

# Rapport de stage 3A: Machine Learning appliqué à la cybersécurité

Mathieu Thomassin

2024-08-31

[Download PDF](#)

## Table des matières

- [Introduction](#)
- [01 - Découverte et Contexte](#)
- [02 - Recherche et Collecte de Données](#)
- [03 - Choix des Problèmes et des Outils de Machine Learning](#)
- [04 - Exploration de Modèles et Outils Avancés](#)
- [05 - Anomaly Detection et Modèles de Classification](#)
- [06 - Exploration d'Articles Scientifiques](#)
- [07 - Modélisation et Évaluation des Modèles](#)
- [08 - Travail avec Splunk et Pipelines](#)
- [09 - Expérimentations et Déploiement](#)
- [10 - Reproductibilité et MLOps](#)
- [Conclusion](#)
- [Annexe](#)

## Présentation du stagiaire et du maître de stage

- **Stagiaire** : Mathieu THOMASSIN
- **Maître de stage** : Michael ORSUCCI, responsable du *Security Operation Center* de l'Insee
- **Titre du stage** : Machine Learning appliqué à la cybersécurité

## **Introduction**

L'Insee a renforcé récemment ses capacités opérationnelles en termes de cybersécurité via notamment la création de son SOC (Security Operation Center) en septembre 2023. Devant la quantité de données et leur diversité et face aux évolutions des techniques et tactiques des attaquants, les méthodes de détection de cyberattaques peuvent avoir des limites. L'application d'algorithmes de Machine Learning peut aider les analystes SOC à repérer des attaques. Se déroulant au sein de l'équipe SOC construite depuis peu, le stage va permettre d'appliquer des techniques de Machine Learning et de deep learning sur des jeux de données réelles, afin de participer à la détection d'incidents de sécurité.

## **Présentation de l'organisation**

L'Insee a renforcé récemment ses capacités opérationnelles en termes de cybersécurité via notamment la création de son SOC (Security Operation Center) en septembre 2023. Cette équipe est répartie entre les sites de la DR de Nantes et la DR de Metz et a pour objectif de renforcer la sécurité du système d'information (SI).

## **Contexte général du stage**

Devant la quantité de données et leur diversité et face aux évolutions des techniques et tactiques des attaquants, les méthodes de détection de cyberattaques peuvent avoir des limites. L'application d'algorithmes de Machine Learning peut aider les analystes SOC à repérer des attaques. Le stage avait donc pour but de permettre d'appliquer des techniques de Machine Learning et de deep learning sur des jeux de données réelles, afin de participer à la détection d'incidents de sécurité.

## **Objectifs du stage**

### **Objectifs principaux**

Il s'agissait d'appliquer des techniques de détection de requêtes malveillantes arrivant dans le SI de l'Insee. Les requêtes arrivant au sein du SI peuvent être centralisées par un SIEM (Security Information and Event Management).

## **Résultats attendus**

- Offrir un service s'ajoutant au SIEM permettant d'identifier des requêtes malveillantes.
- Pouvoir comparer différents meilleurs modèles (au sens d'une recherche dans les hyper-paramètres) entre eux.

- Favoriser les bonnes pratiques du MLOps : reproductibilité, contrôle de version, automatisation, surveillance, collaboration.
- Étendre la méthodologie à des jeux de données publics différents.
- Explorer les algorithmes de détection d'anomalie.

## **Importance de la mission pour l'Insee**

À titre d'exemple, l'Insee détient l'application Elire. Une attaque réussie sur cette application perturberait le bon déroulement de la vie démocratique. Finaliser la formation d'attaché statisticien en s'appliquant à résoudre un problème concret avant une prise de poste dans un domaine proche.

## **Contexte général du stage**

Ce stage s'inscrit pleinement dans le parcours que j'ai progressivement construit tout au long de ma carrière à l'Insee. Débutant il y a plusieurs années sur un poste de contrôleur-programmeur, j'ai pu apprécier les cours d'informatique et de statistiques pendant mon parcours à l'Ensai. J'y ai suivi les options menant progressivement à se spécialiser jusqu'au master en Informatique et traitement des données.

Au cours de ma préparation et de l'obtention de la qualification d'analyste, ma curiosité pour la sécurité des systèmes d'information s'est particulièrement développée. L'analyse des requêtes au sein d'un réseau soulève de nombreuses questions : volume des données, méthode de traitement statistique avec le machine learning, rapidité et efficacité de ce traitement. C'est un domaine où les techniques de machine learning présentent un intérêt certain.

Les enjeux de la cybersécurité peuvent se révéler particulièrement lourds, comme on peut le découvrir dans les journaux pour de nombreuses organisations. Ce stage au SOC représente une opportunité unique de confronter ces intérêts théoriques à des problématiques concrètes de cybersécurité, renforçant ainsi mes compétences et ma compréhension dans un domaine en pleine expansion.

## **Importance de la cybersécurité et du machine learning dans ce domaine**

La cybersécurité et le machine learning sont cruciaux pour les entreprises aujourd'hui pour plusieurs raisons. D'une part, la cybersécurité est essentielle pour protéger les données sensibles contre les cyberattaques qui sont de plus en plus sophistiquées et fréquentes. Les conséquences d'une faille de sécurité peuvent être dévastatrices pour l'Insee, incluant des pertes de données, des dommages à la réputation et la crédibilité de l'Institut, et des impacts sur les utilisateurs.

D'autre part, le machine learning offre des outils puissants pour détecter et prévenir les menaces en temps réel. Grâce à ses capacités de traitement et d'analyse de vastes quantités de

données, le machine learning peut identifier des modèles et des anomalies que les méthodes traditionnelles pourraient manquer. Par exemple, la DSI a déployé sur le poste de chaque agent HarfangLab, un outil de sécurité informatique reposant sur des modèles de machine learning pour détecter la présence de logiciels malveillants, des malware, et permettant d'isoler le poste compromis.

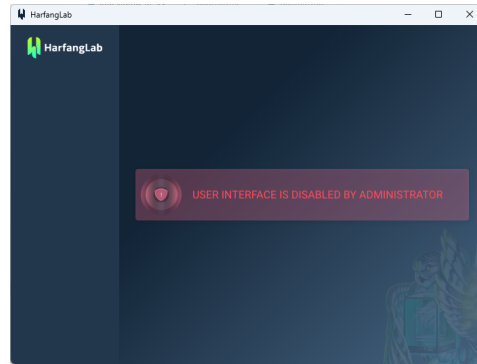


Figure 1: HarfangLab, déployé sur le poste de travail

Le machine learning permettrait également de réduire le travail des analystes du SOC en classifiant automatiquement les requêtes comme bénignes ou malveillante.

## Objectifs du stage

En débutant ce stage, j'avais pour but de pouvoir réaliser des objectifs à la fois académiques, en lien avec mes intérêts développés lors de la scolarité, ainsi que des objectifs professionnels, me permettant de m'insérer au mieux dans la structure de l'Insee pour débiter ma carrière d'attaché.

1. **Alignement avec mes objectifs de carrière** : Ce stage présentait une opportunité unique de développer des compétences à la fois en statistique et en informatique. De plus, au lieu de se dérouler dans un service de développement, il a eu lieu dans un service de production, ce qui m'a permis de découvrir une autre perspective de l'informatique par rapport à l'informatique de type "data science" pratiquée à l'Ensa. En renforçant mes compétences en machine learning et en informatique, j'ai donc pu acquérir de l'expérience pour de futurs rôles au sein de l'Insee.
2. **Développement de compétences spécifiques** : Pour réaliser la tâche principale de ce stage, j'avais besoin de maîtriser des techniques de machine learning et de deep learning appliquées à la cybersécurité, ainsi que de les inscrire dans une approche de type DevOps ou MLOps. Il m'a donc fallu utiliser des outils comme Scikit-learn, MLflow, des techniques de deep learning, et utiliser le cluster Kubernetes pour entraîner et déployer un modèle avec une API.

3. **Application pratique des connaissances théoriques** : Si j’ai appris à faire du machine learning à l’école et y ai découvert des notions de DevOps, ce n’est qu’en arrivant en stage que j’ai pu découvrir l’étendue des problèmes pratiques que cela peut poser. Isolés, dans des environnements de travail bien conçus, voire seulement à travers une présentation théorique, les cours m’ont permis d’acquérir des connaissances. Cependant, c’est seulement en travaillant sur des projets réels que j’ai pu répondre de façon pratique en piochant dans la boîte à outils de mes cours.
4. **Contribution à l’organisation d’accueil** : Débutant dans le domaine de la cybersécurité, je n’avais pas pour objectif d’apporter une solution exploitable en production. En revanche, il était essentiel pour moi de démontrer ma capacité future à prendre une position d’ingénieur capable d’envisager un problème nouveau et d’y apporter une solution pratique exploitable par une équipe. La création d’un socle d’entraînement pour un modèle de détection de requêtes malveillantes, destiné à être utilisé par des utilisateurs, sert cet objectif.
5. **Exploration des défis et opportunités** : Face à un nouveau problème pour moi, j’ai voulu prendre le temps d’explorer certaines spécificités liées à la cybersécurité et les pratiques MLOps. J’ai donc recherché, à travers le web, dans des livres et des articles, les méthodes classiques en machine learning pour améliorer la sécurité des systèmes d’information ainsi que les façons possibles de les mettre en œuvre de manière robuste et reproductible.

Ces objectifs m’ont permis de m’assurer que ce stage serait bénéfique pour mon parcours comme pour l’Insee. Cependant, s’ils apparaissent rétrospectivement relativement clairs, ils ont pourtant été construits au fur et à mesure de mon avancée comme nous allons maintenant l’explorer.

## 1. Découverte et Contexte

### Premiers pas dans le stage : Enthousiasme et incertitude

- **Sentiments initiaux** : Avant de commencer le stage, j’avais pu discuter avec le RSSI (Responsable Sécurité du Système d’Information) et le DSI (Directeur du Système d’Information) de l’Insee au cours d’une formation interne avant l’oral de la qualification d’analyste. J’étais plutôt impressionné par les enjeux, le SI de l’Insee étant régulièrement la cible d’attaques dont j’entendais parler. Cependant, si je savais que je n’étais pas formé comme un “véritable” informaticien (n.d.), n’ayant reçu qu’une “sensibilisation à l’informatique de production ou à la sécurité” (n.d.) p55, j’avais plutôt confiance dans mes récentes capacités à traiter des données et dans mon envie de découvrir le domaine.
- **Raisons de ces sentiments** : Lors de la scolarité, j’avais pris l’habitude d’explorer les rayonnages de la bibliothèque de l’Ensaï, et j’y avais repéré un livre (n.d.) sur le machine learning et la sécurité. Me remémorant ce dont m’avait parlé mon ancien maître de stage

sur le plan de reprise d'activité de l'application critique Elire, et en sachant qu'un système informatique génère une quantité de données sur lesquelles il est possible de travailler pour mieux en comprendre les rouages, la perspective de travailler sur la sécurité du SI de l'Insee m'intéressait beaucoup. Cependant, je n'avais pas encore d'idée sur l'application concrète d'un tel projet, ce qui me plongeait dans une certaine incertitude.

- **Objectifs personnels initiaux** : C'est pourquoi, j'avais tout d'abord comme objectif de mieux comprendre ce qu'il était possible de faire en sécurité informatique au sein de l'Insee. N'y avait-il pas déjà des outils externes très performants ? Qu'était-il possible d'apporter en tant qu'attaché statisticien débutant ?

## Introduction à la cybersécurité : Définition et importance

- **Définition de la cybersécurité** : La cybersécurité consiste à protéger les systèmes, les réseaux et les programmes contre les attaques numériques. (n.d.) Elle vise à garantir la confidentialité, l'intégrité et la disponibilité des informations. Assurer la sécurité du système d'information (SSI) consiste à gérer les risques de sécurité selon une démarche en trois étapes: lister, évaluer et traiter les risques. (n.d.) La notion de SOC, *Security Operation Center*, devient alors essentiel pour mettre en oeuvre une politique de cybersécurité. UN SOC, doit "monitorer l'ensemble des composants d'un système d'information et être capable de détecter et de sélectionner parmi des milliards d'octets des éléments caractéristiques d'une cyberattaque" (2024).
- **Importance de la cybersécurité** : Une attaque informatique a aujourd'hui d'autant plus de valeur que l'activité des organisations est pratiquement toujours menée à l'aide d'outils informatiques. Sans défendre correctement cet outil, la continuité de l'activité est menacée d'interruption plus ou moins forte. La cybersécurité cherche également à protéger les données sensibles et personnelles, notamment au travers de l'obligation légale issue du Règlement Général sur la Protection des Données (RGPD). La confiance que le grand public accorde à l'Insee serait amoindrie en cas d'attaque réussie. C'est pourquoi la cybersécurité doit faire l'objet d'un soin permanent par l'Insee.
- **Liens avec le machine learning** : La cybersécurité est très naturellement un domaine d'application du machine learning. En effet, on peut y obtenir des jeux de données robustes qui permettront "d'annuler certains des progrès les plus complexes dans la compétence des attaquants". Le machine learning peut ainsi améliorer ou remplacer les "solutions basées sur des règles dans des problèmes comme la détection d'intrusion, la classification des logiciels malveillants ou l'analyse réseau" ((n.d., 5–6) ).

## Présentation de l'équipe SOC et du système d'information (SI)

### Structure de l'équipe SOC

L'équipe SOC de l'Insee a été créée en 2023 et est constituée de 6 membres répartis sur deux sites : Nantes et Metz. Son objectif principal est d'établir et de mettre en œuvre la politique de sécurité du système d'information (SI) de l'Insee. L'équipe est dirigée par un responsable unique qui pilote les travaux de manière transversale entre les deux sites. Les membres de l'équipe incluent des analystes de sécurité, des ingénieurs en cybersécurité, et des experts en gestion des incidents.

## Fonctionnement du SOC

L'équipe SOC a pour mission de surveiller en permanence le SI de l'Insee afin de détecter, qualifier et remédier aux incidents de sécurité. Les principales activités du SOC incluent (n.d.):

- **Veille et qualification des vulnérabilités** : Identifier quotidiennement les vulnérabilités affectant le SI de l'Insee et évaluer leur impact potentiel.
- **Conception de solutions de sécurité** : Développer et mettre en place des solutions technologiques pour renforcer la sécurité du SI, notamment par la mise en place de solutions techniques innovantes pour garantir la sécurité du SI.
- **Détection et gestion des incidents** : Utiliser des outils avancés comme le SIEM (Security Information and Event Management) ou des EDR (Endpoint Detection and Response) pour détecter les activités suspectes et les incidents de sécurité, puis élaborer et mettre en œuvre des plans de remédiation en collaboration avec les équipes concernées.
- **Maintien en condition de sécurité** : Assurer la sécurité continue des différents systèmes d'information en surveillant les infrastructures et en veillant à l'application des correctifs nécessaires.
- **Support et expertise** : Apporter une expertise en sécurité aux différentes unités de l'Insee, conseiller sur les meilleures pratiques et aider à la décision en matière de sécurité informatique.

L'équipe SOC joue un rôle crucial dans la préservation de l'intégrité, de la confidentialité et de la disponibilité des données et des systèmes de l'Insee. Leur travail permet de protéger les actifs numériques de l'organisation contre une variété de menaces cybernétiques.

- **Présentation du SI** : Le système d'information de l'Insee est documenté sur un wiki interne (2024). On peut notamment y trouver des schémas sur l'architecture du SI de production, mais pas sur celle du SI d'administration. L'architecture du SIA est très fortement guidée par les recommandations de l'anssi. (n.d.) Pour l'essentiel, on peut retenir une organisation en 3 couches.

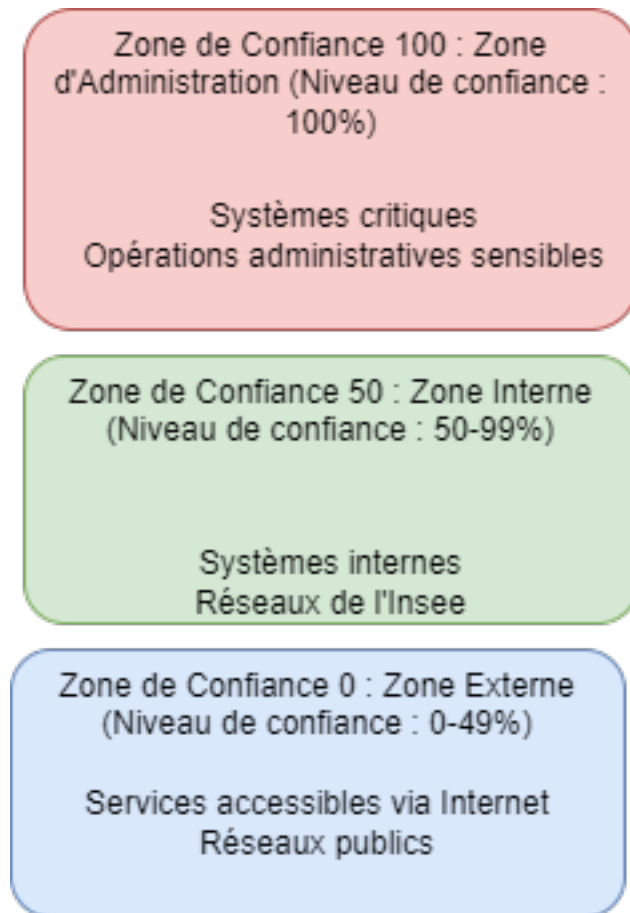


Figure 2: Schéma simplifié du Système d'Information de l'Insee

### Compréhension du SIEM et des logs

- **Rôle du SIEM** : Expliquez ce qu'est un SIEM (Security Information and Event Management) et son rôle dans la cybersécurité.

Un SIEM (Security Information and Event Management) est essentiel dans la conduite d'une politique de sécurité informatique. C'est un outil qui rassemble des données provenant de différentes machines du système d'information, les centralise et effectue des analyses sur elles. Chaque appareil du réseau génère des logs que l'on peut regrouper en deux types :

- **Logs centrés sur l'hôte** : Capturent les événements survenus à l'intérieur de la machine hôte, tels qu'un utilisateur accédant à un fichier ou l'exécution d'un processus.
- **Logs centrés sur le réseau** : Génèrent des logs lorsque les machines communiquent entre elles ou avec l'extérieur, comme le trafic web sur le site [www.insee.fr](http://www.insee.fr), un fichier



récupéré via le protocole FTP, ou encore un utilisateur se connectant au moyen d'un VPN.

L'importance du SIEM provient du fait que le flux des logs est très important, de l'ordre de plusieurs dizaines de Go par jour pour le SI de l'Insee. En cas d'incident, examiner les logs machine par machine est une tâche pratiquement impossible sans SIEM. Ce dernier permet de faire des recherches dans les logs, de corréliser les événements et de répondre rapidement aux incidents.

- **Utilisation des logs** : Décrivez l'importance des logs dans la détection des incidents de sécurité. Log Sources and Log Ingestion Every device in the network generates some kind of log whenever an activity is performed on it, like a user visiting a website, connecting to SSH, logging into his workstation, etc. Some common devices that are found in a network environment are discussed below:

#### Windows Machine

Windows records every event that can be viewed through the Event Viewer utility. It assigns a unique ID to each type of log activity, making it easy for the analyst to examine and keep track of. To view events in a Windows environment, type Event Viewer in the search bar, and it takes you to the tool where different logs are stored and can be viewed, as shown below. These logs from all windows endpoints are forwarded to the SIEM solution for monitoring and better visibility.

#### Linux Workstation

Linux OS stores all the related logs, such as events, errors, warnings, etc. Which are then ingested into SIEM for continuous monitoring. Some of the common locations where Linux store logs are:

```
/var/log/httpd : Contains HTTP Request / Response and error logs.  
/var/log/cron   : Events related to cron jobs are stored in this location.  
/var/log/auth.log and /var/log/secure : Stores authentication related logs.  
/var/log/kern   : This file stores kernel related events.
```

Here is a sample of a cron log:

```
May 28 13:04:20 ebr crond[2843]: /usr/sbin/crond 4.4 dillon's cron daemon, started with  
loglevel notice May 28 13:04:20 ebr crond[2843]: no timestamp found (user root job sys-  
hourly) May 28 13:04:20 ebr crond[2843]: no timestamp found (user root job sys-daily) May 28  
13:04:20 ebr crond[2843]: no timestamp found (user root job sys-weekly) May 28 13:04:20 ebr  
crond[2843]: no timestamp found (user root job sys-monthly) Jun 13 07:46:22 ebr crond[3592]:  
unable to exec /usr/sbin/sendmail: cron output for user root job sys-daily to /dev/null
```

#### Web Server

It is important to keep an eye on all the requests/responses coming in and out of the webserver for any potential web attack attempt. In Linux, common locations to write all apache related logs are /var/log/apache or /var/log/httpd.

Here is an example of Apache Logs:

```
192.168.21.200 - - [21/March/2022:10:17:10 -0300] "GET /cgi-bin/try/ HTTP/1.0" 200 3395
127.0.0.1 - - [21/March/2022:10:22:04 -0300] "GET / HTTP/1.0" 200 2216
```

Log IngestionShows Log Ingestion in SIEM

All these logs provide a wealth of information and can help in identifying security issues. Each SIEM solution has its own way of ingesting the logs. Some common methods used by these SIEM solutions are explained below:

- 1) Agent / Forwarder: These SIEM solutions provide a lightweight tool called an agent (forwarder by Splunk) that gets installed in the Endpoint. It is configured to capture all the important logs and send them to the SIEM server.
- 2) Syslog: Syslog is a widely used protocol to collect data from various systems like web servers, databases, etc., are sent real-time data to the centralized destination.
- 3) Manual Upload: Some SIEM solutions, like Splunk, ELK, etc., allow users to ingest offline data for quick analysis. Once the data is ingested, it is normalized and made available for analysis.
- 4) Port-Forwarding: SIEM solutions can also be configured to listen on a certain port, and then the endpoints forward the data to the SIEM instance on the listening port.

Why SIEM SIEM is used to provide correlation on the collected data to detect threats. Once a threat is detected, or a certain threshold is crossed, an alert is raised. This alert enables the analysts to take suitable actions based on the investigation. SIEM plays an important role in the Cyber Security domain and helps detect and protect against the latest threats in a timely manner. It provides good visibility of what's happening within the network infrastructure. SIEM Capabilities

SIEM is one major component of a Security Operations Center (SOC) ecosystem, as illustrated below. SIEM starts by collecting logs and examining if any event/flow has matched the condition set in the rule or crossed a certain threshold

Some of the common capabilities of SIEM are:

Correlation between events from different log sources.

Provide visibility on both Host-centric and Network-centric activities.

Allow analysts to investigate the latest threats and timely responses.

Hunt for threats that are not detected by the rules in place.

## SOC Analyst Responsibilities

SOC Analysts utilize SIEM solutions in order to have better visibility of what is happening within the network. Some of their responsibilities include:

Monitoring and Investigating.

Identifying False positives.

Tuning Rules which are causing the noise or False positives.

Reporting and Compliance.

Identifying blind spots in the network visibility and covering them.

- **Intégration des logs dans le SIEM** : Mentionnez comment les logs sont intégrés et utilisés par le SIEM pour la surveillance et la détection des menaces.

Analysing Logs and Alerts SIEM tool gets all the security-related logs ingested through agents, port forwarding, etc. Once the logs are ingested, SIEM looks for unwanted behavior or suspicious pattern within the logs with the help of the conditions set in the rules by the analysts. If the condition is met, a rule gets triggered, and the incident is investigated.

## Dashboard

Dashboards are the most important components of any SIEM. SIEM presents the data for analysis after being normalized and ingested. The summary of these analyses is presented in the form of actionable insights with the help of multiple dashboards. Each SIEM solution comes with some default dashboards and provides an option for custom Dashboard creation. Some of the information that can be found in a dashboard are:

Alert Highlights

System Notification

Health Alert

List of Failed Login Attempts

Events Ingested Count

Rules triggered

Top Domains Visited

An example of a Default dashboard in Qradar SIEM is shown below:

## Correlation Rules

Correlation rules play an important role in the timely detection of threats allowing analysts to take action on time. Correlation rules are pretty much logical expressions set to be triggered. A few examples of correlation rules are:

If a User gets 5 failed Login Attempts in 10 seconds - Raise an alert for Multiple Failed Logins  
If login is successful after multiple failed login attempts - Raise an alert for Successful Login After Multiple Failed Attempts  
A rule is set to alert every time a user plugs in a USB (Useful if USB is restricted as per policy)  
If outbound traffic is > 25 MB - Raise an alert to potential Data exfiltration Attempt (Usually during a data breach)

How a correlation rule is created

To explain how the rule works, consider the following Eventlog use cases:

Use-Case 1:

Adversaries tend to remove the logs during the post-exploitation phase to remove their tracks. A unique Event ID 104 is logged every time a user tries to remove or clear event logs. To create a rule based on this activity, we can set the condition as follows:

Rule: If the Log source is WinEventLog AND EventID is 104 - Trigger an alert Event Log Cleared

Use-Case 2: Adversaries use commands like whoami after the exploitation/privilege escalation phase. The following Fields will be helpful to include in the rule.

Log source: Identify the log source capturing the event logs

Event ID: which Event ID is associated with Process Execution activity? In this case, event ID 4688

NewProcessName: which process name will be helpful to include in the rule?

Rule: If Log Source is WinEventLog AND EventCode is 4688, and NewProcessName contains whoami, then Trigger an ALERT WHOAMI command Execution DETECTED

In the previous task, the importance of field-value pairs was discussed. Correlation rules keep an eye on the values of certain fields to get triggered. That is the reason why it is important to have normalized logs ingested.

Alert Investigation

When monitoring SIEM, analysts spend most of their time on dashboards as it displays various key details about the network in a very summarized way. Once an alert is triggered, the events/flows associated with the alert are examined, and the rule is checked to see which conditions are met. Based on the investigation, the analyst determines if it's a True or False positive. Some of the actions that are performed after the analysis are:

Alert is False Alarm. It may require tuning the rule to avoid similar False positives from occurring.

Alert is True Positive. Perform further investigation.

Contact the asset owner to inquire about the activity.

Suspicious activity is confirmed. Isolate the infected host.

Block the suspicious IP.

Let's move on to the next task and explore how SIEM works.

## **2. Recherche et Collecte de Données**

### **Attente de Splunk et recherche de datasets pertinents**

- Types de données : Réseau, logs, HTTP
- Sélection et préparation des datasets



Présentation des premières données obtenues

### **3. Choix des Problèmes et des Outils de Machine Learning**

Définition des problèmes de machine learning en cybersécurité

Introduction à Scikit-learn et choix des modèles

Formation à l'interprétabilité des modèles

### **4. Exploration de Modèles et Outils Avancés**

Essais avec XGBoost pour la classification des malwares

Découverte de MLflow et utilisation de l'API

Introduction au deep learning avec le livre "Deep Learning from Scratch" (DLFS)

Exploration des design patterns en deep learning

### **5. Anomaly Detection et Modèles de Classification**

Détour par la détection d'anomalies : Difficultés et recentrage

Analyses simples avec KNN et clustering

Analyse des données HTTP et détection des attaques (ex. DDOS)

Focalisation sur l'analyse des URL et compréhension des attaques

### **6. Exploration d'Articles Scientifiques**

Lecture et analyse de l'article "Machine Learning for Cybersecurity Applications" de la West Virginia University (WVU)

Étude de "A Comprehensive Review of Anomaly Detection in Web Logs" du Hasso Plattner Institute (HPI)

Analyse de l'article "CRISIS2020\_EasyChair\_PID\_011"

Exploration de "Conf\_SIN2022\_\_SWAF" pour le développement d'un pare-feu applicatif basé sur du machine learning

### **7. Modélisation et Évaluation des Modèles**

Modélisation des URL : KNN, SVM, régression logistique, CNN

Utilisation de ChatGPT et des GPU pour accélérer le développement

Organisation des priorités et gestion des expérimentations

### **8. Travail avec Splunk et Pipelines**

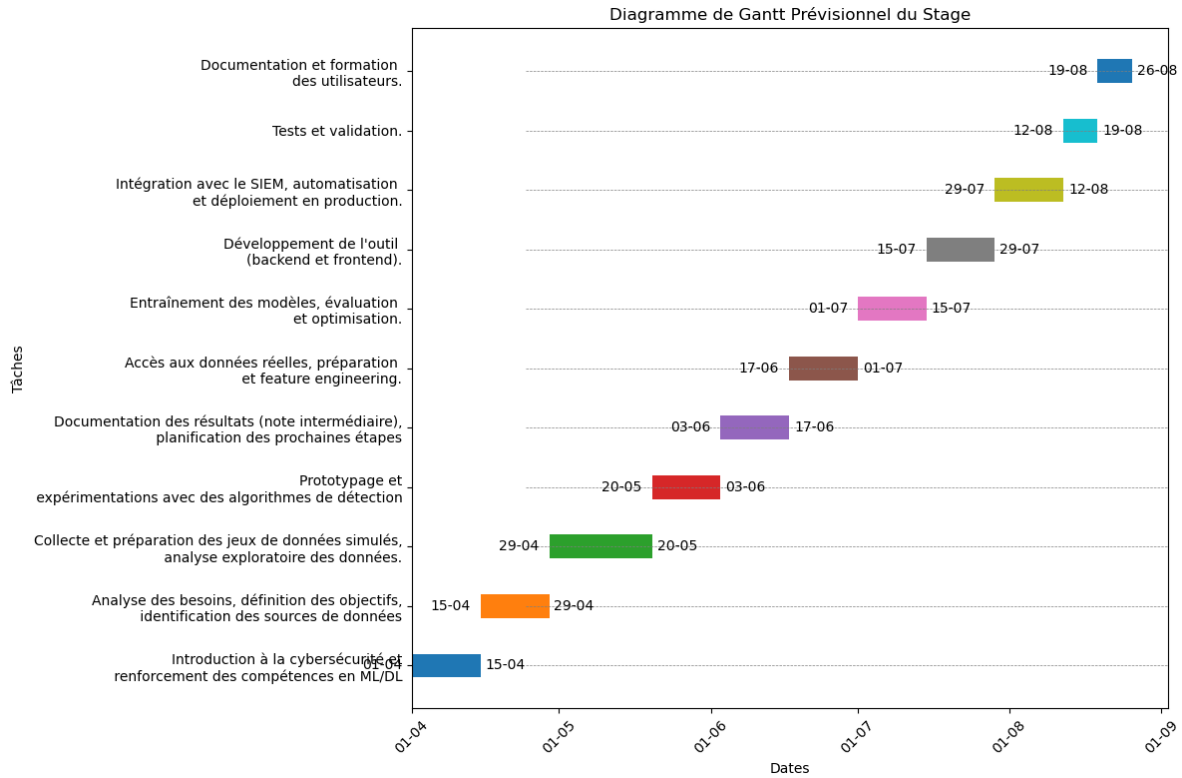


Figure 3: Gant

## Détails techniques des implémentations et configurations

### Bibliographie et références

- “57SNSSI03020Ac - Analyste Du Security Operations Center (SOC).” n.d. *Intranet INSEE*. Accessed August 7, 2024. [https://intranet.insee.fr/jcms/32252877\\_DBFileDocument/fr/57snssi03020ac-analyste-du-security-operations-center-soc?details=true](https://intranet.insee.fr/jcms/32252877_DBFileDocument/fr/57snssi03020ac-analyste-du-security-operations-center-soc?details=true).
- Chio, Clarence, and David Freeman. n.d. “Machine Learning Et Sécurité [Book].” Accessed August 6, 2024. <https://www.oreilly.com/library/view/machine-learning-et/9782412043561/>.
- “Cybersécurité.” 2024. *Wikipédia*. <https://fr.wikipedia.org/w/index.php?title=Cybers%C3%A9curit%C3%A9&oldid=215272230>.
- “Déchiffrer Le Mag 12 (Décembre 2023) - Version Pdf Accessible.” n.d. *Intranet INSEE*. Accessed August 6, 2024. [https://intranet.insee.fr/jcms/29406323\\_DBFileDocument/fr/dechiffrer-le-mag-12-decembre-2023-version-pdf-accessible?details=true](https://intranet.insee.fr/jcms/29406323_DBFileDocument/fr/dechiffrer-le-mag-12-decembre-2023-version-pdf-accessible?details=true).
- “MI-2020-2 Carrieres Informatiques Insee CRCD.” n.d. *Intranet INSEE*. Accessed August 6, 2024. [https://intranet.insee.fr/jcms/58667\\_DBFileDocument/fr/mi-2020-2-carrieres-](https://intranet.insee.fr/jcms/58667_DBFileDocument/fr/mi-2020-2-carrieres-)



- [informatiques-insee-crcd?details=true](#).
- “MI-2020-2 Carrieres Informatiques Insee Rapport.” n.d. *Intranet INSEE*. Accessed August 6, 2024. [https://intranet.insee.fr/jcms/58657\\_DBFileDocument/fr/mi-2020-2-carrieres-informatiques-insee-rapport?details=true](https://intranet.insee.fr/jcms/58657_DBFileDocument/fr/mi-2020-2-carrieres-informatiques-insee-rapport?details=true).
- “Qu’est-Ce Que La Cybersécurité ?” n.d. *Cisco*. Accessed August 6, 2024. [https://www.cisco.com/c/fr\\_fr/products/security/what-is-cybersecurity.html](https://www.cisco.com/c/fr_fr/products/security/what-is-cybersecurity.html).
- “Recommandations Relatives à l’administration Sécurisée Des SI | ANSSI.” n.d. Accessed August 7, 2024. <https://cyber.gouv.fr/publications/recommandations-relatives-ladministration-securisee-des-si>.
- “Reseau · Wiki · Domaine Production Informatique / DPII Pour DSI / Documentation Du SI · GitLab.” 2024. *GitLab*. <https://gitlab.insee.fr/domaine-production-informatique/dpii-pour-dsi/documentation-du-si/-/wikis/Reseau>.