UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

POTENTIEL DE LA FOUILLE DES DONNÉES EN CYBERSÉCURITÉ

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN SCIENCES ET TECHNOLOGIES DE L'INFORMATION

PAR

JEAN-JARCKE MALASI MUKOMBELWA

NOVEMBRE 2023

UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

Département d'informatique et d'ingénierie

Mémoire intitulé :

POTENTIEL DE LA FOUILLE DES DONNÉES EN CYBERSÉCURITÉ

---

présenté par
Jean-Jarcke MALASI MUKOMBELWA

pour l'obtention du grade de maîtrise ès Science (M.Sc.)

a été évalué par un jury composé des personnes suivantes :

Dr. Rokia Missaoui ..... Directrice de recherche

Dr. Stéphane Gagnon ..... Président du jury

Dr. Raphaël Khoury ..... Membre du jury

Mémoire présenté

Remerciements

Ce mémoire a été réalisé au laboratoire de recherche sur l'information multimédia (LARIM) de l'université de Québec en Outaouais, sous la bienveillante direction du professeur Rokia Missaoui. Je tiens à exprimer ma gratitude envers toutes les personnes qui ont contribué à la réalisation de ce travail.

En premier lieu, je remercie Dr. Rokia Missaoui, professeur au département d'informatique de l'Université de Québec en Outaouais, pour son accueil, sa direction rigoureuse, sa disponibilité, sa patience, et ses précieux conseils.

Un grand merci aux membres du jury, Dr. Stéphane Gagnon, président, et Dr. Raphaël Khoury, membre du jury, pour leur expertise et leurs précieux commentaires.

Mes remerciements vont également aux autorités académiques, en particulier à Karine Dufour, technicienne en administration, gestion départementale du département d'informatique et d'ingénierie, ainsi qu'au corps enseignant de l'université du Québec en Outaouais pour leur accueil chaleureux et leur encadrement.

Je souhaite exprimer ma profonde gratitude envers ma chère épouse, Harmonie Malasi, pour son soutien inébranlable.

Un merci spécial à mon cher père, Dr. Frank Kalulumia Pene Numbi, pour son encouragement constant à rechercher l'excellence.

Enfin, je remercie tous les membres de ma famille, mes amis, mes collègues et tous ceux qui ont contribué, de près ou de loin, à ma formation. Je vous dis merci à tous.

Table des matières

Liste des figures

Liste des abréviations, sigles et acronymes

APT *Advanced Persistent Threat*

DDoS *Distributed denial of service*

DCBD Découverte de connaissances à partir des bases de données

DDoS *Distributed Denial of Service attack*

CID Confidentialité, Integrité et Disponibilité

Résumé

Au cours de ces dernières années et au temps du numérique, les attaques de la cyber-sphère de données sont l'une des plus grandes menaces pour la cybersécurité. Leurs conséquences peuvent être catastrophiques, tant pour les entreprises que pour les particuliers. De tels événements peuvent ruiner des organisations et des vies humaines. Le nombre et l'ampleur de ces attaques ont augmenté au fil des ans. La moindre vulnérabilité peut ouvrir des brèches pour les intrusions dans les systèmes et les réseaux, pour les attaques d'initiés et d'autres menaces pour la sécurité des systèmes.

L'objectif de ce mémoire de maîtrise est d'étudier le potentiel de techniques de fouille de données et d'apprentissage machine dans le domaine de la cybersécurité en vue de mieux connaître les caractéristiques des attaques pour leur détection et leur prévention. Plus précisément, il s'agit de recenser en premier lieu les principales formes et particularités des cyberattaques, d'identifier et d'appliquer un ensemble de techniques de fouille de données dans la cyber-sphère permettant ainsi de détecter et d'anticiper des menaces et des changements dans les failles de sécurité.

Nos travaux d'analyse d'un ensemble d'attaques réelles entre 2005 et 2023 sur la plateforme de science des données *RapidMiner* nous ont permis d'extraire des connaissances intéressantes sur les principaux pays commanditaires d'attaques et leurs stratégies, les attaques dominantes, les types de cibles (gouvernement, secteur privé) ainsi que les périodes de temps où les menaces ou incidents sont les plus observés.

Abstract

In recent years and in the digital age, attacks on the cyber-sphere have become one of the greatest threats to cybersecurity. Their consequences can be catastrophic, both for companies and individuals. Such events can ruin organizations and human lives. The number and scale of these attacks have increased over the years. The slightest vulnerability can open the door to system and network intrusions, to insider attacks and other threats to system security.

The aim of this master's thesis is to study the potential of data mining and machine learning techniques in the field of cybersecurity in order to gain a better understanding of the characteristics of attacks with a view to their detection and prevention. More specifically, the aim is first to identify the main forms and characteristics of cyber-attacks, and then to identify and apply a set of data mining techniques in the cyber-sphere, enabling threats and changes in security flaws to be detected and anticipated.

Our work concerned the analysis of a set of real attacks between 2005 and 2023 on the data science platform RapidMiner and has enabled us to extract interesting insights into the main attack-sponsoring countries and their strategies, the dominant attacks, the types of targets (government, private sector) and the time periods when threats or incidents are most observed [1].

1. La traduction en anglais de ce résumé a été possible grâce au logiciel *DeepL*.

Chapitre 1

Introduction

1.1 Problématique

Les cyberattaques ont connu une augmentation significative au cours des dernières années en raison des avancées technologiques et de la numérisation dans presque tous les domaines de la vie sociale. L'avènement du télétravail, du commerce électronique, de l'infonuagique et d'autres activités en ligne a grandement élargi la surface d'attaque. En effet, les systèmes informatiques présentent de nombreuses vulnérabilités exploitables par des pirates, que ce soit au niveau du réseau, des logiciels ou de l'infrastructure.
Cependant, les cyberattaques peuvent prendre diverses formes, telles que le déni de service (surcharger intentionnellement un système ou un réseau afin d'empêcher la satisfaction des requêtes légitimes, c'est-à-dire rendre indisponible le système en l'inondant de requêtes), l'espionnage (utiliser des logiciels malveillants pour collecter discrètement des données confidentielles sur la victime), la destruction de données (l'application de logiciels malveillants pour effacer ou rendre inutilisables les données d'un système informatique), les rançongiciels (avoir des logiciels malveillants qui bloquent l'accès aux ressources afin d'exiger le paiement d'une rançon à la victime pour la restauration de ses données), et l'hameçonnage (pratique frauduleuse utilisée par les pirates pour soutirer des données confidentielles comme le mot de passe ou l'identité de la victime en lui faisant croire qu'elle interagit avec une institution de confiance comme sa banque par exemple et tant d'autres qui peuvent toujours nuire aux institutions).

Les cyberattaques représentent la plus grande menace pour toute institution et tout individu dans un monde hautement numérisé car elles ciblent tous les secteurs de la société, y compris les gouvernements, les secteurs privés ainsi que les organisations civiles et militaires. Les conséquences des cyberattaques peuvent être graves. Par exemple, les institutions victimes d'attaques de rançongiciels subissent des coûts financiers considérables pour récupérer leurs données et cela affecte également la confiance de leurs clients. De plus, ces attaques peuvent entraîner d'importantes perturbations économiques si elles paralysent des infrastructures critiques telles que les réseaux électriques, etc. La manipulation politique, comme la falsification des résultats de votes peut remettre en cause l'intégrité du processus démocratique. Sur le plan social, les cyberattaques ont généré une perte généralisée de confiance envers les

interactions en ligne, ce qui incite les individus à éviter ou à limiter leurs activités numériques par crainte pour la sécurité de leurs données.

Selon Radio Canada [37], le mercredi 12 avril 2023, le Canada a subi une vague de cyberattaques provenant très probablement de pirates russes. Des banques, des entreprises, des autorités portuaires ainsi que le gouvernement fédéral ont été la cible de ces attaques. La Banque TD a subi des attaques par déni de service qui ont paralysé 503 services en ligne de la plus grande banque du Canada. La Banque Laurentienne a été aussi victime d'attaques similaires. Le gouvernement fédéral du Canada et les autorités portuaires, notamment le port de Montréal et le port de Québec, ont également été visés par une cyberattaque similaire qui a bloqué l'accès à leurs sites Web. Les serveurs d'Hydro-Québec ont également été touchés par cette vague de cyberattaques russes, entraînant des dysfonctionnements et des paralysies de leur site Web.

La priorité actuelle est d'assurer la sécurité des données, des réseaux et des infrastructures essentielles contre les cyberattaques. Cependant, les cyberattaques évoluent constamment et les origines de ces attaques restent souvent inconnues ou hypothétiques. Certains états parrainent les pirates pour mener des cyberattaques. Il est donc nécessaire d'identifier, de détecter et prédire ces attaques dans le futur. Ainsi, la cybersécurité revêt une importance capitale de nos jours car elle englobe l'ensemble des stratégies et des méthodes à mettre en place pour lutter contre les cyberattaques. Cela signifie qu'elle assure la protection des données, des systèmes et des réseaux contre les cybermenaces. La question de l'identification, de la détection et de la prédiction des attaques dans le but d'assurer une meilleure gestion de la cybersécurité est cruciale. Dans quelle mesure peut-on parvenir à remédier à ce défi ? Il existe plusieurs approches et techniques qui ont été proposées dans la littérature pour remédier à cette question. Nous les présentons dans le chapitre 3 relatif à l'état de l'art. En effet, dans le domaine de la cybersécurité, l'utilisation de la fouille de données s'est avérée prometteuse pour l'identification et la détection des cyberattaques, ainsi que pour la prédiction. En combinant ces techniques avec l'apprentissage machine, il est possible de résoudre des problèmes complexes de cybersécurité en mettant en place des mesures de détection et de prévention des attaques.

Des études ont montré l'efficacité de la fouille de données et de l'apprentissage machine pour la détection des intrusions, comme l'ont souligné *Buczak et Guven* [5] dans leur travail portant sur la détection des intrusions dans la cybersécurité.
Un autre exemple de l'utilisation de la fouille de données est l'analyse séquentielle des alertes de sécurité. L'étude de *Husák et al.* [18] présente une recherche sur l'exploitation du potentiel de la fouille de données pour la découverte de motifs séquentiels [1] dans les alertes de cybersécurité.

1.2 Contexte

L'intégration du numérique par les institutions augmente la valeur des données. Qu'il s'agisse du domaine médical, financier, gouvernemental ou de la sécurité, toutes les institutions génèrent, traitent

1. c.-à-d. des règles d'association dont la conclusion se produit quelque temps après la prémisse

et stockent chaque jour des quantités croissantes de données. L'analyse de ces données peut fournir des connaissances utiles pour la prise de décision et pour la planification.

Les cyberattaques produisent d'énormes quantités de données qui doivent été analysées pour aider au développement de solutions de cybersécurité. Dans ce contexte, les techniques de fouille de données deviennent de plus en plus importantes pour la recherche de solutions en cybersécurité et en opérations frauduleuses. Elles permettent d'analyser rapidement de vastes ensembles de données provenant de systèmes de sécurité afin de détecter des incidents et des motifs cachés révélant des problèmes de sécurité et de fraude.
Ces techniques sont essentielles pour identifier diverses menaces, telles que les logiciels malveillants, les intrusions dans les systèmes et les réseaux, ainsi que d'autres types d'attaques [19]. Les techniques de fouille de données et d'apprentissage automatique se sont avérées prometteuses au cours des dernières années en tant qu'outils essentiels dans plusieurs domaines, en particulier dans la lutte contre les cybermenaces et les fraudes.
Maintenant que le contexte est défini, nous allons fixer les objectifs de notre travail dans la sous-section suivante.

1.3 Objectifs

Notre projet de mémoire de maîtrise vise à faire une revue de la littérature pour explorer l'utilisation de techniques de fouille de données dans le domaine de la cybersécurité en vue de détecter et prévenir les attaques dans le cyberespace, telles que les intrusions dans les systèmes et les réseaux, les attaques d'initiés et autres menaces pour la cybersécurité.

L'objectif est triple : tout d'abord, recenser les principales formes de cyberattaques et leur degré d'importance ; ensuite, identifier et appliquer un ensemble de techniques de fouille de données pour identifier, anticiper les attaques et les changements dans les failles de sécurité ; et enfin, mettre l'accent sur la visualisation des données et des connaissances pour faciliter la compréhension et l'interprétation des résultats d'analyse par les décideurs.

Dans ce qui suit, ce mémoire est structuré de la manière suivante : le chapitre 2 aborde les concepts fondamentaux liés à notre étude, tels que la cybersécurité, les cyberattaques, la fouille de données et l'apprentissage machine. Le chapitre 3 présente une revue de la littérature ainsi que les méthodes déjà appliquées. Le chapitre 4 se concentre sur notre démarche d'analyse en décrivant le jeu de données choisi ainsi que l'outil d'analyse que nous avons utilisé, à savoir la plate-forme *RapidMiner* et deux autres outils, et en présentant les résultats obtenus. Enfin, le chapitre 5 conclut notre mémoire en mettant l'accent sur les principales connaissances extraites des données et les travaux futurs.

Chapitre 2

Rappels

Ce chapitre est dédié à quelques rappels sur la cybersécurité, les types de cyberattaques ainsi que sur la fouille de données et l'apprentissage machine.

2.1 Cybersécurité

La cybersécurité est un ensemble de pratiques, technologies, mesures et processus qui consistent à protéger les données sensibles, les réseaux et les systèmes critiques contre les attaques numériques telles que les intrusions non autorisées et les interruptions non requises [9, 50]. Elle a pour but principal de garantir la confidentialité, l'intégrité et la disponibilité des données et des systèmes d'information en utilisant diverses méthodes de défense contre les cyberattaques [57].

La confidentialité se rapporte à la protection des données sensibles : c'est-à-dire, la mise en place des mesures de protection des données afin de s'assurer qu'elles demeurent accessibles seulement aux personnes autorisées à les consulter et ne soient pas détournées [50, 57]. Ces mesures peuvent être plus ou moins strictes conformément au niveau de la sensibilité de données et celui des dommages que pourrait entraîner leur accès par des personnes non autorisées et malveillantes.

L'intégrité consiste à garantir la cohérence, l'exactitude et la fiabilité des données tout au long de leur création jusqu'à leur suppression. Il s'agit de la mise en œuvre des mesures pour s'assurer que les données ne puissent pas subir de modifications non autorisées durant leur transmission [58, 50].

La disponibilité signifie que les utilisateurs autorisés d'accéder aux données puissent le faire chaque fois que c'est nécessaire [24, 50]. Cela passe par une mise en œuvre d'un ensemble de dispositions pour maintenir un environnement d'exploitation en état de fonctionnement.

La triade confidentialité, intégrité et disponibilité (CID) est la base de la cybersécurité parce que lorsqu'il y a violation de données ou lorsqu'une attaque se produit, c'est que l'un ou plusieurs des principes mentionnés ci-dessus sont compromis. Outre ces propriétés de base de la cybersécurité, il existe d'autres caractéristiques que nous citons ci-après :

- **La non-répudiation** préserve l'intégrité et la confiance dans les transactions et les opérations d'un système parce qu'elle utilise des mécanismes tels que la traçabilité des actions, les signa-

tures électroniques ou les journaux d'audit permettant de s'assurer qu'un utilisateur ne peut pas contester les opérations qu'il a effectuées légitimement. Comme exemple, un étudiant ayant soumis son examen sur la plate-forme *Moodle* ne peut pas affirmer plus tard qu'il ne l'a pas fait.
- **L'imputation** empêche un tiers de s'approprier les actions d'un autre utilisateur. Cela évite ainsi de fausses accusations.
- **L'authentification** est un processus essentiel pour identifier les utilisateurs et gérer leurs accès aux espaces de travail appropriés, assurant ainsi la sécurité des systèmes d'information [50].

Il est essentiel de mettre en place une stratégie de sécurité robuste comprenant plusieurs niveaux de protection contre les activités malveillantes en ligne afin d'assurer une cybersécurité efficace. En effet, les mesures de sécurité incluent la prévention des cyberattaques visant à obtenir, modifier ou supprimer des données sans autorisation, à voler de l'argent à une organisation privée ou gouvernementale ou aux particuliers ainsi qu'à perturber le bon fonctionnement des systèmes ou des réseaux.

Dans la section suivante, nous donnons la définition des cyberattaques et plusieurs exemples de ces opérations.

2.2 Cyberattaques

Une cyberattaque désigne le fait qu'un acteur malveillant parvient à exploiter les faiblesses d'un système informatique, de réseaux ou logiciels [57]. Cela peut avoir divers objectifs, allant du vol d'argent, de données financières et de propriété intellectuelle à la simple perturbation des opérations d'une institution, etc. De nos jours, certains pays sont également engagés dans plusieurs cas de cyberattaques connus sous le nom de cyberincidents étatiques. Ces attaques visent à obtenir des informations classifiées sur des adversaires géopolitiques, à transmettre un message spécifique, etc.

Nous nous focalisons dans ce travail sur des cyberincidents parrainés par les États. Dans ce qui suit, nous abordons quelques types de cyberattaques et malgré le fait que la liste n'est pas exhaustive, nous estimons avoir couvert l'essentiel dans le cadre de notre projet.

2.2.1 Logiciels malveillants

Les logiciels malveillants (*Malware*) sont des programmes informatiques conçus pour perturber le fonctionnement normal d'un système ou pour causer des dommages à des données [35]. Le but d'un logiciel malveillant est déterminé par l'intention malveillante, agissant contre les exigences de la victime du système.

Ces types de logiciels sont devenus un instrument à la fois des pirates et des gouvernements pour voler des données personnelles, financières ou commerciales. Ces logiciels peuvent chiffrer ou supprimer des données sensibles, modifier ou détourner des fonctions, espionner l'activité des victimes ou encore gagner de l'argent, mais peut aussi être utilisé à des fins de sabotage ou de motivations politiques.

Il existe différents types de logiciels malveillants tels que les logiciels espionnage *(Spywares)*, les rançongiciels *(ransomwares)* etc.

Les pirates disposent de diverses tactiques pour installer et lancer des logiciels malveillants sur des systèmes informatiques, notamment le téléchargement automatique, la clé USB, l'écriture de code malveillant sur des sites Web et des liens ou l'utilisation de pièces jointes de courriels.

Détaillons un peu quelques-unes des pratiques malveillantes.

2.2.2 Hameçonnage

L'hameçonnage (*phishing*) est une attaque par usurpation d'adresse électronique par laquelle le pirate utilise des méthodes frauduleuses incitant la victime à entrer des données personnelles sur un faux site Web dont l'apparence semble identique à celle du site légitime. En utilisant des courriels, le pirate partage des liens malveillants ou des pièces jointes destinées à exécuter une variété de fonctions, y compris l'extraction de données d'authentification de comptes de la victime [6]. Cela peut entraîner un vol d'identité et des pertes financières.

Il est difficile de distinguer le message ou courriel réussi d'hameçonnage et le message ou le courriel authentique (provenant d'une institution bien connue) puisque le premier comprend des logos, des graphiques et données d'identification recueillies auprès de l'institution piratée [6]. Ces liens malveillants dans les messages donnent l'impression qu'ils mènent vers l'institution usurpée.

2.2.3 Rançongiciel

Le rançongiciel (*ransomware*) est un type d'attaque par logiciel malveillant où le pirate accède aux données cruciales de la cible afin de les chiffrer et d'exiger un paiement (souvent sous forme de cryptomonnaie) pour les déchiffrer [6].

Ce type de logiciel est capable d'occasionner des dommages à l'institution grâce à des techniques telles que l'exploitation des vulnérabilités du système, l'ingénierie sociale, les courriels d'hameçonnage et les campagnes ciblées. L'infection des systèmes se produit lorsque l'utilisateur clique sur un lien, visite une page Web ou installe un fichier, une application ou un logiciel contenant un code malveillant spécialement conçu pour télécharger et installer discrètement le logiciel de rançon ; il se propage à travers le réseau de l'institution à la recherche de cibles de valeur qu'il peut chiffrer [6].

2.2.4 Espionnage

L'espionnage est une attaque par logiciel espion qui collecte des données à l'insu de la victime. C'est un acte visant à obtenir de manière illégale un accès à des données confidentielles, souvent détenues par un gouvernement ou une institution, par l'utilisation de réseaux informatiques sans le consentement de la personne ou de l'entité qui en est propriétaire [26].

Contrairement aux virus, les logiciels espions ne se propagent pas, mais ils sont souvent installés en exploitant les vulnérabilités de sécurité. Ils sont parfois dissimulés et regroupés avec d'autres logiciels installés par les utilisateurs.

2.2.5 Déni de service distribué

Une attaque par déni de service est une cyberattaque dont l'objectif est de rendre un service indisponible, priver ainsi les utilisateurs autorisés d'y accéder. La plupart de ces attaques sont actuellement réalisées à partir de multiples sources, ce qui les qualifie d'attaques par déni de service distribué *(Distributed Denial of Service attack-DDoS)* [35].

Ces attaques peuvent prendre différentes formes, telles que l'inondation d'un réseau afin de le rendre inopérant, la perturbation des connexions entre deux machines pour empêcher l'accès à un service spécifique, le blocage de l'accès à un service pour une personne en particulier, ou encore l'envoi de quantités massives de données à une box Internet [29].
Les conséquences d'une attaque par déni de service sont la paralysie d'un serveur de fichiers, l'impossibilité d'accéder à un serveur web ou encore l'interruption de la distribution des courriers électroniques au sein d'une institution.

2.2.6 Sabotage

Le sabotage consiste à perturber un processus physique comme la distribution de l'électricité ou le fonctionnement normal des centrifugeuses nucléaires en utilisant des logiciels malveillants [27].
Un exemple de cybersabotage est l'attaque *Stuxnet* qui est considérée comme l'une des plus complexes et parfaites jamais réalisée. *Stuxnet* a été découvert en 2010 et était apparemment conçu pour saboter les centrifugeuses nucléaires utilisées dans le programme nucléaire iranien. Le logiciel malveillant ciblait les systèmes de contrôle industriels et exploitait des vulnérabilités spécifiques pour prendre le contrôle des centrifugeuses et les manipuler de manière à les endommager ou à perturber leur fonctionnement [25].

2.2.7 Doxage

Le doxage *(Doxing)* est une action malveillante qui consiste à rechercher et à diffuser sur Internet des informations privées ou d'identification sur une personne ou un groupe dans le but de nuire à leur vie privée ou de les exposer à des risques [35]. Par exemple l'attaque contre *Sony Pictures* en 2014 a été un cas d'attaque doxage où des pirates connus sous le nom de *Guardian of Peace* avaient piraté les systèmes informatiques de l'entreprise et menaçaient de publier des informations sensibles le soir même si l'entreprise n'accédait pas à leurs requêtes [17, 20].

2.2.8 Vol financier

Le vol financier *(Financial Theft)* est une cyberattaque qui consiste à s'emparer illégalement d'actifs tels que de l'argent liquide ou des cryptomonnaies dans le but de réaliser un gain financier [11, 35].

L'exemple le plus courant est la fraude bancaire en ligne : les pirates cherchent par diverses méthodes comment obtenir un accès non autorisé à des comptes bancaires en ligne, des méthodes telles que l'hameçonnage, les logiciels malveillants ou l'ingénierie sociale.

2.2.9 Destruction de données

La destruction de données *(Data Destruction)* est une pratique malveillante qui consiste à utiliser un logiciel malveillant afin de supprimer ou corrompre les données pour causer des dommages importants, ou perturber les opérations ou bien compromettre la continuité des activités d'une institution pour les rendre complètement inutilisables [35].

2.2.10 Défiguration

La dégradation *(Defacement)* est un acte non autorisé qui consiste à modifier l'apparence visuelle ou le contenu d'un site Web ou d'un compte de réseau social sans l'autorisation du propriétaire [35].

C'est le cas par exemple d'un pirate qui parvient à prendre le contrôle d'un site Web appartenant à une institution et accède à son système de gestion de contenu. Il peut alors modifier la page d'accueil du site en remplaçant son contenu par des messages choisis par le pirate, tels que des images choquantes, des messages politiques ou des symboles offensants. Les visiteurs du site seront alors confrontés à cette nouvelle apparence du site, ce qui peut nuire à l'image de l'institution et semer la confusion et le trouble parmi les utilisateurs [51].

2.2.11 Attaque de l'abreuvoir

L'attaque de l'abreuvoir *(Watering hole attack)* est une technique utilisée pour compromettre une cible en insérant un logiciel malveillant sur un site Web que la cible est susceptible de visiter [2]. Supposons qu'il existe un site Web très populaire parmi un certain groupe d'utilisateurs ciblés, tels que des étudiants partageant un intérêt commun (suivre une formation). Les pirates identifient ce site comme une opportunité pour atteindre leurs cibles (ex. des étudiants). Ensuite, ils vont compromettre le site ou infecter ses pages avec du code malveillant. Lorsque les étudiants ciblés visitent le site compromis, leur ordinateur ou leur appareil sera infecté par le code malveillant sans qu'ils s'en rendent compte.

2.2.12 Acteurs

L'acteur de la cyberattaque *(actor)* est un groupe d'individus ou d'états qui, avec des intentions malveillantes, visent à tirer profit de faiblesses d'un système ou d'évolutions technologiques pour obtenir un accès non autorisé à ce système afin d'accéder aux données, dispositifs et réseaux des victimes.

Après avoir exploré les bases de la cybersécurité et des cyberattaques, nous allons maintenant nous pencher sur les principales notions de la fouille de données et de l'apprentissage machine (automatique) et fournir une brève description du processus de découverte de connaissances.

2.3 Fouille de données

2.3.1 Définition

La fouille de données désigne un ensemble de techniques d'analyse de données volumineuses permettant d'extraire des connaissances pertinentes à partir de grandes quantités de données existantes [12]. Son but est d'identifier des motifs, modèles et associations qui ne seraient pas évidents par une simple analyse visuelle.
La fouille de données permet de trouver des corrélations et associations entre les variables, de prédire des résultats futurs et de détecter des anomalies. Les techniques de fouille de données incluent la classification, le regroupement, la régression et les règles d'association, souvent combinées avec des outils de visualisation de données pour faciliter la découverte de modèles et de règles [14].

Dans le domaine de la cybersécurité, ces techniques ont été utilisées pour détecter les intrusions, les activités malveillantes, les anomalies et les attaques de logiciels malveillants. Elles contribuent également à l'analyse des journaux de sécurité et à la prévention des cyberattaques [5].
Les techniques de fouille de données sont largement utilisées dans divers domaines où il existe de grandes quantités de données à stocker et à traiter. Par exemple, dans le domaine bancaire, ces techniques sont employées pour détecter la fraude de cartes de crédit, évaluer les risques de crédit et prédire les tendances du marché financier [14]. Dans le domaine médical, elles sont utilisées pour prédire des diagnostics, repérer des motifs dans les données cliniques, prédire des résultats cliniques, identifier des facteurs de risque et surveiller l'efficacité des traitements [14].

La gestion de la relation client bénéficie également de ces techniques, permettant de prédire les comportements des clients, d'anticiper leurs achats futurs, de segmenter les clients en fonction de leurs préférences, etc. Ces applications aident les entreprises à améliorer leurs performances de vente en réduisant les coûts et les risques associés, tout en renforçant les stratégies de fidélisation de la clientèle [14].
Dans ce qui suit, nous présentons le processus de découverte de connaissances.

2.3.2 Processus de découverte de connaissances

Le processus d'identification de motifs à partir des données est une tâche complexe qui vise à découvrir des motifs valides, nouveaux et potentiellement utiles. Ces motifs doivent permettre de prendre des décisions éclairées et doivent être compréhensibles par les humains. Les données utilisées en entrée de ce processus présentent une diversité inhérente. Elles peuvent être de nature numérique, textuelle, symbolique, temporelle ou provenant de diverses sources. De plus, elles peuvent varier en termes de structure, pouvant être ensemblistes, arborescentes, séquentielles, sous forme de graphes ou de textes, etc. Les données peuvent également être incomplètes et contenir des erreurs. Elles peuvent être dynamiques, évoluer dans le temps, ou être reçues sous forme de flux continus.

Les objectifs de ce processus peuvent revêtir différentes formes [34]. Par exemple, l'utilisateur peut chercher à résoudre un problème de classification pour prédire le pays commanditaire ou le type d'une cyberattaque. Cela rend ce processus très complexe et les étapes peuvent varier considérablement en fonction de la nature des données et des objectifs de l'application [34]. La figure 2.1 tirée de notes de cours [30] montre clairement le cycle de développement du processus de découverte de connaissances.

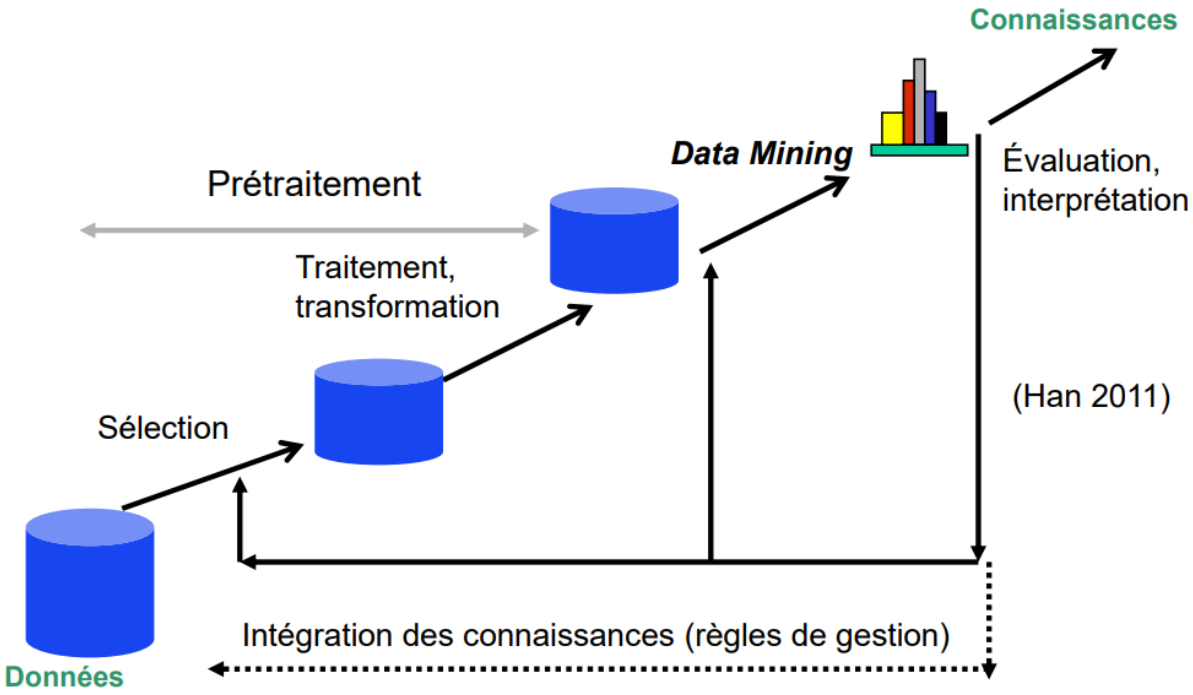


FIGURE 2.1 – Processus de découverte de connaissances.

Il découle de cette illustration que les processus de découverte de connaissances à partir de données couvrent quatre étapes à savoir :

Prétraitement

Les données identifiées sont souvent incomplètes, contiennent du bruit ou des attributs superflus, présentent une qualité hétérogène ou ne sont pas dans le format d'entrée requis par les algorithmes de fouille de données. Par conséquent, il est nécessaire de procéder à une sélection d'instances et d'attributs, un nettoyage et à une mise en forme des données afin de les rendre compatibles avec les techniques de fouille de données [14]. En d'autres termes, il est essentiel de garantir la cohérence des données, de réduire la redondance et l'ambiguïté, d'incorporer de nouvelles données, d'intégrer des connaissances a priori, de gérer les valeurs manquantes, ainsi que d'appliquer des techniques de recodification et d'agrégation des données.

La sélection de données doit être liée aux objectifs d'identification de nouvelles connaissances. Il est nécessaire d'abord d'identifier les sources de données (un ensemble de données) qui pourraient

être utiles ainsi que les attributs et les enregistrements pertinents dans la construction des nouvelles connaissances [14]. Dans le cadre de ce projet de mémoire, nous avons par exemple choisi un ensemble réel de cyberattaques produites entre 2005 et 2023 et avons sélectionné cinq attributs, à savoir le type d'attaque, la date, la catégorie, le pays commanditaire et la description textuelle de l'attaque.

Étape de fouille de données

Le choix de l'algorithme de fouille de données dépend du type des données et des objectifs à atteindre. Une fois que les données ont été sélectionnées et prétraitées, elles sont analysées à l'aide d'un ou de plusieurs algorithmes pour effectuer des tâches telles que la classification ou le regroupement ou la production de motifs comme les règles d'association ou les cas aberrants. [14]. Bien que la fouille de données ne représente qu'une étape du processus global de découverte de connaissances, elle est celle qui suscite le plus d'attention dans la littérature [12, 14, 44].

Interprétation et évaluation des résultats obtenus

Les connaissances extraites ne sont habituellement pas directement compréhensibles. Cette étape vise à manipuler le format de sortie des algorithmes afin de présenter les résultats de manière accessible (la visualisation) et facilement interprétable par les utilisateurs. Cela permet de déterminer la pertinence des motifs et des modèles identifiés, ainsi que leur adéquation par rapport aux objectifs de l'analyse des données [14].

Maintenant que nous avons passé en revue quelques concepts liés aux processus de découverte de connaissances, nous allons ci-après rappeler les principales catégories de fouille de données.

Principales formes de fouille de données

Il existe trois principales formes de fouille de données qui sont la prédiction, la découverte et la détection de déviations :

- **La prédiction** implique l'utilisation de techniques d'apprentissage supervisé pour prédire la valeur d'une variable cible (régression) ou la classe d'appartenance (classification). Les différentes méthodes telles que les réseaux bayésiens, les arbres de décision et les réseaux de neurones sont utilisées dans ce contexte. Par exemple, l'identification des cyberattaques ou des anomalies peut être réalisée en appliquant des techniques de prédiction [15].
- **La découverte** est une méthode d'analyse de données qui utilise des techniques d'apprentissage non supervisé et d'analyse exploratoire pour identifier des structures cachées, des modèles ou des associations au sein des données. Il n'y a pas d'étiquette à prédire dans ce contexte. L'objectif est d'extraire des connaissances à partir des données sans nécessairement avoir d'idées préconçues quant aux résultats attendus [34].
- **La détection de déviations** consiste à identifier des valeurs exceptionnelles *(outliers)* ainsi qu'à analyser les tendances qui se dégagent des données [5].

Dans ce qui précède, nous avons rappelé les notions liées à la fouille de données et au processus de découverte de connaissances. Dans ce qui suit, nous abordons l'apprentissage machine et ses deux principales formes.

2.4 Apprentissage machine

L'apprentissage machine est une branche de l'intelligence artificielle qui utilise des algorithmes et des techniques pour développer des solutions capables d'apprendre à partir de données et d'observations. En utilisant ces solutions, il est possible de prédire des résultats futurs ou d'extraire des modèles grâce à l'analyse de données complexes. Parmi ces techniques, on distingue la classification et le regroupement [43].

Dans notre contexte, l'apprentissage machine permet de détecter les comportements anormaux, de catégoriser les attaques selon leurs attributs et de prédire les menaces futures afin de renforcer la défense et assurer une meilleure protection contre les cyberattaques [33]. Il se divise généralement en deux catégories principales qui sont l'apprentissage supervisé et l'apprentissage non supervisé.

2.4.1 Apprentissage supervisé

L'apprentissage supervisé consiste à entraîner un modèle à partir de données étiquetées d'un ensemble d'apprentissage (*training set*) où chaque exemple de données est associé à une étiquette ou une classe prédéfinie [19]. Le modèle utilise ces exemples étiquetés pour apprendre à effectuer des prédictions ou à prendre des décisions similaires sur un ensemble de test (*validation/test set*).

Par exemple, dans un système de détection d'intrusion, le modèle peut être entraîné à partir d'un ensemble de données où chaque attaque est étiquetée comme malveillante avec divers degrés de gravité. Le modèle apprend à partir de ces exemples étiquetés afin de pouvoir reconnaître et signaler de nouvelles tentatives d'attaques malveillantes violentes.

L'apprentissage supervisé dans le domaine des cyberattaques permet ainsi de développer des modèles de détection efficaces, capables d'identifier différents types d'attaques et de les classer en fonction de leur degré gravité ou de leur impact potentiel [47].

2.4.2 Apprentissage non supervisé

L'apprentissage non supervisé, quant à lui, se concentre sur l'analyse de données non étiquetées de manière autonome [14]. Cela signifie que ces algorithmes explorent le modèle des données sans aucun guide préalable et cherchent à découvrir des structures, des groupes ou des similarités inhérentes aux données. Par exemple, dans le domaine de la cybersécurité, les techniques d'apprentissage non supervisé peuvent être utilisées pour trouver des groupes homogènes d'attaques et d'éventuels cas exceptionnels.

2.4.3 Autres formes d'apprentissage machine

En plus de ces deux catégories principales, il existe d'autres formes d'apprentissage machine, telles que **l'apprentissage semi-supervisé**, qui combine des données étiquetées et non étiquetées pour améliorer les performances du modèle [19, 33], et **l'apprentissage par renforcement**, où le modèle apprend à travers l'interaction avec un environnement en recevant des récompenses ou des sanctions en fonction de ses actions [22].

Dans le cadre de ce projet de mémoire, notre attention sera portée sur deux aspects clés de l'apprentissage machine : l'apprentissage supervisé et l'apprentissage non supervisé. Dans la prochaine sous-section, nous détaillons quelques techniques que nous avons utilisées.

2.5 Techniques de fouille de données et d'apprentissage machine

Les techniques de fouille de données complètent celles de l'apprentissage machine et sont souvent utilisées conjointement dans de nombreux domaines, y compris la cybersécurité. En effet, l'apprentissage machine offre des méthodes et des algorithmes puissants pour explorer et analyser les données, tandis que la fouille de données fournit des outils permettant de découvrir des motifs et des relations utiles dans des données généralement volumineuses. Ensemble, ces deux technologies permettent d'extraire des connaissances précieuses, de faire des prédictions pour une prise de décisions éclairées fondées sur les données.

Ce chapitre commence par présenter les différentes techniques d'exploration de données et d'apprentissage automatique qui peuvent être appliquées à diverses tâches, telles que la classification, le regroupement de données et la génération de règles d'association. Ensuite, il aborde les concepts de base de l'analyse de texte.

2.5.1 Classification

La classification est une technique à la fois de fouille de données et d'apprentissage automatique la plus fréquemment utilisée. Elle est basée sur l'apprentissage supervisé et vise à prédire la classe d'appartenance d'une nouvelle instance [14]. Elle consiste en un processus divisé en deux étapes distinctes. La première étape est l'apprentissage, où un algorithme de classification utilise un ensemble d'apprentissage contenant des données préalablement affectées à des classes connues. L'algorithme apprend à partir de cet ensemble d'apprentissage et construit un modèle. Dans la deuxième étape, appelée test, le modèle ainsi créé est utilisé pour classer de nouvelles instances en leur attribuant la classe qui leur convient [14].

La classification est une forme de modélisation prédictive qui utilise une fonction f pour associer des attributs d'entrée à des attributs de sortie qui représentent des étiquettes [14]. Cette technique a été appliquée aux données structurées et non structurées (comme les images) afin de prédire la classe d'appartenance de nouvelles instances [19].

Il existe une grande variété d'algorithmes de classification [19, 56], tels que les arbres de décision, les réseaux bayésiens, la régression logistique, les machines à vecteurs de support (SVM), les réseaux de neurones artificiels, les forêts aléatoires, la méthode de k-plus proches voisins (k-NN), et bien d'autres procédures.

Arbre de décision

La production d'arbres de décision est une technique d'apprentissage supervisé utilisée pour explorer diverses solutions possibles d'une décision. Le résultat est une structure arborescente où chaque nœud interne (nœud non-feuille) représente un test sur un attribut, chaque branche représente le résultat du test, et chaque nœud feuille (ou nœud terminal) contient une étiquette de classe. L'attribut le plus d'influent sur la décision de l'arbre se trouve dans le nœud racine [14].

L'arbre de décision est construit à partir d'un ensemble de données comprenant des valeurs d'attributs cibles connues et permet d'extraire des règles dites de classification. Les résultats obtenus sont ensuite généralisés à l'ensemble des données de test. Les arbres de décision sont largement reconnus comme étant l'une des structures de classification les plus populaires.

Réseaux bayésiens

Les réseaux bayésiens *(Naive Bayes)* offrent une classification basée sur le théorème de Bayes avec l'hypothèse d'indépendance entre les attributs [8]. L'algorithme calcule la probabilité qu'une instance de données appartienne à chaque classe et affecte la classe avec la probabilité la plus élevée comme classe prédite. Cette technique est connue pour sa simplicité, son évolutivité et son efficacité dans le traitement de données volumineuses. Malgré son hypothèse naïve, elle fonctionne souvent bien dans la pratique et sert de modèle de référence pour de nombreuses tâches de classification notamment l'analyse de texte ou encore la catégorisation de documents [55, 53].

Machine à vecteurs de support

La machine à vecteurs de support (MVS) ou *Support Vector Machine* est un algorithme de classification qui vise à trouver un hyperplan optimal pour séparer les données en différentes classes. En maximisant la marge et les distances entre les hyperplans, la précision de la classification des points de données peut être améliorée. Les points de données situés sur la limite de l'hyperplan sont appelés points de vecteur de support [1, 19].

La technique de MVS est utilisée pour résoudre des problèmes de classification linéaire et non linéaire. De plus, elle peut être utilisée pour la détection d'une ou plusieurs classes selon les exigences spécifiques de l'application. Cependant, il est important de noter qu'elle nécessite des ressources importantes en mémoire et en temps de traitement. Les performances du classificateur sont également influencées par la fonction et les paramètres du noyau [14].

2.5.2 Regroupement

Le regroupement *(clustering)* est une technique d'apprentissage non supervisé qui aide à regrouper des ensembles de données en fonction de leur similarité. L'objectif est de former des groupes *(clusters)*, c'est-à-dire des ensembles d'enregistrements qui sont similaires entre eux et distincts des enregistrements des autres groupes.

La principale différence entre le regroupement et la classification réside dans le fait que le regroupement n'implique pas d'attributs étiquetés. Contrairement à la classification, une tâche de regroupement ne cherche pas à prédire ou à estimer la valeur d'attributs. Au lieu de cela, les algorithmes de regroupement visent à diviser l'ensemble des données en sous-ensembles relativement homogènes en maximisant l'homogénéité à l'intérieur de chaque groupe et en minimisant les similarités entre les différents groupes.

Il existe quatre principales méthodes de regroupement. Il s'agit de méthodes de (i) partitionnement, (ii) hiérarchiques, (iii) basées sur la densité et (iv) basées sur la grille. Ces méthodes offrent différentes approches pour former des groupes en fonction des caractéristiques des données et des objectifs spécifiques de regroupement [14].

Algorithme k-moyennes

L'algorithme de regroupement k-moyennes permet d'identifier k groupes distincts et d'assigner chaque élément à un groupe particulier [49]. Les groupes sont composés d'éléments qui présentent des similitudes entre eux. La mesure de similarité entre les éléments est déterminée par une distance spécifique entre eux [41].

Il est largement utilisé comme méthode principale pour le regroupement de données [49, 44]. Il commence par initialiser les centroïdes (points d'équilibre) en utilisant des données aléatoires provenant de l'ensemble de données, puis il procède en deux étapes pour améliorer les centroïdes et les groupes correspondants.

- Regrouper chaque élément en fonction de sa proximité avec le centroïde le plus proche.
- Replacer chaque centroïde vers la moyenne décrivant son groupe respectif.

Après plusieurs itérations, l'algorithme trouve des partitions stables de l'ensemble de données, indiquant que l'algorithme a convergé.

2.5.3 Matrice de corrélation

Une matrice de corrélation est tout simplement un tableau qui présente une relation linéaire entre des paires d'attributs aléatoires, généralement désignées par A et B. La corrélation est une mesure qui fournit le degré (la force et la direction) d'une relation entre deux attributs. Elle est représentée par un nombre compris entre -1 et 1, appelé coefficient de corrélation. Lorsque ce nombre est égal à -1, il indique une relation inverse (négative) entre les attributs. Cela signifie que si l'attribut A augmente, l'attribut B tend à diminuer, et vice versa. Un nombre égal à 0 indique l'absence de corrélation (et donc

indépendance) entre les attributs A et B, ce qui signifie que les changements de l'attribut A n'ont pas d'effets sur B et inversement. Un nombre égal à 1 indique une relation positive, où les deux attributs A et B ont tendance à augmenter ou à diminuer ensemble.

Une corrélation est dite positivement forte, si elle est proche de 1, mais si elle est proche de -1, il s'agit d'une corrélation négativement étroite entre les attributs, tandis qu'une corrélation proche de 0 est dite faible [48].

La corrélation est souvent utilisée dans l'analyse des données pour identifier les attributs qui sont étroitement liés et peuvent avoir un impact les uns sur les autres. Elle permet également de détecter les relations linéaires entre les attributs et d'explorer les dépendances entre les attributs [14]. À titre d'exemple, il peut exister une corrélation positive entre le nombre de cyberattaques et les ventes des logiciels de sécurité au cours d'un mois.

Maintenant que nous avons présenté des notions sur le regroupement et sur la méthode de k-moyennes ainsi que la matrice de corrélation, nous allons aborder dans la sous-section suivante quelques concepts relatifs aux règles d'association.

2.5.4 Règles d'association

L'analyse des règles d'association permet de découvrir des modèles et des associations cachées entre différents attributs. Une règle d'association est une déclaration de la forme $X \rightarrow Y$, où X et Y représentent des ensembles d'attributs. Elle indique qu'il existe une association entre X et Y. Cela signifie que la présence de certaines caractéristiques de X (antécédent) est souvent liée à la présence d'autres caractéristiques de Y (conclusion) [14].

La robustesse d'une règle d'association est déterminée par trois mesures importantes, à savoir : le support, la confiance et l'intérêt *(lift)*.

- **Le support** représente la fréquence ou la proportion d'observations qui contiennent à la fois les caractéristiques de X et les caractéristiques de Y. Un support élevé indique que la règle est fréquente.
- **La confiance** mesure la probabilité d'observer Y lorsque X se produit. Une confiance élevée indique que la règle est pertinente et qu'elle fournit une inférence fiable. Cette mesure peut induire en erreur car elle ne représente pas une relation de cause à effet
- **L'intérêt** est la confiance divisée par la probabilité de la conclusion. Il mesure le degré de corrélation ou de dépendance entre X et Y. Un intérêt positif (respectivement négatif) indique une corrélation positive (respectivement négative) entre X et Y alors qu'une valeur nulle indique une indépendance entre X et Y.

Plusieurs algorithmes d'extraction des règles d'association ont été utilisés dans la littérature, notamment l'algorithme Apriori et l'algorithme FP-growth *(frequent pattern growth)* [19].

Algorithme Apriori

L'algorithme *Apriori*, développé par Agrawal et Srikant en 1994 [13], fonctionne en deux étapes et est généralement basé sur la création d'ensembles d'items *(itemsets)* fréquents et de règles d'association. Il utilise des opérations de jointure et d'élagage de manière itérative jusqu'à identifier de tous les ensembles de motifs les plus fréquents. Ainsi, pour évaluer la fréquence d'apparition des motifs, un seul minimum du support doit être défini.

Tout d'abord, l'algorithme compte le nombre d'occurrences de chaque élément(ex. un produit, un type d'attaque). Dans un premier temps, les *itemsets* à un seul élément dont l'occurrence dépasse le seuil minimum de support sont identifiés. Ensuite, seuls les candidats ayant un support supérieur ou égal au seuil fixé sont conservés pour l'itération suivante. Par la suite, lors de l'étape à laquelle l'algorithme a retenu des *itemsets* de k éléments, une opération de jointure est effectuée pour générer des motifs de $k+1$ éléments. Après cette étape, l'algorithme effectue une opération d'élagage pour supprimer les motifs ayant un support inférieur au seuil défini, ne conservant ainsi que les motifs fréquents. Ces itérations se poursuivent jusqu'à ce qu'aucun nouvel *itemset* ne puisse être généré. Pour chaque *itemset* fréquent S, l'algorithme Apriori crée autant de règles d'association qu'il existe de sous-ensembles (autre que l'ensemble vide) $X \subset S$ sous la forme $\mathrm{X} \rightarrow \mathrm{S} \setminus \mathrm{X}$.

Vers les années 2000, *Han et Pei* [36] ont proposé une amélioration de l'algorithme *Apriori* par un nouvel algorithme appelé algorithme *FP-Growth* pour la production de motifs fréquents permettant la génération efficace de règles d'association. Ils ont introduit une structure de données appelée arbre FP *(FP-tree)* avec FP pour *Frequent pattern*, qui aide à simplifier et à accélérer le processus d'extraction des motifs fréquents ainsi qu'à réduire la redondance dans les règles.

Ainsi, grâce à cette méthode, l'algorithme a gagné en temps de traitement et en espace mémoire du fait que le nombre de consultations de l'ensemble de données est de deux seulement. Il est donc plus efficace que l'algorithme *Apriori (Zhang et al* [59]).

En plus de la fouille des données numériques, il existe une demande croissante d'analyse de documents textuels qui conduit à accroitre l'importance d'un autre domaine similaire appelé la fouille de texte. En effet, avec l'expansion d'Internet, du Web, des réseaux sociaux, et d'autres sources de données, une quantité croissante de données textuelles est produite chaque jour nécessitant une attention particulière et une analyse parfois complexe.

Dans la sous-section suivante, nous abordons quelques notions sur ce sujet.

2.5.5 Fouille de textes

La fouille de texte, faisant partie de la fouille de données et de l'intelligence artificielle, est un ensemble de techniques visant à découvrir des connaissances à partir de données textuelles (corpus). Elle utilise des méthodes analytiques de fouille de données et d'apprentissage automatique telles que la classification et le regroupement et des algorithmes d'apprentissage en profondeur pour explorer et mettre en évidence des relations cachées dans les données non structurées [28, 14].

Un corpus est un ensemble de documents similaires (documents liés entre eux soit par période ou par entité). Par exemple, la liste des cyberattaques du mois de mai ou tous les avis sur un produit acheté ou un service reçu. Comparativement à une table dans une base de données, un corpus peut être défini en fonction du besoin.

La fouille de textes est un domaine multidisciplinaire parce qu'elle intègre des outils issus de divers domaines tels que la recherche d'informations, les statistiques, la linguistique informatique, la sémantique et le traitement du langage naturel. C'est en effet un domaine capital en pleine croissance à cause de la quantité de données textuelles produites en plus grand nombre chaque jour par diverses sources telles que les réseaux sociaux, le Web, des appareils spécialisés, etc.

Le processus d'extraction de connaissances à partir de corpus de texte comprend deux aspects essentiels : la recherche d'informations et le traitement automatique de la langue.

Le prétraitement de texte

La préparation de texte consiste à nettoyer, normaliser et préparer les données textuelles afin de les rendre appropriées pour l'analyse et la fouille de texte [28]. Citons quelques étapes courantes de prétraitement de texte avant son analyse :

- **Le nettoyage des données** : Il s'agit de supprimer les caractères indésirables, la ponctuation, les chiffres et autres éléments non pertinents du texte. Cela peut être réalisé en utilisant des opérations telles que la correction des erreurs d'encodage, la suppression des caractères spéciaux, des balises HTML, etc.
- **La tokenisation** est une étape qui consiste à diviser le texte en unités significatives appelées « tokens ». Les jetons peuvent être des mots, des phrases ou même des caractères individuels.
- **Suppression des mots vides** : Il s'agit d'éliminer des mots courants qui n'apportent pas beaucoup d'informations dans l'analyse textuelle, tels que « la », « et », « à », etc. Ils sont généralement supprimés pour réduire le bruit et la taille des données.
- **La normalisation** vise à mettre les mots dans une forme normalisée. Cela peut inclure la conversion en minuscules ou en majuscules, la réduction des mots à leur forme racine *(stemming)* ou encore la lemmatisation (obtention des formes lexicalement corrigées des mots).
- **Le traitement des entités nommées**. Dans certains cas, il peut être nécessaire de reconnaître et de traiter les entités nommées telles que les noms de personnes, d'organisations, de lieux, etc.
- **Le traitement des synonymes** : la résolution des synonymes peut être effectuée pour regrouper les mots ayant des significations similaires. Par exemple, « acteur » et « comédien » peuvent être regroupés en tant que synonymes, ce qui simplifie l'analyse.

Ces étapes de prétraitement avant la fouille de texte peuvent varier en fonction des besoins spécifiques de l'analyse ou du projet en cours. L'objectif principal est de nettoyer, normaliser et structurer les données textuelles afin de faciliter l'extraction d'informations et l'analyse ultérieure.

Algorithme TF-IDF

Le TF-IDF *(term frequency-inverse document frequency)* est une méthode de pondération largement utilisée en fouille de texte dans la recherche d'information pour évaluer l'importance d'un terme dans un document par rapport à un corpus D de documents [54].

Le poids d'un terme dans un document est calculé en fonction de deux mesures : la fréquence TF du terme t dans le document d et IDF qui est l'inverse de la fréquence du terme dans le corpus.

La formule générale pour le calcul du TF-IDF par rapport au terme t dans le document d est la suivante :

TF-IDF$(t, d) = TF(t, d) \times IDF(t)$.

Ainsi, les termes qui sont fréquents dans le document mais rares dans le corpus auront un poids élevé, ce qui les rendra plus importants dans l'analyse.

La fréquence $TF(t, d)$ du terme t dans le document d est le rapport entre le nombre d'apparitions du terme t dans le document d et le nombre total de mots dans ce même document. Ainsi, plus un terme apparaît fréquemment dans un document, plus son poids TF sera élevé.

La fréquence inverse de document (IDF) *(inverse document frequency)* mesure la rareté du terme dans l'ensemble du corpus. Les termes qui apparaissent rarement dans le corpus ont une valeur IDF plus élevée, ce qui augmentera leur poids et signifie qu'ils sont considérés comme plus importants dans la pondération TF-IDF [54].

Cette mesure est calculée en prenant le logarithme de l'inverse de la proportion de documents du corpus D qui contiennent le terme t : $IDF(t) = \log \frac{|D|}{|\{d_j \in D : t \in d_j\}|}$.

Dans le contexte de la recherche d'information, le TF-IDF est souvent utilisé pour ordonner les documents par ordre de pertinence. Dans les modèles vectoriels, chaque document et la requête sont représentés par des vecteurs de poids TF-IDF. La similarité entre la requête et chaque document peut ensuite être calculée à l'aide de la distance cosinus entre les vecteurs [21].

Chapitre 3

État de l'art

Les cyberattaques ont généré une quantité importante de données qu'il est crucial de prendre en considération pour détecter les vulnérabilités et les menaces. Pour les chercheurs en cybersécurité, la fouille de données est un outil précieux qui leur permet de collecter et d'analyser ces données massives afin de repérer les menaces, d'identifier les modèles de comportement des attaquants, de détecter les signaux faibles et de découvrir de nouvelles menaces.

De nombreux auteurs ont utilisé des techniques de fouille de données dans le domaine de la cybersécurité, ce qui en fait un domaine de recherche important et dynamique.

Husák et al. [18] examinent l'utilisation de la fouille de données dans le domaine de la cybersécurité, en particulier de la fouille de motifs séquentiels (c'est-à-dire des règles dans lesquelles Y se produit plus tard après X dans la règle $X \rightarrow Y$) et de règles dans l'analyse des alertes de sécurité. Ils soulignent que bien que la fouille de données soit capable d'extraire des motifs cachés dans les données, son potentiel n'est pas exploité dans la communauté de la cybersécurité. Ils présentent une étude de cas sur l'utilisation de la fouille de données pour étudier la corrélation entre les alertes et la prédiction des attaques. Différentes méthodes de fouille de motifs séquentiels et de règles d'association sont appliquées afin de trouver celle qui est à la fois rapide et qui fournit des résultats pertinents pour les alertes de sécurité [18]. Une expérience est réalisée avec des données réelles provenant d'une plateforme de partage d'alertes, et les résultats sont analysés et comparés pour tirer des conclusions et des recommandations. Ils soulignent également l'importance de l'exploration de motifs séquentiels pour détecter des attaques complexes et avancées comme les attaques en plusieurs étapes ou les menaces persistantes [18]. Les optimisations à considérer lors de l'utilisation de la fouille de motifs séquentiels dans le domaine de la cybersécurité sont évoquées, notamment en raison de la variabilité et de l'incomplétude des données. Les auteurs recommandent le calcul de règles d'association séquentielles pour l'analyse de la corrélation entre les alertes et la prédiction d'attaques puisque les résultats sont plus complets et significatifs qu'avec l'application d'algorithmes de production de simples règles d'association. Ils soulignent également l'importance de la sélection des caractéristiques, de la gestion des séquences longues et des optimisations spécifiques pour améliorer les performances et la pertinence des résultats. Enfin, ils mettent en garde contre les défis liés à la formalisation de la validité des résultats et recommandent d'approcher ce problème avec une bonne connaissance du domaine d'application, en gardant à l'esprit les spécificités de la cybersécurité.

D'autres auteurs appliquent à la fois la fouille de données et l'apprentissage automatique en cybersécurité car ces approches fournissent des outils puissants pour identifier et prédire des attaques agressives et des comportements indésirables.

Buczak et Guven [5] présentent une revue de la littérature sur les différentes méthodes de fouille de données et d'apprentissage machine pour la détection d'intrusions en cybersécurité. Ils présentent des techniques de prétraitement de données, des algorithmes de classification, des méthodes d'apprentissage non supervisé et supervisé, ainsi que des approches hybrides pour la détection d'intrusions. Les auteurs soulignent les avantages et les inconvénients de chaque méthode, ainsi que les défis liés à leur mise en œuvre dans un environnement réel. Ils donnent aussi des recommandations pour l'utilisation de ces méthodes. Cependant, ils soulignent également qu'il n'y a pas de méthode unique efficace pour toutes les applications de cybersécurité et que les critères tels que la précision, la complexité et l'interprétabilité doivent être pris en compte lors du choix d'une méthode. Les auteurs suggèrent de collecter de nouvelles données étiquetées en cybersécurité pour permettre de formuler et tester des méthodes de fouille de données et d'apprentissage machine afin de contribuer aux avancées significatives dans ce domaine. Par ailleurs, la recherche de cas exceptionnels tels des comportements anormaux ou des activités dans des schémas de comportements et d'événements [5]. Par exemple, il peut s'agir d'une tentative d'intrusion inhabituelle ou d'un trafic réseau anormalement élevé. Ces valeurs exceptionnelles peuvent indiquer la présence d'une activité malveillante ou d'une tentative d'attaque en cours.

Shaukat et al. [48] offrent un aperçu sur les défis auxquels les techniques d'apprentissage machine sont confrontées dans la protection de systèmes contre les attaques. Ils proposent une revue des techniques d'apprentissage machine utilisées dans la détection des intrusions, de *spams* et des logiciels malveillants sur les réseaux informatiques et mobiles au cours de la dernière décennie. Ils décrivent brièvement chaque méthode d'apprentissage machine, les ensembles de données de sécurité fréquemment utilisées, les outils et les mesures d'évaluation d'un modèle. En mettant en évidence l'importance croissante de la sécurité dans un monde de plus en plus connecté, ils mesurent les performances en temps de calcul de plusieurs algorithmes d'apprentissage machine.

Sarker [46] fournit un aperçu complet des algorithmes d'apprentissage machine pour l'analyse intelligente des données et de leur applicabilité dans divers domaines. Il a mis en évidence l'importance des données de qualité et des performances des algorithmes d'apprentissage pour le succès des solutions basées sur l'apprentissage machine. L'article identifie également des défis et des opportunités de recherche dans le domaine.

Dans *Bhuyan et al.* [4], les auteurs examinent différentes approches de détection d'intrusion basées sur les cyberattaques et discutent de plusieurs critères d'évaluation des performances d'une méthode ou d'un système de détection. *Abdel-Fattah et Al.* [1] ont montré que les méthodes de fouille de données et d'apprentissage machine en cybersécurité aident à réduire et à atténuer les attaques, l'intimidation et les accès non autorisés.

L'analyse des motifs dans les données des cyberattaques permet de détecter les modèles récurrents ou les évolutions significatives dans les attaques [7]. Par exemple, en observant une augmentation régulière des motifs d'intrusion provenant d'une certaine région géographique, il est possible de prendre des mesures préventives spécifiques pour renforcer la sécurité dans cette zone.

Salem et al. [45] soulignent l'importance de la fouille de données pour améliorer la détection des intrusions et renforcer les mesures de cybersécurité. L'objectif principal de cet article est de sensibiliser aux multiples menaces auxquelles les réseaux informatiques d'un pays peuvent être exposés. Cette sensibilisation est cruciale pour comprendre l'approche adoptée afin de protéger les systèmes et les réseaux nationaux.

L'article présente également la structure d'un système de détection d'intrusion, illustrée dans la figure 1. La détection d'intrusion logicielle est définie comme le processus de repérage d'intrusions en utilisant diverses méthodes, notamment la collecte de données comportementales des utilisateurs à l'intérieur et à l'extérieur du système. De plus, ce processus d'analyse complète des données relatives aux activités des utilisateurs internes et externes et vise à identifier les comportements anormaux du système.

Ils explorent comment la fouille de données est appliquée dans divers aspects de la sécurité, notamment la création de modèles, l'analyse en temps réel, la fouille de données distribuée et la visualisation des données, pour identifier et répondre efficacement aux menaces en matière de sécurité.

Il est essentiel de connaître l'origine d'une cyberattaque, les cibles potentielles et le type d'attaques menées au cours d'une certaine période. Cela peut aider à se préparer pour l'avenir, car avoir des informations sur le passé facilite également la prise des précautions et renforce ainsi les mesures de la lutte contre les cyberattaques.

Dans le prochain chapitre, nous faisons appel à des techniques d'apprentissage machine et de fouille de données pour l'identification, la détection et la prédiction de cyberattaques.

Chapitre 4

Application

Dans ce chapitre, nous nous concentrons sur l'utilisation de différentes techniques et algorithmes d'exploration de données et d'apprentissage automatique pour analyser les données relatives aux cyberattaques. Nous appliquons principalement des méthodes de classification, de regroupement, de règles d'association, et de fouille de texte. Nous commençons par présenter notre ensemble de données ainsi que l'outil d'analyse utilisé, qui est la plateforme *RapidMiner*. Ensuite, après avoir rappelé brièvement nos objectifs, nous entamons l'analyse en utilisant des approches d'apprentissage supervisé et non supervisé, pour ensuite examiner la fouille de texte.

4.1 Données

4.1.1 Sources de données

Il s'agit d'un ensemble de données téléchargées sur Kaggle le 25 août 2023 [23], une plateforme de collections de données pour des analyses par des chercheurs. L'ensemble de données s'appelle *State-Sponsored Cyber-Operations* développé par le *Council on Foreign Relations* qui est un groupe de réflexion impartial américain fondé en 1921 et axé sur la politique étrangère et les relations internationales. Le programme de politique numérique du groupe maintient un ensemble de données complètes des cyberincidents parrainées par des états depuis 2005. Cet ensemble de données sur les cyberincidents fut développé en recueillant des informations auprès de diverses sources telles que *APT Groups and Operations de Florian Roth, Center for Strategic and International Studies, Kaspersky Lab*, et des entreprises de cybersécurité [35].

Les données sont mises à jour trimestriellement et les utilisateurs peuvent contribuer à l'ensemble de données via une fonctionnalité de *crowdsourcing* (plate-forme de collecte d'informations, d'opinions ou de travaux auprès d'un groupe de personnes, généralement via Internet). Les informations sont publiées via le site *Net Politics* et les mises à jour peuvent inclure l'ajout d'incidents, d'acteurs de menace, ou des modifications des données déjà présentes. Cet ensemble de données concerne les pirates soutenus par des états et permet d'identifier les opérations de cyberincidents liées à leur politique étrangère. Les données

se limitent aux attaques par déni de service, espionnage, défiguration, destruction de données, sabotage et doxage [35]. Les définitions de ces termes sont présentées dans la deuxième section du chapitre 2 relatif à quelques rappels sur les concepts de base.

4.1.2 Description de données

L'ensemble de données que nous allons traiter couvre la période allant de 2005 à 2023. Il est composé de 845 lignes et 12 colonnes (attributs) qui sont : le titre de la cyberattaque, la date de son apparition, le groupe (affiliation) à l'origine de l'attaque, la description de l'événement, la réponse à la cyberattaque, les victimes de l'attaque, le pays commanditaire de la cyberattaque, le type de cyberattaque menée, la catégorie d'institutions ciblées et la ou les sources qui ont signalé la cyberattaque [23, 35].

La description détaillée de chaque colonne est présentée en annexe.

Dans ce travail, nous visons à répondre, entre autres, aux questions suivantes :

- Quels sont les pays commanditaires du plus grand nombre de cyberattaques et quels types d'organismes qui y sont visés ?
- Existe-t-il de fortes associations entre des attributs décrivant des attaques ?
- Est-ce qu'il y a des pays où l'on applique beaucoup plus un type d'attaques qu'un autre ?
- Est-ce qu'il y a des périodes de l'année et des jours de la semaine où il y a plus de cyberattaques ?
- Quels sont les groupes d'attaques et leurs spécificités ?

Maintenant que nous avons identifié les questions, présentons l'outil que nous utilisons pour le traitement de nos données.

4.2 Outil *RapidMiner*

Présentation

RapidMiner est l'une des plateformes de science des données les plus populaires. Elle intègre des techniques d'apprentissage machine et de fouille de données, ainsi que plusieurs modules de prétraitement et de visualisation. Elle renferme l'ensemble du processus d'analyse de données et utilise une approche sans code mettant l'accent sur la visualisation des données, des flux de travail *(workflows)* et des résultats, ce qui facilite les interprétations [39, 16, 40].

Particularités

La plateforme *RapidMiner* propose des outils puissants pour le prétraitement des données qui aident à des tâches différentes d'importation et de fouille de données, de sélection, de fusion, d'épuration et de réorganisation des données. En outre, elle met à disposition 167 modèles d'analyse telle que les arbres de décision, le regroupement, les règles d'association, l'apprentissage profond, etc. Il est également possible de créer des macros pour automatiser des tâches répétitives [34, 40].

La plateforme *RapidMiner* offre une variété d'opérations et d'algorithmes pour répondre à différents besoins d'analyse de données. Elle facilite l'intégration des données à partir de différentes bases de données et fichiers, notamment *Excel, Access, Oracle SQL Server, MySQL*, etc. De plus, elle est capable de gérer plusieurs types de données, y compris le texte.

Composants

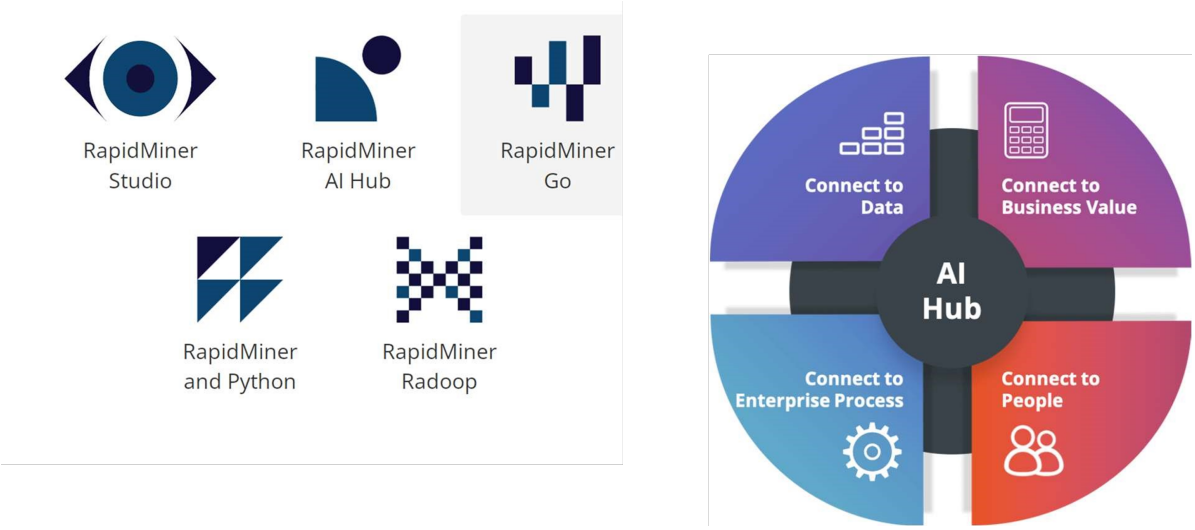


FIGURE 4.1 – Modules *RapidMiner*.
(source :[41])

- *RapidMiner Studio* est un logiciel doté d'un concepteur de flux de travaux visuels. Il comprend également des extensions telles qu'*Auto Model* facilitant la création et la validation de modèles d'analyse [40], ainsi que *Turbo Prep* offrant des fonctionnalités de prétraitement de données flexibles [42].
- *Go* est une version spécifique de *RapidMiner* qui propose des fonctionnalités de prototypage rapide d'applications.
- *AI Hub* est une fonctionnalité qui permet la planification et l'exécution des processus d'analyse de données.
- *RapidMiner & Python* offre une intégration avec Python, permettant l'exécution de codes Python au sein des processus.
- *Radoop* permet d'effectuer l'analyse de données massives sur un *cluster Hadoop* [39, 34].

RapidMiner présente plusieurs points forts qui offrent une large gamme de possibilités pour le prétraitement et l'analyse de données de différents types. Les interfaces sont intuitives et conviviales et il existe des mécanismes variés de visualisation des données et des résultats de la fouille de données et de texte. Une version gratuite de RapidMiner est disponible pour une utilisation éducative. Malgré tous ces avantages, *RapidMiner* présente quelques limites, à savoir : il n'y a pas de version disponible en français de *RapidMiner* et la documentation en français est rare.

Nous avons brièvement pris connaissance de la plateforme utilisée pour l'analyse de données dans ce projet. Nous abordons ci-après le prétraitement des données pour l'analyse.

4.3 Prétraitement de données

Le prétraitement des données dans le processus de découverte des connaissances est effectivement tel qu'il a été mentionné dans la littérature : coûteux et long, mais il reste utile pour le processus de fouille de données et la production de connaissances utiles à la prise de décision [19]. Il consiste à effectuer l'extraction, l'épuration, l'intégration, la transformation, la réduction et la discrétisation des données [16].

4.3.1 Préparation de données

RapidMiner assure la qualité et la cohérence des données utilisées dans notre processus d'analyse en contrôlant les erreurs, en remplaçant les valeurs incohérentes et en éliminant éventuellement certains enregistrements lors de l'importation des données.

Initialement, l'outil détecte et signale les erreurs potentielles dans le jeu de données, réalisant une analyse dès l'importation pour repérer des anomalies telles que des valeurs non numériques dans des colonnes déclarées comme étant numériques, des valeurs manquantes, etc. Nous avons identifié par exemple huit attaques contenant des erreurs de syntaxe. Nous avons également identifié des erreurs de type au niveau de l'attribut date comme, par exemple, la valeur « France » qui ne peut pas être interprétée comme une date valide. Cette détection précoce permet à l'utilisateur d'apporter des corrections avant de poursuivre le processus d'analyse, assurant ainsi la fiabilité des résultats finaux.

En outre, pour résoudre ces anomalies et permettre la poursuite de l'importation malgré la présence d'erreurs, *RapidMiner* offre l'option obligatoire « *Replace errors by missing values* ,» qui remplace automatiquement les valeurs erronées par des valeurs manquantes. Cette approche donne à l'utilisateur la possibilité de traiter ultérieurement les valeurs manquantes au lieu d'interrompre complètement l'importation suite aux erreurs.

Enfin, dans le but de maintenir l'intégrité du jeu de données, des enregistrements contenant des erreurs graves pouvant compromettre l'analyse peuvent être supprimés. Dans notre cas, au moins 42 attaques contenant des erreurs susceptibles de compromettre nos analyses ont été supprimées. Cette mesure garantit que le jeu de données utilisé pour l'analyse est de haute qualité, sans valeurs incohérentes significatives.

Il est important de noter que c'est l'examen exhaustif de l'ensemble des données par *RapidMiner* qui assure une détection complète des erreurs, même celles qui pourraient ne pas être apparentes lors de la prévisualisation.

Dans notre liste, seuls 803 enregistrements, soit un total de 42 instances écartées, ont pu être analysés. Une partie des données est présentée dans la figure 4.2.

Row No.	Type	Title	Date	Affiliations	Description	Response	Victims	Sponsor	Category
92	Espionage	Targeting of U...	Jun 29, 2022	Believed to be ...	The Israeli defe...	Confirmation ...	UN peacekeep...	Iran (Islamic R...	Civil society, M...
93	Espionage	Targeting of U...	Sep 19, 2022	Sandworm	A Russian thre...	Unknown	Ukrainian telec...	Russian Feder...	Private sector,...
94	Espionage	Targeting of or...	Dec 6, 2022	Calisto	As part of a phi...	Unknown	Six private co...	Russian Feder...	Civil society, P...
95	Espionage	Targeting of Io...	Jan 23, 2022	Believed to be ...	The signals int...	Unknown	Mayors and ot...	Israel	Government
96	Espionage	Targeting of As...	Mar 17, 2022	Sandworm	Sandworm tar...	Criminal charg...	Asus Routers i...	Russian Feder...	Private sector
97	Sabotage	Targeting of go...	Apr 27, 2022	APT 28	APT 28 condu...	?	Government n...	Russian Feder...	Government
98	Espionage	Targeting of E...	Jul 23, 2022	APT 37	A North Korea...	Unknown	Czech Republi...	Korea (Democ...	Government
99	Espionage	Targeting of P...	Nov 10, 2022	Sandworm	Russian threat...	Unknown	Ukrainian and ...	Russian Feder...	Private sector
100	Espionage	Antlion	Feb 2, 2022	?	Antlion is a Chi...	?	Financial depar...	?	Private sector
101	Espionage	Targeting of Isr...	Apr 6, 2022	APT-C-23	Hamas-linked ...	?	Israeli individua...	Palestine, Stat...	Government
102	Financial Theft	Targeting of B...	Jun 1, 2022	Believed to be ...	Iranian govern...	Denouncemen...	Boston Childre...	Iran (Islamic R...	Private sector
103	Espionage	Targeting of in...	Aug 8, 2022	APT 18	Chinese hacke...	Unknown	Industrial plant...	China	Government, ...
104	Espionage	Targeting of U...	Feb 25, 2022	UNC1151	Belarusian thre...	Confirmation ...	Ukrainian and ...	Belarus, Russi...	Military
105	Espionage	Targeting of R...	Apr 27, 2022	Mustang Panda	Chinese threat...	?	Russian official...	China	Government, ...
106	Espionage	Targeting of H...	Mar 29, 2022	Believed to be ...	The Federal S...	Confirmation ...	Hungarian For...	Russian Feder...	Government
107	Espionage	Targeting of go...	Sep 29, 2022	APT 10	APT 10 installe...	Unknown	Two Middle Ea...	China	Government
108	Espionage	Earth Lusca	Jan 17, 2022	?	Earth Lusca is ...	?	Academic instit...	China	Government, ...
109	Espionage	Targeting of U...	Mar 18, 2022	Gamaredon	Russian APTâ...	Confirmation ...	Ukrainian state...	Russian Feder...	Government

ExampleSet (803 examples, 1 special attribute, 8 regular attributes)

FIGURE 4.2 – Affichage des données.

La plateforme offre la possibilité de changer le type des données dès le processus d'importation de données et d'écarter des attributs qui ne peuvent être utiles pour l'analyse. Dans notre cas, nous avons donc écarté les attributs nommés « source 1, source 2 et source 3 » parce qu'ils ne sont composés que des liens HTTP et avons défini l'attribut « type » comme étiquette à la figure 4.2. La deuxième étape a consisté à gérer les valeurs manquantes dans nos données.

4.3.2 Gestion des valeurs manquantes

La gestion des valeurs manquantes est une des opérations du prétraitement des données lequel est une étape cruciale du processus de découverte de connaissances. Les données incomplètes peuvent introduire des biais et des erreurs dans nos analyses, ce qui peut avoir un impact négatif sur les résultats. Heureusement, *RapidMiner* offre une gamme de méthodes et de techniques pour traiter efficacement les valeurs manquantes.

- **Suppression des lignes ou colonnes** contenant des valeurs manquantes. L'une des approches les plus simples consiste à éliminer les lignes ou les colonnes contenant des valeurs manquantes. Cela peut être approprié lorsque les données manquantes sont rares ou que les lignes ou les colonnes concernées ne sont pas essentielles pour l'analyse. Cependant, cela peut entraîner une perte d'informations si les données manquantes sont nombreuses et concernent des attributs importants [14].
- **Déclaration de valeurs manquantes**. Parfois, les valeurs manquantes sont intentionnelles ou déjà connues. Dans de tels cas, on peut déclarer explicitement ces valeurs manquantes dans RapidMiner pour éviter qu'elles ne soient interprétées comme des données valides [40]. Par exemple, dans notre ensemble de données, un commanditaire d'une attaque est un état tel que la Chine, la Russie, le Canada, etc. Si, par erreur, un enregistrement indique « espionnage » ou « gouvernement » comme commanditaire, nous pouvons alors spécifier (déclarer) que ces valeurs,

telles que « espionnage » ou « gouvernement », dans cet attribut doivent être traitées comme des valeurs manquantes plutôt que comme des commanditaires valides. Cependant, il est important de noter que cela ne correspond pas à notre cas.

- **Remplacement des valeurs manquantes**. RapidMiner offre diverses options pour remplacer les valeurs manquantes. Dans cette plate-forme, les valeurs manquantes sont représentées par un point d'interrogation (?) [40]. On a plusieurs options pour gérer les valeurs manquantes dans nos données. On peut décider de ne rien faire du tout, d'utiliser des statistiques comme la moyenne, le minimum ou le maximum pour les données numériques et les dates existantes, ou encore de les remplacer par zéro ou une valeur spécifique. En ce qui concerne les données catégorielles, on peut opter pour l'utilisation du mode, qui est la valeur la plus fréquemment observée, ou simplement choisir une valeur spécifique pour remplacer les valeurs manquantes.
- **L'imputation des valeurs manquantes** . L'imputation est un processus utilisé dans l'analyse de données pour remplacer les valeurs manquantes par des valeurs estimées en se basant sur d'autres données disponibles. Cela permet de garantir que les données sont complètes et prêtes pour l'analyse ou la modélisation [40, 14]. RapidMiner offre plusieurs méthodes d'imputation pour gérer les valeurs manquantes dans vos ensembles de données, notamment l'imputation par la moyenne, la médiane, la régression, les k plus proches voisins (knn), etc. Ces méthodes d'imputation diffèrent dans la manière dont elles estiment les valeurs manquantes, et le choix dépendra du type des données quel'on traite, de la distribution des données et de la complexité de la relation entre les attributs.
  - **Imputation par la moyenne**. Cette méthode remplace les valeurs manquantes par la moyenne des données non manquantes dans la même colonne. Elle est appropriée pour les données numériques et convient lorsque les données suivent une distribution normale.
  - **Imputation par la médiane**. Elle remplace les valeurs manquantes par la médiane (valeur centrale) des données non manquantes dans la même colonne. Elle est robuste aux valeurs aberrantes et peut être préférable lorsque les données sont fortement asymétriques.
  - **Imputation par la régression**. Elle utilise un modèle de régression pour prédire les valeurs manquantes en fonction des autres attributs de la même ligne. Elle est plus sophistiquée, car elle tient compte des relations entre les variables, mais elle peut être sensible aux erreurs du modèle.
  - **Imputation par les k plus proches voisins (*knn*)**. Elle repose sur la similarité entre les instances de données. Elle recherche les k exemples les plus proches (en fonction d'une mesure de similarité) qui ont des valeurs non manquantes similaires, puis impute la valeur manquante en fonction des valeurs de ces exemples voisins. Elle est adaptée à la fois pour les données numériques et catégorielles et tient compte de la structure sous-jacente des données [14].

    Il est recommandé d'expérimenter différentes méthodes d'imputation et de choisir celle qui convient le mieux à un ensemble de données et aux objectifs d'analyse ou de modélisation.
- **Utilisation de modèles de prédiction**. Une autre approche intéressante consiste à utiliser des modèles de prédiction pour estimer les valeurs manquantes. Les arbres de décision et d'autres algorithmes d'apprentissage automatique peuvent être exploités à cet effet. RapidMiner facilite

la création et l'évaluation de ces modèles pour traiter les valeurs manquantes de manière robuste [40, 14].

Dans l'analyse de données, il est recommandé de veiller à la qualité des données afin d'assurer la fiabilité et la pertinence des résultats obtenus. Cela implique de vérifier si les données sont complètes (c'est-à-dire qu'elles ne présentent pas des valeurs manquantes), cohérentes (en excluant les valeurs aberrantes, qui sont des données incohérentes) et qu'elles ne contiennent pas des doublons (lorsqu'une même attaque est enregistrée avec trois identifiants différents par exemple).

Commençons cette partie par la vérification des valeurs manquantes comme le montre la figure 4.3 qui présente les statistiques avant les remplacements des valeurs manquantes indiquées par le « ? » dans la figure précédente.

Label					
Type	Nominal	43	Least: Doxing (6)	Most: Espionage (644)	Values: Espionage (644), Sabotage (33), ...[5 more]
Title	Nominal	0	Least: Ã (1)	Most: Targetin [...] oshan (2)	Values: Targetin [...] rm Roshan (2), Targetin [...] companies (2), ...[796 more]
Date	Date-time	22	Earliest date: Jul 11, 2006	Latest date: Jun 16, 2023	Duration: 6184 days
Affiliations	Nominal	143	Least: Ã and th [...] ence. (1)	Most: Lazarus Group (28)	Values: Lazarus Group (28), APT 28 (17), ...[380 more]
Description	Nominal	5	Least: believed [...] dent. (1)	Most: Chinese [...] vity. (2)	Values: Chinese [...] activity. (2), A China- [...] dustries. (1), ...[795 more]
Response	Nominal	582	Least: Unknown [...] qmAkw (1)	Most: Unknown (88)	Values: Unknown (88), Denounce [...] ctivities (4), ...[122 more]
Victims	Nominal	34	Least: â€œThe h [...] neâ€□ (1)	Most: United States (21)	Values: United States (21), South Korea (7), ...[687 more]
Sponsor	Nominal	47	Least: Uzbekistan (1)	Most: China (268)	Values: China (268), Russian Federation (167), ...[53 more]
Category	Nominal	32	Least: Private [...] ciety (1)	Most: Private sector (204)	Values: Private sector (204), Government (187), ...[29 more]

FIGURE 4.3 – Statistiques avant le remplacement des valeurs manquantes.

Ayant identifié les valeurs manquantes de chaque colonne de nos données qui sont actuellement constituées de neuf attributs au total, dont un attribut spécial et 8 attributs réguliers, nous avons constaté que huit d'entre elles présentent des valeurs manquantes, à l'exception de l'attribut titre *(title)*.

Les valeurs manquantes pour la plupart des attributs importants dans notre analyse sont en général peu fréquentes, généralement inférieures ou égales à 5 %. Cependant, nous notons des taux plus élevés pour les attributs « Affiliation » (17 %) et « Response » (67 %), bien que ces derniers ne soient pas particulièrement pertinents pour notre analyse.

Nous nous lançons désormais dans le traitement de ces valeurs manquantes, une étape cruciale pour garantir la fiabilité de notre analyse. Dans un premier temps, nous avons opté pour la méthode du remplacement par des valeurs spécifiques. Cependant, les résultats ont révélé une limitation importante de cette approche. Elle avait tendance à renforcer une classe de données au détriment d'une autre, ce qui pouvait avoir un impact significatif sur nos conclusions. Pour illustrer cela, prenons l'exemple du commanditaire d'attaques avec 47 valeurs manquantes, dont 20 étaient associées au Canada et 30 au Mexique. En remplaçant les valeurs manquantes par « Canada », cela aurait créé une forte prédominance

du Canada par rapport au Mexique, bien que dans la réalité, le Mexique a dû être plus représenté. Cette distorsion aurait pu conduire à des conclusions biaisées.

Ensuite, nous avons expérimenté le remplacement par la valeur la plus fréquente, en calculant le mode (« average ») pour les attributs nominaux, avec une adaptation pour les attributs de type date où nous avons utilisé les valeurs maximales et minimales. Cette méthode a également montré des résultats similaires à la précédente, renforçant davantage la classe la plus fréquente, ce qui risque de créer un déséquilibre significatif entre les classes. En reprenant l'exemple précédent, cela signifierait que le mode du Mexique serait plus important que celui du Canada, ce qui peut non seulement accentuer la proportion du Mexique, mais également déséquilibrer la relation entre les deux commanditaires d'attaques.

Pour surmonter ces problèmes, nous avons ensuite opté pour l'imputation par K plus proches voisins (KNN). Cette approche a permis de prédire les valeurs manquantes et de les attribuer de manière aléatoire, ce qui a abouti à des résultats plus pertinents et équilibrés. Cette méthode a pris en compte la structure sous-jacente des données pour produire des estimations plus justes et éviter les biais potentiels observés dans les méthodes précédentes.

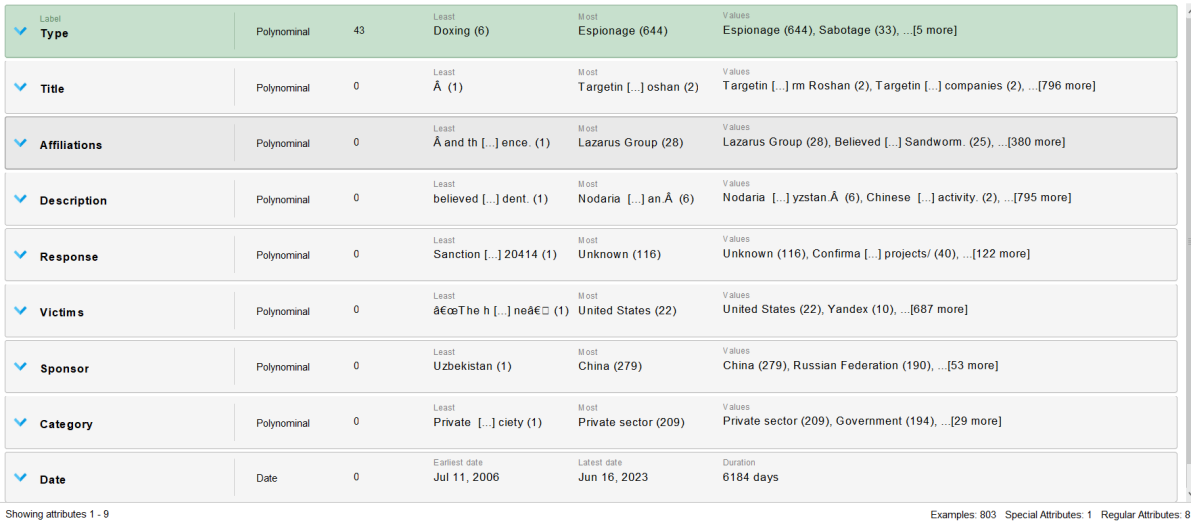


FIGURE 4.4 – Statistiques après le remplacement des valeurs manquantes.

Pour l'attribut « type d'attaques », qui joue un rôle essentiel en tant qu'étiquette dans notre analyse, nous entamons notre processus en éliminant les enregistrements incomplets. Cette étape est nécessaire pour la préparation de nos modèles de classification, car des données manquantes pourraient entraîner des résultats imprécis.

Par la suite, nous mettons en œuvre des modèles de classification, notamment des arbres de décision, pour prédire les types d'attaques. Ces modèles exploitent les données complètes pour élaborer des prédictions précises. Une fois que nous avons obtenu ces prédictions, nous les utilisons pour remplacer les valeurs manquantes dans cet attribut. Cette démarche permet de compléter notre ensemble de données tout en préservant la cohérence et la fiabilité de nos résultats.

4.3.3 Ajout d'un nouvel attribut

L'attribut « date », tel qu'il était initialement présenté, ne fournissait pas des informations pertinentes. Afin d'obtenir des résultats plus significatifs, nous avons décomposé cet attribut en quatre nouveaux éléments : le jour de la semaine, le mois de l'année, le trimestre et l'année. Pour ce faire, nous avons mis en place un processus distinct pour dupliquer la colonne de la date. Chacun des nouveaux attributs a été renommé comme suit : *day* pour le jour de la semaine, *Month* pour le mois de l'année, *Quarter* pour le trimestre et *Year* pour l'année.

4.3.4 Transformation des données

À ce stade, nous avons maintenant 10 attributs dans notre ensemble de données, avec 803 lignes, ce qui nécessite quelques transformations. Ensuite, tous ces attributs sont convertis en valeurs numériques en spécifiant le format pour chacune d'elles. Nous avons aussi la possibilité de faire cette transformation à l'aide des fonctionnalités de *Turbo Prep*. Dans l'option *Transform*, nous pouvons copier la colonne *date* et utiliser l'opération *change type*, pour extraire le jour de la semaine, le mois de l'année, le trimestre ou l'année. Diverses autres transformations sont effectuées au cours du développement de chaque modèle selon les besoins afin d'améliorer sa performance et son analyse.

Dans la section suivante, nous appliquons différentes méthodes de fouille de données et d'apprentissage machine pour analyser nos données. Cette étape constitue une phase de notre étude dans laquelle nous analysons les données relatives à quatre aspects principaux : le type d'attaque, la date, la catégorie et le commanditaire des cyberattaques. **Nous commençons par appliquer les algorithmes d'apprentissage supervisé.**

4.4 La classification

La classification est une méthode qui consiste à construire un modèle en regroupant des données similaires en classes, concepts et groupes de variables prédéfinis. Elle permet également d'analyser de nouvelles variables ajoutées à l'ensemble de données et de les classifier en fonction de leur correspondance avec les classes existantes [19].

Dans cette partie, nous avons effectué un prétraitement particulier sur nos données. Comme nous l'avions mentionné précédemment, notre analyse pour cette phase du projet porte sur cinq attributs principaux que nous avons sélectionnés : le type, la catégorie, la date, le commanditaire et la description.

Nous avons observé que la classe dominante était celle de l'espionnage avec 643 enregistrements sur les 803 retenus.

Afin de nous concentrer sur les autres types d'attaques, nous avons exclu la classe dominante (espionnage). Ainsi, notre analyse porte uniquement sur les 159 enregistrements restants. Pour cela, nous avons créé un nouveau fichier Excel sans la classe d'espionnage. Afin d'avoir des attributs équilibrés, nous avons effectué des agrégations avec les autres types d'attaques. Étant donné que la nouvelle classe

dominante est le déni de service, nous avons à nouveau regroupé les types de vol financier et de sabotage pour former une nouvelle classe que nous avons appelée « Rançongiciel ». De plus, les autres types d'attaques tels que le doxage, la dégradation et la destruction des données ont été regroupés sous le terme « hameçonnage ». Ainsi, notre nouvel ensemble de données comprend les classes « Rançongiciel » et « hameçonnage » et le déni de service, ce qui nous permettra de nous concentrer sur l'analyse de ces trois différents types d'attaques.

Nous avons aussi considéré dans cette analyse que la cible « société civile » peut être combinée avec le privé pour former la catégorie nommée P. De même, la fusion des cibles « militaire » et « gouvernement » a pour label G. En ce qui concerne l'attribut « commanditaire », nous avons regroupé les pays qui sont rares en utilisant un seuil relatif de 0,02, c'est-à-dire les pays ayant commandité moins de 10 attaques dans l'ensemble de données.

Nous souhaitions prédire le type d'attaque en fonction des autres attributs décrivant l'attaque, tels que la catégorie d'institutions ciblées et le pays commanditaire de l'attaque. Nous allons utiliser ces attributs comme variables indépendantes pour entraîner notre modèle de prédiction.

En analysant ces attributs et leur relation avec de type d'attaque, nous espérons identifier des schémas et les tendances qui nous permettront de prédire le type d'attaque en fonction des autres caractéristiques disponibles.

Pour étudier la prédiction du type d'attaque en fonction des caractéristiques disponibles, nous allons utiliser deux algorithmes : les arbres de décision et les réseaux bayésiens.

4.4.1 Arbres de décision

Nous considérons le type d'attaque comme attribut de classification dans notre nouvel ensemble de données. Ensuite, nous avons divisé l'ensemble de données en un ensemble d'entraînement (80 %) et un ensemble de test (20 %). L'ensemble d'entraînement (127 enregistrements sur les 159 retenus) sera utilisé pour entraîner le modèle, tandis que l'ensemble de test (32 enregistrements sur les 159 retenus) sera utilisé pour évaluer sa performance.

Dans cet arbre de décision, l'attribut racine est « sponsors », ce qui signifie que le pays commanditaire de la cyberattaque a le plus d'impact sur la décision ou le type d'attaque. Ensuite, sur les branches de cet arbre, nous retrouvons la catégorie d'institution visée par l'attaque.

Le parcours de l'arbre est représenté par une ligne allant de la racine aux feuilles, comme illustré dans la figure 4.5 ci-dessous. Il exprime une règle de classification. À titre d'exemple, lorsque le pays commanditaire est la Chine et que la catégorie d'institution visée est le secteur privé, il s'agit d'un déni de services dans près de 100 % (21/21) des cas ou encore si le commanditaire est la Russie et que la catégorie d'institution visée est le gouvernement, il s'agit d'un hameçonnage dans près de 100 % (16/16)

des cas, alors que si le commanditaire est la Corée du Nord ou l'Iran et que la catégorie d'institution visée est le gouvernement, il s'agit d'un rançongiciel dans près de 100 % (12/12) ou (6/6) des cas.

```
Sponsor = China
|   Category = G: Ransomware {Ransomware=11, Denial of service=6, Phishing=0}
|   Category = G, P: Denial of service {Ransomware=0, Denial of service=4, Phishing=0}
|   Category = P: Denial of service {Ransomware=0, Denial of service=21, Phishing=0}
|   Category = P, G: Denial of service {Ransomware=0, Denial of service=3, Phishing=0}
Sponsor = Iran
|   Category = G: Phishing {Ransomware=2, Denial of service=0, Phishing=3}
|   Category = G, P: Denial of service {Ransomware=0, Denial of service=5, Phishing=0}
|   Category = P: Ransomware {Ransomware=6, Denial of service=0, Phishing=0}
|   Category = P, G: Ransomware {Ransomware=1, Denial of service=1, Phishing=0}
Sponsor = North Korea
|   Category = G: Ransomware {Ransomware=5, Denial of service=0, Phishing=0}
|   Category = G, P: Phishing {Ransomware=0, Denial of service=0, Phishing=1}
|   Category = P: Ransomware {Ransomware=12, Denial of service=0, Phishing=0}
|   Category = P, G: Ransomware {Ransomware=1, Denial of service=0, Phishing=0}
Sponsor = Other
|   Category = G: Denial of service {Ransomware=1, Denial of service=5, Phishing=1}
|   Category = G, P: Denial of service {Ransomware=2, Denial of service=3, Phishing=1}
|   Category = P: Denial of service {Ransomware=2, Denial of service=6, Phishing=3}
|   Category = P, G: Ransomware {Ransomware=1, Denial of service=1, Phishing=0}
Sponsor = Pakistan: Denial of service {Ransomware=0, Denial of service=3, Phishing=0}
Sponsor = Russia
|   Category = G: Phishing {Ransomware=0, Denial of service=0, Phishing=16}
|   Category = G, P: Phishing {Ransomware=1, Denial of service=1, Phishing=5}
|   Category = P: Ransomware {Ransomware=7, Denial of service=0, Phishing=4}
|   Category = P, G: Denial of service {Ransomware=0, Denial of service=1, Phishing=0}
Sponsor = United States
|   Category = G: Denial of service {Ransomware=2, Denial of service=3, Phishing=0}
|   Category = P: Ransomware {Ransomware=2, Denial of service=1, Phishing=0}
|   Category = P, G: Denial of service {Ransomware=0, Denial of service=1, Phishing=0}
```

FIGURE 4.5 – Arbre de décision

Cela signifie que pour les cyberattaques ayant pour pays commanditaire la Chine et ciblant le secteur privé, la prédiction du type d'attaque est principalement un déni de services avec une probabilité de près de 100 %, la Russie et ciblant le gouvernement, la prédiction du type d'attaque est principalement un hameçonnage avec une probabilité de près de 100 %, la Corée du Nord ou encore l'Iran et ciblant le secteur privé, la prédiction du type d'attaque est principalement un rançongiciel avec une probabilité de près de 100 %.

L'arbre de décision est présenté dans la figure 4.6.

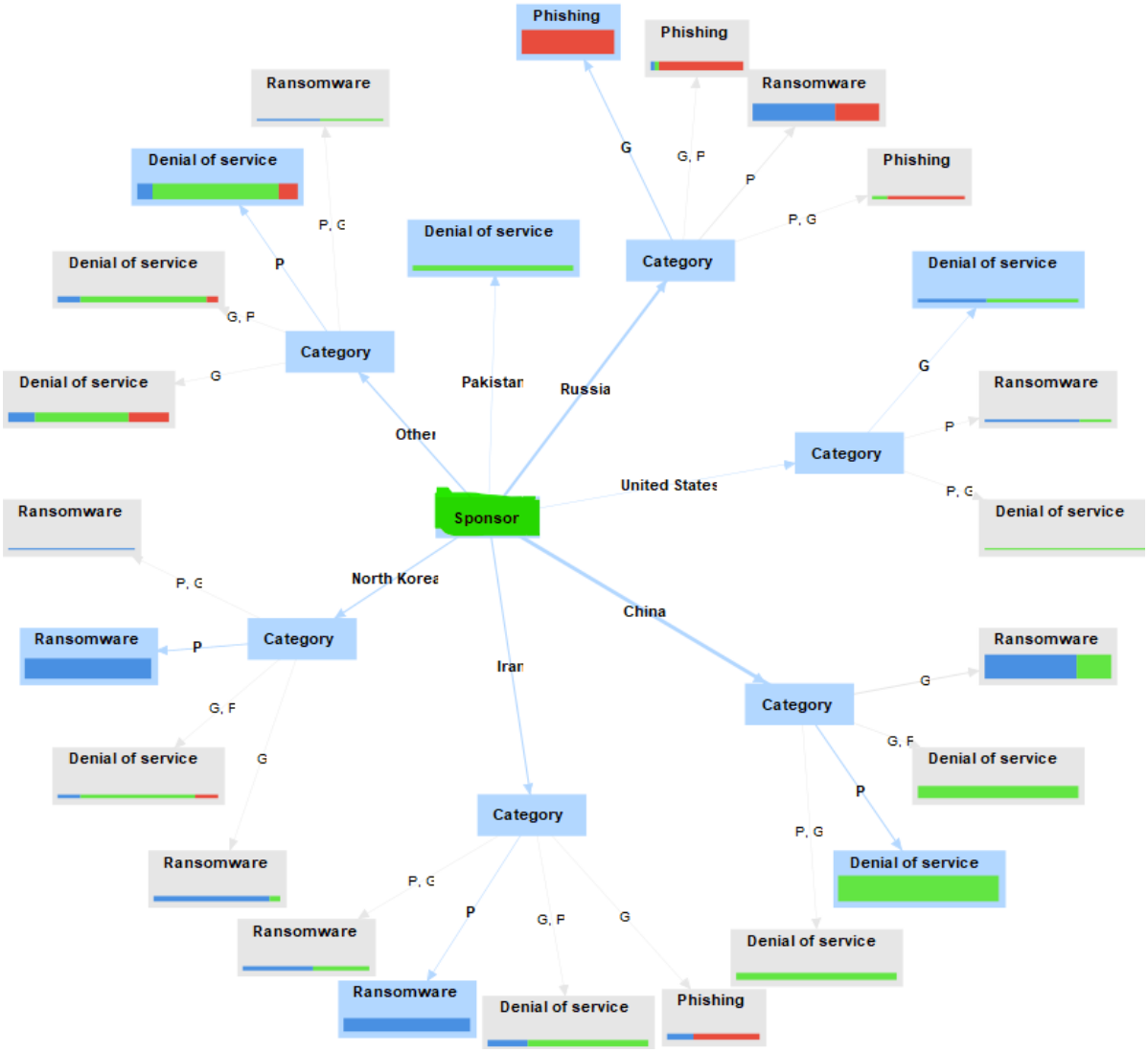


FIGURE 4.6 – Graphique de l'arbre de décision

Nous remarquons ici l'utilisation de trois couleurs différentes pour représenter chaque type d'attaque en fonction de leur proportion. La couleur verte est attribuée à la classe du déni de services, qui est la classe dominante avec (41,1 %). Ensuite, nous avons la couleur bleue pour représenter la classe des rançongiciels (36,5 %), et enfin la couleur rouge pour représenter la classe de l'hameçonnage qui est la classe la plus faible en termes de proportion (22,4 %).

Un autre détail important à noter ici est l'intensité des couleurs qui représente l'importance des nœuds en fonction de leur proportion d'instances sur les branches de l'arbre. Comme nous pouvons l'observer, les nœuds ayant une forte intensité de couleurs ont une proportion élevée d'instances. Par exemple, les nœuds terminaux marqués en vert relatif au déni de service (Chine et secteur privé), en rouge relatif aux hameçonnages (Russie et le gouvernement) et en bleue relative aux rançongiciels (Corée du Nord ou

Iran et secteur privé) dans des couleurs vives, avec un nombre de 21 pour le premier, 16 le second, 12 et 6 pour le troisième et le quatrième, indiquent une proportion importante d'instances correspondantes.

Prenons l'exemple de ces deux nœuds de feuilles dans la figure 4.6 de l'arbre et essayons de donner des explications additionnelles.

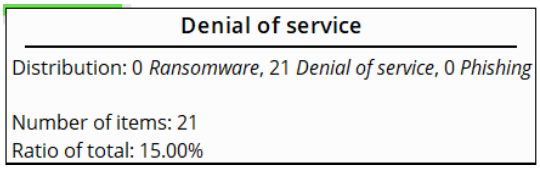


FIGURE 4.7 – Noeud feuille déni de service (Chine et secteur privé)

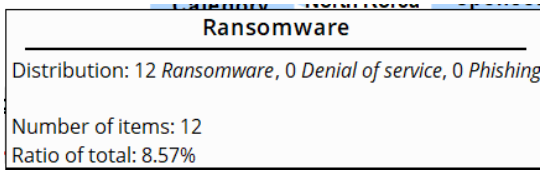


FIGURE 4.8 – Noeud feuille rançongiciel (Corée du nord et secteur privé)

Le nœud correspondant au déni de service 4.7 nœud feuille, déni de service (Chine et secteur privé) qui présente un ratio de 15,00 %, ce qui en fait un nœud important en termes de proportion. Sur les 21 cas inclus dans ce nœud, tous sont classés comme déni de services, tandis que l'hameçonnage et le rançongiciel sont absents. Cela montre la prédominance des attaques par déni de service représentant 100 % des cas. Par conséquent, nous pouvons conclure que lorsque le pays commanditaire est la Chine et que la catégorie ciblée est le secteur privé, il s'agit très probablement d'une attaque de déni de service.

Le nœud correspondant au rançongiciel (Corée du Nord et secteur privé) qui présente un ratio de 8,57 %, soit moins que celui de déni de service. Il reste quand même non négligeable en termes de proportion. Sur les 12 cas inclus dans ce nœud, tous sont des rançongiciels, tandis que l'hameçonnage et le déni de service sont absents. Cela montre la prédominance des attaques par rançongiciel qui constitue le 100,0 % des cas. Nous pouvons alors conclure que lorsque le pays commanditaire est la Corée du Nord et que la catégorie ciblée est le secteur privé, il s'agit très probablement d'une attaque de déni de services.

4.4.2 Règles de classification

Une attaque est de type rançongiciel dans les cas suivants :

- Si le pays commanditaire est la Russie et que la cible est le secteur privé. Cela se produit dans 7 cas sur 11, soit une probabilité de 63,64 %.
- Si le pays commanditaire est la Russie et que la cible est le gouvernement et le secteur privé. Cela se produit dans 1 cas sur 7, soit une probabilité de 14,29 %.
- Si le pays commanditaire est la Chine et que la cible est le gouvernement. Cela se produit dans 11 cas sur 17, soit une probabilité de 64,70 %.
- Si le pays commanditaire est l'Iran et que la cible est le secteur privé. Cela se produit dans 6 cas sur 6, soit une probabilité de 100,00 %.
- Si le pays commanditaire est l'Iran et que la cible est le gouvernement. Cela se produit dans 2 cas sur 5, soit une probabilité de 40,00 %.

- Si le pays commanditaire est l'Iran et que la cible est à la fois le secteur privé et le gouvernement. Cela se produit dans 1 cas sur 2, soit une probabilité de 50,00 %.
- Si le pays commanditaire est les États-Unis et que la cible est le secteur privé. Cela se produit dans 2 cas sur 3, soit une probabilité de 66,67 %.
- Si le pays commanditaire est les États-Unis et que la cible est le gouvernement. Cela se produit dans 2 cas sur 5, soit une probabilité de 40,00 %.
- Si le pays commanditaire est la Corée du Nord et que la cible est le secteur privé. Cela se produit dans 12 cas sur 12, soit une probabilité de plus de 100 %.
- Si le pays commanditaire est la Corée du Nord et que la cible est le gouvernement. Cela se produit dans 5 cas sur 5, soit une probabilité de plus de 100 %.
- Si le pays commanditaire est la Corée du Nord et que la cible est le secteur privé et le gouvernement. Cela se produit dans 1 cas sur 1, soit une probabilité de plus de 100 %.
- Si le pays commanditaire est un autre pays et que la cible est le secteur privé et le gouvernement. Cela se produit dans 1 cas sur 2, soit une probabilité de plus de 50 %.

Ces observations nous permettent de déterminer les conditions dans lesquelles une attaque peut être classée comme de type rançongiciel en se basant sur le pays commanditaire et la cible spécifique de l'attaque.

4.4.3 Performance du modèle de classification par arbres de décision

accuracy: 99.19%

	true Phishing	true Ransomware	true Denial of service	class precision
pred. Phishing	29	1	0	96.67%
pred. Ransomware	0	39	0	100.00%
pred. Denial of service	0	0	54	100.00%
class recall	100.00%	97.50%	100.00%	

FIGURE 4.9 – Performance du modèle

Le modèle des arbres de décision que nous avons entraîné présente une précision de 99,19 %. Cela signifie que le modèle a réussi à prédire correctement le type d'attaque dans 99,19 % des cas évalués. Une précision aussi élevée indique que le modèle est performant dans sa capacité à classifier les attaques de manière précise.

Nous allons maintenant incorporer le type d'attaque « espionnage », puis effectuer un rééquilibrage des classes en augmentant la valeur des autres catégories. Pour commencer, nous remplaçons le commanditaire (pays) par la région (continent) dans un souci de généralisation. À cet effet, nous avons récupéré un jeu de données depuis Kaggle en date du 27 août 2023, nommé *country-mapping-iso-continent-region*. Ce fichier contient 11 attributs, mais nous avons jugé que seuls 3 d'entre eux étaient nécessaires : le nom du pays, la région ou le continent, et la sous-région. Nous effectuons une jointure interne avec les attributs « commanditaires d'attaques *(sponsor)* » de notre ensemble de données initial et « nom du pays *(contries)* » du nouveau jeu de données comme critères de liaison (attributs clés), car ils correspondent.

Ensuite, nous avons abordé le traitement de l'attribut « institution ciblée ». Nous avons regroupé sous l'instance « autres », que nous symbolisons par « O », toutes les cibles simultanées, c'est-à-dire toutes les fois où une attaque a visé deux ou plusieurs catégories d'institutions en même temps. Par exemple, les installations militaires et la société civile, ou le secteur privé et le gouvernement, civil et privé, etc., sont la cible d'une même attaque. Nous avions déjà symbolisé les autres catégories de la manière suivante : « G » pour gouvernement, « C » pour société civile, « P » pour secteur privé et « M » pour installations militaires.

Ainsi, nous avons quatre types d'attaques, cinq catégories, cinq régions ou continents, et un total de données d'entrées de 1130 attaques pour les données d'entrée de notre modèle. Nous avons construit un arbre de décision avec ces nouvelles données en les divisant en un ensemble d'entraînement de 80 % (904) et un ensemble de tests de 20 % (226).

```
Continent = Africa: Espionage {Ransomware=0, Phishing=0, Denial of service=0, Espionage=9}
Continent = Americas
|   Category = C: Denial of service {Ransomware=0, Phishing=0, Denial of service=4, Espionage=0}
|   Category = G: Espionage {Ransomware=0, Phishing=0, Denial of service=0, Espionage=1}
|   Category = O: Espionage {Ransomware=0, Phishing=0, Denial of service=0, Espionage=2}
|   Category = P: Denial of service {Ransomware=0, Phishing=0, Denial of service=1, Espionage=0}
Continent = Asia
|   Category = C: Denial of service {Ransomware=0, Phishing=0, Denial of service=94, Espionage=0}
|   Category = G: Espionage {Ransomware=0, Phishing=0, Denial of service=0, Espionage=132}
|   Category = M: Espionage {Ransomware=0, Phishing=0, Denial of service=0, Espionage=26}
|   Category = O: Denial of service {Ransomware=28, Phishing=0, Denial of service=99, Espionage=69}
|   Category = P: Denial of service {Ransomware=0, Phishing=0, Denial of service=190, Espionage=0}
Continent = Europe
|   Category = C: Phishing {Ransomware=0, Phishing=26, Denial of service=0, Espionage=0}
|   Category = G: Phishing {Ransomware=0, Phishing=100, Denial of service=0, Espionage=0}
|   Category = M: Phishing {Ransomware=0, Phishing=8, Denial of service=0, Espionage=0}
|   Category = O: Phishing {Ransomware=0, Phishing=34, Denial of service=1, Espionage=19}
|   Category = P: Ransomware {Ransomware=56, Phishing=0, Denial of service=0, Espionage=0}
Continent = Oceania
|   Category = M: Ransomware {Ransomware=1, Phishing=0, Denial of service=0, Espionage=0}
|   Category = P: Phishing {Ransomware=0, Phishing=2, Denial of service=0, Espionage=0}
```

FIGURE 4.10 – Arbres de décision

Lorsque nous parcourons notre arbre de décision à la figure 4.10, nous observons ce qui suit (les lignes jaunes) :

- Si le continent commanditaire est l'Afrique, il s'agit d'une attaque de type espionnage.
- Si le continent commanditaire est l'Amérique et que la catégorie d'institutions visée est la société civile, il s'agit d'une attaque de type déni de service dans presque 100 % des cas (4/4).
- Si le continent commanditaire est l'Asie et que l'institution ciblée est le gouvernement, alors il s'agit d'un espionnage dans près de 100 % des cas (132/132).
- Si le continent commanditaire est l'Asie et que la catégorie d'institutions visée est « autres », il s'agit d'un déni de service dans près de 50,51 % des cas (99/196).
- Si le commanditaire est l'Europe et que la catégorie d'institutions visée est l'installation militaire, il s'agit d'un hameçonnage dans près de 100 % des cas (8/8).

– Si le commanditaire est l'Europe et que la catégorie d'institutions visée est le secteur privé, il s'agit d'un rançongiciel dans près de 100 % des cas (56/56).

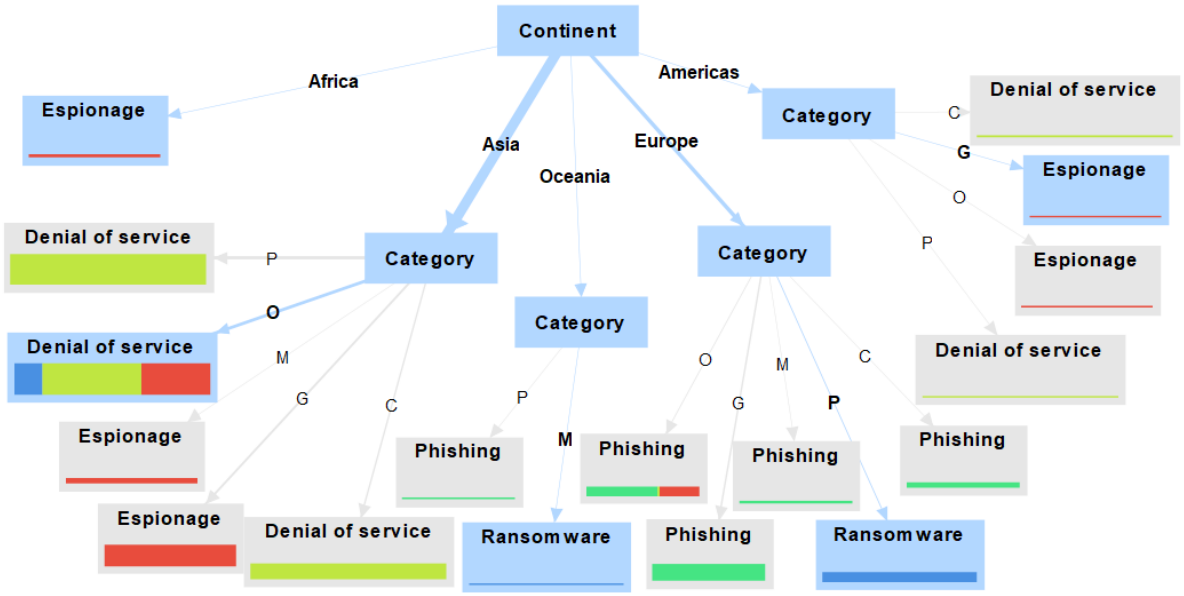


FIGURE 4.11 – Schéma de l'arbres de décision

Il y a quatre couleurs différentes pour représenter chaque type d'attaque en fonction de leur proportion sur l'histogramme du schéma de l'arbre de décision à la figure 4.11. La couleur rouge est attribuée à la classe « espionnage », qui est la classe dominante avec une proportion de 33,8 %. La couleur jaune représente la classe « déni de service » (26,8 %), la couleur bleue représente la classe « rançongiciels » (20,7 %), et enfin la couleur verte représente la classe « hameçonnage », qui est la classe avec la proportion la plus faible (18,8 %).

Un autre détail important à noter ici est l'intensité des couleurs qui représente l'importance des nœuds en fonction de leur proportion d'instances sur les branches de l'arbre. Comme nous pouvons l'observer, les nœuds ayant une intensité de couleur élevée ont une proportion élevée d'instances. Par exemple, les nœuds terminaux, marqués en rouge (Asie et gouvernement) en relation avec l'espionnage, et en jaune en relation avec les dénis de service (Asie et le secteur privé), avec un nombre de 132 pour le premier et 190 pour le second, indiquent une proportion importante d'instances correspondantes.

Nous pouvons extraire certaines règles de classification. Par exemple, on peut dire qu'une attaque est de type espionnage si et seulement si :

– Le commanditaire est l'Afrique dans les 9 cas sur 9, soit une probabilité de 100 %.
– Le commanditaire est l'Amérique, et l'institution ciblée est le gouvernement dans 1 cas sur 1, soit une probabilité de 100 %.
– Le commanditaire est l'Amérique, et l'institution ciblée est « autres » (deux catégories d'institutions ciblées ou plus) dans 2 cas sur 2, soit une probabilité de 100 %.

- Le commanditaire est l'Asie, et l'institution ciblée est le gouvernement dans presque 132 cas sur 132, soit une probabilité de 100 %.
- Le commanditaire est l'Asie, et l'institution ciblée est militaire dans 26 cas sur 26, soit une probabilité de 100 %.
- Le commanditaire est l'Europe, et l'institution ciblée est « autres » (deux catégories d'institutions ciblées ou plus) dans 19 cas sur 54, soit une probabilité de 35,19 %.
- Le commanditaire est l'Asie, et l'institution ciblée est « autres » (deux catégories d'institutions ciblées ou plus) dans 69 cas sur 196, soit une probabilité de 35,20 %.

accuracy: 95.00% +/- 9.85% (micro average: 95.03%)

	true Ransomware	true Phishing	true Denial of service	true Espionage	class precision
pred. Ransomware	37	0	0	0	100.00%
pred. Phishing	0	127	0	0	100.00%
pred. Denial of service	8	1	228	2	95.40%
pred. Espionage	1	24	0	296	92.21%
class recall	80.43%	83.55%	100.00%	99.33%	

FIGURE 4.12 – Performance du modèle

Le modèle d'arbres de décision que nous avons entraîné présente une précision de 95,00 %. Cela signifie que le modèle a réussi à prédire correctement le type d'attaque dans 95,00 % des cas évalués. Une précision aussi élevée, de 100 % pour le rançongiciel et l'hameçonnage, indique que le modèle est performant dans sa capacité à classifier ces attaques de manière précise, et presque parfaite pour les autres. Le rappel est également excellent pour le déni de service et presque parfait pour les autres.

Ensuite, nous avons remplacé le continent par pays et regroupé les pays représentant un seuil relatif de moins de 0,04 dans l'attribut que nous appelons « autres pays ». Avec les mêmes données d'entrée que dans le modèle précédent, nous avons construit un nouvel arbre de décision.

```
Sponsor = China
|   Category = C: Denial of service {Ransomware=0, Denial of service=46, Espionage=0, Phishing=0}
|   Category = G: Espionage {Ransomware=0, Denial of service=0, Espionage=104, Phishing=0}
|   Category = M: Espionage {Ransomware=0, Denial of service=0, Espionage=18, Phishing=0}
|   Category = O: Denial of service {Ransomware=0, Denial of service=140, Espionage=0, Phishing=0}
|   Category = P: Denial of service {Ransomware=0, Denial of service=117, Espionage=0, Phishing=0}
Sponsor = Iran
|   Category = C: Ransomware {Ransomware=20, Denial of service=0, Espionage=0, Phishing=0}
|   Category = G: Espionage {Ransomware=0, Denial of service=0, Espionage=20, Phishing=0}
|   Category = M: Ransomware {Ransomware=2, Denial of service=0, Espionage=0, Phishing=0}
|   Category = O: Espionage {Ransomware=0, Denial of service=0, Espionage=56, Phishing=0}
|   Category = P: Ransomware {Ransomware=38, Denial of service=0, Espionage=0, Phishing=0}
Sponsor = North Korea
|   Category = C: Ransomware {Ransomware=19, Denial of service=0, Espionage=0, Phishing=0}
|   Category = G: Espionage {Ransomware=0, Denial of service=0, Espionage=20, Phishing=0}
|   Category = M: Ransomware {Ransomware=3, Denial of service=0, Espionage=0, Phishing=0}
|   Category = O: Ransomware {Ransomware=23, Denial of service=0, Espionage=0, Phishing=0}
|   Category = P: Ransomware {Ransomware=67, Denial of service=0, Espionage=0, Phishing=0}
Sponsor = Other
|   Category = C: Espionage {Ransomware=0, Denial of service=0, Espionage=60, Phishing=0}
|   Category = G: Espionage {Ransomware=0, Denial of service=0, Espionage=38, Phishing=0}
|   Category = M: Denial of service {Ransomware=0, Denial of service=7, Espionage=0, Phishing=0}
|   Category = O: Espionage {Ransomware=0, Denial of service=0, Espionage=49, Phishing=0}
|   Category = P: Espionage {Ransomware=0, Denial of service=0, Espionage=26, Phishing=0}
Sponsor = Russia
|   Category = C: Phishing {Ransomware=0, Denial of service=0, Espionage=0, Phishing=30}
|   Category = G: Phishing {Ransomware=0, Denial of service=0, Espionage=0, Phishing=112}
|   Category = M: Phishing {Ransomware=0, Denial of service=0, Espionage=0, Phishing=7}
|   Category = O: Phishing {Ransomware=0, Denial of service=0, Espionage=0, Phishing=68}
|   Category = P: Ransomware {Ransomware=67, Denial of service=0, Espionage=0, Phishing=0}
```

FIGURE 4.13 – Arbres de décision

Nous allons parcourir une ligne allant de la racine aux feuilles de notre modèle illustré dans la figure 4.13 ci-dessous. Cela exprime une règle de classification. À titre d'exemple, si le pays commanditaire est la Chine et que la catégorie d'institution visée est le secteur privé, il s'agit d'un déni de service dans près de 100 % des cas (117/117). De même, si le commanditaire est la Russie et que la catégorie d'institution visée est le gouvernement, il s'agit d'un hameçonnage dans près de 100 % des cas (112/112). En revanche, si le commanditaire est la Corée du Nord ou l'Iran et que la catégorie d'institution visée est le secteur privé, il s'agit d'un rançongiciel dans près de 100 % des cas (67/67) ou (38/38).

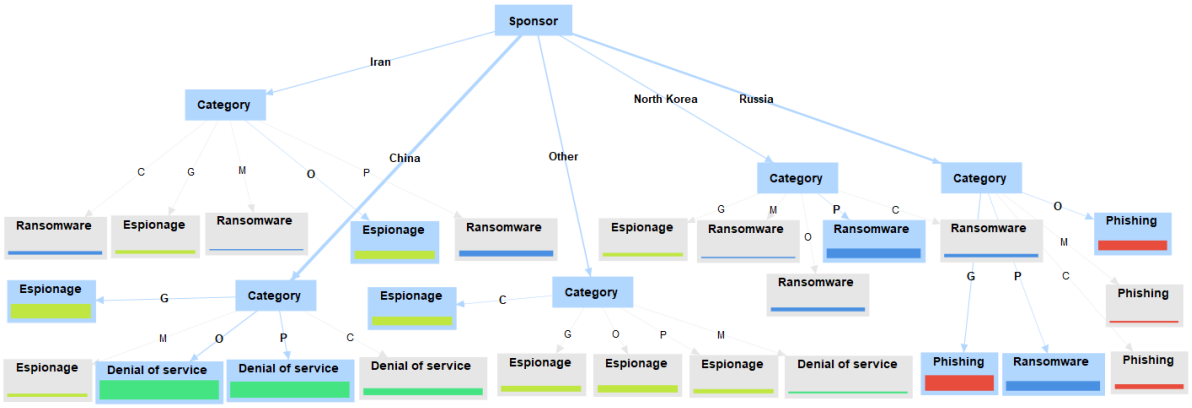


FIGURE 4.14 – Graphique de l'arbre de décision

Nous remarquons quatre couleurs différentes pour représenter chaque type d'attaque en fonction de leur proportion sur l'histogramme du schéma de l'arbre de décision à la figure 4.11. La couleur jaune est attribuée à la classe « espionnage », qui est la classe dominante avec une proportion de 33,8 %. La couleur verte représente la classe « déni de service » (26,8 %), la couleur bleue représente la classe « rançongiciels » (20,7 %), et enfin, la couleur rouge représente la classe « hameçonnage », qui est la classe avec la proportion la plus faible (18,8 %).

Nous pouvons extraire des règles de classification de notre arbre de décision illustré dans la figure 4.14. Par exemple, une attaque est de type déni de service dans les cas suivants :

- Si le pays commanditaire est la Chine et que la cible est la société civile, cela se produit dans 46 cas sur 46, soit une probabilité de 100 %.
- Si le pays commanditaire est la Chine et que la cible est 'autres' (deux ou plusieurs institutions à la fois), cela se produit dans 140 cas sur 140, soit une probabilité de 100 %.
- Si le pays commanditaire est la Chine et que la cible est le secteur privé, cela se produit dans 117 cas sur 117, soit une probabilité de 100 %.
- Si le pays commanditaire est 'autres' (le reste du monde excluant les quatre pays présents dans l'arbre 4.14) et que la cible est une installation militaire, cela se produit dans 7 cas sur 7, soit une probabilité de 100 %.

La performance du modèle est bonne telle que présentée à la figure 4.15

accuracy: 99.39%

	true Ransomware	true Denial of service	true Espionage	true Phishing	class precision
pred. Ransomware	239	0	0	0	100.00%
pred. Denial of service	0	303	0	0	100.00%
pred. Espionage	0	7	391	0	98.24%
pred. Phishing	0	0	0	217	100.00%
class recall	100.00%	97.74%	100.00%	100.00%	

FIGURE 4.15 – Performance du modèle d'arbres de décision

Le modèle d'arbres de décision que nous avons entraîné présente une précision de 99,37 %, ce qui représente une amélioration significative par rapport aux modèles précédents. Cela signifie que le modèle a réussi à prédire correctement le type d'attaque dans 99,37 % des cas évalués. Une précision aussi élevée indique que le modèle est performant dans sa capacité à classifier les attaques de manière précise. La précision est excellente (100 %) pour le rançongiciel, le déni de service et l'hameçonnage, et elle est presque parfaite pour l'espionnage. Le rappel est également excellent (100 %) pour le rançongiciel, l'espionnage et l'hameçonnage, et il est presque parfait pour le déni de service.

4.5 Réseaux bayésiens

Pour ce qui est de l'application de la méthode des réseaux bayésiens, nous conservons les mêmes données que lors de l'utilisation des arbres de décision ayant inclus l'espionnage à la figure 4.13. La précision du modèle *(accuracy)* est de 99 %, les données sont alors bien classées. Voici les résultats obtenus.

Nous tenons à rappeler que, pour l'application de la méthode des réseaux bayésiens, nous utilisons les mêmes données que lors de l'application des arbres de décision, comme indiqué dans la figure 4.13, car nous avions déjà inclus le type espionnage à ces données. La figure 4.16 affiche la performance du modèle ; La précision du modèle (accuracy) est de 99 %, ce qui signifie que les données sont correctement classées.

accuracy: 99.05% +/- 2.02% (micro average: 99.05%)

	true Ransomware	true Denial of service	true Espionage	true Phishing	class precision
pred. Ransomware	228	0	0	0	100.00%
pred. Denial of service	0	310	0	0	100.00%
pred. Espionage	11	0	391	0	97.26%
pred. Phishing	0	0	0	217	100.00%
class recall	95.40%	100.00%	100.00%	100.00%	

FIGURE 4.16 – Performance du modèle réseau bayésien

Le modèle réseau bayésien que nous avons entraîné présente une précision de 99 %. Cela signifie que le modèle a réussi à prédire correctement le type d'attaque dans 99 % des cas évalués. Une précision aussi élevée indique que le modèle est performant dans sa capacité à classifier les attaques de manière précise. Le rappel est excellent (100 %) pour le déni de service, espionnage et hameçonnage ; il est presque parfait pour le rançongiciel. La précision est excellente (100 %) pour le rançongiciel, déni de service et hameçonnage ; elle est presque parfaite pour l'espionnage.

4.5.1 Distribution simple

La distribution des résultats du modèle réseau bayésien pour les classes donne une vue probabiliste des prédictions du modèle, ce qui peut être utile pour la prise de décision, l'évaluation de l'incertitude et la compréhension de la fiabilité des prédictions. Pour notre modèle, elle se présente comme suit :

```
Class Ransomware (0.207)
2 distributions

Class Denial of service (0.268)
2 distributions

Class Espionage (0.338)
2 distributions

Class Phishing (0.188)
2 distributions
```

FIGURE 4.17 – Classes

En observant cette distribution, nous constatons que les classes sont assez équilibrées malgré l'écart qui les sépare. La classe d'espionnage représente la proportion la plus élevée avec 0,338, suivie par la classe du déni de service avec 0,268, puis la classe rançongiciel avec 0,207, tandis que la classe de l'hameçonnage a la proportion la plus faible avec 0,188. Cela signifie qu'il existe un écart d'au moins 15 % entre la classe la plus élevée et la classe la plus faible.

4.5.2 Représentations graphiques

Nous avons retenu quelques graphiques simples pour visualiser la répartition des attaques par catégorie à la figure 4.18 et par commanditaire à la figure 4.19 :

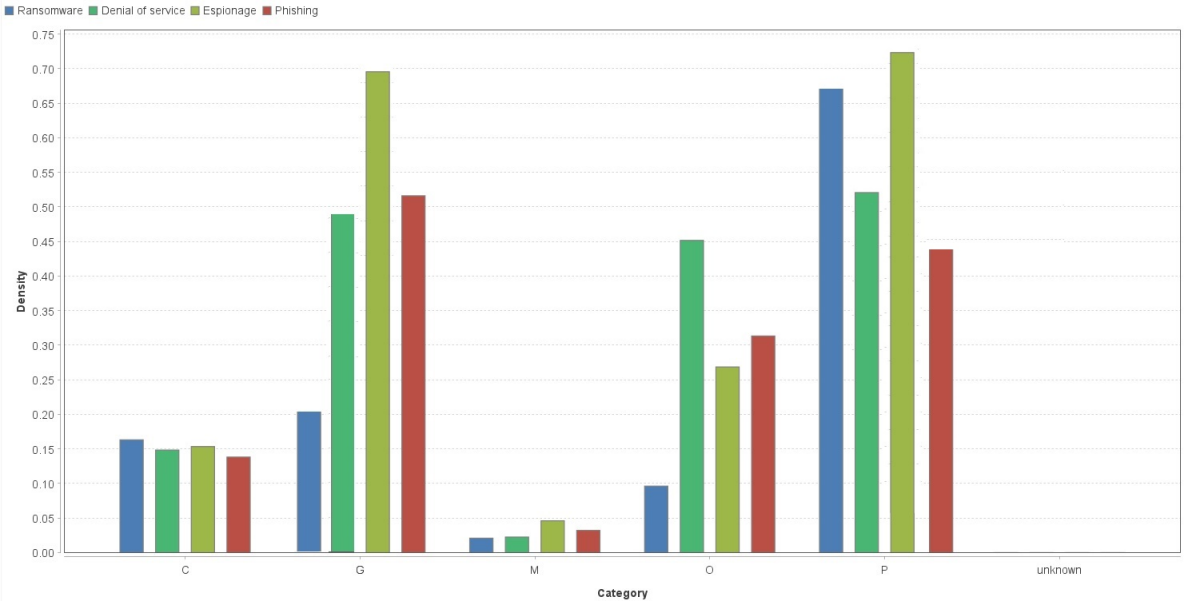


FIGURE 4.18 – Graphique catégorie par type d'attaques

Ce graphique représente les catégories d'institutions ciblées par les attaques. Le bleu représente les attaques de rançongiciel, le vert concerne les attaques de déni de service, le marron couvre les attaques d'espionnage, et le rouge représente les attaques d'hameçonnage.

Nous pouvons observer que le secteur privé (P) est la catégorie la plus ciblée, en commençant par les attaques de type espionnage (marron), suivies des attaques de rançongiciel (bleu), puis de déni de service (vert) et d'hameçonnage (rouge). En ce qui concerne le gouvernement (G), il est principalement visé par l'espionnage, avec notamment une proportion importante d'hameçonnage, suivi des attaques de déni de service, et une proportion non négligeable de rançongiciels.

Nous pouvons tirer de ce graphique la mise en évidence des différences de ciblage entre les secteurs privés et gouvernementaux. En effet, en plus de l'espionnage, les attaques de rançongiciel et l'hameçonnage semblent être plus fréquents dans le secteur privé, tandis que le déni de service et l'hameçonnage sont plus susceptibles de se produire dans le gouvernement.

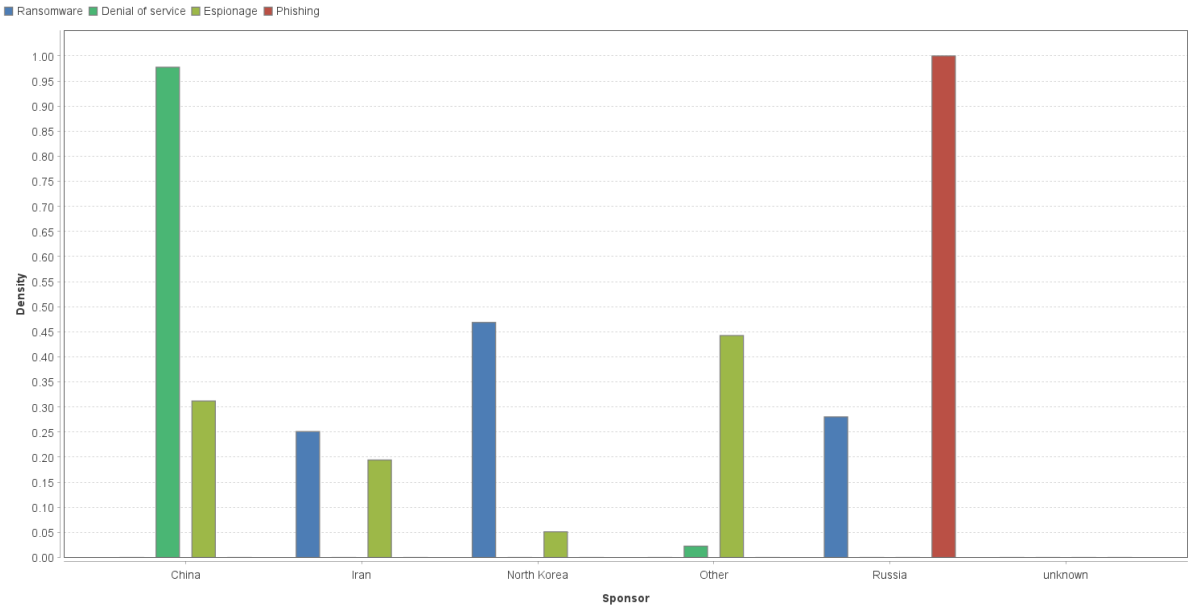


FIGURE 4.19 – Graphique commanditaire par type d'attaques

Sachant que l'espionnage est dominant dans presque tous les pays, nous constatons que la barre la plus haute pour la Chine est verte (déni de service), tandis que pour la Russie, elle est rouge (hameçonnage), et pour la Corée du Nord et l'Iran, la barre la plus haute est bleue (rançongiciel).

En d'autres termes, pour la Chine, le type d'attaque le plus courant après l'espionnage est le déni de service, symbolisé par la couleur verte. En Russie, c'est l'hameçonnage qui est le type d'attaque le plus courant après l'espionnage, symbolisé par la couleur rouge. En ce qui concerne la Corée du Nord et l'Iran, le type d'attaque le plus fréquent après l'espionnage est le rançongiciel, symbolisé par la couleur bleue.

Ces observations montrent que les pays commanditaires ont des préférences et des stratégies différentes en termes de types d'attaques. La Chine se concentre principalement sur les attaques de déni de service, la Russie privilégie l'hameçonnage, tandis que la Corée du Nord et l'Iran sont davantage associés aux attaques de rançongiciel.

Nous allons maintenant appliquer des algorithmes d'apprentissage non supervisé, tels que le regroupement *(clustering)* et les règles d'association.

4.6 Regroupement

En utilisant les méthodes de regroupement, nous cherchons à découvrir des groupes d'attaques similaires. Pour cela, nous avons appliqué un algorithme de regroupement largement utilisé dans la fouille de données et connu sous le nom de k-moyennes *(k-means)*. Cette approche nous aide à mieux comprendre les différentes cyberattaques en fonction de leurs caractéristiques (type, catégorie et commanditaires impliqués).

4.6.1 Méthode k-moyennes

Comme nous l'avons mentionné précédemment, dans la première phase de notre analyse, nous avons choisi de nous concentrer sur cinq attributs, tandis que pour le regroupement, nous retenons trois attributs spécifiques : le type d'attaque, les commanditaires et la catégorie des institutions victimes des cyberattaques. Nous avons conservé l'ensemble complet de données, soit un total de 803 entrées.

Pour faciliter le traitement, nous avons converti toutes les valeurs de ces attributs nominaux en valeurs numériques. Dans le premier regroupement que nous avons effectué, nous avons utilisé k=3, ce qui a permis de former trois groupes distincts. On peut observer cette représentation graphique dans la figure 4.20 avec les groupes ci-dessous.

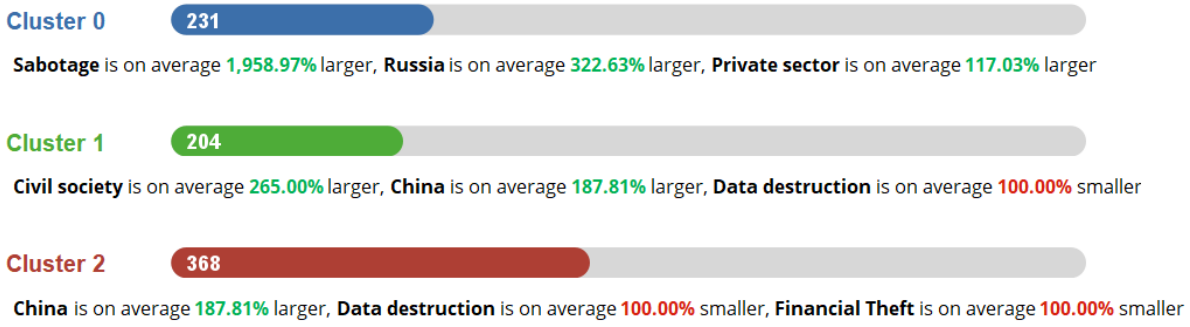


FIGURE 4.20 – Visualisation de groupes

Nous constatons ici que le groupe 2 se distingue en étant majoritaire, avec un total de 368 attaques représentées en rouge à la figure 4.20 :

- **Groupe 0** (en bleu) est composé de 231 attaques. Il se distingue par les caractéristiques suivantes :
  - Type d'attaque : Principalement des attaques de sabotage.
  - Commanditaires : Implication notable de la Russie en tant qu'acteur.
  - Cibles : Principalement des attaques dirigées vers le secteur privé.

  En effet, ce groupe comporte des attaques caractérisées par du sabotage, souvent associées à des acteurs russes parrainés par l'état, et ciblant principalement des entités du secteur privé.
- **Groupe 1** (en vert) comprend 204 attaques et présente les caractéristiques suivantes :
  - Type d'attaque : Une fréquence plus élevée d'attaques contre la société civile.

— Commanditaires : Une forte présence de la Chine en tant qu'acteur.
— Cibles : Moins d'attaques de destruction de données.
Ce groupe suppose que les attaques regroupées ici sont davantage orientées vers des entités de la société civile, impliquent souvent des acteurs chinois parrainés par l'était et ont tendance à éviter la destruction de données.

– **Groupe 2** (en rouge) : Il s'agit du groupe le plus volumineux, comprenant 368 attaques, et il se caractérise par :
— Commanditaires : forte implication de la Chine en tant qu'acteur.
— Type d'attaque : Moins d'attaques de destruction de données.
— Cibles : Moins de vols financiers.
Ce groupe indique que les attaques regroupées ici sont principalement le fait d'acteurs chinois, qu'elles sont moins axées sur la destruction de données et qu'elles sont moins liées à des vols financiers.

Le centroïde[1] de chaque attribut au sein d'un groupe est représenté par la figure 4.21.

Attribute	cluster_0	cluster_1	cluster_2
Data destruction	0	0	0
Espionage	0	1	1
Financial Theft	0	0	0
Sabotage	1	0	0
Denial of service	0	0	0
Russia	1	0	0
China	0	1	1
North Korea	0	0	0
Vietnam	0	0	0
Iran	0	0	0
South Korea	0	0	0
Ukraine	0	0	0
Israel	0	0	0
Civil society	0	1	0
Private sector	1	0	0
Government	0	1	1
Military	0	0	0

FIGURE 4.21 – Centroïde

Le centroïde représente des profils caractéristiques au sein de chaque groupe de notre ensemble de données sur les cyberattaques. Le centroïde est la valeur moyenne ou centrale des attributs au sein de chaque groupe. Voici une interprétation de ces centroïdes f(figure 4.21) :

1. Ici, le centroïde d'un attribut représente le point d'équilibre ou de gravité du cluster k.

- Groupe 0 : Le centroïde montre des valeurs d'attributs définies à 1 pour les attributs « sabotage, » « Russie, » et « secteur privé. ». Cela signifie que ce groupe est principalement caractérisé par des attaques de sabotage, avec la Russie comme commanditaire majeur, et des cibles principalement axées sur le secteur privé. En d'autres termes, ce groupe couvre des attaques de sabotage ciblant des entreprises privées et commanditées par la Russie.
- Groupe 1 : Le centroïde présente des valeurs d'attributs définies à 1 pour les attributs « espionnage, » « société civile, » « Chine, » et « gouvernement. » Cela indique que ce groupe est principalement caractérisé par des attaques d'espionnage impliquant la Chine en tant que commanditaire majeur, et ayant pour cibles des entités liées à la société civile et au gouvernement. En d'autres termes, ce groupe représente des attaques d'espionnage visant la société civile et le gouvernement, avec la Chine comme commanditaire principal.
- Groupe 2 : Le centroïde montre des valeurs d'attributs définies à 1 pour les attributs « espionnage, » « Chine, » et « gouvernement. ». Cela indique que ce groupe partage des caractéristiques similaires avec le groupe 1, à savoir des attaques d'espionnage avec la Chine comme commanditaire principal, mais avec le gouvernement comme cible unique.

La figure 4.22 illustre la répartition des groupes sous forme d'une carte. L'intensité des couleurs indique l'impact de l'attaque sur le secteur ciblé.

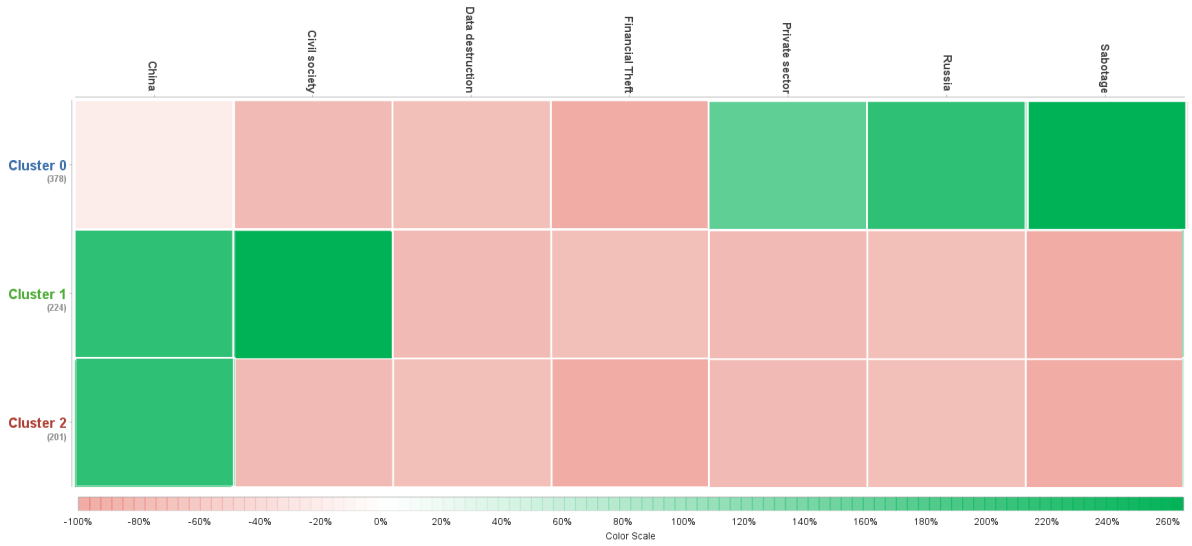


FIGURE 4.22 – Carte graphique

Grâce à la carte de chaleur *(Heat map)* [38], RapidMiner permet de visualiser les groupes avec leurs principales caractéristiques et les différences significatives entre eux. Par exemple, le groupe 0 présente en moyenne des valeurs beaucoup plus élevées pour les variables « sabotage », « Russie », et « secteur privé », affichées en vert. Le groupe 1 contient en moyenne des valeurs beaucoup plus élevées pour la Chine et la société civile, tandis que pour le groupe 2, nous observons des valeurs élevées uniquement pour la variable Chine.

Dans le second regroupement que nous avons effectué, nous avons utilisé k=4, ce qui a permis de former quatre groupes distincts. On peut observer cette représentation graphique des groupes à la figure 4.23.

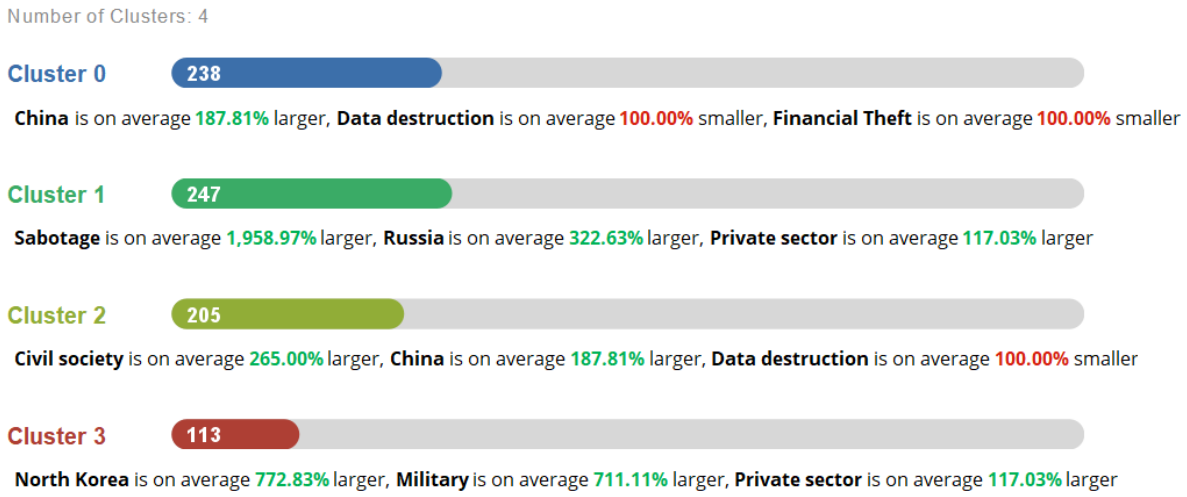


FIGURE 4.23 – Visualisation des groupes

- Le groupe-0 est prédominé par les attaques commanditées par la Chine qui peuvent en minorité être du vol financier ou la destruction de données.
- Dans le groupe-1, les attaques sont attribuées principalement à la Russie qui conduit majoritairement du sabotage et cible principalement le secteur privé.
- Dans le groupe-2, les attaques ciblent principalement la société civile, en minorité de type destruction de données, attribuées principalement à la Chine.
- Dans le groupe-3, la Corée du Nord prédomine les attaques dont la cible est majoritairement soit les installations militaires ou la société civile.

Les caractéristiques de chaque groupe sont représentées dans la figure 4.24.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Data destruction	0	0	0	0
Financial Theft	0	0	0	0
Sabotage	0	1	0	0
Denial of service	0	0	0	0
Russia	0	1	0	0
China	1	0	1	0
North Korea	0	0	0	1
South Korea	0	0	0	0
Ukraine	0	0	0	0
Civil society	0	0	1	0
Private sector	0	1	0	1
Military	0	0	0	1

FIGURE 4.24 – Centroïde pour k=4

- Le centroïde du groupe 0 présente une valeur d'attributs définie à 1 pour l'attribut « Chine ». Cela suppose que ce groupe est composé d'attaques provenant majoritairement de la Chine.
- Le centroïde du groupe 1 présente des valeurs d'attributs définies à 1 pour les attributs « sabotage, » « Russie, » et « secteur privé », . Cela indique que ce groupe est constitué majoritairement par les attaques attribuées à la Russie, de type sabotage et ayant un majeur impact sur le secteur privé.
- Dans le groupe 2, le centroïde présente des valeurs d'attributs définies à 1 pour les attributs « Chine, » et « société civile », ce qui signifie qu'il est composé d'attaques attribuées minoritairement à la Chine ayant un impact beaucoup plus significatif sur la société civile.
- Le centroïde du groupe 3 présente des valeurs d'attributs définies à 1 pour les attributs « Corée du Nord, » « installations militaires, » et « secteur privé, ». En d'autres mots, ce groupe est prédominé par les attaques attribuées à la Corée du Nord et ayant un majeur impact sur les installations militaires et aussi le secteur privé.

La figure 4.25 illustre la répartition des groupes sous forme d'une carte et montre le degré d'importance que peut avoir un attribut dans le groupe. Cette figure illustre seulement ce qui est déjà présenté au niveau du centroïde à la figure 4.24.

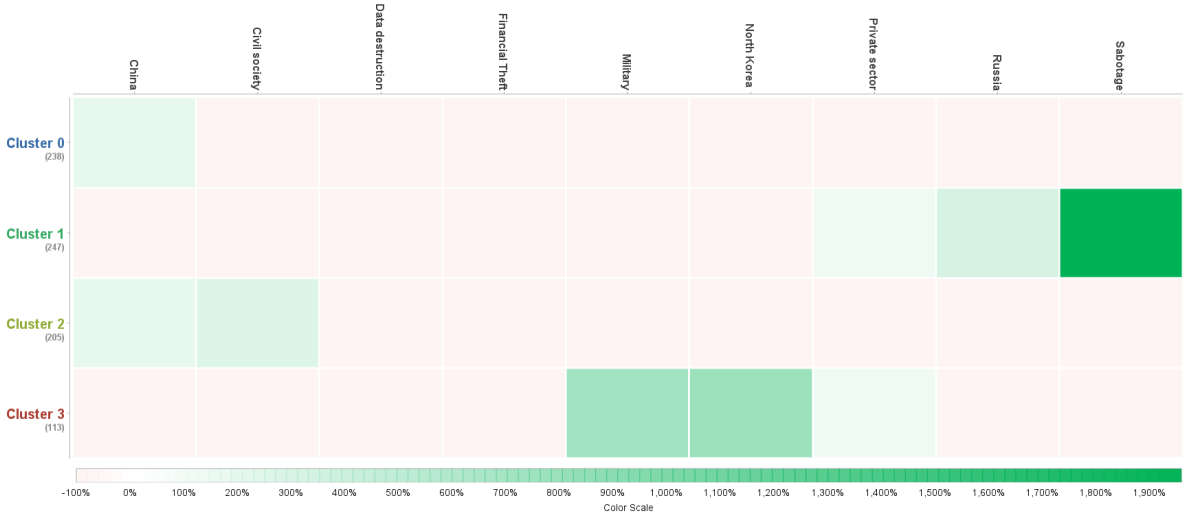


FIGURE 4.25 – Carte graphique

On peut observer que la forte présence d'attaques se trouve dans le groupe 1 des attributs « sabotage et Russie » en couleur verte et la couleur marron marque l'absence de la Chine par exemple dans ce groupe.

Nous allons maintenant changer le pays commanditaire par la sous-région commanditaire. Cela nous a permis d'effectuer un regroupement sous-régional des 789 attaques présenté à la figure 4.26.

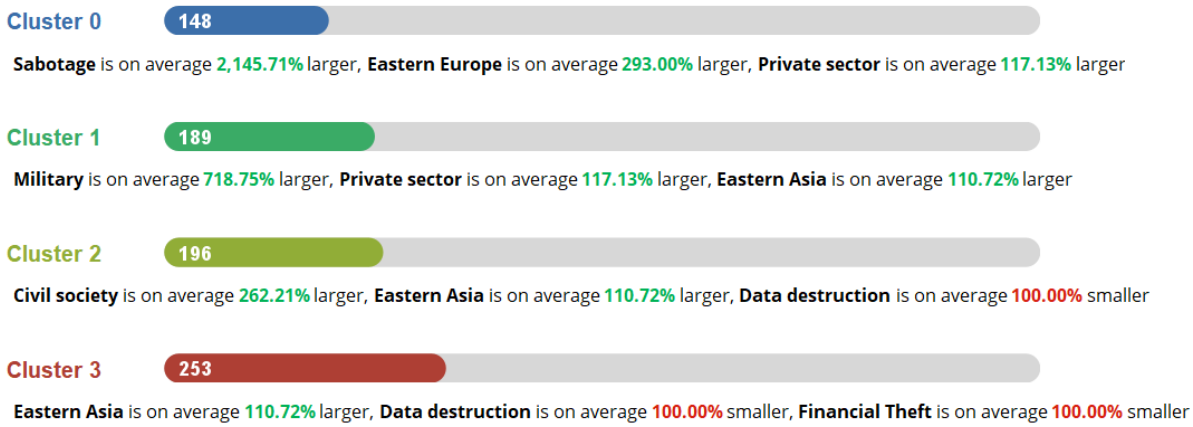


FIGURE 4.26 – Le regroupement sous-régional k=4

On observe ici que le groupe dominant est le groupe 3 avec 253 attaques alors que les autres détiennent une centaine chacun (figure 4.26).

- Le groupe 0 est constitué des attaques en majorité de type sabotage, provenant de la sous-région Est-Européen, et visant principalement le secteur privé.
- Le groupe 1 est constitué des attaques provenant en majeur partie de la sous-région est-asiatique et ciblant principalement les installations militaires et le secteur privé.
- Le groupe 2 est constitué des attaques provenant en majeur partie de la sous-région est-asiatique, de type destruction de données en majorité et ciblant principalement les entités de la société civile.
- Le groupe 3 est constitué des attaques provenant en majeur partie de la sous-région est-asiatique pouvant être majoritairement de type destruction de données ou de vol financier.

Cette réalité est montrée par le centroïde qui est présenté à la figure 4.27.

Cluster	Data dest...	Espionage	Financia...	Sabotage	Eastern Euro...	Denial of ser...	Southern Asia	Northern ...	Southern Eur...	Civil so...	Private s...	Govern...	Military	Eastern Asia	Northern Am
Cluster 0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
Cluster 1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	0
Cluster 2	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0
Cluster 3	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0

FIGURE 4.27 – Carte graphique pour le regroupement sous-régional

Nous remarquons que :

- Le centroïde du groupe 0 présente des valeurs d'attributs définies à 1 pour les attributs « sabotage, » « Est-Europe, » et « secteur privé ». Cela signifie que ce groupe est composé d'attaques de type sabotage provenant majoritairement de la sous-région Est-Européenne ayant un impact significatif sur le secteur privé.
- Le centroïde du groupe 1 présente des valeurs d'attributs définies à 1 pour les attributs « espionnage, » « secteur privé, » « installations militaires, » et « Asie de l'est, ». Cela signifie que ce groupe est

composé d'attaques de type espionnage commanditées en majeure partie dans la sous-région de l'Asie de l'Est et ciblant principalement le secteur privé ainsi que les installations militaires.

- Le centroïde du groupe 2 présente des valeurs d'attributs définies à 1 pour les attributs « espionnage, » « société civile, » « gouvernement, » et « Asie de l'Est, ». Cela indique que ce groupe est composé d'attaques de type espionnage commanditées en majeure partie dans la sous-région de l'Asie de l'Est et ayant pour cibles principales le gouvernement et la société civile.
- Dans le groupe 3, le centroïde présente des valeurs d'attributs définies à 1 pour les attributs « espionnage, » « gouvernement, » et « Asie de l'est, ». Cela indique que le groupe 3 partage des caractéristiques similaires au groupe 2, à savoir des attaques de type espionnage, dont le principal commanditaire est la sous-région de l'Asie de l'Est, mais la simple nuance est que dans ce dernier groupe la cible principale est tout simplement le gouvernement.

On peut valider toutes ses informations en observant le graphique présenté à la figure 4.28 nommé graphique sous-régional.

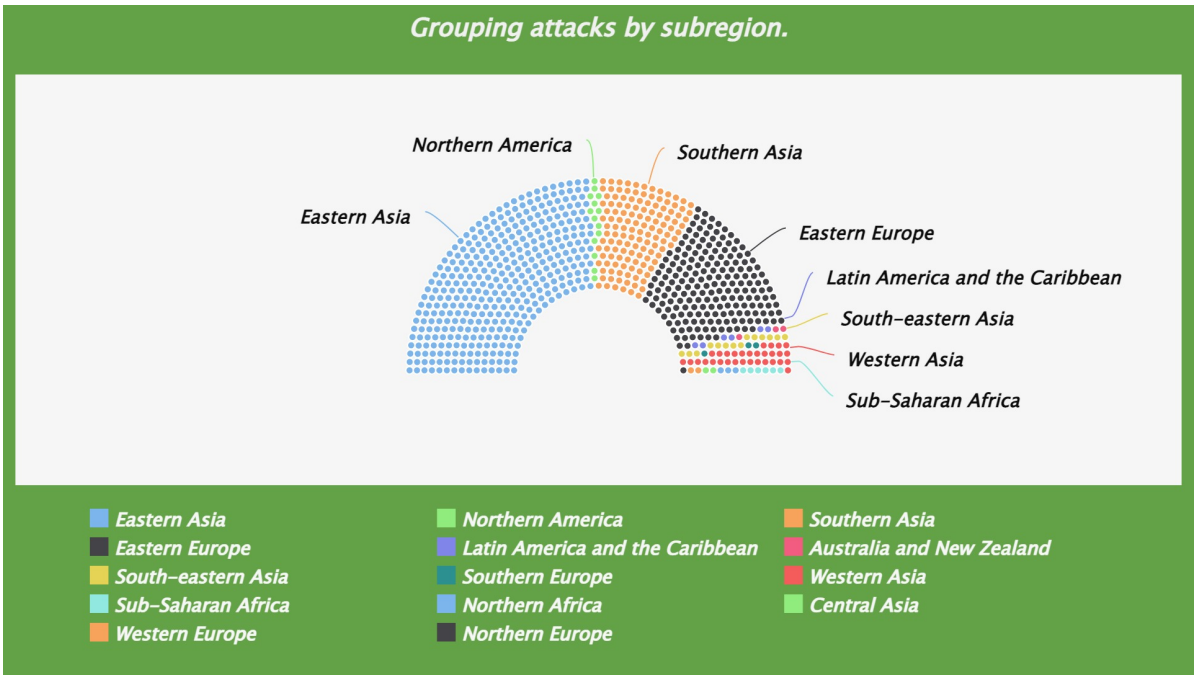


FIGURE 4.28 – Graphique sous régional

Notre objectif est de déterminer la sous-région qui commandite le plus grand nombre d'attaques dans le monde. Le processus de regroupement des attaques et le graphique sous-régional (voir la figure 4.28) révèlent les faits suivants :

- La grande majorité des attaques commanditées en Asie proviennent de la sous-région de l'est, occupant ainsi la plus grande proportion mondiale avec un total d'au moins 381 attaques sur 789.
- En Europe, la sous-région de l'est se démarque comme étant la plus active en termes de cyberattaques commanditées, se classant deuxième à l'échelle mondiale avec 192 attaques sur 789.

- L'Asie du Sud joue un rôle significatif en tant que l'un des principaux commanditaires occupant la troisième place mondiale avec 131 attaques.
- Même si son nombre d'attaques commanditées est relativement faible, seulement 14, l'Amérique du Nord continue de commanditer des attaques, ce qui est surprenant étant donné sa position en bas du classement mondial, derrière la sous-région de l'Ouest asiatique, qui en compte 31. Néanmoins, l'Amérique du Nord conserve la cinquième place au classement mondial.
- L'Amérique latine et l'Afrique subsaharienne sont les moins actives, avec chacune moins de 10 attaques sur les 789.

Il est important de noter que la vigilance à l'égard de chaque sous-région est essentielle, car les tendances peuvent rapidement évoluer en raison des avancées technologiques actuelles.

Ainsi, une fois le regroupement terminé, nous allons présenter la matrice de corrélation avant d'aborder les règles d'association.

4.7 Matrice de corrélation

Nous avons effectué une transformation de nos données d'entrée pour l'obtention de la matrice de corrélation ci-après. Nous avons regroupé les attaques de type vol financier et sabotage sous le type « Rançongiciel », et les types « doxage », « destruction de données » et « dégradation » sous le type « Hameçonnage ».

En ce qui concerne l'attribut « catégorie », nous avons combiné la catégorie « société civile » avec la catégorie « secteur privé », tandis que la catégorie « militaire » a été ajoutée à la catégorie « gouvernement ». Ainsi, nous avons obtenu deux catégories distinctes : le secteur privé et le gouvernement. Cela nous permet d'effectuer une analyse sur un nombre réduit d'attributs par un processus de généralisation et d'obtenir un autre type de règles d'association dites généralisées comme indiqué par la suite.

Attributes	Russian...	Korea	United States	Iran	Ukraine	China	Israel	Government	Private sector	Gouv-Private	Private-Gouv	Phishing	Ransomware	Denial of service
Russian Federation	1	-0.262	-0.160	-0.248	-0.118	-0.329	-0.074	0.231	-0.240	0.078	-0.037	0.286	-0.019	-0.226
Korea	-0.262	1	-0.102	-0.158	-0.075	-0.209	-0.047	-0.177	0.156	0.029	-0.075	-0.182	0.232	-0.064
United States	-0.160	-0.102	1	-0.096	-0.046	-0.128	-0.029	0.195	-0.071	-0.107	-0.046	0.059	0.064	-0.110
Iran	-0.248	-0.158	-0.096	1	-0.071	-0.198	-0.044	-0.117	0.051	-0.006	0.147	0.015	0.021	-0.032
Ukraine	-0.118	-0.075	-0.046	-0.071	1	-0.094	-0.021	-0.012	-0.069	0.123	-0.034	-0.101	-0.049	0.132
China	-0.329	-0.209	-0.128	-0.198	-0.094	1	-0.059	-0.163	0.172	0.001	-0.094	-0.169	-0.253	0.383
Israel	-0.074	-0.047	-0.029	-0.044	-0.021	-0.059	1	0.075	-0.020	-0.049	-0.021	-0.063	0.043	0.013
Government	0.231	-0.177	0.195	-0.117	-0.012	-0.163	0.075	1	-0.643	-0.233	-0.099	0.176	-0.078	-0.077
Private sector	-0.240	0.156	-0.071	0.051	-0.069	0.172	-0.020	-0.643	1	-0.510	-0.217	-0.126	0.219	-0.099
Gouv-Private	0.078	0.029	-0.107	-0.006	0.123	0.001	-0.049	-0.233	-0.510	1	-0.079	0.016	-0.183	0.159
Private-Gouv	-0.037	-0.075	-0.046	0.147	-0.034	-0.094	-0.021	-0.099	-0.217	-0.079	1	-0.101	-0.049	0.132
Phishing	0.286	-0.182	0.059	0.015	-0.101	-0.169	-0.063	0.176	-0.126	0.016	-0.101	1	-0.383	-0.491
Ransomware	-0.019	0.232	0.064	0.021	-0.049	-0.253	0.043	-0.078	0.219	-0.183	-0.049	-0.383	1	-0.617
Denial of service	-0.226	-0.064	-0.110	-0.032	0.132	0.383	0.013	-0.077	-0.099	0.159	0.132	-0.491	-0.617	1

FIGURE 4.29 – Matrice de corrélation

Nous rappelons que la matrice de corrélation fournit le degré de corrélation (compris entre -1 et 1) linéaire entre chaque paire de variables. À titre d'exemple, la valeur de corrélation de 0.383 entre

le type d'attaque « Déni de service » et le commanditaire « Chine » est la plus élevée dans la matrice et indique une corrélation linéaire positive assez faible illustrant le fait que les attaques commanditées par la Chine et de type « Déni de service » ont tendance à augmenter et à diminuer légèrement dans le même sens. Par contre, une valeur négative du coefficient de corrélation de -0.226 entre ce même type d'attaque et la « Russie » indique qu'il existe une corrélation linéaire négative faible et signifie que l'augmentation d'attaques par la Russie implique une baisse d'attaques de type « Déni de service ». En outre, il existe une corrélation linéaire positive faible entre le type d'attaque « Rançongiciel » et le commanditaire « Corée » tandis qu'il y a un coéfficient de corrélation linéaire négative faible entre ce type d'attaque et le commanditaire « Chine ».

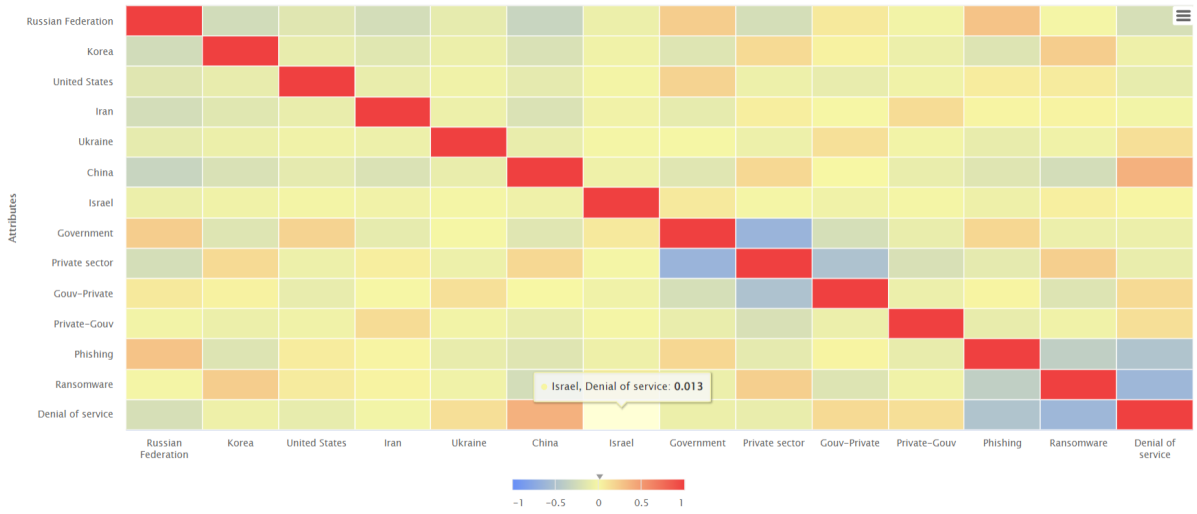


FIGURE 4.30 – La visualisation de la matrice

La visualisation de cette matrice présente quatre couleurs différentes dont le bleu, le jaune, le marron et le rouge. Ces couleurs ont différentes intensités qui représentent les degrés de corrélation entre deux variables. Tout en se référant à la matrice de corrélation à la figure 4.29, on peut observer ce qui suit :

- La couleur jaune indique une corrélation faible et le 0 au centre de la fine barre signifie qu'il n'y a pas de corrélation. Par exemple, la valeur 0.013 entre Israël et le déni de service signifie qu'Israël n'effectue pas d'attaques de déni de services.
- La couleur bleue représente une corrélation linéaire négative. Plus elle est foncée, plus elle se rapproche de -1. Par exemple, plus le nombre d'attaques de type déni de service augmente, plus le nombre d'attaques de type rançongiciel avec une valeur de corrélation de -0.62.
- La couleur marron représente une corrélation positive. Un coefficient de corrélation de 1 indique une corrélation positive parfaite, tandis qu'une valeur proche de 0 indique une très faible corrélation.

  Bien qu'elle soit relativement faible (0.389), il existe une corrélation linéaire positive entre le commanditaire Chine et le déni de service comme expliqué précédemment.
- La couleur rouge indique une corrélation de 1 au niveau de la diagonale de la matrice.

4.7.1 Règles d'association

Pour les règles d'association, nous utilisons les données transformées précédemment, c'est-à-dire les mêmes transformations que celles utilisées pour la matrice de corrélation.

FIGURE 4.31 – Arbre de croissance des motifs fréquents

Les règles d'association que nous avons obtenues se lisent de la manière suivante (cf. les figures 4.31 et 4.32). La règle 27 indique que si l'attaque est commandée par la Chine et qu'elle cible à la fois le gouvernement et le secteur privé, alors il s'agit d'une attaque de déni de service, avec un support faible de 3 % et une confiance de 100 %. La règle 26 stipule que si l'attaque est de type vol financier et qu'elle est commandée par la Corée du Nord, alors elle vise le secteur privé, avec un support de 8,4 % et une confiance de 100 %. Ces deux règles sont donc des implications.

4.7.2 Description de règles d'association

```
[Denial of service, Korea] --> [Private sector] (confidence: 0.500)
[Denial of service,  Iran ] --> [Private sector] (confidence: 0.500)
[Sabotage] --> [Russian Federation] (confidence: 0.528)
[Denial of service] --> [Private sector] (confidence: 0.529)
[ United States] --> [Government] (confidence: 0.556)
[Private sector, Denial of service] --> [ China] (confidence: 0.556)
[Financial Theft, Russian Federation] --> [Private sector] (confidence: 0.571)
[Korea] --> [Financial Theft] (confidence: 0.591)
[Korea] --> [Private sector, Financial Theft] (confidence: 0.591)
[Sabotage, Government] --> [Russian Federation] (confidence: 0.615)
[ Government, Private sector] --> [Denial of service] (confidence: 0.625)
[ China] --> [Private sector, Denial of service] (confidence: 0.625)
[ Iran ] --> [Private sector] (confidence: 0.650)
[Sabotage,  Government, Private sector] --> [Russian Federation] (confidence: 0.667)
[Financial Theft,  Iran ] --> [Private sector] (confidence: 0.714)
[Financial Theft] --> [Private sector] (confidence: 0.740)
[ China] --> [Private sector] (confidence: 0.750)
[Private sector, Korea] --> [Financial Theft] (confidence: 0.765)
[Denial of service,  China] --> [Private sector] (confidence: 0.769)
[Korea] --> [Private sector] (confidence: 0.773)
[ Ukraine] --> [Denial of service] (confidence: 0.800)
[Private sector, Government] --> [Denial of service] (confidence: 0.800)
[Sabotage,  Iran ] --> [Private sector] (confidence: 0.800)
[ China] --> [Denial of service] (confidence: 0.812)
[Private sector,  China] --> [Denial of service] (confidence: 0.833)
[Financial Theft, Korea] --> [Private sector] (confidence: 1.000)
[ China,  Government, Private sector] --> [Denial of service] (confidence: 1.000)
```

FIGURE 4.32 – Règles d'association

Ici, nous pouvons voir qu'une règle d'association est simplement représentée par une ligne dans la description, contenant sa prémisse, sa conclusion et sa confiance. Dans le cas de la dernière règle 27, la prémisse est « La Chine, gouvernement, secteur privé » et la conclusion est « Le déni de service ».

Dans ce qui suit, nous présentons la qualité de quelques règles d'association.

No.	Premises	Conclusion	Support	Confide...	Lift
26	Financial Theft, Korea	Private sector	0.084	1	1.711
27	China, Government, Private sector	Denial of service	0.032	1	2.265
25	Private sector, China	Denial of service	0.130	0.833	1.887
24	China	Denial of service	0.169	0.812	1.840
21	Ukraine	Denial of service	0.026	0.800	1.812
22	Private sector, Government	Denial of service	0.026	0.800	1.812
23	Sabotage, Iran	Private sector	0.026	0.800	1.369
20	Korea	Private sector	0.110	0.773	1.322
19	Denial of service, China	Private sector	0.130	0.769	1.316
18	Private sector, Korea	Financial Theft	0.084	0.765	2.355
17	China	Private sector	0.156	0.750	1.283
16	Financial Theft	Private sector	0.240	0.740	1.266
15	Financial Theft, Iran	Private sector	0.032	0.714	1.222
14	Sabotage, Government, Private sector	Russian Federation	0.026	0.667	2.281
13	Iran	Private sector	0.084	0.650	1.112
11	Government, Private sector	Denial of service	0.097	0.625	1.415
12	China	Private sector, Denial of service	0.130	0.625	2.674

FIGURE 4.33 – Données de sortie pour le règles d'association

Nous avons utilisé la mesure appelée l'intérêt pour le calcul des règles d'association, tel que nous pouvons le constater dans le tableau de données présenté dans la figure 4.33. Nous prenons par exemple les règles 26 et 27 que nous avons examinées dans la sous-section « Règles d'association » (voir Figure 4.31).

Pour la règle 27, le *lift* ou l'intérêt est de 2. 265, ce qui signifie qu'il existe une corrélation positive entre le fait que la Chine commande une attaque qui cible à la fois le gouvernement et le secteur privé et le fait que cette attaque soit de type déni de service. Quant à la règle 26, l'intérêt est de 1. 711, ce qui signifie qu'il existe une corrélation positive moins forte entre le fait qu'une attaque de type vol financier soit commanditée par La Corée du Nord et le fait que la cible soit le secteur privé.

4.8 Implications avec négation

La production des règles d'association avec négation dans le contexte de l'analyse des cyberattaques permet de définir des règles telles que : « Si le commanditaire d'une attaque est A, alors il n'y a pas de vol financier ». La création de ces règles contenant des éléments négatifs permet de repérer de nouveaux motifs indiquant l'absence de propriétés d'instances en présence d'autres propriétés comme par exemple l'achat de pain n'entraîne jamais l'achat de craquelins. Dans cette section, nous allons nous servir des outils *Lattice Miner* [31] et *conExp* (*Concept Explorer*) [2] qui, partant d'un contexte (tableau) binaire

2. Disponible à l'adresse suivante : http ://conexp.sourceforge.net/

décrivant des objets et leurs attributs, permettent de produire des groupes (*clusters*) sous forme de nœuds d'un treillis de concepts formels ainsi que des règles d'association. Ensuite, nous présentons les résultats obtenus.

4.8.1 *Lattice Miner* et *ConExp*

Lattice Miner est un logiciel d'analyse formelle de concepts [10] disponible sur GitHub et développé en Java au laboratoire LARIM. Tout comme ConExp [3], il permet, entre autres, la construction, la visualisation et l'exploration d'un treillis de concepts ainsi que la production de règles d'association (y compris des implications).

Les treillis de concepts sont souvent complexes, même pour des contextes de petite échelle. Ces outils visent à simplifier l'utilisation et la visualisation de ces treillis en mettant en évidence les nœuds jugés pertinents pour l'utilisateur.

Dans la figure 4.34 , nous exposons une partie des données sous forme d'un tableau binaire. Cette portion de données comprend 789 incidents d'attaques avec trois variables de base et leurs modalités : (i) le commanditaire, où nous avons remplacé le pays par la région ou le continent ; (ii) la cible avec quatre catégories d'institutions ciblées, à savoir le secteur privé, le gouvernement, la société civile et la cible militaire, et finalement (iii) le type d'attaques avec le vol financier et sabotage regroupés sous le libellé « Rançongiciel », tandis que les types « doxage », « destruction de données » et « dégradation » sont rassemblés sous l'intitulé « Hameçonnage » et à cela nous ajoutons l'espionnage et le déni de service.

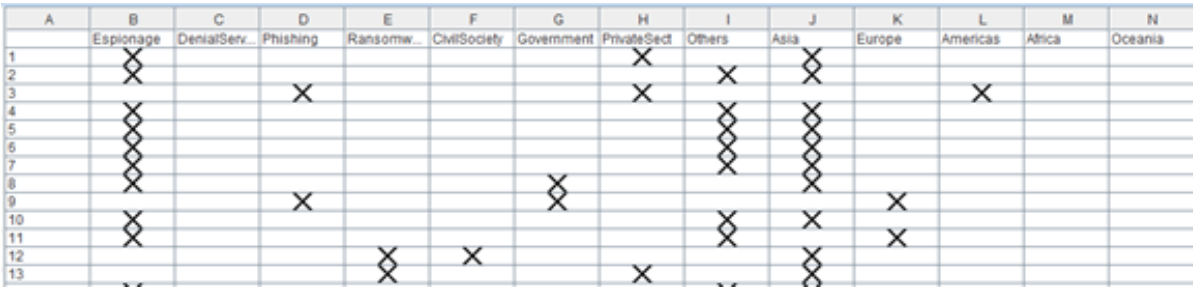


FIGURE 4.34 – Création d'un contexte binaire avec *Lattice Miner*

Les données binaires extraites du tableau présenté dans la figure 4.34 ont été en premier lieu manipulées avec *Concept Explorer*.

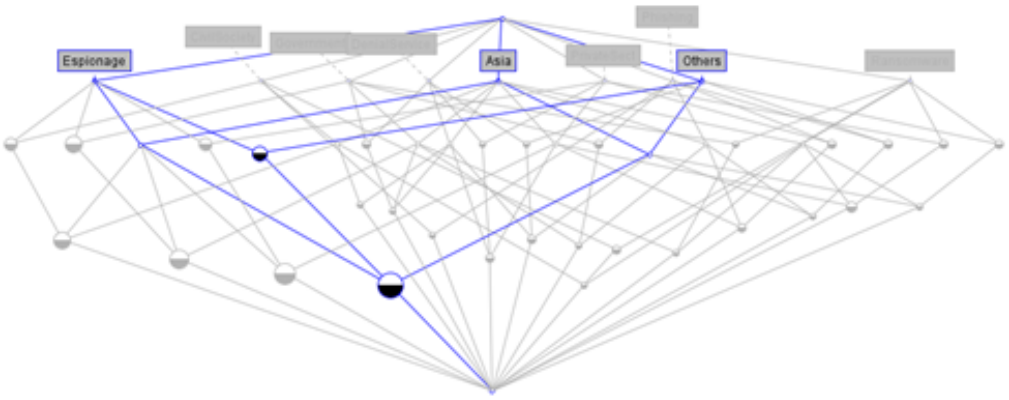


FIGURE 4.35 – Génération du treillis de concepts

Nous pouvons observer dans la figure 4.35 qu'il existe un nœud qui est le plus grand possible, correspondant à 195 attaques de type « espionnage ». Cela concerne la catégorie d'institution ciblée « autre (Other) » comprenant les installations militaires et les cibles simultanées, c'est-à-dire deux ou plusieurs catégories d'institutions à la fois) et impliquant comme commanditaire la région « Asie. »

Avant d'aborder les implications avec négation, expliquons trois implications (parmi cinquante) produites par *ConExp* ayant un support absolu non nul mis entre <> :

- $< 9 >$ *Africa ==> Espionage Civil society* . Cette implication signifie que chaque fois que la région commanditaire de l'attaque est l'Afrique, il s'agit de l'espionnage ciblant la société civile dans 9 cas parmi l'ensemble des 789 attaques.
- $< 2 >$ *DenialService Americas ==> Government*. Cela indique que chaque fois que le type d'attaque est un déni de service et que la région commanditaire est l'Amérique, il s'agit du gouvernement qui est visé avec un support absolu très faible de 2 sur 789.
- $< 29 >$ *Asia Civil society Government Private secteur ==> Espionage*. Cela signifie qu'à chaque fois que la région communautaire est « Asie » et que les cibles sont à la fois société civile, gouvernement et secteur privé, il s'agit d'une attaque de type espionnage dans 29 cas sur 789.

Le fait d'avoir une implication dont le support est nul signifie l'absence du groupe de propriétés dans la prémisse de l'implication. Cela permet alors d'extraire |A| implications avec négation à partir de $A \rightarrow B$. À titre d'exemple, si $A = \{a, b, c\}$, alors on obtient $a, b \rightarrow \neg c$, $a, c \rightarrow \neg b$, et $b, c \rightarrow \neg a$.

Voici quelques exemples :

- L'implication $< 0 >$ *Ransomware Asia Military ==> Phishing DenialService Espionage Europe Africa Americas Civil society Government Private secteur* indique qu'il n'y a jamais eu d'attaques de type vol financier dont la région commanditaire est l'Asie et qui vise une installation militaire. Cela permet de générer trois implications avec négation. L'une d'elles signifie que le vol financier commandité par l'Asie ne concerne jamais une cible militaire.
- L'implication $< 0 >$ *Ransomware Americas Government ==> Phishing DenialService Espionage Asia Europe Africa Civil society Private secteur Military* indique qu'il n'y a pas eu d'attaques

de type rançongiciel, commanditées par l'Amérique et ciblant le gouvernement. Elle permet de générer trois implications avec négation. L'une d'elles spécifie que l'attaque par rançongiciel par l'Amérique **ne cible pas** le gouvernement. Une autre implication indique que s'il y a des attaques par rançongiciel ciblant le gouvernement, cela n'est pas commandité par l'Amérique.

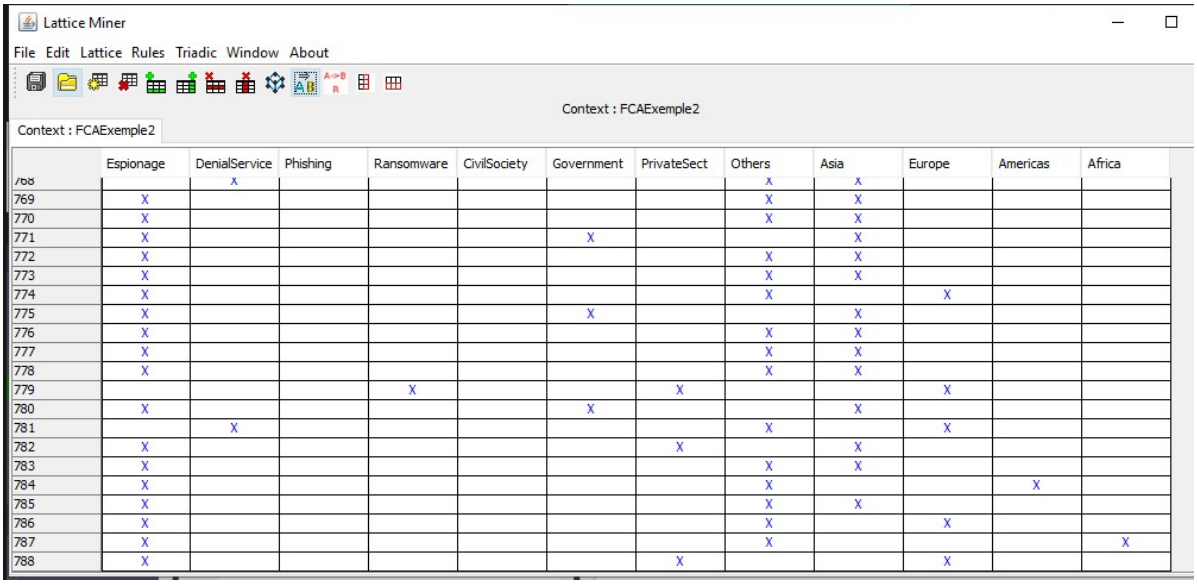
Prenons maintenant une confiance d'au moins 74 % et un support d'au moins 10 % et reprenons les processus de génération des règles d'association. Nous trouvons 14 règles :

- *1 < 76 > CivilSociety Asia =[96 %]=> < 73 > Espionage ;*
- ***2 < 109 > Government Asia =[94 %]=> < 103 > Espionage ;***
- *3 < 208 > Others Asia =[94 %]=> < 195 > Espionage ;*
- *4 < 559 > Asia =[90 %]=> < 503 > Espionage ;*
- *5 < 275 > Others =[90 %]=> < 247 > Espionage ;*
- *6 < 112 > CivilSociety =[88 %]=> < 99 > Espionage ;*
- *7 < 189 > Government =[87 %]=> < 164 > Espionage ;*
- ***9 < 162 > Espionage PrivateSect =[81 %]=> < 132 > Asia ;***
- *10 < 166 > PrivateSect Asia =[80 %]=> < 132 > Espionage ;*
- *11 < 58 > Others Europe =[79 %]=> < 46 > Espionage ;*
- *12 < 29 > Ransomware PrivateSect =[79 %]=> < 23 > Asia ;*
- *13 < 76 > Government Europe =[79 %]=> < 60 > Espionage ;*
- *14 < 247 > Espionage Others =[79 %]=> < 195 > Asia ;*

Nous expliquons ci-après les règles identifiées en gras :

- *< 109 > Government Asia =[94 %]=> < 103 > Espionage* signifie qu'à chaque fois que nous avons une attaque ciblant le gouvernement et que la région commanditaire est l'Asie, il s'agit d'espionnage avec une confiance de 94 % (= 103/109) et un support de 15 % (= 109/789).
- *< 162 > Espionage PrivateSect =[81 %]=> < 132 > Asia* indique qu'à chaque fois qu'il y a une attaque de type espionnage qui cible le secteur privé , alors la région commanditaire est l'Asie avec une confiance de 81 % (= 132/162) et un support de 20 % (= 162/789).

Revenons à *Lattice Miner* et affichons de nouveau nos données sous format binaire.



	Espionage	DenialService	Phishing	Ransomware	CivilSociety	Government	PrivateSect	Others	Asia	Europe	Americas	Africa
768		X						X	X			
769	X							X	X			
770	X							X	X			
771	X					X			X			
772	X							X	X			
773	X							X	X			
774	X							X		X		
775	X					X			X			
776	X							X	X			
777	X							X	X			
778	X							X	X			
779				X			X			X		
780	X					X			X			
781		X						X		X		
782	X						X		X			
783	X							X	X			
784	X							X			X	
785	X							X	X			
786	X							X		X		
787	X							X				X
788	X						X			X		

FIGURE 4.36 – Création d'un contexte binaire avec Lattice Miner

Dans l'exemple de la figure 4.36, le contexte est composé de 788 objets et de 13 attributs lesquels sont nommés respectivement : espionnage, déni de service, hameçonnage, rançongiciel, société civile, gouvernement, secteur privé, autres, Asie, Europe, Amérique et Afrique.

Avec un support minimum de 10% et une confiance d'au moins 70%, nous obtenons 14 règles (comme avec *ConExp*) indiquées par la figure 4.37.

Context : FCAExemple2

Min. support : 10.0%

Min. confidence : 70.0%

Rule count : 14

#	Antecedent	=>	Consequence	Support	Confidence
1.	{Asia, Government}	=>	{Espionage}	13.07%	94.49%
2.	{Asia, Others}	=>	{Espionage}	24.74%	93.75%
3.	{Asia}	=>	{Espionage}	63.83%	89.98%
4.	{Others}	=>	{Espionage}	31.34%	89.81%
5.	{CivilSociety}	=>	{Espionage}	12.56%	88.39%
6.	{Government}	=>	{Espionage}	20.81%	86.77%
7.	{Espionage, PrivateSect}	=>	{Asia}	16.75%	81.48%
8.	{Asia, PrivateSect}	=>	{Espionage}	16.75%	79.51%
9.	{Espionage, Others}	=>	{Asia}	24.74%	78.94%
10.	{PrivateSect}	=>	{Asia}	21.06%	78.3%
11.	{PrivateSect}	=>	{Espionage}	20.55%	76.41%
12.	{Europe}	=>	{Espionage}	19.03%	75.75%
13.	{Others}	=>	{Asia}	26.39%	75.63%
14.	{Espionage}	=>	{Asia}	63.83%	74.85%

FIGURE 4.37 – Génération des règles d'association dans *Lattice Miner*

Nous remarquons que la règle 3 : *Asia => Espionage* a un support 63,83 %, le plus élevé de tous, et une confiance de 89.98 %. Cela signifie que la probabilité qu'une attaque commanditée par la région d'Asie est de près de 90% de type espionnage et que l'espionnage effectué par l'Asie se produit près de 64% parmi l'ensemble des attaques. Inversement, la règle 14 : *Espionage => Asia* a le même support que la précédente association mais signifie que la probabilité est de 74.85% que le commanditaire d'une attaque de type espionnage soit l'Asie.

4.9 Analyse de l'attribut date

Comme décrit dans la phase de prétraitement, nous allons maintenant utiliser chaque nouvel attribut dérivé de l'attribut date particulièrement pour découvrir des informations et des connaissances. Nous tenons à rappeler que nous utilisons l'ensemble de données d'origine, sans transformation de l'attribut « type d'attaque », a part le faite d'avoir fixé un seuil absolu égal à 10 pour cet attribut. Cela signifie que nous conservons les types d'attaques qui ont au moins 10 enregistrements, ce qui nous donne 5 types d'attaques, à savoir : la destruction de données (qui englobe le doxage et la défiguration), l'espionnage, le déni de service, le rançongiciel et l'hameçonnage.

4.9.1 Attaques par jour de la semaine

Nous avons effectué une analyse pour déterminer les jours de la semaine où il y a le plus d'attaques, en utilisant la colonne de date et le type d'attaque. À partir de la colonne de date de chaque attaque, nous avons extrait le jour de la semaine correspondant à chaque attaque. Ensuite, nous avons regroupé les attaques par jour de la semaine et compté le nombre d'attaques qui se produisent chaque jour. Les résultats obtenus sont présentés dans l'histogramme de la figure 4.39.

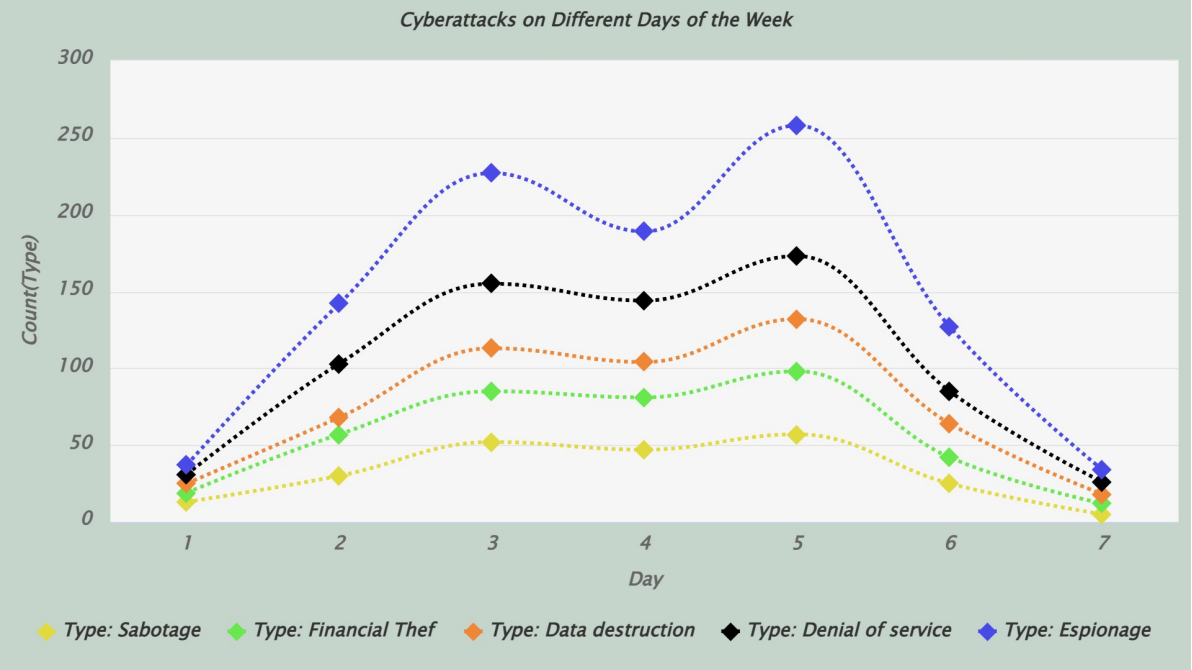


FIGURE 4.38 – Nombre d'attaques par jour de la semaine

En examinant la figure 4.38, nous pouvons observer que la semaine débute relativement calme le lundi (1) avec une montée progressive de la courbe qui se poursuit le mardi (2) pour atteindre son sommet le mercredi (3). Nous notons une légère descente le jeudi, tandis que le vendredi représente le sommet le plus élevé, suivi d'une diminution durant le weekend jusqu'au dimanche, où elle atteint le point le plus bas.

Nous pouvons affirmer cette tendance en observant un autre diagramme qui est présenté sous forme de bars à la figure 4.39.

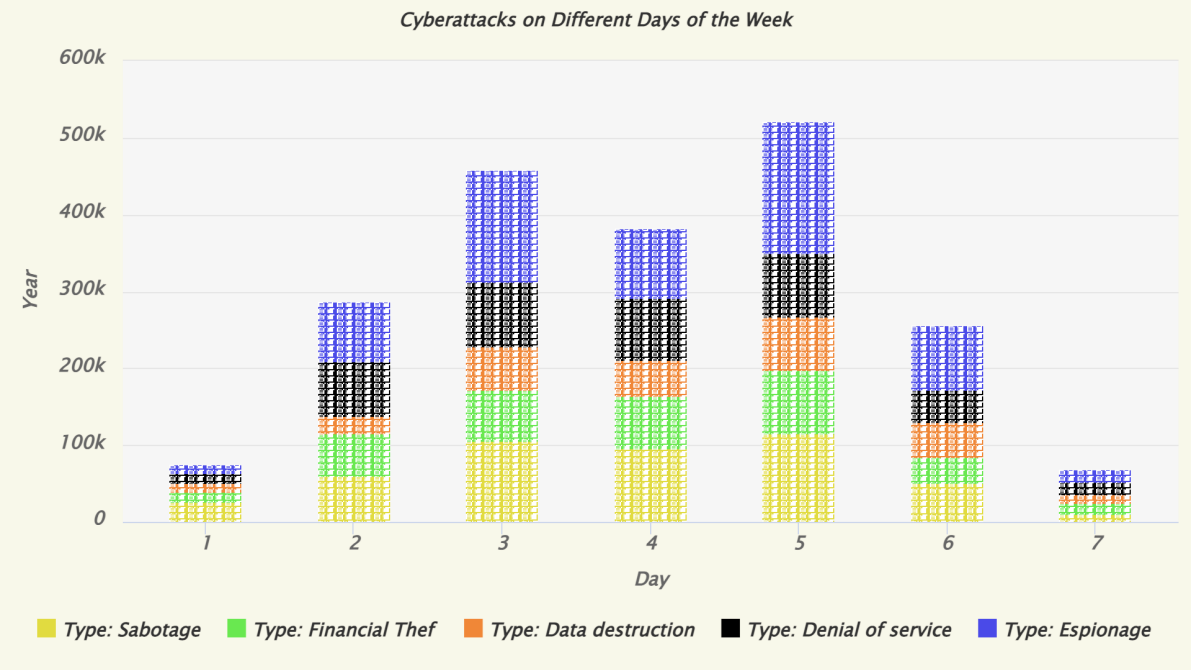


FIGURE 4.39 – Nombre d'attaques par jour de la semaine

Ce graphique met en évidence les jours de la semaine où le nombre d'attaques est le plus élevé. Il permet d'identifier les tendances et les fluctuations des attaques tout au long de la semaine.

Ici, le lundi correspond à la valeur 1, le mardi à 2, le mercredi à 3, et ainsi de suite jusqu'au dimanche qui correspond à la valeur 7. Les résultats présentés dans la figure 4.38 et la figure 4.39 permettent de dégager une tendance générale en termes de fréquence et de types d'attaques observées selon les jours de la semaine.

La barre 5 (vendredi) est le jour de la semaine où il y a le plus d'attaques, avec plus de 400 attaques observées au cours des 18 dernières années. La barre 3 (mercredi) se place en deuxième position avec plus de 300 attaques, suivie de près par la barre 4 (jeudi) avec environ 200 attaques. Les barres 2 (mardi) et 6 (samedi) sont presque à égalité, avec environ 142 attaques pour le premier et 127 pour l'autre, la barre 1 (lundi) est avant-dernière avec environ 37. Enfin, la barre 7 (dimanche) clôture le cycle avec moins de 34 attaques, ce qui en fait le jour de la semaine où il y a le moins d'attaques enregistrées.

Une analyse de la barre 5 (vendredi) montre que les attaques de type espionnage ont atteint un pic le plus élevé de toutes plus de 250 attaques, ce qui en fait le type d'attaque le plus fréquent ce jour-là. Ensuite, nous observons les attaques de déni de service qui s'élèvent à plus de 150 et le sabotage plus de 100, viennent ensuite les attaques de destruction de données qui se situent autour de 90, suivies des attaques de vol financier avec environ 60 cas.

Ces informations nous permettent de visualiser la répartition des différents types d'attaques selon les jours de la semaine. Il est intéressant de constater que le déni de service est le type d'attaque le plus prévalent, suivi par le sabotage, le doxage, la destruction de données, le vol financier et enfin la dégradation. Cette analyse peut aider à mieux cibler les mesures de sécurité et à anticiper les types d'attaques les plus probables selon le jour de la semaine.

4.9.2 Attaques par mois de l'année

Les figures 4.40 et 4.40 représentent le nombre d'attaques par mois au cours d'une année. Nous remarquons que les attaques se produisent beaucoup au début de l'année : le mois de février suivi du mois de mars , et puis janvier et avril représentent les périodes où on constate une augmentation des attaques. La même tendance a été observée en 2023 au cours de 5 premiers mois. Aussi, vers la fin de l'année, nous constatons qu'il y a une augmentation des attaques au cours du mois d'octobre alors qu'il y a une baisse relativement faible pendant les préparatifs des festivités tandis que, du mois de juillet jusqu'au mois de septembre puis novembre il y a le même degré de risques ou d'attaques, mais avec une spécificité qui est la variété de types d'attaques.

Au mois de septembre par exemple, il y a deux types d'attaques : le déni de services et le vol financier alors qu'au mois de mai, il y a quatre types d'attaques dont les plus importantes sont le déni de services et le sabotage.

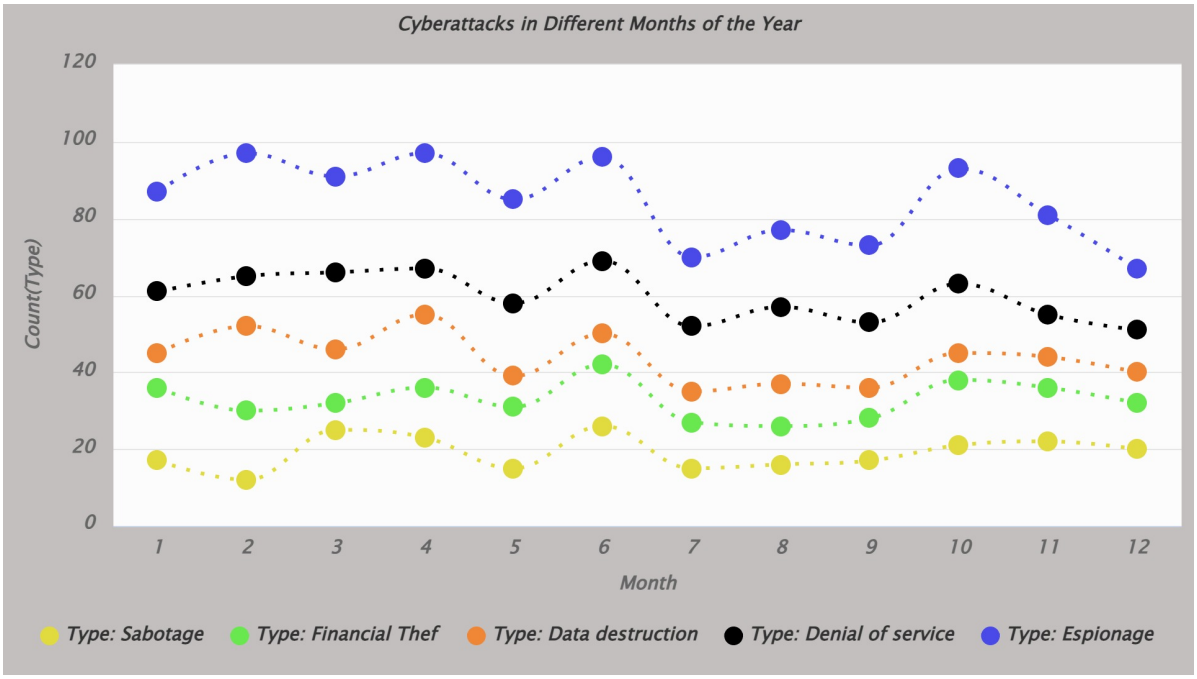


FIGURE 4.40 – Nombre d'attaques par mois

Chaque courbe à la figure 4.40 représente l'évolution d'un type d'attaque. Prenons par exemple la ligne bleue qui correspond aux attaques d'espionnage. Nous remarquons qu'à partir de janvier, le nombre d'attaques d'espionnage demeure élevé, augmentant encore davantage en février, avec une légère baisse en mars, suivie d'une forte augmentation en avril et juin. En juillet, malgré une diminution et un calme relatif observés jusqu'en septembre, le nombre d'attaques d'espionnage reste le plus élevé parmi toutes les attaques. En octobre, ce nombre connaît une hausse avant une légère diminution qui se prolonge jusqu'en décembre.

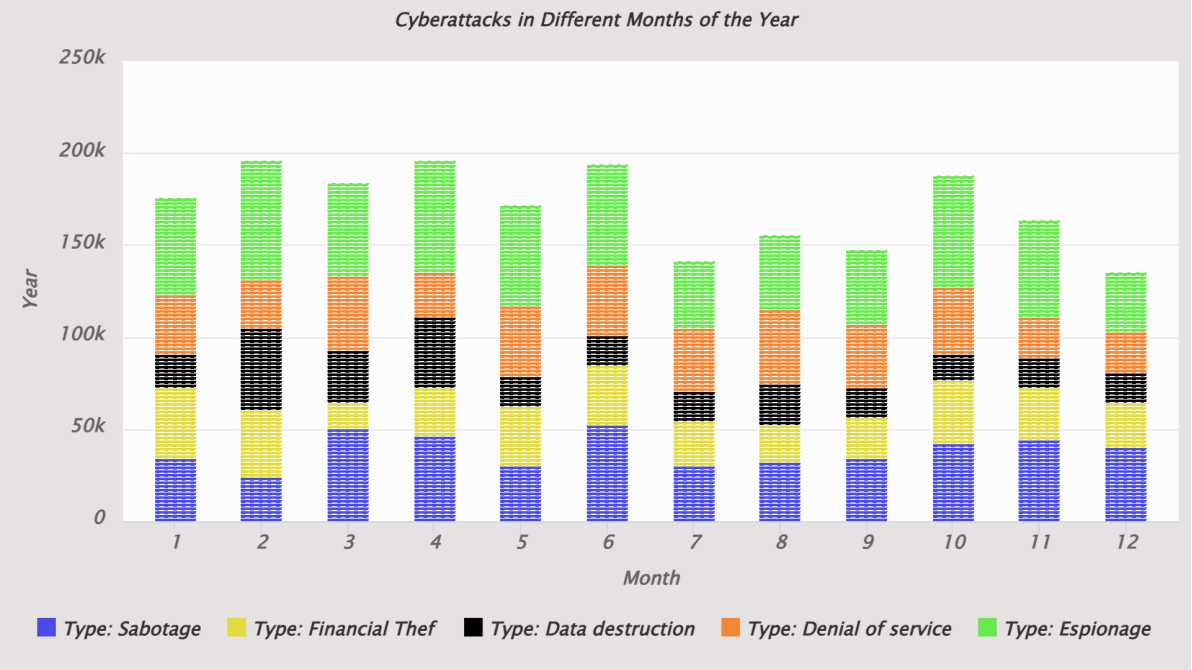


FIGURE 4.41 – Nombre d'attaques par mois

Nous allons analyser le mois d'avril (barre 4), l'un des mois où l'on observe le plus d'attaques au cours de l'année, à la figure 4.40. L'espionnage (représenté en vert sur la barre 4) arrive en tête avec plus de 100 attaques, suivi du sabotage (en bleu) avec près de 70 attaques, puis du déni de service (en orange) avec plus de 60 attaques, et enfin de la destruction de données (en noir) et du vol financier (en jaune).

4.9.3 Attaques par trimestre de l'année

Nous avons également trouvé intéressant de comprendre les trimestres pendant lesquels le nombre d'attaques est le plus élevé. Les résultats des figures 4.42 et 4.43 révèlent que l'intensité des attaques est plus élevée au cours des premier et deuxième trimestres de l'année. Les attaques d'espionnage (représentées en ligne verte) et les attaques par déni de service (en ligne orange) constituent la menace la plus prédominante pendant cette période. Au deuxième trimestre, l'intensité des attaques demeure

élevée, avec les attaques par sabotage, devenant la principale menace, suivie des attaques par déni de service.

Le troisième trimestre enregistre le plus faible nombre d'attaques de l'année, bien que le déni de service demeure une menace significative. Finalement, au quatrième trimestre, le nombre et la diversité des attaques augmentent, avec des incidents de destruction de données et des attaques par déni de service devenant plus fréquents.

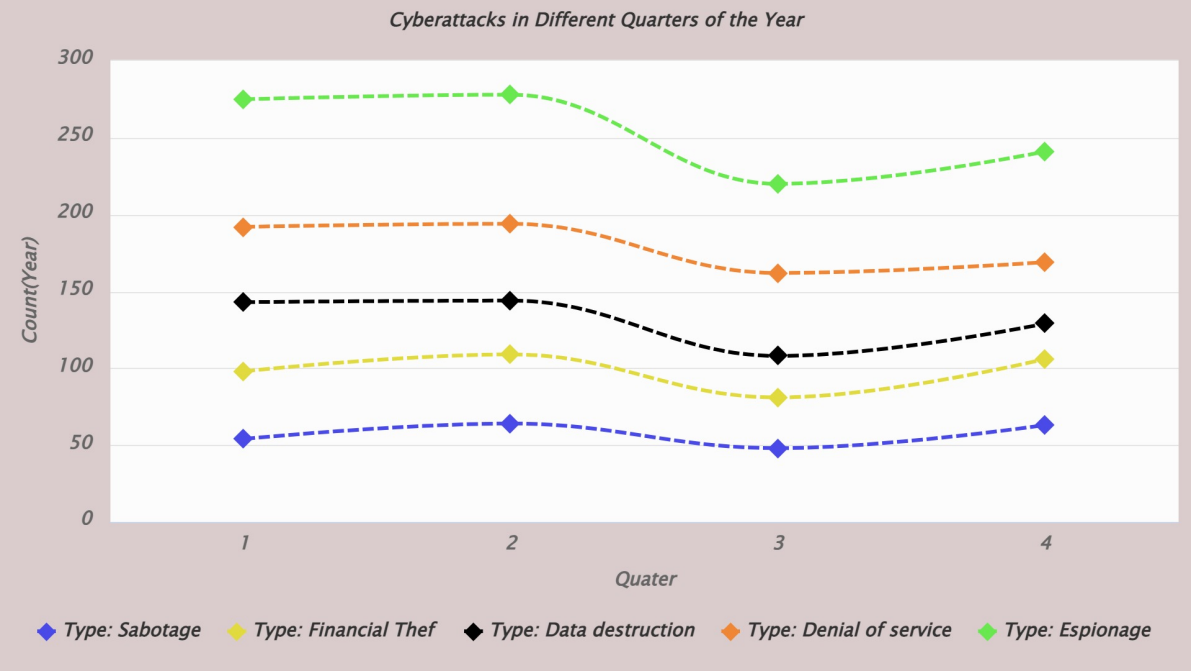


FIGURE 4.42 – Nombre d'attaques par trimestre sous forme linéaire

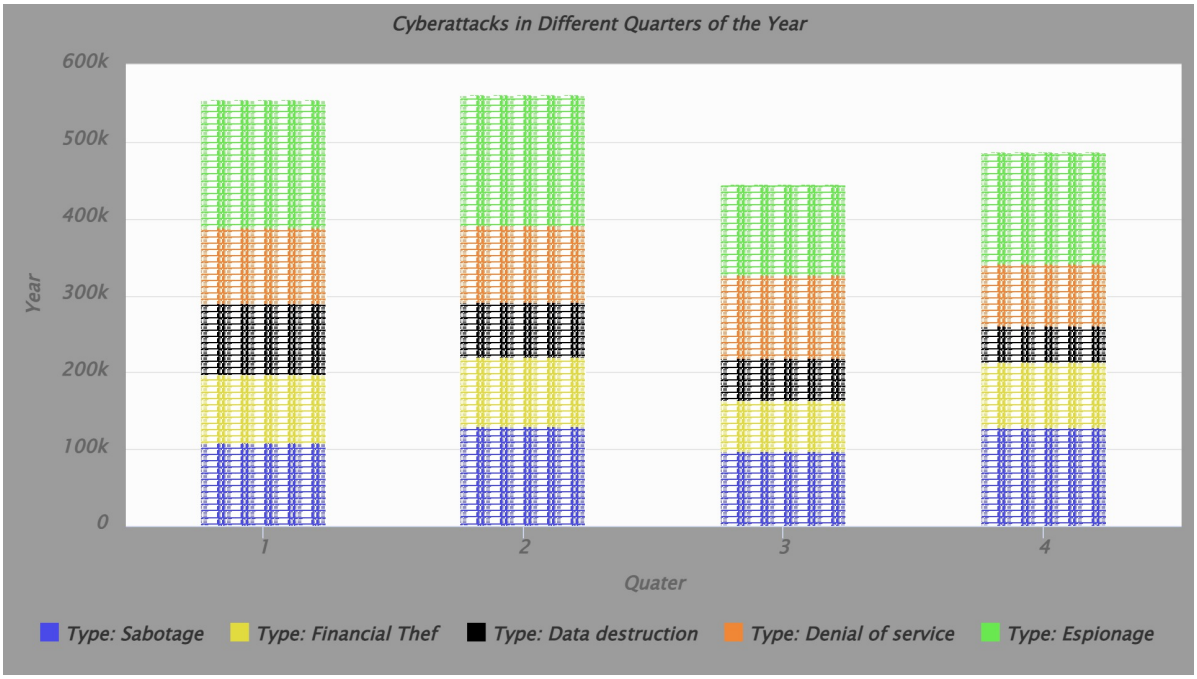


FIGURE 4.43 – Nombre d'attaques par trimestre

En examinant le graphique présenté à la figure 4.43, nous pouvons constater à titre d'exemple que la barre 4 est dominée par les attaques d'espionnage qui occupent la première position, suivies par le sabotage, le déni de service et le vol financier.

4.9.4 Attaques au cours de l'année

Nous désirons comprendre l'évolution du nombre d'attaques entre 2005 et 2023.

La figure 4.44 montre qu'en 2022, on a observé une quantité significative d'attaques jamais observée au paravent. Cette croissance a démarré depuis 2016 jusqu'à aujourd'hui, mais on remarque une petite baisse du nombre d'attaques au cours de l'année 2021, tandis que les années 2020 et 2022 ont connu une augmentation plus importante des attaques. Les types d'attaques varient d'une année à l'autre. Cette tendance se poursuit notamment pour l'année en cours, 2023, où l'on observe une diversité de types d'attaques qui continuent de causer des victimes.

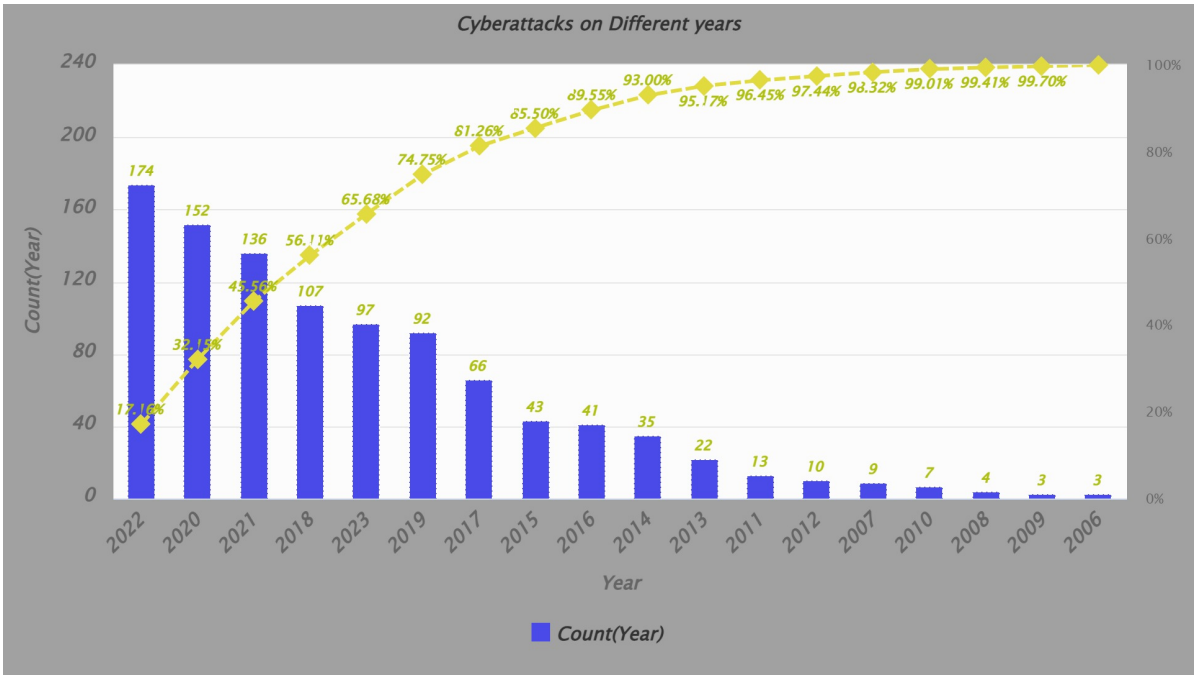


FIGURE 4.44 – Évolution du nombre des cyberattaques au cours des années

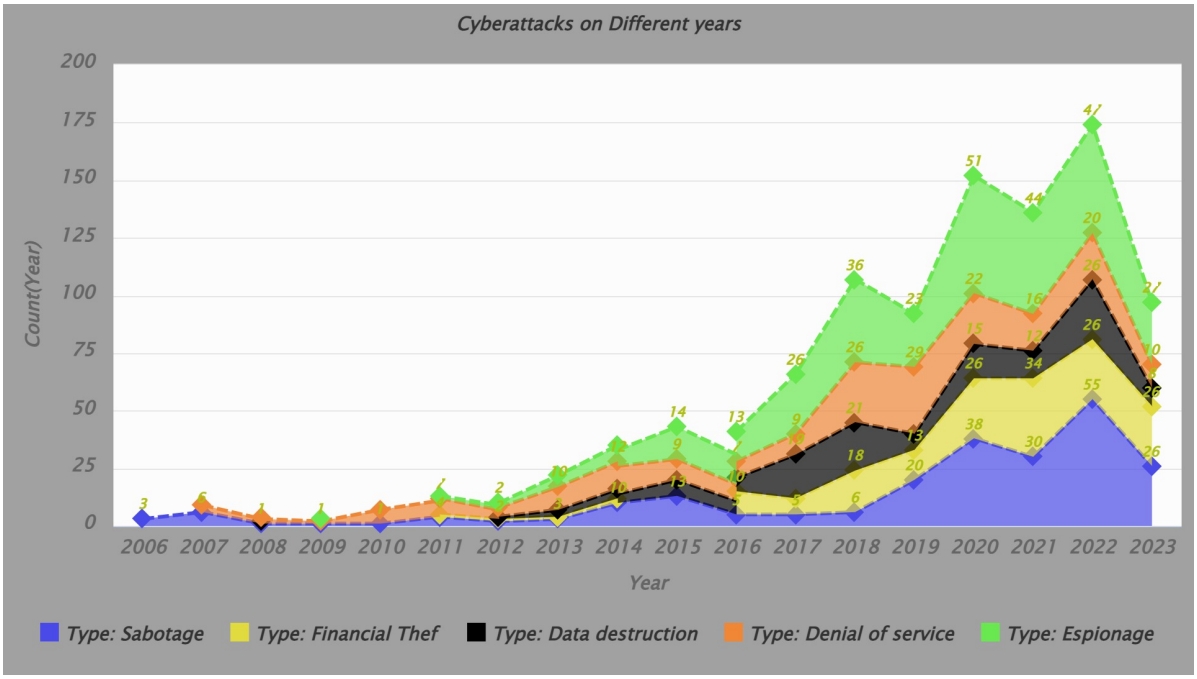


FIGURE 4.45 – Évolution du nombre des cyberattaques au cours des années.

La figure 4.45 offre une vue d'ensemble de la répartition de ces types d'attaques au fil du temps, ce qui nous permet de mieux comprendre les tendances et les évolutions dans le paysage des cybermenaces. Les principaux types d'attaques pris en compte sont l'espionnage, le déni de service, la destruction de données et le sabotage. La connaissance de l'évolution d'attaques revêt une importance cruciale pour l'élaboration de stratégies de sécurité efficaces et la protection contre les attaques potentielles.

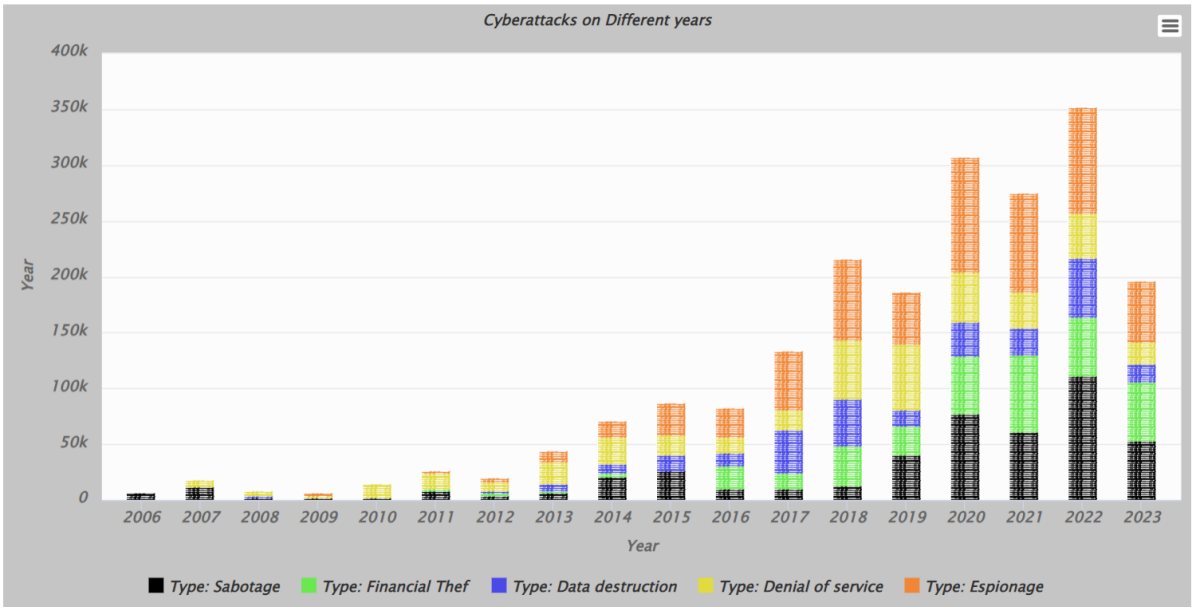


FIGURE 4.46 – Évolution du nombre des cyberattaques au cours des années.

La figure 4.46 révèle qu'entre 2005 et aujourd'hui, il existe une prédominance constante des attaques de type espionnage. Cela est suivi par les attaques de déni de service, de destruction de données et de sabotage. Ces attaques représentent les menaces les plus préoccupantes en cybersécurité.

Nous pouvons observer cette tendance pour une période donnée, par exemple au cours de l'année X et du mois Y, où il y a le plus d'attaques. Cela permet aux institutions d'accroître leur vigilance pendant cette période.

4.9.5 Période d'attaques (mois-année)

Nous avons ajouté un nouvel attribut « mois-année » afin d'analyser le mois de l'année où le nombre d'attaques est le plus élevé. En concaténant les attributs « mois » et « année », nous pouvons visualiser les résultats de manière plus claire et approfondie.

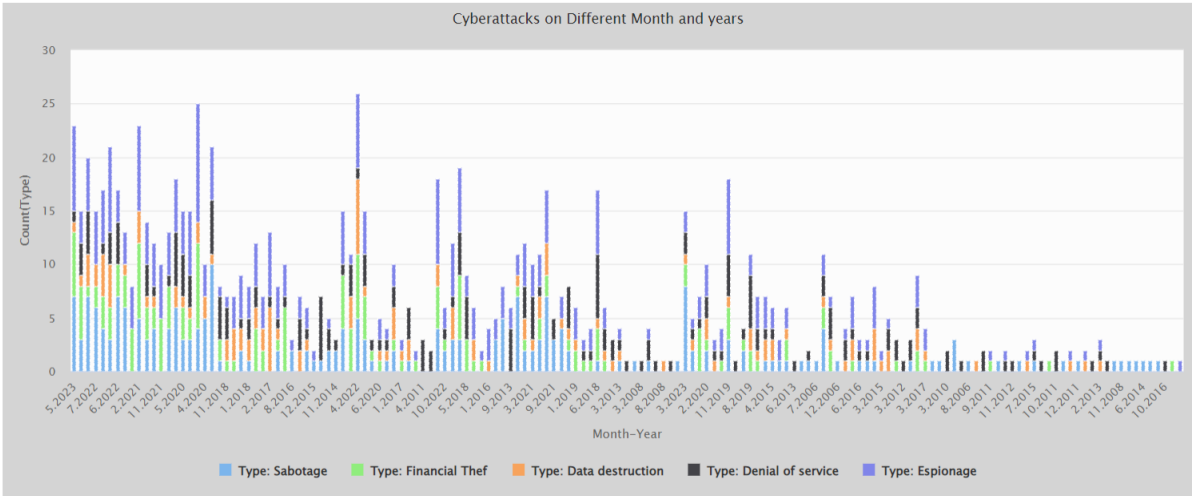


FIGURE 4.47 – Attaques par période mois-année

En 2023, le mois de mai a été le mois le plus touché par les attaques dans le monde, suivi de près par le mois de mars. Les types d'attaques les plus fréquents étaient l'espionnage en premier, suivi du sabotage et du déni de service.

Par ailleurs, en remontant dans le temps, le mois d'avril 2022 s'est avéré être le mois le plus critique en termes d'attaques au cours des 18 dernières années. Les trois types d'attaques dominants étaient le déni de service, le sabotage et le vol financier. Plus particulièrement, le mois d'avril 2022 s'est caractérisé par un nombre exceptionnellement élevé d'attaques de type vol financier.

En revanche, si l'on examine l'année 2016, le mois de décembre de cette année se distingue par le nombre le plus élevé d'attaques de type destruction de données.

4.9.6 Période d'attaques : trimestre-année

Nous avons effectué une concaténation des attributs « année » et « trimestre » pour obtenir l'attribut « période », ce qui nous permet d'analyser la période de l'année où le nombre d'attaques est le plus élevé. Comme nous l'avons déjà observé, le premier trimestre de chaque année est une période où les organisations doivent prendre des mesures de sécurité renforcées pour faire face à la vague d'attaques qui survient généralement en hiver. La figure 4.48 illustre clairement cette tendance, et l'année en cours ne fait pas exception. En fait, les attaques se sont intensifiées davantage pendant l'hiver de l'année 2023.

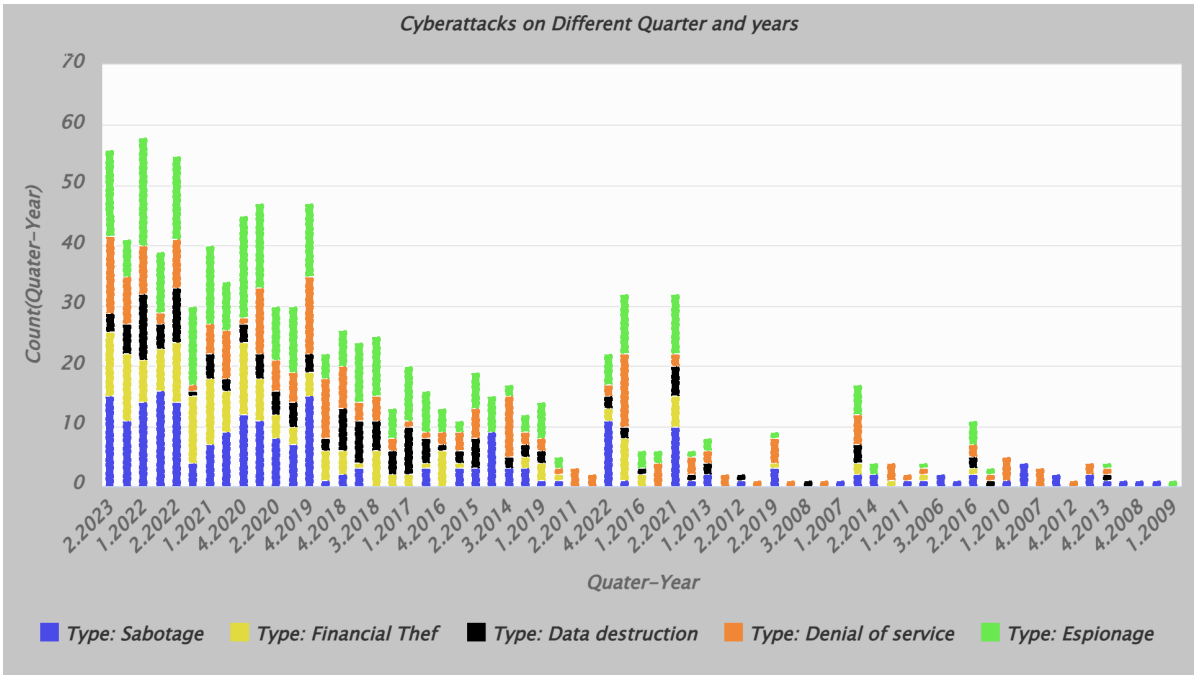


FIGURE 4.48 – Trimestres les plus menaçants de l'année

Au cours du deuxième trimestre de 2023, on a observé une augmentation significative des attaques de déni de service par rapport aux années précédentes. Parallèlement, des actes de sabotage et de vol financier ont également été signalés. De manière remarquable, les attaques de destruction de données ont atteint un niveau comparable à celui enregistré au troisième et au quatrième trimestre de 2022, dépassant ainsi les niveaux observés les années précédentes. Quant aux attaques de vol financier, elles sont restées au même niveau qu'au troisième trimestre de 2022, mais ont augmenté au quatrième trimestre de la même année.

Il est important de noter que le deuxième trimestre de l'année 2022 se démarque comme le trimestre ayant enregistré le plus grand nombre d'attaques au cours des 18 dernières années, soit sur une décennie et demie.

Nous pouvons visualiser cette tendance de manière textuelle pour mieux comprendre le sujet, comme présenté dans la figure 4.49

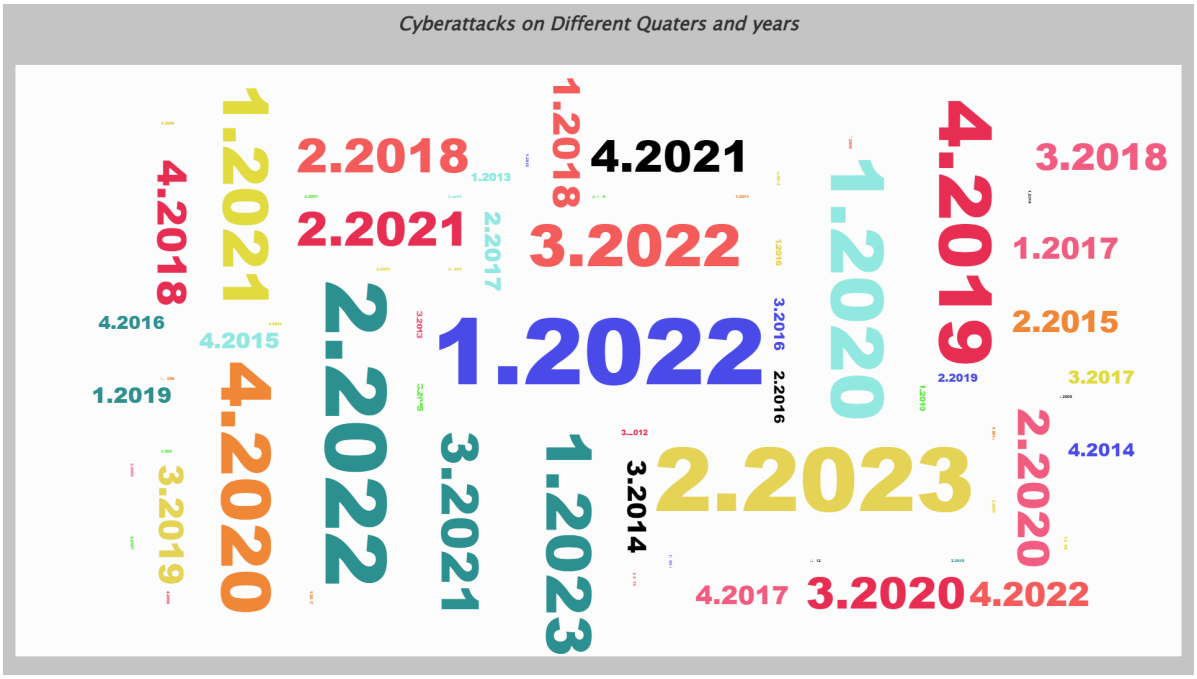


FIGURE 4.49 – Visualisation des trimestres où il y a le plus d'attaques

De 2020 jusqu'à aujourd'hui, la période hivernale qui s'étend du 21 décembre (la fin du quatrième trimestre) au 18 mars (premier trimestre) de l'année suivante regroupe un grand nombre d'attaques. C'est une période où les cyberattaques sont généralement en augmentation et elles diminuent vers la fin du deuxième trimestre, ce qui en fait la période la plus dangereuse de l'année en termes d'attaques.

En ce qui concerne les institutions ciblées par les cyberattaques, nous avons examiné ce facteur dans la deuxième étape de notre analyse. Nous voulions identifier les institutions les plus visées pendant certaines périodes de l'année, comme illustrées dans la figure 4.50.

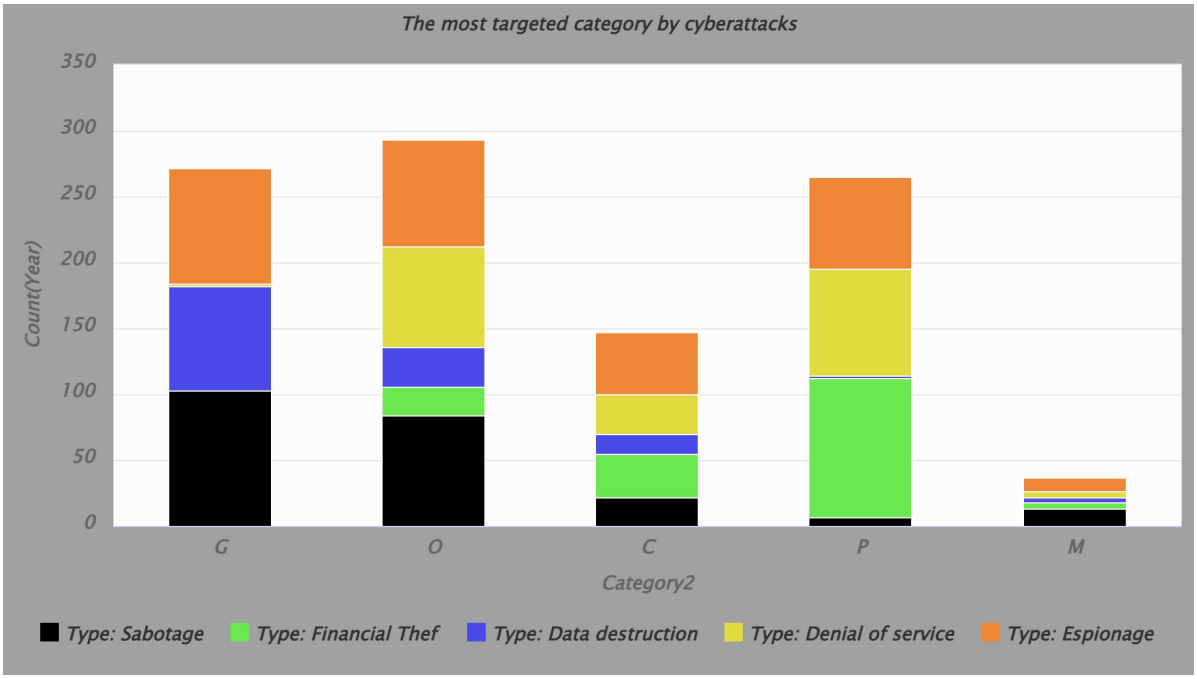


FIGURE 4.50 – Les organismes sont le plus ciblés

On peut observer ici que le gouvernement (G) et le secteur privé (P) sont les principales cibles des cyberattaques. Entre 2007 et 2023, les attaques visant ces institutions ne cessent d'augmenter. Il y a un grand nombre d'attaques ciblant plus au moins deux catégories d'institutions (O) à la fois. De plus en plus la société civile (C) est ciblée par les attaques parrainées par leurs propres états. L'année 2021 est particulière pour les attaques ciblant le site militaire (M)

4.10 Fouille de texte

Lors de la recherche de mots-clés dans l'attribut « description », nous avons effectué des prétraitements tels que la tokenisation, la suppression des mots vides, la filtration des jetons, et la normalisation des données. Ces étapes sont essentielles pour préparer les données textuelles avant de les analyser et de les exploiter dans le cadre de fouille de texte. Pour le nettoyage des données, par l'opération transformation, nous éliminons le bruit tels que les caractères *« â€™s »* par exemple. L'opération de tokenisation a divisé nos textes en jetons, par le filtrage, les caractères indésirables tels que la ponctuation et les chiffres et mots vides *(stopwords)* fréquents sans signification tels que *« the », « and », « â », « ââ »* ont été supprimés. Nous avons également supprimé les mots-clés contenant moins de quatre caractères. De plus, nous avons traité les mots synonymes.

Pour affecter un poids aux mots-clés extraits des titres, nous avons calculé leur valeur IDF et TF-IDF. À titre d'exemple, le nombre total de documents dans notre corpus est de 500, et le nombre d'apparitions du mot « acteur » dans les différents titres est 443 fois : $\text{idf}_{\text{actor}} = \log \frac{|D|}{|\{d_j : t \in d_j\}|} = \log \frac{500}{443} = 1.13$. Pour

le mot attaque, il apparaît 148 fois, son $\text{idf}_{\text{attack}} = \log \frac{500}{148} = 3.38$. On peut observer que le mot « attaque » a un poids élevé par rapport à « acteur », cela signifie que lors de l'analyse, le mot attaque est à privilégier.

La figure 4.51 montre un échantillon de termes et leur valeur TF-IDF.

Row No.	id	cluster	text	phishing	companies	targets	threat	spear	russian	accessing	malware	chinese	governm...
694	694	cluster_1	chinese linked ...	0	0.037	0	0	0	0.490	0	0	0.447	0.334
235	235	cluster_2	threat compani...	0	0.074	0	0.137	0	0	0	0	0.445	0.333
378	378	cluster_1	russia backed ...	0.434	0.036	0	0.132	0.548	0.471	0	0	0	0.322
737	737	cluster_1	calisto russian ...	0.430	0.106	0	0.131	0.543	0.467	0	0	0	0.318
605	605	cluster_3	computers ass...	0	0	0	0	0	0	0	0	0.841	0.315
677	677	cluster_3	chinese threat ...	0	0.103	0.171	0.128	0.527	0	0.486	0.403	0.414	0.310
17	17	cluster_2	using unusual r...	0	0.067	0	0	0	0	0.951	0	0	0.303
354	354	cluster_3	russian threat ...	0.407	0.067	0.333	0.124	0.514	0.442	0	0.393	0	0.301
358	358	cluster_1	russian military...	0	0	0.163	0	0	0.864	0	0	0	0.295
560	560	cluster_3	gamaredon ru...	0.390	0.064	0	0.119	0	0.424	0	0.753	0	0.289
561	561	cluster_2	gamaredon ru...	0.380	0.094	0.156	0.116	0.480	0.413	0.442	0.367	0	0.282
172	172	cluster_3	threat compani...	0.378	0.062	0	0.116	0.478	0	0	0.730	0	0.280
200	200	cluster_3	threat compani...	0	0.179	0.149	0.332	0	0	0	0.350	0.719	0.269
364	364	cluster_1	russian compa...	0.361	0.030	0.148	0	0.455	0.392	0.420	0	0	0.267
329	329	cluster_2	sandworm tar...	0	0	0.143	0	0	0	0.406	0.337	0	0.259
663	663	cluster_3	chinese threat ...	0	0.026	0.128	0.095	0	0.679	0	0.603	0.310	0.232
104	104	cluster_3	threat compani...	0	0.229	0.127	0.189	0	0	0	0	0.918	0.229

ExampleSet (758 examples, 3 special attributes, 14 regular attributes)

FIGURE 4.51 – Pondération TF-IDF des mots-clés

Cette fréquence de mots est aussi représentée graphiquement dans la figure 4.52 en fonction de la fréquence d'apparition dans les documents et du nombre de documents dans lesquels le mot apparaît.

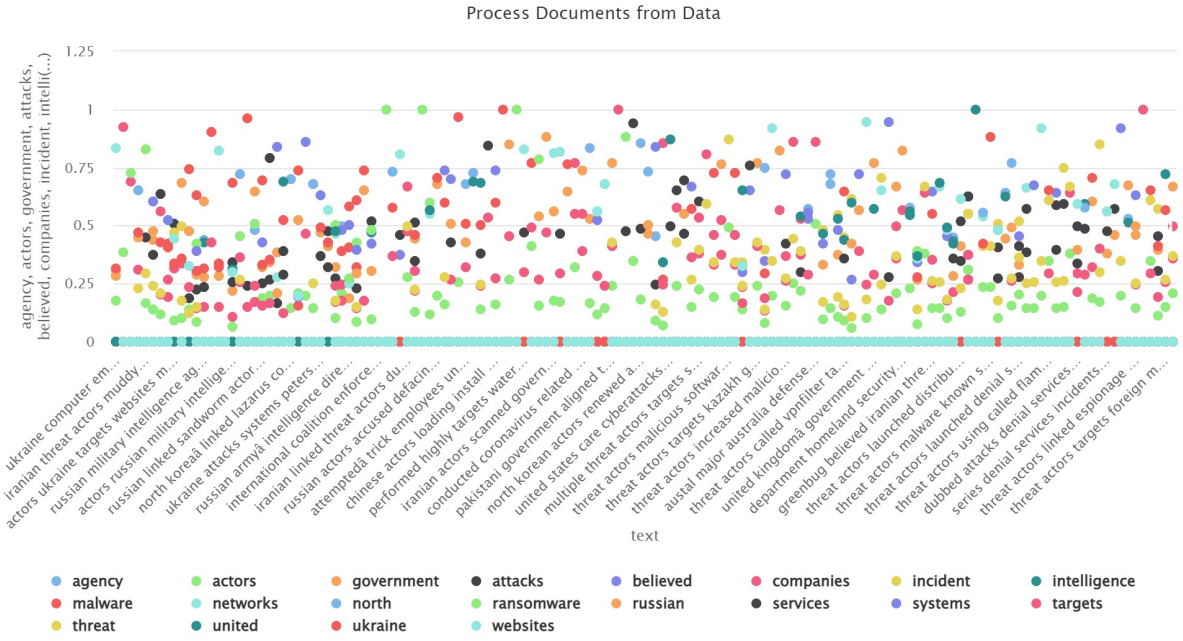


FIGURE 4.52 – Graphique IDF de mots-clés

La ligne verte est le nombre d'apparitions d'un mot dans le corpus (fréquence) alors que la ligne bleue indique le nombre de documents dans lesquels le mot est mentionné.

Nous avons trouvé les mots pertinents avec lesquels nous pouvons appliquer certaines techniques de fouille de données. Nous avons toutefois écarté le mot acteur qui est largement repris dans plusieurs documents mais qui véhicule une information peu pertinente pour notre analyse.

4.10.1 Les règles d'association

Nous avons ensuite généré des règles d'association comme indiqué à la figure 4.53 à partir de corpus.

```
[targets] --> [threat] (confidence: 0.590)
[targets] --> [companies, threat] (confidence: 0.590)
[targets, attacks] --> [companies] (confidence: 0.615)
[companies, targets, attacks] --> [threat] (confidence: 0.625)
[companies] --> [threat] (confidence: 0.628)
[companies, attacks] --> [threat] (confidence: 0.667)
[attacks] --> [companies] (confidence: 0.677)
[companies, targets] --> [threat] (confidence: 0.742)
[targets] --> [companies] (confidence: 0.795)
[threat] --> [companies] (confidence: 1.000)
[threat, targets] --> [companies] (confidence: 1.000)
[threat, attacks] --> [companies] (confidence: 1.000)
[threat, targets, attacks] --> [companies] (confidence: 1.000)
```

FIGURE 4.53 – Régles d'association

Nous pouvons identifier les règles d'association, comme l'exemple suivant : la règle « campagnies », « cibles » et « attaques » implique « menace » avec une confiance de 62.5 % et un support faible de 5 %. Cela indique que lorsque nous cherchons une attaque qui comprend les termes « compagnies », « cibles » et « attaques », il est très probable qu'elle soit une « menace ». La règle 11 (cf. Figure 4.54) indique que les termes « menace » et « cibles » impliquent « campagnies » avec une confiance de 100 % et un support de 23 %. Cela signifie que chaque fois que nous avons les termes « menace », « cibles » dans la description d'une attaque, alors cette attaque vise une compagnie.

Le graphique des règles d'association est illustré à la figure 4.54.

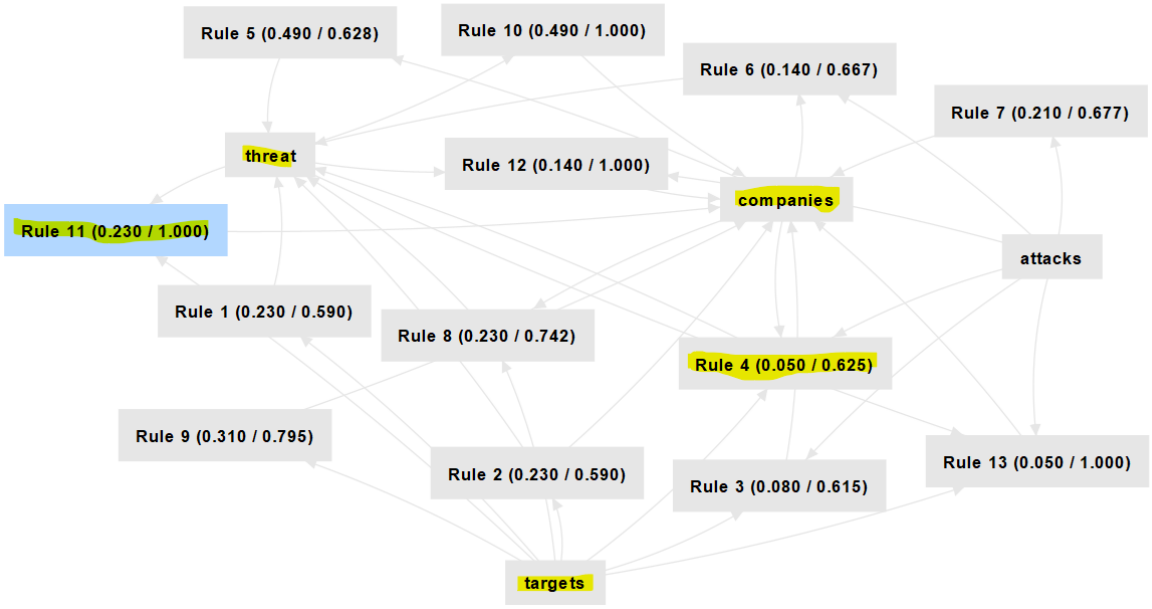


FIGURE 4.54 – Arbre de croissance de motifs (FP)

La règle 4 avec une confiance de 62.5 % et un support de 5 % signifie qu'à chaque fois qu'une attaque contient les mentions : « campagnies », « cibles » et « attaques », il s'agit à 62.5 % d'une ménace.

Nous examinons également l'intérêt *(lift)* des règles illustré à la figure 4.55.

No.	Premises ↓	Conclusion	Support	Confidence	Lift
13	threat, targets, attacks	companies	0.050	1	1.282
11	threat, targets	companies	0.230	1	1.282
12	threat, attacks	companies	0.140	1	1.282
10	threat	companies	0.490	1	1.282
9	targets	companies	0.310	0.795	1.019
7	attacks	companies	0.210	0.677	0.868

FIGURE 4.55 – Tableau de règles d'association

L'intérêt pour cette règle 4 est de 1.276, ce qui indique qu'il existe une corrélation positive entre la prémisse et la conclusion de la règle.

Nous pouvons procéder maintenant à la phase de regroupement des mots-clés.

4.10.2 Le regroupement

Nous prenons un échantillonnage de 100 enregistrements pour k=3 afin de créer trois groupes de données textuelles à la figure 4.56.

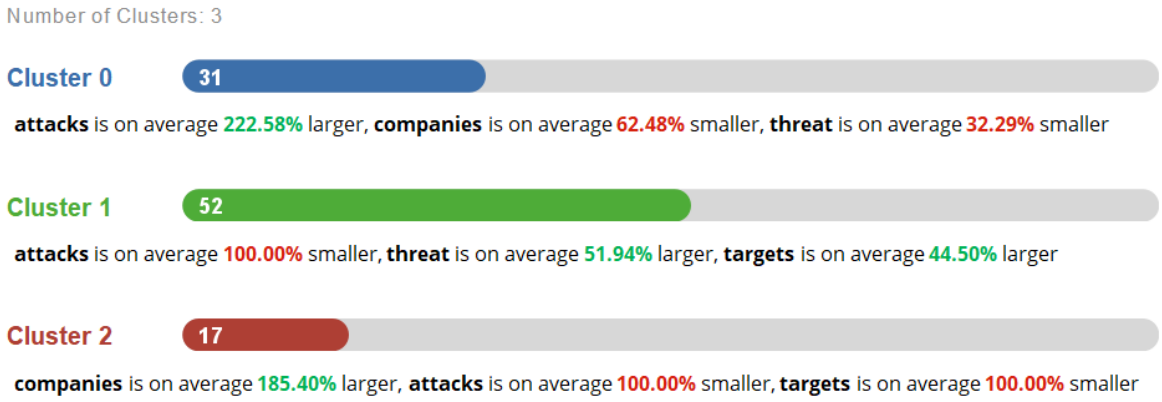


FIGURE 4.56 – Regroupement de données textuelles

On peut observer dans la figure 4.56 que les groupes sont formés et se répartissent de la manière suivante : le groupe-0 est dominé par le mot "attaques" et une faible représentation des mots "compagnies" et "menaces". Dans le groupe-1, il y a une faible représentation du mot "attaques", mais une forte représentation des termes "menaces" et "cibles". Le groupe-2 est largement dominé par le mot "compagnies" et une faible représentation des mots "attaques" et "cibles".

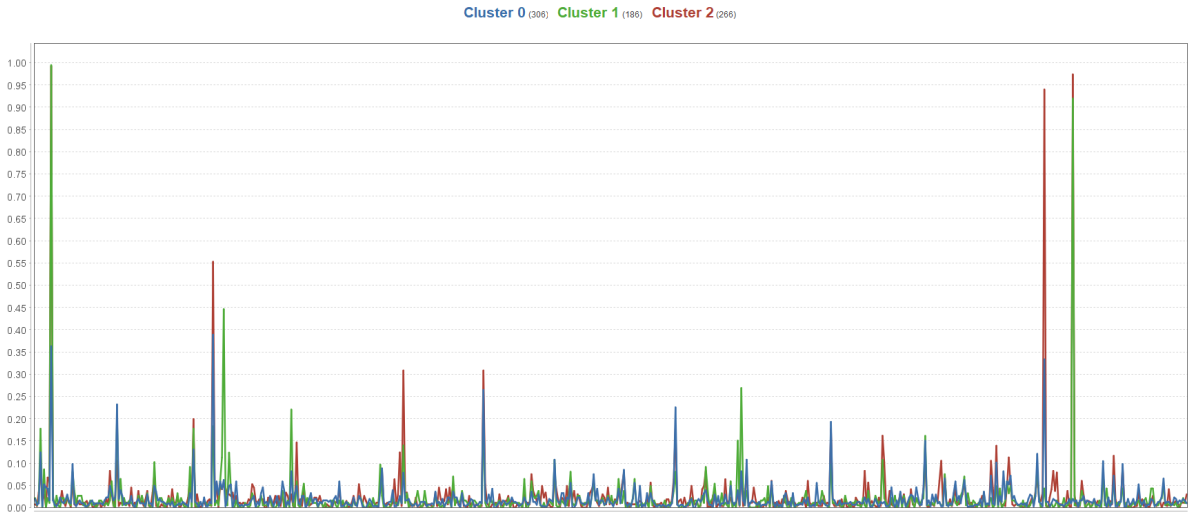


FIGURE 4.57 – Visualisation des groupes

Dans cette visualisation, nous pouvons identifier les caractéristiques des groupes selon la présence des termes dans leur description. Les groupes 1 et 2 respectivement en vert et en rouge semblent être souvent dominants.

4.10.3 Le regroupement révisé

Le tableau de mots importants par classe d'attaques

Row No.	word	in docum...	total	in class (Espionage)	in class (Phishing)	in class (Ransomware)	in class (Denial of service)
1	government	208	231	193	12	8	18
2	compromis	173	199	174	8	10	7
3	attack	147	178	126	11	15	26
4	network	135	157	123	9	19	6
9	espionage	128	128	122	0	1	5
5	malware	126	149	119	10	10	10
6	hacker	126	138	103	6	17	12
7	company	121	137	116	3	8	10
8	group	115	136	113	7	10	6
10	organisation	111	121	106	4	2	9

FIGURE 4.58 – Tableau de mots-clés

Prenons l'exemple du mot « Ukraine». Il apparaît au total dans 43 documents sur un total d'occurrences égal à 51. Le corpus a 754 documents. On peut alors calculer son TFI-DF en multipliant la valeur de son TF (fréquence du terme) et celle de son IDF (l'inverse de la fréquence de document). Ainsi, le TF-IDF pour le mot « Ukraine » est : $\text{tfidf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_i = \frac{43}{51} \cdot \log \frac{754}{43} = 1.043 = 1$.

Le calcul du TF-IDF est affiché à la figure 4.59 et montre l'importance de quelques mots.

Row No.	Attribute	Importance
1	website	1
2	believe	0.814
3	ukraine	0.806
4	hacker	0.755
5	launch	0.726
6	office	0.611
7	servers	0.593
8	entity	0.591
9	russian	0.551
10	services	0.538
11	operation	0.436
12	state	0.411

FIGURE 4.59 – Les mots pertinents (TF-IDF)

Nous avons retenu 20 mots-clés que nous affichons dans un histogramme d'importance de mots à la figure 4.60

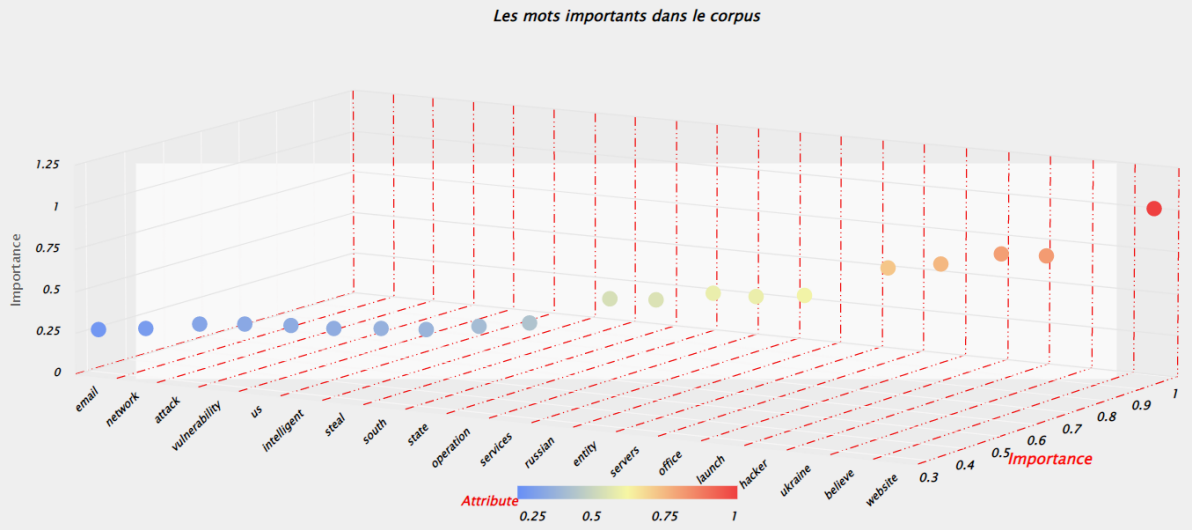


FIGURE 4.60 – Visualisation de la pertinence des mots

Nous pouvons remarquer que l'importance d'un mot est déterminée par deux caractéristiques principales : la taille de la boulette et sa couleur, laquelle varie progressivement vers le rouge (valeur comprise entre 0,75 et 1). C'est le cas, par exemple, du mot « Ukraine » qui présente une boulette de grande taille et une couleur se rapprochant du rouge. Les boulettes de couleur jaune ont une valeur comprise entre 0,5 et 0,75, tandis que les plus petites boulettes de couleur bleue ont une valeur inférieure à 0,5.

Nous prenons un échantillonnage de 756 enregistrements pour k=4 afin de créer quatre groupes pour les données textuelles à la figure 4.61.

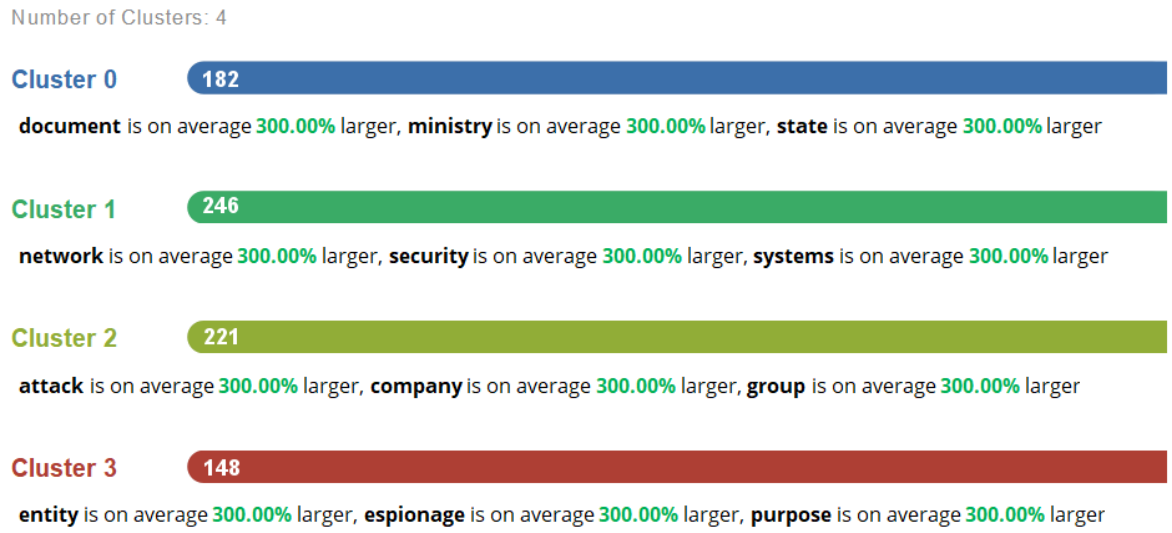


FIGURE 4.61 – Regroupement de données textuelles

Les groupes formés sont caractérisés par des valeurs dominantes de mots, respectivement de la manière suivante (cf. la figure 4.61) :

Le groupe 0 comprend les mots-clés suivants : campagne, Chine et données.

Le groupe 1 comprend les mots-clés suivants : document, logiciels malveillants et ministère.

Le groupe 2 comprend les mots-clés suivants : attaques, entreprises (institution) et groupe.

Le groupe 3 comprend les mots-clés suivants : entité, espionnage et objectif.

Cluster	attack	company	docume...	entity	espiona...	govern...	group	include	malware	ministry	network	organis...	purpose	russ...	sector	security	state	syster
Cluster 0	0	0	0.495	0	0	0	0	0	0.294	0.495	0	0	0	0	0	0	0.384	0
Cluster 1	0	0	0	0	0	0	0	0	0.288	0	0.277	0	0	0	0	0.419	0	0.684
Cluster 3	0	0	0	0.529	0.335	0.244	0	0	0	0	0	0.362	0.380	0	0.521	0	0	0
Cluster 2	0.278	0.311	0	0	0	0.218	0.319	0.464	0	0	0	0.325	0	0.345	0	0	0	0

FIGURE 4.62 – Centroïde

Dans les figures 4.62 et 4.63, nous pouvons observer que les groupes formés sont répartis de la manière suivante :

Le groupe 0 est principalement caractérisé par des attaques mentionnant les mots suivants : « document », « logiciels malveillants », « ministère » et « pays » *(state)*, tandis que les autres mots ont une influence moins importante.

Le groupe 1 est principalement caractérisé par des attaques mentionnant les mots suivants : « logiciels malveillants », « internet », « sécurité », « systèmes » et « outil » *(tool)*, tandis que les autres mots ont une influence moins importante.

Le groupe-2 est principalement caractérisé par des attaques mentionnant les mots suivants : « entité », « espionnage », « gouvernement », « organisation », « objectif » et « secteur », tandis que les autres mots ont une influence moins importante.

Le groupe-3 est principalement caractérisé par des attaques mentionnant les mots suivants : « attaques », « entreprise », « gouvernement », « organisation », « groupe », « inclure », « Russie » et « télécommunications », tandis que les autres mots ont une influence moins importante.

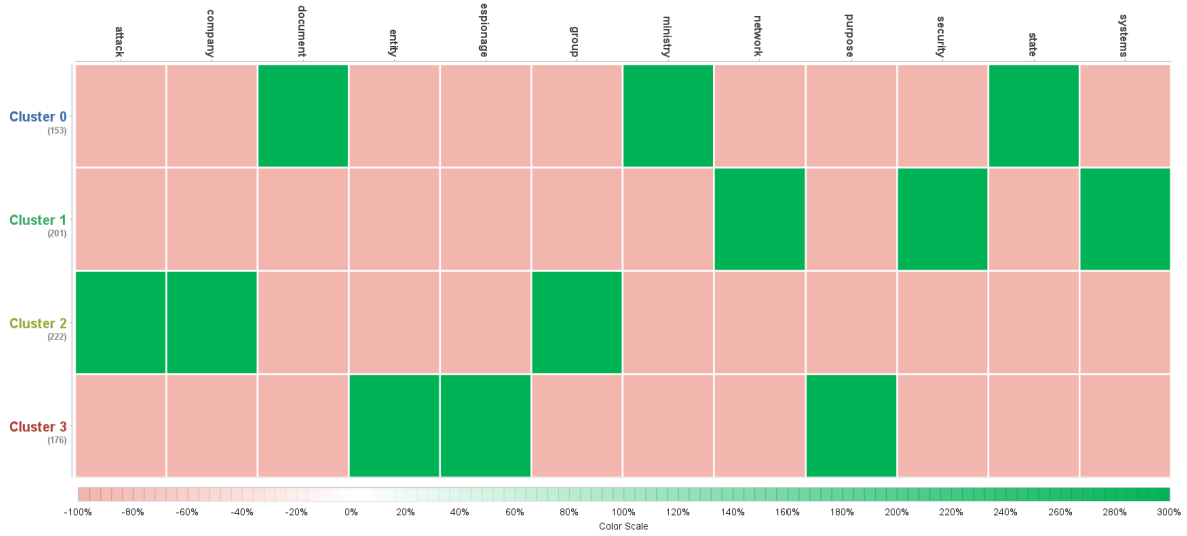


FIGURE 4.63 – Cartes des groupes

La couleur utilisée dans la représentation indique la proportion de l'influence du mot dans le groupe auquel il appartient. La couleur verte est utilisée pour montrer des proportions largement positives, par exemple, le mot « attaque » dans le groupe 3 ou « espionnage » dans le groupe 4. En revanche, la couleur rouge est utilisée pour représenter les proportions négatives, comme le mot « attaque » dans les groupes 1 et 2.

Nous découvrons, grâce à l'analyse textuelle de l'attribut « description » d'attaques, que les mots-clés indiquent souvent les menaces ou les types d'attaques, telles que « menaces », « incidents », « attaques », « hameçonnage », « rançongiciel », le « déni de service », « espionnage », « destruction de données », etc. Ils nous renseignent également sur les catégories d'institutions visées, telles que la société civile, le gouvernement, le secteur privé, et même les installations militaires. De plus, ils nous informent sur les commanditaires d'attaques, comme la Corée du Nord, la Chine, la Russie, les États-Unis d'Amérique, etc., ainsi que d'autres détails importants, comme les entités atteintes par l'attaque telles que Microsoft ou système, etc. Tous ceux-ci servent à décrire l'attaque.

Cela confirme que les attributs analysés dans ce projet revêtent une importance cruciale dans la description d'une cyberattaque.

Chapitre 5

Conclusion

Pour conclure, nous rappelons que l'objectif principal de ce mémoire est d'utiliser des techniques de fouille de données et d'apprentissage automatique dans le domaine de la cybersécurité pour identifier, caractériser et prévenir des cyberattaques. Nous avons analysé un ensemble de données regroupant les cyberincidents parrainés par les états entre 2005 et 2023 en utilisant principalement la plateforme de science des données *RapidMiner*. Nous avons également utilisé les prototypes *Lattice Miner* et *Concept Explorer* de construction du treillis de concepts (groupes conceptuels) et de production de règles d'association, y compris les implications avec négation.

Dans un premier temps, nous avons rappelé les concepts de base de la cybersécurité, de la fouille de données et de l'apprentissage automatique, et nous avons effectué un survol de la littérature pour identifier des travaux importants liés à notre analyse.

Les résultats de notre analyse indiquent que les états ont investi massivement dans la guerre cybernétique en finançant des cyberattaques contre des gouvernements, des entités privées, la société civile et même des installations militaires. Le cyberespionnage prend de l'ampleur dans le monde et est souvent accompagné d'attaques de sabotage, de destruction de données et de déni de service, entre autres.

Nous avons constaté que la Chine représente 33,8 %, la Russie 21,1 %, l'Iran 12,5 % et la République de Corée près de 10 % des principaux commanditaires de cyberattaques dans le monde. Les attaques visent principalement les organisations privées, suivies des organisations gouvernementales et civiles, mais les sites militaires ne sont pas épargnés.

Des résultats d'analyse de type regroupement ont permis de trouver des groupes (*clusters*), dont un groupe qui couvre des attaques de sabotage ciblant des entreprises privées et commanditées par la Russie et un autre qui représente des attaques d'espionnage visant la société civile et le gouvernement, avec la Chine comme commanditaire principal.

Ces observations révèlent que les pays commanditaires adoptent des préférences et des stratégies différentes en ce qui concerne les types d'attaques. Par exemple, la Chine se concentre principalement sur les attaques par déni de service en plus de l'espionnage, la Russie privilégie l'hameçonnage en plus de l'espionnage, tandis que la Corée du Nord, en plus de l'espionnage, est plus souvent associée aux attaques par rançongiciel.

En passant du pays vers la région commanditaire, nous constatons que les attaques proviennent principalement de l'Asie de l'Est et de l'Europe de l'Est, tandis que l'Amérique latine et l'Afrique subsaharienne sont moins actives. En général, l'Asie et l'Europe sont les régions les plus touchées, avec une importance croissante de la cybersécurité.

L'application d'algorithmes de production des règles d'association a permis d'identifier des associations et même des corrélations entre les attributs décrivant les attaques. Par exemple, on a trouvé que chaque fois que la région commanditaire de l'attaque est l'Afrique, il s'agit nécessairement de l'espionnage ciblant la société civile dans les neuf cas observés parmi l'ensemble des 789 attaques. De même, si la cible d'une attaque est le gouvernement et la région commanditaire est l'Asie, alors il s'agit d'espionnage avec un support de $15\tilde{\%}$ et une confiance de $94\tilde{\%}$.

La production d'implications avec négation nous a permis de constater, à titre d'exemple, qu'il n'y a pas eu d'attaques de type vol financier, commanditées par l'Amérique et ciblant le gouvernement.

Dans le but de contrecarrer d'éventuelles cyberattaques en prévoyant des ressources supplémentaires tant humaines que matérielles, nous avons effectué des transformations des dates d'attaques et produit des graphiques montrant les jours de la semaine et les périodes de l'année les plus dangereuses. Ainsi, le vendredi est le jour de la semaine où l'on observe le plus grand nombre d'attaques avec environ 400 cyberattaques. Parmi celles-ci, on compte près de 180 attaques de type déni de service et environ 80 attaques de type sabotage. Le jeudi se classe en deuxième position. Bien que dominé aussi par le déni de service, le nombre d'attaques de type vol financier est significatif par rapport aux autres jours de la semaine. En revanche, le lundi semble être le jour de la semaine où l'on observe le moins d'attaques.

La période hivernale est propice aux cyberattaques qui débutent vers la fin du quatrième trimestre de l'année, soit le 21 décembre de l'année en cours et se poursuivent jusqu'au début du deuxième trimestre, en avril ou mai de l'année suivante.

Les premier et deuxième trimestres de 2023 confirment cette tendance à la hausse des cyberattaques, dépassant ainsi la période allant de 2020 à 2022.

Dans le cadre de ce mémoire, nous avons effectué une analyse de nos données en examinant de nombreuses méthodes de fouille de données telles que la classification, le regroupement, les règles d'association ainsi que la fouille de texte. Toutes ces méthodes confirment l'importance de la fouille de données en cybersécurité en mettant en évidence des connaissances essentielles à la prise de décision et au renforcement des mesures de sécurité. À titre d'exemple, le fait de connaître l'origine d'une cyberattaque aiderait les responsables dans leur comportement envers le pays ou l'entité commanditaire potentielle. Si la période propice à une cyberattaque est connue, cela permettrait une prise rapide de mesures à l'approche de cette période.

Les résultats de notre analyse nous permettent d'émettre quelques observations dont certaines ont des implications importantes pour la communauté de la cybersécurité.

- Les incidents cybernétiques et principalement l'espionnage sont en hausse, nécessitant des mesures de protection des données et de détection précoce accrues.
- Les stratégies d'attaques varient, exigeant une adaptation des techniques de défense.
- La saisonnalité des cyberattaques doit être prise en compte pour renforcer la sécurité pendant les périodes critiques.

- La vigilance est de mise pour chaque région, sous-région et pays car les résultats d'analyse et des tendances peuvent changer rapidement en fonction des avancées technologiques et des conflits géopolitiques de l'ère.
- L'analyse de données avancée tenant compte d'ontologies est importante pour la cybersécurité.

L'avenir de la cybersécurité repose sur une analyse plus avancée des cyberattaques, intégrant des techniques de traitement du langage naturel de pointe et une exploration approfondie des attributs pertinents tels que les dégâts et les victimes. Il s'agit de renforcer la prévention et la réponse aux cybermenaces, offrant ainsi une sécurité accrue dans un monde de plus en plus numérique et interconnecté. Nos travaux futurs couvrent les aspects suivants :

- Utilisation de techniques avancées de traitement de la langue naturelle pour une analyse plus précise des descriptions textuelles des cyberattaques. C'est le cas en particulier de *SetFit (Sentence Transformer Fine-tuning)* [52] qui représente une méthode d'apprentissage automatique prometteuse utilisant un transformateur de phrase pré-entraîné dont l'ajustement à un problème choisi est effectué avec un faible corpus
- Intégration d'attributs additionnels liés aux cyberattaques et prise en compte d'ontologies du domaine comme D3FEND liée à ATT&CK, une base de connaissances sur les cybermenaces et les tactiques, maintenue par Mitre [32].
- Surveillance et adaptation continues pour suivre l'évolution des menaces.

Annexe A

Description de l'ensemble de données

Il s'agit d'un ensemble de données qui retrace les cyberopérations parrainées par des États dans le monde entier depuis 2005. Les 12 colonnes composantes de l'actuel ensemble de données sont décrites de la manière suivante :

- Le **Titre** présente un bref résumé décrivant la cyberopération.
- La **Date** fait référence à la date à laquelle la cyberopération a été menée.
- L'**Affiliation** renferme les groupes ou entités impliquées dans la cyberopération, si connus.
- La **Description** présente du texte décrivant la cyberopération menée.
- La **Réponse** donne une précision sur les suspects impliqués dans la cyberopération, si applicable.
- Les **Victimes** : variable qui recense les personnes ou organisations qui ont été ciblées par la cyberopération, si connues.
- Le **Sponsor** ou commanditaire recense l'état ou le pays suspecté d'être à l'origine de la cyberopération.
- Le **Type** donne le type de cyberattaque.
- La **Catégorie** concerne les institutions visées par la cyberopération.
- La **Sources-1** est la source ayant signalé la cyberopération.
- La **Sources-2** : une deuxième source ayant rapporté la cyberopération, si applicable.
- La **Sources-3** : une troisième source ayant signalé la cyberopération, si applicable.

Bibliographie

[1] Abdel-Fattah, F., AlTamimi, F., and Farhan, K. A. Machine learning and data mining in cybersecurty. In *2021 International Conference on Information Technology (ICIT)* (2021), IEEE, pp. 952–956.

[2] Allen, J., Yang, Z., Landen, M., Bhat, R., Grover, H., Chang, A., Ji, Y., Perdisci, R., and Lee, W. Mnemosyne : An effective and efficient postmortem watering hole attack investigation system. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (2020), pp. 787–802.

[3] Bazhanov, K., and Obiedkov, S. Optimizations in computing the duquenne–guigues basis of implications. *Annals of mathematics and artificial intelligence 70* (2014), 5–24.

[4] Bhuyan, M. H., Bhattacharyya, DhrubaKumar, and Kalita, J. Network anomaly detection : methods, systems and tools. *Ieee communications survey et tutorials 16*, 1 (2013), 303–336.

[5] Buczak, A. L., and Guven, E. A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Commun. Surv. Tutorials 18*, 2 (2016).

[6] Cao, Y., Han, W., and Le, Y. Anti-phishing based on automated individual white-list. In *Proceedings of the 4th ACM workshop on Digital identity management* (2008), pp. 51–60.

[7] Denning, D. E. An intrusion-detection model. *IEEE Transactions on software engineering*, 2 (1987), 222–232.

[8] Domingos, P. A few useful things to know about machine learning. *Communications of the ACM 55*, 10 (2012), 78–87.

[9] Dua, S., and Du, X. *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[10] Ganter, B., and Wille, R. *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag New York, Inc., 1999. Translator-C. Franzke.

[11] Gordon, L. A., Loeb, M. P., and Sohail, T. A framework for using insurance for cyber-risk management. *Communications of the ACM 46*, 3 (2003), 81–85.

[12] H, W. I., WittenIan, F., Eibe, H., and MarkAndrew. *Data Mining : Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington, MA : Morgan Kaufmann, 2011.

[13] Han, J., Cheng, H., Xin, D., and Yan, X. Frequent pattern mining : current status and future directions. *Data mining and knowledge discovery 15*, 1 (2007), 55–86.

[14] Han, J., Pei, J., and Tong, H. *Data mining : concepts and techniques*. Morgan kaufmann, 2022.

[15] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. *ACM sigmod record 29*, 2 (2000), 1–12.

[16] Hofmann, Markus et Klinkenberg, R. *RapidMiner : Cas d'utilisation de l'exploration de données et applications d'analyse commerciale*. CRC Press, 2016.

[17] Houser, W. Could what happened to sony happen to us ? *IT Professional 17*, 2 (2015), 54–57.

[18] Husák, M., Kašpar, J., Bou-Harb, E., and Čeleda, P. On the sequential pattern and rule mining in the analysis of cyber security alerts. In *Proceedings of the 12th International Conference on Availability, Reliability and Security* (2017), pp. 1–10.

[19] Jiawei, H., Micheline, K., and Jian, P. *Data Mining : Concepts and Techniques.-3rd*. Morgan kaufmann, 2012.

[20] Jr, M. F. Sony hacker paralysis reaches day two – update. https://deadline.com/2014/11/sony-computers-hacked-skull-message-1201295288/. Consulté le 05/05/2023.

[21] Juan-Manuel, T.-M. *Résumé automatique de documents : une approche statistique*. Lavoisier, 2011.

[22] Kaelbling, L. P., Littman, M. L., and Moore, A. W. Reinforcement learning : A survey. *Journal of artificial intelligence research 4* (1996), 237–285.

[23] Kaggle. State-sponsored cyber operations (2005-present). https://www.kaggle.com/Datasets/justin2028/state-sponsored-cyber-operations-2005-present. Consulté le 17/03/2023.

[24] Khoury, R. La sécurité logicielle, une approche défensive. *Les Presses de l'Université Laval* (2021), p. 6,7.

[25] Langner, R. Stuxnet : Dissecting a cyberwarfare weapon. *IEEE Security & Privacy 9*, 3 (2011), 49–51.

[26] Libicki, M. The coming of cyber espionage norms. In *2017 9th International Conference on Cyber Conflict (CyCon)* (2017), IEEE, pp. 1–17.

[27] Maasberg, M., Zhang, X., Ko, M., Miller, S. R., and Beebe, N. L. An analysis of motive and observable behavioral indicators associated with insider cyber-sabotage and other attacks. *IEEE Engineering Management Review 48*, 2 (2020), 151–165.

[28] Manning, C. D. *An introduction to information retrieval*. Cambridge university press, 2009.

[29] Mirkovic, J., and Reiher, P. A taxonomy of ddos attack and ddos defense mechanisms. *ACM SIGCOMM Computer Communication Review 34*, 2 (2004), 39–53.

[30] Missaoui, R. Intelligence d'affaires. In *Notes de cours, Université du Québec en Outaouais* (2022).

[31] Missaoui, R., and Emamirad, K. Lattice miner 2.0 : A formal concept analysis tool. In *Supplementary Proceedings of ICFCA'2017* (Dordrecht, 2017), K. Bertet, D. Borchmann, P. Cellier, and S. Ferré, Eds., Université de Rennes, pp. 91–94.

[32] Mitre. D3fend ontology resources. https://d3fend.mitre.org/resources/ontology/. Consulté le 11/09/2023.

[33] Mohammed, M., Khan, M. B., and Bashier, E. B. M. *Machine learning : algorithms and applications*. Crc Press, 2016.

[34] North, M. *Data mining for the masses*, vol. 615684378. Global Text Project Athens, 2012.

[35] on Foreign Relations, C. Cyber operations tracker. https://www.cfr.org/cyber-operations/. Consulté le 05/05/2023.

[36] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. Mining sequential patterns by pattern-growth : the prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering 16*, 11 (2004), 1424–1440.

[37] Radio-Canada. Le site web d'hydro-québec paralysé. https://ici.radio-canada.ca/nouvelle/1971255/hydro-quebec-panne-cyberattaque?depuisRecherche=true. Consulté le 10/05/2023.

[38] RapidMiner. Cluster model visualizer. https://docs.rapidminer.com/10.3/studio/operators/modeling/segmentation/cluster_model_visualizer.html/. Page consultée le 16 juin 2023.

[39] RapidMiner. Description en français de rapidminer. https://edutechwiki.unige.ch/fr/RapidMiner_Studio. Page consultée le 16 juin 2023.

[40] RapidMiner. Documentation et téléchargement de rapidminer studio. https://docs.rapidminer.com/latest/studio/getting-started/. Page consultée le 16 juin 2023.

[41] RapidMiner. Getting started with rapidminer studio. https://docs.rapidminer.com/latest/studio/getting-started/. Consulté le 04/05/2023.

[42] RapidMiner. Le module turbo prep de rapidminer studio. https://docs.rapidminer.com/latest/studio/guided/turbo-prep/. Page consultée le 16 juin 2023.

[43] Raschka, S. *Python machine learning*. Packt publishing ltd, 2015.

[44] Rokach, L. A survey of clustering algorithms. *Data mining and knowledge discovery handbook* (2010), 269–298.

[45] Salem, I. E., Mijwil, M., Abdulqader, A. W., Ismaeel, M. M., Alkhazraji, A., and Alaabdin, A. M. Z. Introduction to the data mining techniques in cybersecurity. *Mesopotamian journal of cybersecurity 2022* (2022), 28–37.

[46] Sarker, I. H. Machine learning : Algorithms, real-world applications and research directions. *SN computer science 2*, 3 (2021), 160.

[47] Sarker, I. H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., and Ng, A. Cybersecurity data science : an overview from machine learning perspective. *Journal of Big data 7* (2020), 1–29.

[48] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., and Xu, M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access 8* (2020), 222310–222354.

[49] Sinaga, K. P., and Yang, M.-S. Unsupervised k-means clustering algorithm. *IEEE access 8* (2020), 80716–80727.

[50] Singer, P. W., and Friedman, A. *Cybersecurity : What everyone needs to know*. oup usa, 2014.

[51] Tariq, M., Aslam, B., Rashid, I., and Waqar, A. Cyber threats and incident response capability-a case study of pakistan. In *2013 2nd National Conference on Information Assurance (NCIA)* (2013), IEEE, pp. 15–20.

[52] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. Efficient few-shot learning without prompts, 2022.

[53] Virmani, C., Choudhary, T., Pillai, A., and Rani, M. Applications of machine learning in cyber security. In *Handbook of research on machine and deep learning applications for cyber security*. IGI Global, 2020, pp. 83–103.

[54] Wikipedia. Tf-idf. https://fr.wikipedia.org/wiki/TF-IDF. Page consultée le 16 juillet 2023.

[55] Witten, I. H., and Frank, E. Data mining : practical machine learning tools and techniques with java implementations. *Acm Sigmod Record 31* (2002), 76–77.

[56] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. Practical machine learning tools and techniques. In *Data Mining* (2005), vol. 2.

[57] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., and Wang, C. Machine learning and deep learning methods for cybersecurity. *Ieee access 6* (2018), 35365–35381.

[58] Yavanoglu, O., and Aydos, M. A review on cyber security datasets for machine learning algorithms. In *2017 IEEE International Conference on Big Data (Big Data)* (2017), pp. 2186–2193.

[59] Zhang, M., and He, C. Survey on association rules mining algorithms. *Advancing computing, communication, control and management* (2010), 111–118.