																
3106	1																																																																
...	...																																																																
Étiquettes de ligne...	Nombre sur Characte...																																																																
271	189																																																																
258	173																																																																
259	60																																																																
33166	47																																																																
33167	41																																																																
...	...																																																																
<div>13 ImageBase</div> <div>The preferred address of the first byte of the image when it is loaded in memory. This value is a multiple of 64K bytes. The default value for DLLs is 0x10000000 (=268435456 en décimale). The default value for applications is 0x00400000(=4194304 en décimale), except on Windows CE where it is 0x00100000.</div>	<div>« ImageBase » -&gt; 4194304 est mentionné le plus souvent.</div> <div>Nombre sur ImageBase (Milliers)</div> 	<div>Compte de « ImageBase »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur ImageBase...</th></tr></thead><tbody><tr><td>4194304</td><td>95310</td></tr><tr><td>268435456</td><td>671</td></tr><tr><td>16777216</td><td>218</td></tr><tr><td>65536</td><td>34</td></tr><tr><td>5366709120</td><td>31</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur ImageBase...	4194304	95310	268435456	671	16777216	218	65536	34	5366709120	31	...	...	<div>Compte de « ImageBase »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur ImageBase...</th></tr></thead><tbody><tr><td>5366709120</td><td>184</td></tr><tr><td>6442450940</td><td>117</td></tr><tr><td>268435456</td><td>18</td></tr><tr><td>4194304</td><td>7</td></tr><tr><td>6,88E+12</td><td>1</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur ImageBase...	5366709120	184	6442450940	117	268435456	18	4194304	7	6,88E+12	1	...	...	<div>Compte de « ImageBase »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur ImageBase...</th></tr></thead><tbody><tr><td>4194304</td><td>560</td></tr><tr><td>268435456</td><td>9</td></tr><tr><td>16777216</td><td>7</td></tr><tr><td>2097152</td><td>6</td></tr><tr><td>50331648</td><td>5</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur ImageBase...	4194304	560	268435456	9	16777216	7	2097152	6	50331648	5	...	...	<div>Potentiellement pertinent pour détecter Malware</div> <div>4194304(dec)= 0x00400000(hex)</div> <div>Grande majorité de malware Malware</div>																		
Étiquettes de ligne...	Nombre sur ImageBase...																																																																
4194304	95310																																																																
268435456	671																																																																
16777216	218																																																																
65536	34																																																																
5366709120	31																																																																
...	...																																																																
Étiquettes de ligne...	Nombre sur ImageBase...																																																																
5366709120	184																																																																
6442450940	117																																																																
268435456	18																																																																
4194304	7																																																																
6,88E+12	1																																																																
...	...																																																																
Étiquettes de ligne...	Nombre sur ImageBase...																																																																
4194304	560																																																																
268435456	9																																																																
16777216	7																																																																
2097152	6																																																																
50331648	5																																																																
...	...																																																																
<div>16 MajorOperatingSystemVersion</div> <div>The major version number of the subsystem</div>	<div>Étiquettes de lignes -&gt; Nombre de MajorOperatingSystemVersion</div> <table><tbody><tr><td>25180</td><td>6</td></tr><tr><td>9627</td><td>5</td></tr><tr><td>6338</td><td>4</td></tr><tr><td>112</td><td>10</td></tr><tr><td>39</td><td>0</td></tr><tr><td>25</td><td>1</td></tr><tr><td>2</td><td>7</td></tr><tr><td>Total général</td><td>41323</td></tr></tbody></table>	25180	6	9627	5	6338	4	112	10	39	0	25	1	2	7	Total général	41323	<div>Compte de « MajorOperatingSystemVersion »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur MajorO...</th></tr></thead><tbody><tr><td>5</td><td>67983</td></tr><tr><td>4</td><td>23935</td></tr><tr><td>1</td><td>4659</td></tr><tr><td>6</td><td>129</td></tr><tr><td>Total général</td><td>8</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur MajorO...	5	67983	4	23935	1	4659	6	129	Total général	8	<div>Compte de « MajorOperatingSystemVersion »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur MajorO...</th></tr></thead><tbody><tr><td>10</td><td>298</td></tr><tr><td>4</td><td>23</td></tr><tr><td>6</td><td>9</td></tr><tr><td>5</td><td>2</td></tr><tr><td>Total général</td><td>332</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur MajorO...	10	298	4	23	6	9	5	2	Total général	332	<div>Compte de « MajorOperatingSystemVersion »</div> <table><thead><tr><th>Étiquettes de ligne...</th><th>Nombre sur MajorO...</th></tr></thead><tbody><tr><td>4</td><td>387</td></tr><tr><td>5</td><td>186</td></tr><tr><td>1</td><td>14</td></tr><tr><td>6</td><td>11</td></tr><tr><td>Total général</td><td>598</td></tr></tbody></table>	Étiquettes de ligne...	Nombre sur MajorO...	4	387	5	186	1	14	6	11	Total général	598	<div>Potentiellement pertinent pour détecter Legit et Malware</div> <div>Majorité &gt;=6 = Legit</div> <div>Majorité &lt; 6 = Malware</div>								
25180	6																																																																
9627	5																																																																
6338	4																																																																
112	10																																																																
39	0																																																																
25	1																																																																
2	7																																																																
Total général	41323																																																																
Étiquettes de ligne...	Nombre sur MajorO...																																																																
5	67983																																																																
4	23935																																																																
1	4659																																																																
6	129																																																																
Total général	8																																																																
Étiquettes de ligne...	Nombre sur MajorO...																																																																
10	298																																																																
4	23																																																																
6	9																																																																
5	2																																																																
Total général	332																																																																
Étiquettes de ligne...	Nombre sur MajorO...																																																																
4	387																																																																
5	186																																																																
1	14																																																																
6	11																																																																
Total général	598																																																																

<b>16 MajorOperatingSystemVersion</b> The major version number of the subsystem.	Compte de « MajorOperatingSystemVersion » Étiquettes de lignes... Nombre sur MajorOperatingSystemVersion... 5 25185 6 9627 4 6338 10 112 0 39 1 25 7 2 Total général 41323	Compte de « MajorOperatingSystemVersion » Étiquettes de lignes... Nombre sur MajorO... 5 67983 4 23915 1 4659 6 129 0 6 Total général 96724	Compte de « MajorOperatingSystemVersion » Étiquettes de lignes... Nombre sur MajorO... 10 298 4 23 6 9 5 2 Total général 332	Compte de « MajorOperatingSystemVersion » Étiquettes de lignes... Nombre sur MajorO... 4 387 5 186 1 14 6 11 Total général 598	Potentiellement pertinent pour détecter Legit et Malware Majorité >=6 = Legit Majorité < 6 = Malware
<b>20 MajorSubsystemVersion</b> The major version number of the subsystem. Executables being built with Microsoft Visual Studio 2010 which sets the MajorSubsystemVersion and MinorSubsystemVersion in the PE header to 5 and 1	Compte de « MajorSubsystemVersion » Étiquettes de lignes... Nombre sur MajorSubsystemVer... 6 18600 5 11817 4 8803 10 51 1 38 3 13 Total général 41323	Compte de « MajorSubsystemVersion » Étiquettes de lignes... Nombre sur MajorSu... 5 67748 4 28846 3 73 6 57 Total général 96724	Compte de « MajorSubsystemVersion » Étiquettes de lignes... Nombre sur MajorSu... 10 282 6 44 5 4 4 2 Total général 332	Compte de « MajorSubsystemVersion » Étiquettes de lignes... Nombre sur MajorSu... 4 405 5 185 6 8 Total général 598	Potentiellement pertinent pour détecter Legit et Malware Majorité >=6 = Legit Majorité < 6 = Malware
<b>25 Subsystem</b> The subsystem required to run this image 2: IMAGE_SUBSYSTEM_WINDOWS_GUI Windows graphical user interface (GUI) subsystem. 3: IMAGE_SUBSYSTEM_WINDOWS_CUI Windows character-mode user interface (CUI) subsystem.	Compte de « Subsystem » Étiquettes de lignes... Nombre sur Subst... 3 22114 2 17743 1 1427 16 39 Total général 41323	Compte de « Subsystem » Étiquettes de lignes... Nombre sur Subst... 2 96168 3 526 1 30 Total général 96724	Compte de « Subsystem » Étiquettes de lignes... Nombre sur Subst... 3 198 2 132 1 2 Total général 332	Compte de « Subsystem » Étiquettes de lignes... Nombre sur Subst... 2 582 3 16 Total général 598	Potentiellement pertinent pour détecter Malware Subsystem = 2 (Majoritairement)
<b>26 DICharacteristics</b> The DLL characteristics of the image.	Compte de « DICharacteristics » Étiquettes de lignes... Nombre sur DI... 320 16397 0 4800 1344 4354 64 3818 ... ...	Compte de « DICharacteristics » Étiquettes de lignes... Nombre sur DIChara... 33088 57591 0 11988 32768 11783 34112 8099 33024 3040 ... ...	Compte de « DICharacteristics » Étiquettes de lignes... Nombre sur DIChara... 49504 165 16736 106 34144 21 352 15 49632 7 ... ...	Compte de « DICharacteristics » Étiquettes de lignes... Nombre sur DIChara... 0 255 34112 110 32768 74 33088 53 33024 44 ... ...	Pour les malware ce sont très souvent les mêmes valeurs qui reviennent

On constate pour les features que l'IA à sélectionner une répartition des variables assez facile à visualiser. (Commentaires dans le tableau)

Bien qu'individuellement, une seule feature ne permet pas de déterminer si un fichier est légitime ou malicieux, la combinaison des observations statistique peut s'avérer être un outil redoutable pour déterminer le résultat final.

On comprend ainsi mieux le fonctionnement de l'IA dans sa phase d'apprentissage et ainsi que sa conception des modèles, de plus elle nous apprend certaines choses qu'il aurait été assez difficile de déceler sans elle.

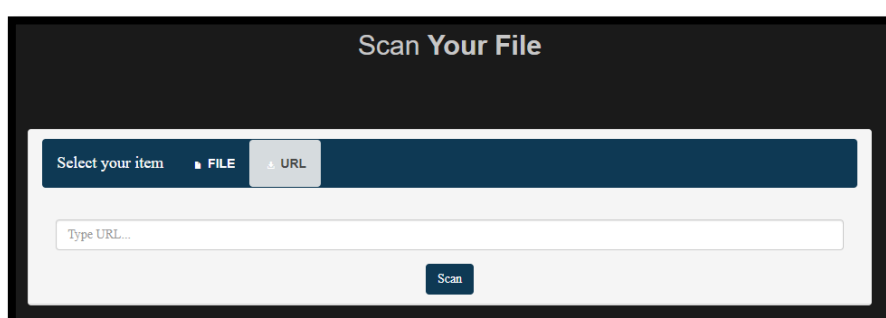
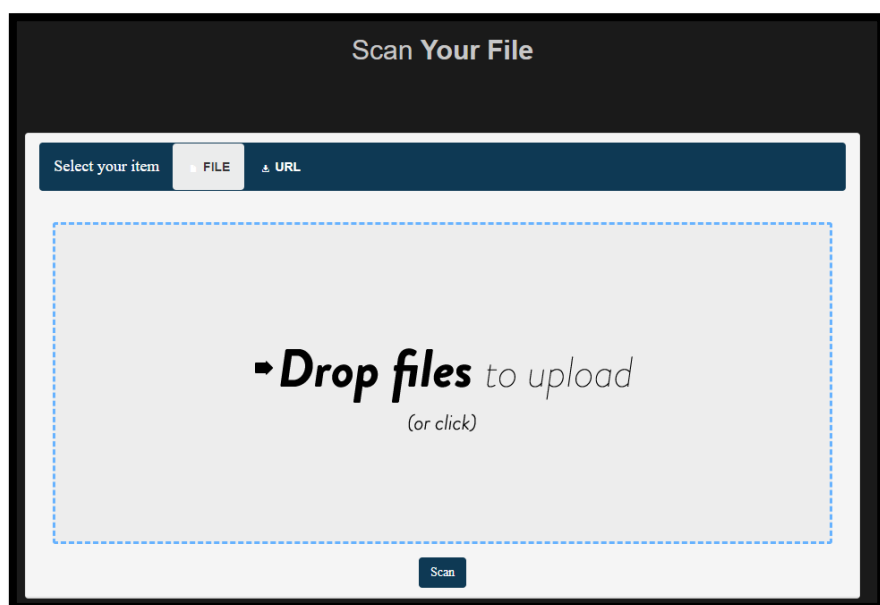
Cette démarche ne permet bien entendu pas de comprendre de manière exhaustive le fonctionnement général de l'IA mais nous donne un bel aperçu de ce que peut accomplir cet outil. La classification s'avère donc être un outil très efficace pour ce genre de problématique.

# POC

Nous avons mis en place une plateforme Web, en l'hébergeant avec le service Cloud d'Azure.

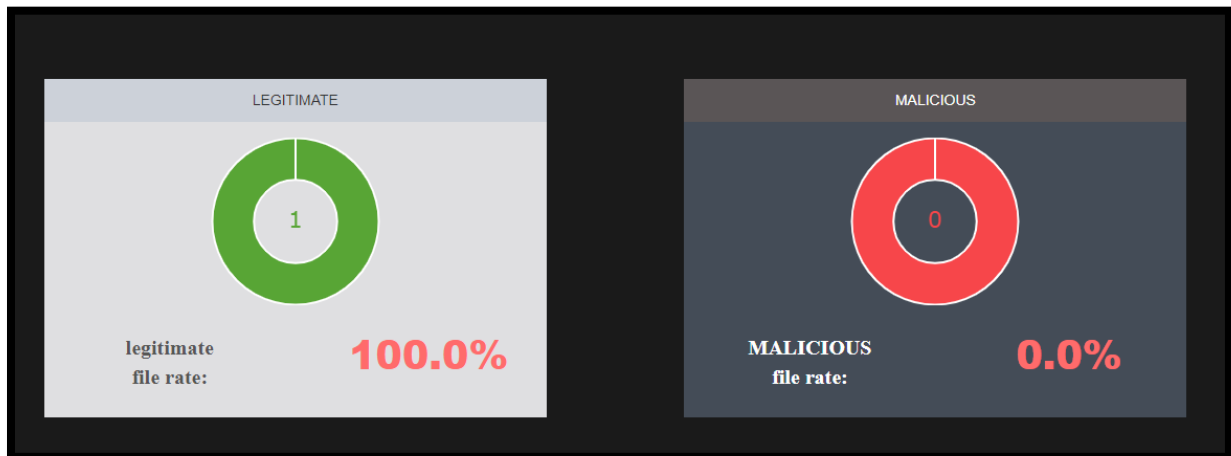
Voici une présentation rapide de l'interface et en parallèle une preuve de concept de la solution.

Nous avons la première page, où nous pouvons déposer un ou plusieurs fichiers, mais également un lien URL vers un fichier, avant de lancer le scan :



Une fois que le scan est lancé, nous passons en deuxième page du site internet ou nous avons de multiples informations à notre disposition.

Voici un exemple sur un fichier légitime :

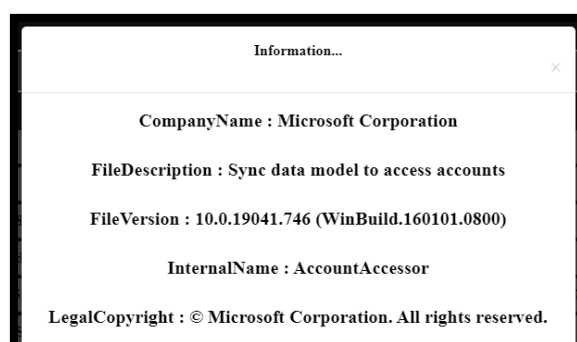


Nous avons d'abord le pourcentage de fichiers légitimes et le pourcentage de fichiers malicieux, pour présentation générale des résultats.

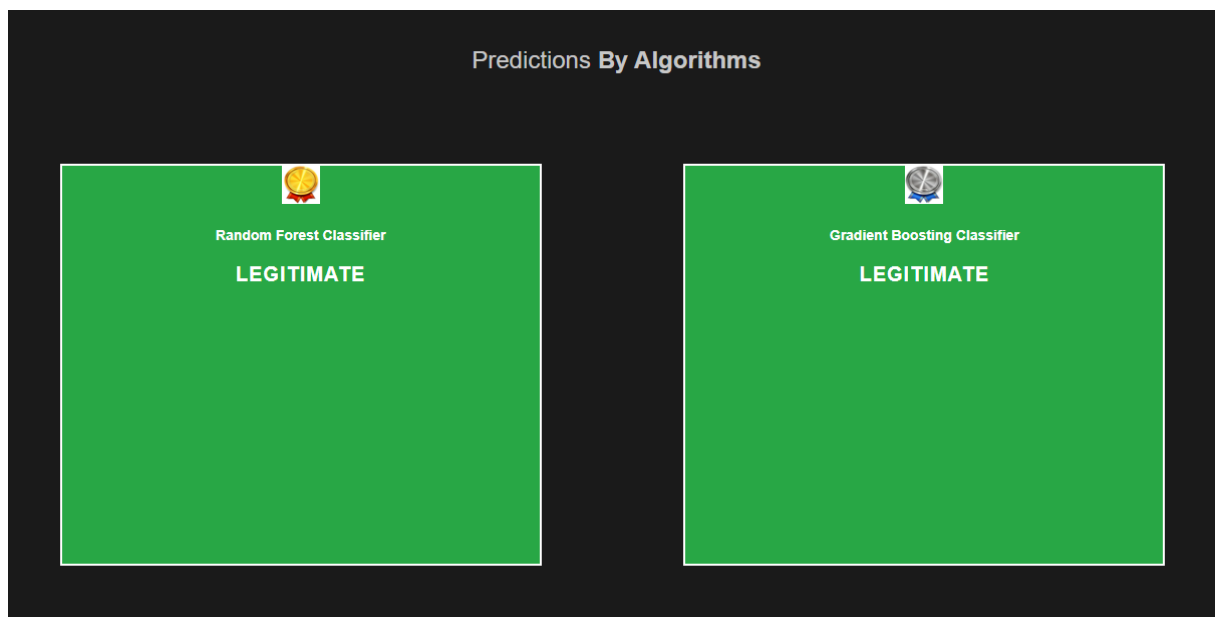
accountaccessor.dll
<b>Bulletin File</b>
Size : 0.27 Mo
Type : DLL
md5 : def8975296868ab8a9e4afc182985f6a
sha256 : 6aaefbee859b0a1859d999b28f8ae28dbca12f1bb4685c4d3f4cc5eb7215cb13
sha512 : 665f0498653158dbef5e6c7f7e8bc4e15b7b0509b064e1e4c955108a72fdf4dd00872a6d0796e5094d43e2f451417cb51732d7f8beb6f2d701cdc0868a3055f7
Signature : File signed
<b>LEGITIMATE</b>

Puis nous avons des informations sur le(s) fichier(s) analysé(s), comme la taille, le type et les différents hashes, avec la prédiction final de l'IA, s'appuyant sur l'algorithme avec le score de précision le plus élevé.

Lorsque le fichier est signé « File signed », nous pouvons voir les informations de l'autorité qui a signé ce fichier en cliquant sur l'écriture « Signature : File signed ».



Les prédictions des deux algorithmes utilisés pour l'apprentissage sont disposés de cette manière, par ordre avec le score le plus élevé à gauche pour **Random Forest Classifier (#1)**, et à droite le score le moins élevé pour **Gradient Boosting Classifier (#2)**.



En cliquant sur l'un des algorithmes (sur l'écriture « LEGITIMATE » ou « MALICIOUS »), une pop-up contenant toutes ses statistiques s'affiche, avec le pourcentage de précision d'entraînement et celui des tests, avec les faux positifs etc....

(Il faut cliquer sur « Information... » en haut de la pop-up pour que les informations s'affichent, et survoler les ronds colorés pour voir le pourcentage)



Pour ne pas dépasser la limitation financière relative à l'abonnement étudiant que nous avons pris sur Azure Cloud pour pouvoir déployer le serveur, nous avons éteint la machine hébergeant le service WinSecure.

Lorsque vous arriverez à cette partie, vous pouvez nous envoyer un mail, et nous pourrons par la suite rallumer le serveur pour que vous puissiez visiter le site, si toutefois vous en ressentez le besoin.

Voici le lien du site :

<https://win-secure.westeurope.cloudapp.azure.com/>

# CONCLUSION

Difficultés rencontrées :

- Mise en production du serveur web.
- Difficultés de compréhension de l'IA.
- Obtention des malwares.

Axes d'améliorations :

- Ajout d'une mauvaise prédiction dans la base de données.
- Conteneurisé les malwares sur le serveur.
- Augmenter les données pour l'apprentissage de l'IA/phase de test.
- Design de l'application web/Meilleur architecture web.
- Ajouter d'autres types de fichiers dans la détection.



