Ministère de l’Enseignement Supérieur, de la Recherche et de l’Innovation

**ECOLE POLYTECHNIQUE DE THIES**



**DEPARTEMENT GENIE INFORMATIQUE ET TÉLÉCOMMUNICATIONS**

**Projet de Fin d’Études en vue de l’obtention du Diplôme d’Ingénieur de Conception**

Présenté par : Papa Omar DIOP

**Sujet :**

**Conception et Réalisation d'un système de détection et de gestion des anomalies liées aux dysfonctionnements des applications**

\

Soutenu publiquement le ……. devant le jury composé de :

|  |  |  |  |
| --- | --- | --- | --- |
| Président |  |  |  |
| Examinateurs |  |  |  |
|  |  |  |
| Encadrant |  |  |  |
|  |  |  |

**PRÉFACE**

**DEDICACES**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**REMERCIEMENTS**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**Résumé**

Quand il s’agit de la détection d’anomalies, combien de fois on a pu entendre un collègue dire si tu rencontres tel ou tel bug fais ceci ou bien il y a seulement le lead qui connait comment résoudre le problème. Ne pas documenter, normaliser ou partager la façon d'analyser les données pour détecter les intrusions potentielles dans un réseau est plus courant qu’on ne le pense, surtout lorsqu’une équipe est très diverse du point de vue technique et de l'expertise. Cela n'affecte pas seulement les stratégies de détection mais aussi la dynamique de l’équipe ainsi que le bon avancement des projets.

Maintenant, combien de fois avons-nous également pensé à un moyen plus efficace, intuitif ou créatif d'analyser les événements de sécurité que notre organisation collecte, mais nous nous sentons limité aux capacités d'une barre de recherche dépendant d'une seule langue ?

Dans le développement de logiciels, il est absolument nécessaire de s'assurer qu'un système, une fois développé, fonctionne au mieux pendant toute sa durée de vie. Avec le temps, il génère des ressources de données à grande échelle impliquant une grande quantité d'informations de logs, dont le contenu est très significatif. Les données des journaux d'applications sont essentielles au maintien des performances des applications. Les techniques d'analyse, de compréhension et de détection des anomalies dans les journaux d'applications sont donc essentielles pour garantir l'efficacité du développement logiciel. Bien qu'initialement entravées par un matériel limité et un manque d'ensembles de données de qualité, les techniques de détection d'anomalies ont récemment bénéficié d'un regain d'intérêt grâce aux progrès de la technologie d'apprentissage automatique et notamment des réseaux neuronaux.

Dans cette étude, nous explorons la détection d'anomalies, les techniques historiques de détection d'anomalies et les progrès récents des méthodes d’apprentissage, qui promettent de révolutionner la détection d'anomalies dans les données des journaux d'applications. En outre, nous analysons les techniques de détection d'anomalies les plus prometteuses et proposons un modèle basé principalement sur les méthodes d’apprentissage automatique qui améliore les techniques existantes.

Nous verrons par la suite que les résultats de l'évaluation sont implémentés à travers un système d’alertes qui répond aux besoins de nos applications actuellement en production. Ce qui améliore considérablement la précision de la fiabilité en temps réel et fournit ainsi une base de données précise et une base de localisation des anomalies pour l'intelligence artificielle pour les opérations informatiques. L'ensemble du processus a été établi depuis le journal original jusqu'à la mesure de la fiabilité en temps réel. De nouvelles méthodes ont été adoptées pour améliorer la précision, le rappel et la valeur F1 de la détection des anomalies. En outre, comme il s'agit d'une approche axée sur l'apprentissage, il est possible de mettre à jour le modèle de manière incrémentielle afin qu'il puisse s'adapter aux nouveaux modèles de journaux qui apparaissent au fil du temps.

Termes clés – **AIOps**, **Détection d'anomalies, journaux d'applications, apprentissage automatique.**

**Abstract**

When it comes to anomaly detection, how many times have we heard a colleague say that if you encounter such-and-such a bug, do this or that, or that only the boss knows how to fix the problem. Not documenting, standardizing, or sharing how to analyze data to detect potential network intrusions is more common than you might think, especially when a team is very diverse in terms of technical expertise. This affects not only detection strategies but also team dynamics and project progress.

How many times have we thought of a more efficient, intuitive, or creative way to analyze the security events collected by our organization, but we feel limited to the capabilities of a single language-dependent search bar?

In software development, it is absolutely necessary to ensure that a system, once developed, works at its best throughout its lifetime. Over time, it generates large-scale data resources involving a large amount of log information, the content of which is very important. Application log data is critical to maintaining application performance. Therefore, techniques for analyzing, understanding, and detecting anomalies in application logs are critical to effective software development. Although initially hampered by limited hardware and a lack of quality datasets, anomaly detection techniques have recently received renewed interest due to advances in machine learning technology, particularly neural networks.

In this paper, we explore anomaly detection, historical anomaly detection techniques, and recent advances in learning methods, which promise to revolutionize anomaly detection in application log data. In addition, we analyze the most promising anomaly detection techniques and propose a model based primarily on machine learning methods that improve on existing techniques.

We will then see that the evaluation results are implemented through an alerting system that meets the needs of our applications currently in production. This significantly improves the accuracy of real-time reliability and thus provides an accurate database and anomaly location basis for artificial intelligence for IT operations. The entire process has been established from the original log to the real-time reliability measurement. New methods were adopted to improve the accuracy, recall, and F1 value of anomaly detection. In addition, because this is a learning approach, the model can be incrementally updated to account for new log patterns that emerge over time.

Index Terms – **AIOps, Anomaly Detection, Application logs, Machine Learning.**

**TABLE DES MATIERES**

**Résumé**

**Abstract**

**Liste des Figures**

**Liste des Tableaux**

**Liste des Sigles et Abréviations**

**Introduction générale**

1. Présentation de la structure d’accueil …………………………………………………
   1. Structure d’accueil ………………………………………………………….
   2. Contexte de l’immersion ……………………………………………………………
2. Problématique …………………………………………………………………
3. Objectifs du projet ……………………………………………………………………
4. Annonce du plan ……………………………………………………………………

**Chapitre 1 : Généralités sur la détection d’anomalies**

1. Mise en contexte ……………………………………………………………
   1. Apprentissage automatique ………………………………………………….
   2. Apprentissage supervisé ………………………………………………
   3. Apprentissage non supervisé ………………………………………………….
   4. Clustering …………………………………………………….
   5. Logs …………………………………………………………………….
   6. Anomalies ……………………………………………………………………
   7. Isolation Forest …………………………………………………….
   8. One Class SVM …………………………………………………….
   9. K-Means …………………………………………………….
   10. Principal Component Analysis (PCA) …………………………………
   11. Accuracy …………………………………………………….
   12. Precision …………………………………………………….
   13. Recall …………………………………………………….
   14. F1Score …………………………………………………….
2. Les types d’anomalies ……………………………………………………………
   1. Les anomalies ponctuelles…………………………………………………
   2. Les anomalies contextuelles …………………………………………………
   3. Les anomalies collectives ………………………………………………………
3. Les challenges
   1. Données non structurées ………………………………………………………
   2. Informations d'exécution redondantes ………………………………………………
   3. Données déséquilibrées …………………………………………………………….

**Chapitre 2 : État de l’art**

1. Évolution des techniques de détection d’anomalies ………………………………………
   1. Détection d'anomalies basée sur la statistique ………………………………………
   2. Détection d'anomalies basée sur la profondeur ……………………………………
   3. Détection d'anomalies basée sur le regroupement (apprentissage non supervisé)
   4. Détection d'anomalies basée sur la distance ……………………………………
   5. Détection d'anomalies basée sur la densité (apprentissage non supervisé) ………
   6. Décomposition spectrale …………………………………………………………
   7. Détection d'anomalies basée sur l'apprentissage automatique supervisé …………
2. Analyse des données fournies …………………………………………………
   1. Nature des données ……………………………………………………………
   2. Types d’anomalies ………………………………………………………………
   3. Etiquettes des données ……………………………………………………………
   4. Sortie de la détection d'anomalies ………………………………………………
3. Discussion sur la meilleure approche pour ces données ……………………………

(Discussion expliquer le pourquoi on choisit telle méthode d’apprentissage par rapport à un autre ect regarder dans tes notes)

**Chapitre 3 : Étude conceptuelle de la solution**

1. Analyse des besoins ………………………………………
   1. Les besoins fonctionnels ………………………………………
   2. Les besoins non fonctionnels ………………………………………………………
2. Modélisation UML ………………………………………………………………
3. Diagramme des cas d’utilisation (Uses Cases) ……………………………………
4. Diagramme des classes ……………………………………………………………
5. Architecture du projet ………………………………………………………………
6. Workflow global du système ……………………………………………………………
7. Collecte de données ………………………………………………………………
8. Description du jeu de données ……………………………………………………………
9. Workflow de la détection ………………………………………………………………
10. Conclusion ………………………………………………………………

**Chapitre 4 : Implémentation et déploiement**

1. Environnement de développement et technologies ………………………………………
   1. Machine Learning Pipeline ………………………………………………
   2. Docker ………………………………………………………………………
   3. Containers and Virtual Machines ………………………………………
   4. Images …………………………………………………………………
   5. Containers Lifecycle …………………………………………………
   6. Networking ……………………………………………………………
   7. Docker Compose …………………………………………………………
2. Apache Kafka …………………………………………………
   1. Kafka architecture …………………………………………………
   2. Topic structure …………………………………………………
   3. Clustered deployment …………………………………………………
   4. Parallel processing of the topic ……………………………………
   5. Zookeeper (talk about it) ……
3. Logstash …………………………………………………
   1. Data extraction …………………………………………
   2. Data transformation …………………………………………………
   3. Data load …………………………………………………
   4. Parallel processing …………………………………………………
   5. Grok (talk about it) …….
4. Elasticsearch ………………………………………………
   1. Elasticsearch architecture …………………………………………………
   2. Difference from a database ………………………………………………
   3. Clustered deployment …………………………………………………
   4. Search and aggregation ………………………………………………
5. Kibana ………………………………………………
   1. Widget and Dashboard …………………………………………………
   2. Discover ………………………………………
   3. ELK stack monitor and management ………………………………………
6. Beats ………………………………………………
   1. Filebeat and Metricbeat…………………………………………
7. Apache Spark ………………………………………………
   1. Challenges in big data processing …………………………………………
   2. Spark cluster …………………………………………
   3. Handling distributed data …………………………………………
   4. Broadcast variable …………………………………………
   5. Low-level unstructured Spark API ………………………………………
   6. Structured Spark API …………………………………………
   7. Distributed execution of jobs …………………………………………
   8. Mleap (talk about it) ……
8. Jupyter Notebook ………………………………………………
9. Summary ……………………………………………
10. Présentation de la solution ……………………………………………………………
11. Analyse et évaluation des résultats obtenus ………………………………………………
12. Déploiement de la solution ………………………………………………………………

**Chapitre 5: Conclusion Générale**

1. Synthèse des travaux
2. Apport du stage
3. Difficultés rencontrées
4. Perspectives

**Références**

## LISTE DES FIGURES

**FIGURES**

## LISTE DES TABLEAUX

**TABLEAU**

## LISTES DES ABBRÉVIATIONS

**ABBREV**

## INTRODUCTION GÉNÉRALE

La détection des anomalies est l'une des tâches les plus importantes pour construire un système fiable et sûr. L'objectif de la détection d'anomalies est de détecter une déviation significative du comportement du système par rapport au comportement normal. La plupart des systèmes nécessitent une détection en temps réel des anomalies dans le but spécifique de réduire les dommages qui peuvent être causés par l'ignorance d'une anomalie ou d'une détection ultérieure. Un exemple d’anomalie qu’on pourrait citer dans une application pourrait être une faille de sécurité, ou bien une panne de service fournit par un tiers. Les implémentations récentes de la détection des anomalies sont principalement basées sur des méthodes d'auto-apprentissage.

L'application des méthodologies d'apprentissage machine ML aux opérations informatiques (AIOps), grâce au nombre croissant de frameworks disponibles, souvent open-source, devient omniprésente dans de nombreux domaines. De plus, l'énorme disponibilité de données d'entraînement, due à la grande capacité de stockage des centres de données et aux systèmes modernes de collecte de données, permet de développer des modèles d'apprentissage automatique plus précis et moins sujets aux erreurs. La diffusion de l'apprentissage automatique correspond à un nombre croissant d'utilisateurs et de systèmes qui accèdent aux prédictions fournies par le modèle d'apprentissage automatique. Cela pose de nouveaux défis, car le modèle ML développé doit être placé dans un environnement dit opérationnel. Les environnements opérationnels ne sont généralement pas sous le contrôle des développeurs, contrairement aux environnements de développement, qui sont entièrement sous contrôle. La principale différence est qu'il est difficile de prévoir l'ampleur des données que le système pourrait être amené à traiter et que ces valeurs peuvent changer à tout moment. En outre, les systèmes matériels et logiciels peuvent tomber en panne à tout moment, mais aucun temps d'arrêt ne peut être toléré dans les environnements opérationnels, car les utilisateurs et les systèmes réels dépendent des services fournis par l'infrastructure ML. Pour résoudre ces deux problèmes, l'infrastructure qui alimente le modèle ML doit être mise à l'échelle et répliquée en conséquence.

Un autre défi des environnements opérationnels est que nous sommes soumis à des exigences temporelles strictes, souvent en temps réel, pour éviter les goulots d'étranglement dans le système supporté par le modèle ML. La plupart du temps, l'application des modèles ML n'est qu'un élément d'une chaîne de traitement plus large, appelée pipeline logiciel. Les pipelines sont utilisés pour traiter les flux de données en continu, l'échelle de l'infrastructure résout en partie ces problèmes, il convient également d'exécuter les modèles d'apprentissage automatique sur des logiciels spécialisés qui permettent une prédiction distribuée à faible latence sur de grandes quantités de données. La surveillance des performances et de l'état des logiciels dans le pipeline devient fondamentale pour répondre aux exigences strictes imposées par la nature de l'environnement opérationnel : les administrateurs doivent être en mesure de détecter les applications défectueuses et de les remplacer rapidement, de reconnaître rapidement les éventuels goulets d'étranglement et de les résoudre dès que possible. La surveillance nécessite une application spécialisée et un stockage dédié pour prendre en charge la collecte des métriques [1], ainsi qu'une interface dédiée pour analyser les données collectées.

Un autre aspect important des environnements opérationnels est la gestion de la sécurité, un problème souvent négligé dans les phases de développement. Les attaques de pirates sur les systèmes opérationnels sont de plus en plus fréquentes, ce qui fait de la sécurité l'une des préoccupations les plus importantes pour les environnements opérationnels où des informations sensibles sont traitées. Nous devrions être en mesure de garantir ce que l'on appelle la triade de la CIA : La confidentialité, c'est-à-dire que toutes les communications entre les étapes du traitement ne doivent pas être lues ou mises à la disposition d'entités non autorisées, l'intégrité, les données doivent être conservées et ne pas être modifiées par des utilisateurs malveillants et la disponibilité, c'est-à-dire la capacité du système à fonctionner en permanence. Il est essentiel de garantir cette propriété dans un environnement opérationnel afin d'éviter les fuites de données potentielles ou la mise hors service du système par un attaquant. La sécurisation de la pile d'applications doit donc être effectuée dès le début du processus de développement en appliquant les meilleures pratiques de sécurité telles que le cryptage et l'authentification entre les parties. Le logiciel choisi doit être bien connues pour avoir été développées avec la sécurité comme première préoccupation, ce que l'on appelle la sécurité dès la conception. Les applications à code source ouvert ont un plus grand avantage que les applications à code source fermé en ce qui concerne la sécurité : étant gratuites, elles sont largement adoptées et testées contre les vulnérabilités et les bogues, et comme leur code est accessible au public, il est généralement examiné par des experts en sécurité.

Pour incorporer les exigences susmentionnées dans notre cadre, nous proposons une architecture logicielle en pipeline supportant l'application de modèles ML dans un environnement opérationnel d'une manière évolutive et tolérante aux pannes, et permettant le traitement en temps quasi réel de données en continu. La conception proposée peut prendre en charge jusqu'à des milliers de sources de données dans un environnement réel et imprévisible et est construite à l'aide de modules logiciels de pointe et de sources ouvertes qui sont enchaînés pour former un pipeline logiciel. Tous les composants prennent en charge le déploiement distribué et sont conformes aux meilleures pratiques de sécurité : toutes les applications sont conçues pour fonctionner dans un cluster afin d'effectuer la distribution du travail et le traitement parallèle, et elles incluent le cryptage des données échangées et l'authentification entre les parties. L'architecture de pipeline proposée est composée de quatre parties :

* Ingestion de données : Applications qui prennent en charge l'ingestion des données dans le système. Elle comporte quatre phases : mise en mémoire tampon, prétraitement et stockage. C'est le point d'entrée des sources de données en continu.
* Application de modèle d'apprentissage automatique : les données stockées sont traitées par lots avec le modèle d'apprentissage automatique et les résultats sont stockés en retour.
* Visualisation des données : permet d'examiner les données d'entrée et les résultats du traitement des données ainsi que la prédiction dans une interface.
* Surveillance de la santé : surveille l'état des applications du pipeline afin de détecter les défaillances logicielles et d'y réagir rapidement.

Afin de réaliser une implémentation de référence et d'avoir un cas d'utilisation réel pour tester le pipeline développé, nous considérons l'application de la ML pour la gestion des anomalies dans le projet DER actuellement en développement à TeamX et utilisons un modèle ML prédéveloppé qui va effectuer la prédiction d’une panne potentielle dans l’application.

Ainsi, nous avons structuré le document en quatre grandes parties. La première partie de ce mémoire portera sur la présentation la structure d’accueil. Ensuite, la deuxième partie va décrire la réalisation d’une étude de l’état de l’art technique dans ce domaine. Dans la troisième on va la proposer d’une architecture appropriée pour le système. Ensuite la quatrième partie traitera la mise en œuvre agile de la solution allant de l’implémentation jusqu’au déploiement et enfin proposer des pistes d’évolution pour la solution.

1. **Présentation de la structure d’accueil**

A remplir later

1. **Problématique**

L’entreprise TeamX est une start-up créée récemment en 2018. Elle évolue dans le domaine de la prestation de service avec des projets tournant autour de domaines assez diversifiés (Fintech, Healthtech, E-commerce, …).

Dans le cadre de la création des projets, les technologies utilisées sont à peu près similaires avec un système de logger permettant de logger les **erreurs** qui surviennent au niveau des apps.

Cette tâche de détection étant effectuée manuellement, il n’existe pas un système de monitoring pour notifier l’équipe en cas de soucis au niveau des applications:

1. un endpoint (api) qui ne marche plus normalement comme l’api Wizzal (cas réel rencontré en entreprise)
2. un service tiers qui n’est plus valide faute de certificat
3. bugs en continu suite au déploiement de nouvelles fonctionnalités

Du coup souvent les développeurs sont obligés d’aller consulter les logs au niveau du serveur pour tracer certaines erreurs relancer par les clients.

1. **Objectifs du projet**

En réponse à cela, l’idée a été émise de créer un système d’alerte basé sur l’IA (Machine Learning) qui fait des prédictions sur les logs pour anticiper les pannes de services sur les applications en production.

1. **Annonce du plan**

Annoncer le plan

## GÉNÉRALITÉ SUR LA DÉTECTION D’ANOMALIES

* 1. **Mise en contexte**

1. **Apprentissage automatique**

L'apprentissage automatique est une forme d'intelligence artificielle qui donne aux ordinateurs la capacité d'apprendre sans être spécifiquement programmés. Il se concentre sur la création de programmes informatiques qui sont susceptibles de changer lorsqu'ils sont exposés à de nouvelles données. Il peut être classé comme supervisé ou non supervisé. Elle consiste à utiliser les bonnes caractéristiques pour construire les bons modèles qui accomplissent les bonnes tâches. Ces tâches comprennent la classification binaire et multi-classes, le regroupement par régression et la modélisation descriptive.

1. **Apprentissage supervisé**

L'apprentissage supervisé nécessite la disponibilité de données étiquetées, car les algorithmes relevant de cette catégorie peuvent appliquer des connaissances antérieures à de nouvelles données.

1. **Apprentissage non supervisé**

D'autre part, l'apprentissage à partir de données non étiquetées est appelé apprentissage non supervisé. Par exemple, pour évaluer des données particulières en clusters, on peut calculer la distance moyenne des centres de clusters. D'autres formes d'apprentissage non supervisé comprennent l'apprentissage d'associations et l'identification de variables cachées telles que les genres de films. La suradaptation est un problème dans l'apprentissage supervisé ; par exemple, l'attribution d'un cluster à chaque point de données réduira la distance moyenne au centre du cluster à zéro, mais ne sera pas très utile. Ces algorithmes tirent des conclusions à partir des ensembles de données.

1. **Clustering**

La tâche consistant à regrouper des données sans information préalable sur les groupes est appelée "clustering". Un algorithme de clustering typique fonctionne en évaluant la similarité entre les instances et en plaçant les instances similaires dans le même cluster et les instances dissemblables dans des clusters différents.

1. **Logs**

Les journaux des grands systèmes de données sont généralement des données non structurées imprimées en séquence temporelle. Normalement, chaque entrée de journal (ligne) peut être divisée en deux parties différentes : constante et variable. La partie constante est constituée des messages imprimés directement par les instructions du code source. Les clés de journal sont souvent extraites de ces parties constantes, où les clés de journal sont les messages constants communs à toutes les entrées de journal similaires. Un journal de test ressemble à ceci :

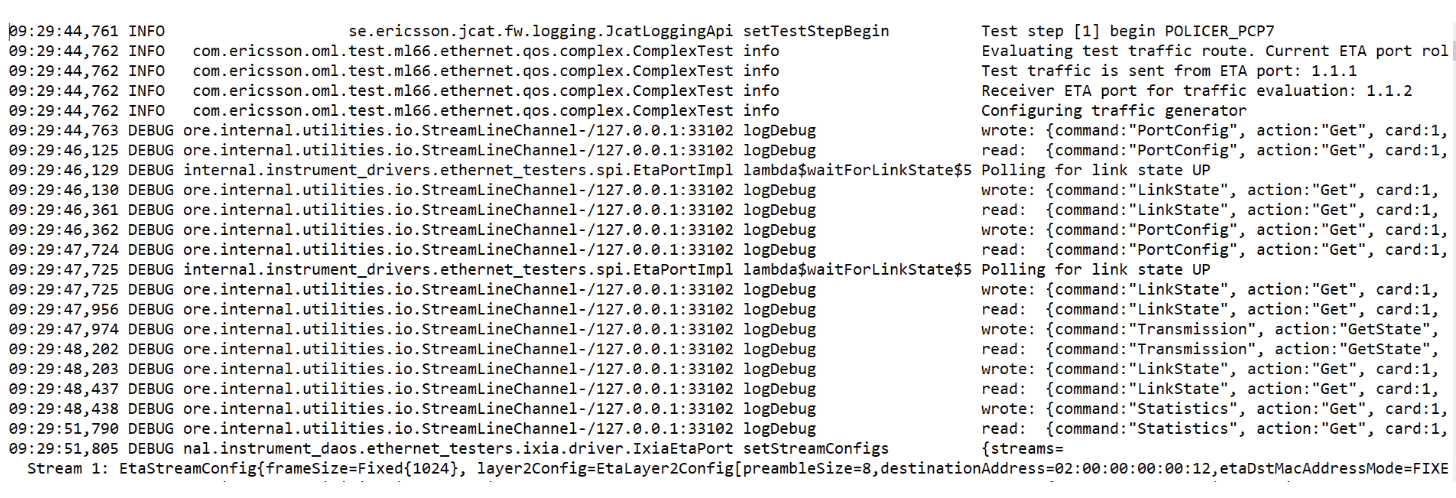
****

Figure 1.1: Un journal de test

Dans la figure ci-dessus, le journal commence par un horodatage suivi d'un événement de type INFO ou DEBUG. Ceci indique le type de message que l'événement génère dans le fichier journal. Après l'événement, une description de l'événement est générée, suivie des commandes et des étapes de test effectuées lors de l'exécution des tests.

1. **Anomalies**

Une anomalie ne cadre pas avec le reste du modèle. Le mot anomalie vient du mot grec "anomolia" qui signifie inégal ou irrégulier. Lorsque quelque chose est inhabituel par rapport aux choses qui l'entourent, on parle d'anomalie.

1. **Isolation forest**

L’isolation forest est une méthode non supervisée basée sur la construction d’arbres. L’idée derrière cet algorithme c’est qu’une donnée atypique sera plus facile à isoler qu’une donnée standard. Cet algorithme calcule, pour chaque donnée du jeu, un score d’anomalie, c’est-à-dire une mesure qui reflète à quel point la donnée en question est atypique. Afin de calculer ce score, l’algorithme isole la donnée en question de manière récursive : il choisit un descripteur et un “seuil de coupure” au hasard, puis il évalue si cela permet d’isoler la donnée en question ; si tel est le cas, l’algorithme s’arrête, sinon il choisit un autre descripteur et un autre point de coupure au hasard, et ainsi de suite jusqu’à ce que la donnée soit isolée du reste. Au final, l’algorithme classe les données suivant leur score d’anomalie. Un score > 0.5 témoigne d’une anomalie.

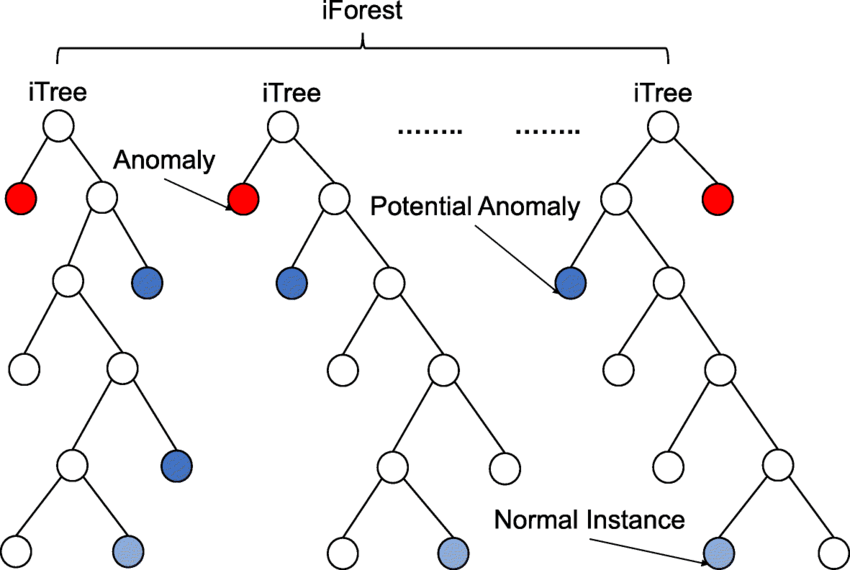


Figure 1.2 : Construction d'iForest pour l'ensemble des données

1. **One Class SVM**

Le SVM à classe unique est une variante du SVM qui peut être utilisée dans un cadre non supervisé pour la détection des anomalies. Lors de la modélisation d'une classe, l'algorithme capture la densité de la classe majoritaire et classe les exemples situés aux extrêmes de la fonction de densité comme des valeurs aberrantes. Cette modification du SVM est appelée SVM à une classe.

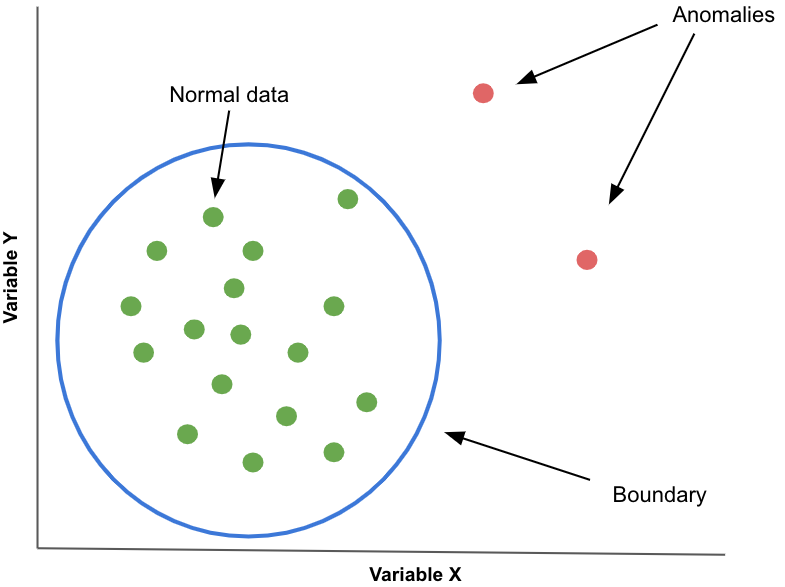


Figure 1.3 : Un exemple visuel d'un simple SVM à une classe

1. **K-Means**

Le clustering K-means peut être une méthode de quantification vectorielle, issue initialement du traitement du signal, qui vise à partitionner n observations en k clusters au cours desquels chaque observation appartient au cluster dont la moyenne est la plus proche, servant de prototype du cluster. Cet algorithme regroupe les données en essayant de séparer les échantillons en n groupes de variances égales, en minimisant un critère appelé somme des carrés d'inertie ou intra-groupe. Il nécessite que le nombre de clusters soit spécifié. Cet algorithme divise un ensemble de N échantillons X en K grappes disjointes C, chacune décrite par la moyenne des échantillons de la grappe. Les moyennes sont communément appelées les "centroïdes" des clusters. Cet algorithme vise à choisir des centroïdes qui minimisent l'inertie, ou le critère de la somme des carrés à l'intérieur d'une grappe. L'inertie est la mesure de la cohérence interne des clusters.

K-means est souvent appelé l'algorithme de Lloyd. En termes simples, l'algorithme comporte trois étapes. La première étape choisit les centroïdes initiaux, la méthode la plus basique étant de strier sur k échantillons de l'ensemble de données X. Après l'initialisation, K-means consiste à boucler entre les deux autres étapes. La première étape assigne chaque échantillon à son centroïde le plus proche. La deuxième étape crée de nouveaux centroïdes en prenant la moyenne de tous les échantillons affectés à chaque centroïde précédent. La différence entre l'ancien et le nouveau centroïde est calculée, et l'algorithme répète ces deux dernières étapes jusqu'à ce que cette valeur soit inférieure à un seuil. En d'autres termes, il répète jusqu'à ce que les centroïdes ne se déplacent plus de manière significative.

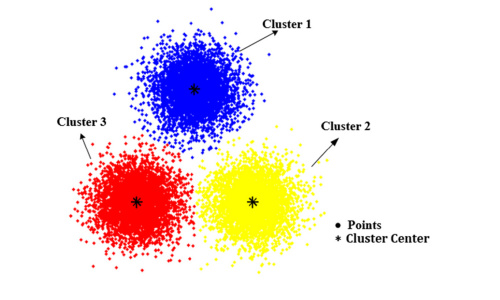


Figure 1.4 : Schéma conceptuel du clustering k-means.

1. **Principal Component Analysis (PCA)**

L'ACP est une méthode statistique qui permet de saisir des modèles dans des données hautement dimensionnelles en choisissant automatiquement un ensemble de coordonnées les composantes principales qui reflètent la covariation entre les coordonnées originales. Nous utilisons l'ACP pour séparer les événements répétitifs dans les vecteurs de caractéristiques, ce qui facilite la détection des modèles de messages anormaux. L'ACP a un temps d'exécution linéaire par rapport au nombre de vecteurs de caractéristiques, de sorte que la détection d'anomalies peut s'étendre à des fichiers journaux volumineux.

L'ACP est l'une des techniques les plus couramment utilisées en analyse multivariée pour la réduction de la dimension et l'extraction de caractéristiques, et elle est particulièrement bien adaptée aux cas où les données sont hautement dimensionnelles.

L'ACP a un large éventail d'applications allant de la compression des données au regroupement.  
Les métriques utilisées pour évaluer les performances des approches utilisées sont mentionnées ci-dessous :

1. **Accuracy**

L’accuracy est la mesure de performance la plus intuitive. Il s'agit simplement d'un rapport entre les observations correctement prédites et le total des observations. Si nous avons une précision élevée pour un modèle particulier, alors ce modèle est le meilleur.

Accuracy = T P + T N/(TP + F P + F N + T N)

1. **Precision**

La précision est le rapport entre les observations positives correctement prédites et le total des observations positives prédites. Une précision élevée est liée à un faible taux de faux positifs.

Precision = TP/TP + FP

1. **Recall**

Le rappel est le rapport entre les observations positives correctement prédites et l'ensemble des observations de la classe actuelle. Il a des valeurs comprises entre 0 et 1. Plus le rappel est élevé, plus le modèle est performant.

Recall = TP/TP + FN

1. **F1Score**

Le score F1 est la moyenne pondérée de la Précision et du Rappel. Ce score prend donc en compte les faux positifs et les faux négatifs. Il a des valeurs entre 0 et 1, plus le score F1 est élevé, plus le modèle est performant.

F 1Score = 2 ∗ (Recall ∗ P recision)/(Recall + Precision)

* 1. **Les types d’anomalies**

1. **Les anomalies ponctuelles**

Une anomalie ponctuelle est une donnée qui dévie de manière significative de la distribution moyenne ou normale du reste des données. Ces données sont souvent générées par un système, et la déviation significative est limitée à des points de données spécifiques et partage peu de contexte avec le reste des données moyennes ou normales.

Les anomalies ponctuelles sont les plus simples à détecter et de nombreuses techniques existent pour automatiser la détection des anomalies ponctuelles. Les anomalies ponctuelles peuvent être rapidement découvertes et corrigées, et ont donc rarement un effet néfaste significatif sur les applications. La figure 1.5 montre des exemples d'anomalies ponctuelles.

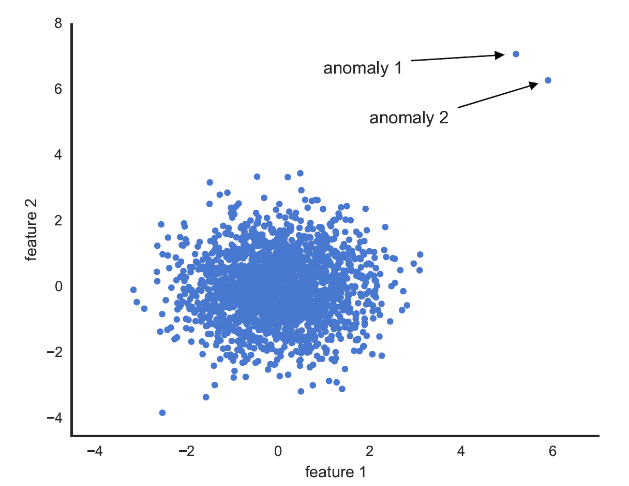


Figure 1.5 : Anomalie ponctuelle

1. **Les anomalies contextuelles**

Une anomalie contextuelle est identifiée comme un comportement anormal restreint à un contexte spécifique, et normal selon d'autres contextes. Ce type d'anomalie, également appelé anomalie conditionnelle, est souvent difficile à détecter car il nécessite une connaissance approfondie du domaine pour comprendre le contexte dans lequel l'anomalie se produit. La figure 6 montre une anomalie contextuelle dans un relevé de température. Les valeurs t1 et t2 sont identiques, mais t2 devient une anomalie une fois replacée dans son contexte : l’été.

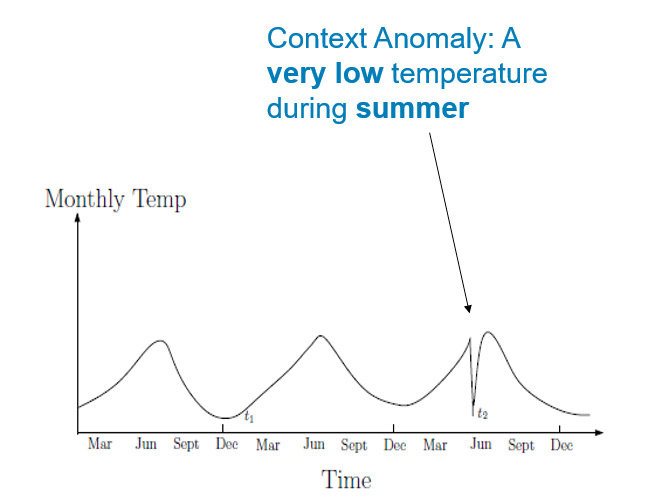


Figure 1.6 : Anomalie contextuelle

1. **Les anomalies collectives**

Contrairement aux anomalies contextuelles et ponctuelles, les anomalies collectives apparaissent comme un groupe de valeurs anormales dans les données. Les anomalies collectives sont des comportements anormaux d'une collection d'instances de données par rapport à l'ensemble des données. Les instances de données individuelles peuvent ne pas représenter une anomalie, mais la présence d'instances de données dans l'anomalie collective est un indicateur de comportement anormal. Il convient toutefois de noter qu'à elle seule, une instance de données ne représente pas une anomalie collective et qu'elle doit apparaître dans une collection de données pour être collectivement anormale. La figure 1.7 montre un électrocardiogramme. L’absence brève d’activité entre deux pulsations du cœur est normale. Si cette valeur se répète sur une plus longue période en revanche, c’est une anomalie.



Figure 1.7 : Électrocardiogramme

* 1. **Les challenges**

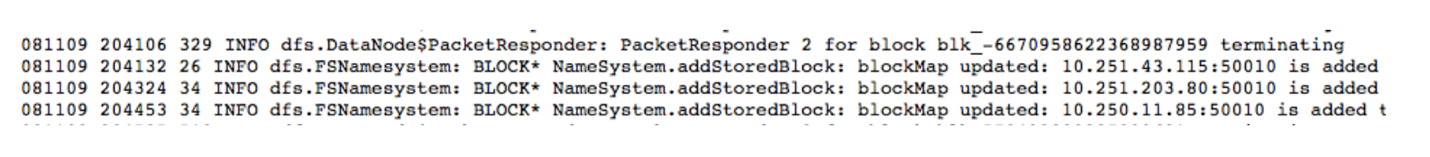
Les anomalies dans les données du journal des applications sont considérées comme des modèles ou des caractéristiques qui ne suivent pas la moyenne ou le comportement normal lors d'un fonctionnement parfait. Comme le décrit Grubbs [20], "une observation aberrante, est une observation qui semble dévier de façon marquée des autres membres de l'échantillon dans lequel elle se produit". De telles anomalies peuvent être initiées par des acteurs malveillants, des bugs au niveau du système ou une utilisation incorrecte de la part de l'utilisateur et sont souvent des symptômes d'une panne ou d'une violation imminente du système. La détection d'anomalies dans les logs est particulièrement difficile, qu'elle soit automatisée ou manuelle, pour les raisons suivantes : des données non structurées, des informations d'exécution redondantes ou bien des données déséquilibrées.

1. **Données non structurées**

Le journal d'une application, comme le montre la figure 2, n'est pas structuré et est stocké en texte brut. Ce manque de structure complique l'analyse des données, qui est encore exacerbée par les formats de journalisation qui varient complètement entre les applications.

1. **Informations d’exécution redondantes**

Comme on peut le voir sur la Fig. 1.8, les journaux d'application contiennent des informations d'exécution telles que l'adresse IP des serveurs. Ces données changent pendant l'exécution et varient d'un serveur à l'autre ; elles sont donc redondantes pour la détection des anomalies. Comme le montre également la figure 1.8 les données des journaux d'application contiennent des données spécifiques au domaine, telles que "blockMap updated" pour les journaux, qui, combinées aux informations d'exécution redondantes, augmentent la complexité de la détection des anomalies.

Figure 1.8 : Journal avec des données redondantes

1. **Données déséquilibrées**

Les données du journal des applications sont conçues pour enregistrer tous les changements apportés à une application et contiennent donc des données fortement déséquilibrées en faveur d'une exécution non anormale. Par exemple, pour des données générées par une application peut contenir moins de 5% par rapport aux événements de journal. La taille et la nature déséquilibrée des données de journal compliquent donc le processus de détection des anomalies.

**ÉTAT DE L’ART**

1. Évolution des techniques de détection d’anomalies ………………………………………
   1. Détection d'anomalies basée sur la statistique ………………………………………
   2. Détection d'anomalies basée sur la profondeur ……………………………………
   3. Détection d'anomalies basée sur le regroupement (apprentissage non supervisé)
   4. Détection d'anomalies basée sur la distance ……………………………………
   5. Détection d'anomalies basée sur la densité (apprentissage non supervisé) ………
   6. Décomposition spectrale …………………………………………………………
   7. Détection d'anomalies basée sur l'apprentissage automatique supervisé …………
2. Analyse des données fournies …………………………………………………
   1. Nature des données ……………………………………………………………
   2. Types d’anomalies ………………………………………………………………
   3. Etiquettes des données ……………………………………………………………
   4. Sortie de la détection d'anomalies ………………………………………………
3. Discussion sur la meilleure approche pour ces données ……………………………

(Discussion expliquer le pourquoi on choisit telle méthode d’apprentissage par rapport à un autre ect regarder dans tes notes)

**ÉTUDE** **CONCEPTUELLE DE LA SOLUTION**

Analyse des besoins

Les besoins fonctionnels

Les besoins non fonctionnels

Modélisation UML

Diagramme des cas d’utilisation (Uses Cases)

Diagramme des classes

Architecture du projet

Workflow global du système

Collecte de données

Description du jeu de données

Workflow de la détection

**IMPLÉMENTATION ET DÉPLOIEMENT**

1. Environnement de développement et technologies ………………………………………
   1. Machine Learning Pipeline ………………………………………………
   2. Docker ………………………………………………………………………
   3. Containers and Virtual Machines ………………………………………
   4. Images …………………………………………………………………
   5. Containers Lifecycle …………………………………………………
   6. Networking ……………………………………………………………
   7. Docker Compose …………………………………………………………
2. Apache Kafka …………………………………………………
   1. Kafka architecture …………………………………………………
   2. Topic structure …………………………………………………

**CONCLUSION GÉNÉRALE**

1. Synthèse des travaux
2. Apport du stage
3. Difficultés rencontrées
4. Perspectives

## RÉFÉRENCES

1. Anomaly Detection [En Ligne] Disponible :   
   <https://www.gavstech.com/anomaly-detection-in-aiops/>
2. What is Elasticsearch? [En Ligne] Disponible: <https://www.elastic.co/guide/en/elasticsearch/reference/8.1/elasticsearch-intro.html>
3. Elastic Machine Learning [En Ligne] Disponible: <https://www.elastic.co/guide/en/machine-learning/current/machine-learning-intro.html#machine-learning-intro>
4. Nginx [En Ligne]. Disponible: <http://nginx.org/en/docs/http/ngx_http_log_module.html#log_format>
5. Mohiuddin Solaimani, “Anomaly Detection for Application Log Data” (2015). Thèse de doctorat, University of Texas at Dallas, Disponible: <https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1640&context=etd_projects>
6. Machine learning on Elastic Search using Apache Spark and ES-Hadoop [En Ligne]. Disponible: <https://excelerate.systems/machine-learning-elasticsearch/>
7. Apache Spark Tutorial [En ligne]. Disponible: <https://www.tutorialkart.com/apache-spark>
8. User Agent Anomaly Detection [En Ligne] Disponible: <https://slideplayer.com/slide/13693768/>
9. <https://posts.specterops.io/threat-hunting-with-jupyter-notebooks-part-1-your-first-notebook-9a99a781fde7>
10. <https://opensource.com/article/19/5/log-data-apache-spark>
11. Spark MLib [En Ligne] Disponible: <https://spark.apache.org/docs/latest/ml-features>
12. Machine Learning Pipeline [En Ligne] Disponible: ​​<https://valohai.com/machine-learning-pipeline/>
13. Machine Learning Model [En Ligne] Disponible:   
    <https://www.youtube.com/watch?v=Om7G2qRBSVw>
14. Embed a Spark ML Model [En Ligne] Disponible:   
    <https://towardsdatascience.com/how-to-embed-a-spark-ml-model-as-a-kafka-real-time-streaming-application-for-production-deployment-933aecb79f3f>
15. Complete Data Science Project Template  (Tres utile pour la presentation) [En Ligne] Disponible:   
    <https://towardsdatascience.com/complete-data-science-project-template-with-mlflow-for-non-dummies-d082165559eb>
16. XGBoost [En Ligne] Disponible:   
    <https://towardsdatascience.com/pyspark-and-xgboost-integration-tested-on-the-kaggle-titanic-dataset-4e75a568bdb>
17. Reason to embed model in kafka instead of remote [En Ligne] Disponible: <https://www.kai-waehner.de/blog/2020/10/27/streaming-machine-learning-kafka-native-model-server-deployment-rpc-embedded-streams/>
18. Creating Apache Kafka Connector [En Ligne] Disponible: <https://www.confluent.io/blog/create-dynamic-kafka-connect-source-connectors/>
19. Send data from kafka to elasticsearch [En Ligne] Disponible: <https://selectfrom.dev/how-to-send-data-from-a-kafka-topic-to-elasticsearch-c623fba84a84>
20. Grubbs, Frank E. "Procedures for detecting outlying observations in samples." Technometrics 11, no. 1 (1969): 1-21.