

# Rapport de stage 3A

Mathieu Thomassin

2024-08-31

[Download PDF](#)

## Table des matières

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

## **Introduction**

**Contexte général du stage**

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

**Objectifs du stage**

### **1. Découverte et Contexte**

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

**Introduction à la cybersécurité : Définition et importance**

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

**Compréhension du SIEM et des logs**

### **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)

- Synthèse des méthodes et résultats principaux ### Étude de "A Comprehensive Review of Anomaly Detection in Web Logs" du Hasso Plattner Institute (HPI)
- Principales techniques d'anomaly detection discutées ### Analyse de l'article "CRISIS2020\_\_EasyChair\_PID\_011"
- Points clés et implications pour la cybersécurité ### Exploration de "Conf\_SIN2022\_\_SWAF" pour le développement d'un pare-feu applicatif basé sur du machine learning
- Innovations et applications potentielles

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

**Modélisation des URL : KNN, SVM, regression 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**

**Traitement des données Splunk en préproduction**

**Tokenization des URL et limitations**

**Développement de pipelines et utilisation de GridsearchCV**

**Évaluation des performances des modèles**

## **9. Expérimentations et Déploiement**

**Utilisation de MLFlow pour la gestion des expérimentations**

**Apprentissage à requêter et déployer un modèle via une API**

**Déploiement sur un cluster Kubernetes et gestion du preprocessing**

**Introduction à Metaflow (Netflix) et tests préliminaires**

## **10. Reproductibilité et MLOps**

**Importance de la reproductibilité et utilisation des cours de l'ENSAE**

**Réalisation d'un projet MLOps :**

- Déploiement d'une API sur Kubernetes
- Système d'entraînement et stockage des modèles ### Exploration de Spark pour le calcul et le streaming

## **Conclusion**

**Bilan du stage et accomplissements**

**Perspectives futures : Améliorations et extensions possibles**

**Remerciements et réflexions personnelles**

## **Annexe**

**Ressources supplémentaires : Notebooks, articles, tutoriels, dépôts Github**

**Détails techniques des implémentations et configurations**

**Bibliographie et références**